生成式搜索与品牌可见性工程

GEO 高级研修项目

研究方法、证据与可复现工程

独立研究工作稿・中文全译本 0.6 ・研究文献截止：2026 年 8 月 24 日

目录

阅读约定

目的与范围

用户尚未访问信源，搜索界面就已综合生成答案，这种情况正变得越来越普遍。对组织而言，由此产生了一个颇具吸引力、却尚未被充分定义的问题：“如何针对 AI 答案进行优化?”有用的回答不能只是一份写作技巧清单。我们需要信息处理流程的模型，需要说明究竟在测量什么的理论，也需要一套判断准则，以明确哪些观察能够支持哪些主张。

本讲义的英文原版与课程配套，但不照搬课程的周次安排和评分结构；本书是该讲义的完整中文译本。它面向信息检索、机器学习、营销科学、传播、产品、数据分析和治理领域的读者。能够消除歧义时，我们使用数学；某项主张需要检验而非反复转述时，我们就为它设计实作任务。

证据、主张与版本

初始论文库 GEO-42 来自首版冻结时用户提供的 Zotero 导出文件，共有 42 条记录：记录的出版年份为 2024 年的有 4 条，2025 年的有 13 条，2026 年的有 25 条；其中 41 条被归类为预印本，1 条被归类为期刊论文。每条记录均含 DOI 和本地 PDF 附件，但有 1 条缺少 URL 和摘要。5 条记录带有元数据警告，最常见的问题是导出的年份与 arXiv 标识符所编码的年份不一致。导出文件本身是可审计的来源对象，因此这些警告不会被悄悄“修正”。规范化目录、BibTeX 文件和异常报告是彼此独立的成果文件。

论文库只是证据体系的一层。官方文档、标准、源代码、基准数据、博客、白皮书、教程和视频，各自具有不同的证据支持上限。厂商白皮书可以记录一种工作流程，却未必证明其效果；预印本可以建立测试环境中的结果，却未必揭示生产平台的实现；官方帮助页面可以定义政策，却未必说明排序模型。

贡献、未作出的主张与评审状态

本稿提出四项方法设计。第一，将生成式搜索的可见性分解为具有版本信息的条件事件，而非一个排序分数。第二，以“主张—证据—信源”控制层连接信源身份、原子主张、证据片段和溯源信息。第三，给出测量与干预协议，使其中的单位、分母、不确定性、失败状态和证据支持上限始终可见。第四，将这些方法与可逆工程、安全措施和治理检查相连接。

这些仍是有待评议的学术设计，并不证明该框架最优，不证明某个具名平台按这些阶段实现，也不证明任何干预能够改善现实世界中的可见性或业务结果。项目受控材料包含 LaTeX 源文件、参考文献、主张与符号登记表、42 张证据卡、图片来源记录和合成实作规范。自动构建与渲染检查不能替代人工科学评审、独立复现、无障碍审查或出版权利确认。

信源身份、版本锁定、数值修正和负面发现都会保留，而不会被无声覆盖。公开代码仓库、经确认的作者信息、引用格式、许可和勘误渠道，将在责任作者确认后公布；本稿不会用虚构隶属关系或联系人信息填补这些尚未确定的事项。

阅读路径

研究路径

阅读第 1–7 章，再阅读论文地图与可复现性附录。重点关注目标估计量、协议、不确定性和外部有效性。

工程路径

阅读第 1、3–5、8、9 章。沿着信源生命周期，依次理解可用性、检索、生成、归因、纠正和受控干预。

品牌与内容路径

阅读第 1、2、6–10 章。先建立事实基础，再优化表达，并区分提及、引证、吸收、显著程度与业务结果。

治理路径

阅读第 2、7、11、12 章。结合主张台账、威胁模型、发布条件、标准登记表和事件处置协议开展工作。

符号体系的约定

本稿尽可能用符号表示可观察的单位。封闭系统内部的某个阶段不可观察时，我们不会把代理指标伪装成那个隐藏变量。平台观察按日期、界面、地区、语言、账户状态和重复次数记录版本。若没有额外的识别论证，任何结果都不能从“在本协议中观察到”升级为“平台就是这样工作的”。

1 从排序链接到生成式答案

核心思想

生成式引擎优化并非单一的排序问题。它研究一系列事件并对其实施负责任的干预：信源可能被发现、被检索、进入上下文、参与生成、获得归因、被用户理解，最终引发行动。只有明确了测量所针对的具体事件，测量才有意义。

1.1 本章回答的问题

读完本章，应能够回答以下四个问题，而不将它们压缩为一个可见性分数：

1. 搜索界面返回生成式答案而非排序列表时，可观察的研究对象是什么？
2. 报告中的指标究竟统计了哪些事件：提及、检索、信源使用、归因、忠实性，还是行动？
3. 从回答、显示的引证和交互轨迹中，能够与不能够推断哪些隐藏流程阶段？
4. 在追求信源方可见性的同时，干预目标应如何维护用户效用、证据、忠实性与安全？

1.2 界面变化尚不等于机制变化

传统网页搜索把排序列表呈现在用户面前。选择信源之前，用户可以查看标题、摘要、域名、位置和链接。生成式搜索界面则可能直接给出综合答案、一组引证、后续对话或一项行动。这改变了注意力的分配，却不意味着可见界面披露了产出它的全部机制。系统可能改写查询、执行多次搜索、结合词法检索与神经检索、对候选项重排序、因上下文预算不足而丢弃段落、生成答案，再单独附上引证。有些产品展示其中的片段，另一些只展示最终回答。

早期 GEO 表述将创作者在生成回答中的可见性明确设为优化对象，并引入使用固定信源集合的基准 [1]。这一表述具有奠基意义，但后续工作指出了关键区别：在预先选定的上下文中优化，并不能确定修改后的文档能否通过大规模语料中的检索与重排序。SAGEO Arena 专门评估检索—重排序—生成流程，并报告某一阶段受益的修改，可能损害另一个阶段 [23]。这两种情境回答不同问题，均不应被说成专有平台的通用模型。

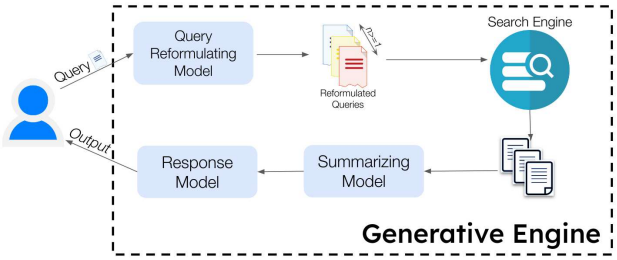


图 1： 生成式引擎的一种基础研究抽象：查询重构、检索、信源摘要和回答生成。

来源说明。取自 Aggarwal et al. [1, 图 2]，PDF 实际第 3 页的矩形矢量裁图，许可为 CC BY 4.0。该图是概念模型，并不披露任何具名生产系统的实现。

定义 1.1: 生成式搜索系统

对于用户状态 u 和查询 q，用随机核表示生成式搜索系统：

$$(Y, \mathcal{A}, L) \sim \mathcal{G}_\theta(\cdot \mid q, u, \mathcal{W}_t)\,, \tag{1}$$

其中，$\mathcal{W}_t$ 为时刻 t 可访问的信息环境，θ 为协议能够识别或记录版本的系统状态，Y 为生成回答，$\mathcal{A}$ 为界面显示的归因或引证集合，L 为任何可观察的交互轨迹。一次实现的观察记为 $(y, \dashv, \ell)$。该定义不假设每个系统都会搜索开放网页、展示引证或只调用一次检索。

显式写出时间参数很重要。网页文档会变化，索引会滞后，模型版本会更新，平台政策也会变化。一次回答不是有关品牌或信源的永恒事实，而是特定系统状态下的一

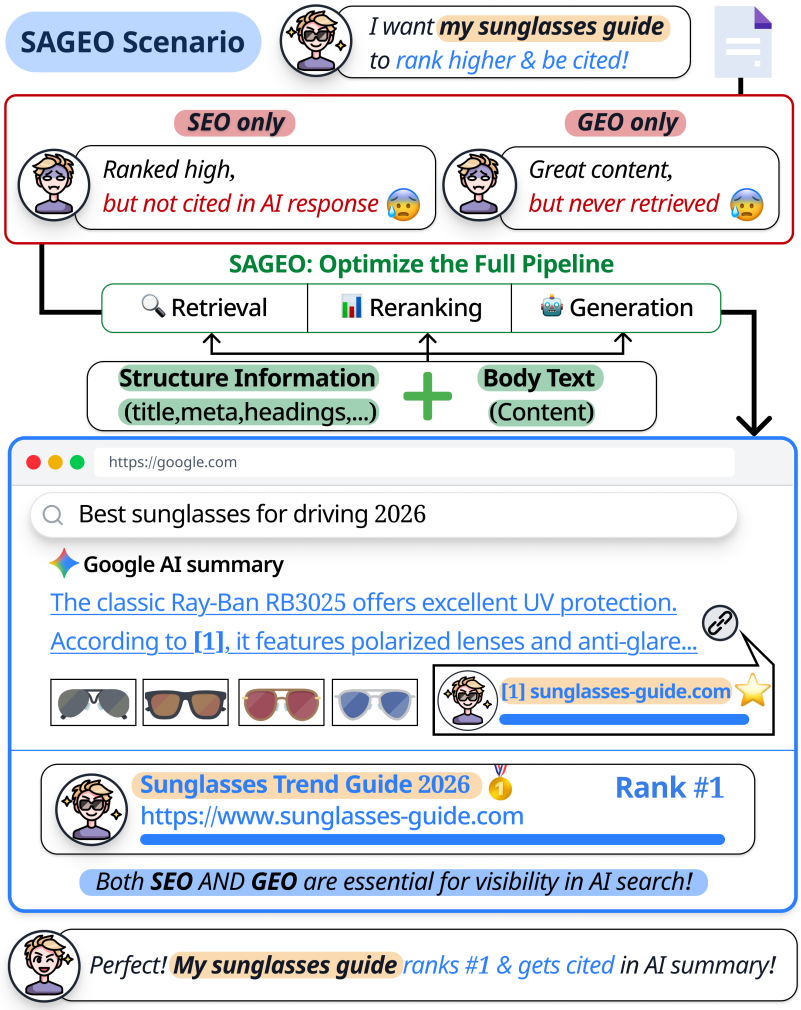


图 2: 重建的搜索增强 GEO 环境中的阶段错配：只优化排序可能在生成阶段失败，只优化生成也可能在检索阶段之前失败。

来源说明。取自 Kim et al. [23, 图 1]，PDF 实际第 2 页的矩形裁图，许可为 CC BY 4.0。它概括了该基准的分阶段测试环境，不代表商业平台架构，也不是普遍性能规律。

次观察。

1.3 贯穿案例：兰湾仪器

本书以一个合成组织为线索，将抽象流程与可检查的任务联系起来。兰湾仪器（*Orchid Bay Instruments*）是一家虚构的便携式水质光谱仪制造商。其受治理的信源集合包含产品规格、独立出具的校准报告和安全通告。采购分析人员提出如下问题：

> 当样本浓度可能超过 40 mg/L，且设备必须在低于 5 °C 的环境下运行时，哪款便携式光谱仪适合现场硝酸盐筛查？

这个案例有意设置了难点。产品页面宣传的标称测量范围为 0–50 mg/L，校准报告只覆盖 10–25 °C 下的 5–35 mg/L，而安全通告要求低于 5 °C 时使用加温外罩。如果流畅的答案仅凭产品页推荐设备，它可能正确提到标称范围，却省略温度条件，并超出独立证据验证的范围。因此，可见性、事实忠实性、信源选择和决策效用可能向不

表 1: 合成贯穿案例中的受控信源对象。数值均为教学输入，不是对现实组织或平台的主张。

受控对象	证据范围	在后续章节中的用途
产品页面	标称范围 0–50 mg/L；没有低温验证。	候选生成与主张切分。
校准报告	独立验证范围为 10–25 °C 下的 5–35 mg/L；列明协议与不确定性。	在受测范围内提供证据支持。
现行安全通告	低于 5 °C 时须使用加温外罩；本通告取代上市时的使用手册。	限定、冲突与替代关系检验。

表 2: 不同可观察对象需要不同的单位与解释。

可观察对象	单位	正事件意味着	不能据此确立
提及	实体—回答	实体字符串或已消歧别名出现	信源使用、准确性或说服效果
引证	信源—回答	可见归因能够解析到该信源	答案忠实使用了某项特定主张
吸收	主张—信源—回答	信源特有证据反映在答案中	正确归因或业务价值
忠实性	主张—回答	答案保留了限定范围的命题	哪个信源或流程阶段导致了这一结果
显著程度	片段—回答	实体或信源具有指定的位置或份额	正面情感或用户注意力
行动	用户—任务	预先声明的行为发生	在没有识别设计时解释其发生原因

同方向变化。

后续章节将记录这些对象的版本，检索其中的段落，分配上下文预算，评分主张—引证支持关系，设计重复测量面板，并评估可逆的证据呈现干预。该案例的任何结果都不是有关生产系统的实证证据。

1.4 经常被混为一谈的单位

精确的 GEO 研究至少区分六种对象：

实体

用户所询问的组织、产品、人物、地点或概念。

主张

有关实体的最小命题，具有时间、地区及其他范围条件。

信源

能够支持或质疑主张的受治理信息对象，例如报告、产品页、数据集或监管记录。

段落

从信源切分得到的可检索片段。切分方式会改变排序器或生成器可获得的证据。

答案

向用户展示的生成语言，包括不确定性表述、顺序和建议。

结果

预先声明的后果：提及、引证、事实吸收、正确描述、访问、合格线索、购买或其他行动。

实体可能被提及却没有引证；信源可能获得引证，但其独有证据没有进入答案；主张可能被吸收，却归给错误信源；品牌可能得到醒目却事实有害的提及；用户也可能未访问任何引用页面就采取行动。因此，在协议同时定义事件及其价值之前，“AI 可见性”并不是一个标量。

近期测量文献进一步强调了这种区分。有研究明确区分引证选择与引证吸收：平台可能在搜索层选择某网页，却不让它以相同程度影响最终答案 [65]。关于引证失败的工作也将未获引证视为分布于整条流程的一组失败，而非单一内容质量缺陷 [47]。这些论文为有用的概念与协议提供了证据，但其中的平台比较仍受各自数据集、日期和访问方法的约束。

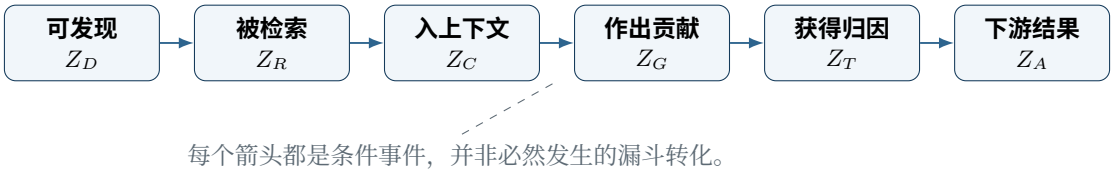


图 3: 条件可见性链。这是原创的概念分解；具体平台可能合并、重复或省略可观察阶段。

1.5 条件可见性链

设 s 为受治理信源，q 为从预先声明的分布 Q 中抽取的查询。定义二元或分级阶段变量：

$$
\begin{aligned}
Z_D &: \text{信源可以被发现并抓取},\\
Z_R &: \text{信源或其某个段落被检索到},\\
Z_C &: \text{信源进入有效生成上下文},\\
Z_G &: \text{信源对生成内容作出贡献},\\
Z_T &: \text{该贡献获得忠实归因},\\
Z_A &: \text{回答支持预先声明的下游结果}.
\end{aligned}
$$

如果某个结果要求六个阶段全部通过，链式法则给出

$$
\begin{aligned}
\Pr(Z_D, Z_R, Z_C, Z_G, Z_T, Z_A \mid q, s) = \Pr(Z_D \mid q, s)\\
\cdot \Pr(Z_R \mid Z_D, q, s)\\
\cdot \Pr(Z_C \mid Z_R, Z_D, q, s)\\
\cdot \Pr(Z_G \mid Z_C, Z_R, Z_D, q, s)\\
\cdot \Pr(Z_T \mid Z_G, Z_C, Z_R, Z_D, q, s)\\
\cdot \Pr(Z_A \mid Z_T, Z_G, Z_C, Z_R, Z_D, q, s). \qquad (2)
\end{aligned}
$$

这里不需要独立性假设。恰恰相反，相依性才是核心问题：一次改写可能提高某个条件概率，同时降低另一个。

这条链是诊断工具，不是在声称封闭系统实现了六个同名模块。当检索不可观察时，引证或许可以代理某些事件，但报告必须明确称其为代理。当平台执行多次搜索时，Z_R 可能是一组序列，而非单次事件。如果生成器输出未引证的参数知识，Z_G 可以发生，却不显示 Z_T。

1.5.1 阶段留存审查算例

考虑兰湾案例中一个含 240 个冻结“查询—系统—时间”单元的教学审查。采集协议在开放索引中直接观察获取过程，在受控系统中记录检索段落与上下文段落，并使用双人编码的主张和归因标签。表 3 中的嵌套计数是合成数据，其目的在于让目标估计量和算术可复现，而不暗示任何具名平台的结果。

观察到的条件比率连乘后相互约去：

$$
\frac{216}{240}\frac{158}{216}\frac{91}{158}\frac{58}{91}\frac{42}{58}\frac{11}{42} = \frac{11}{240} \approx 0.0458. \qquad (3)
$$

这个等式是计数关系，不是独立性论证；它也不能识别哪个内部模块造成损失。最大计数降幅出现在检索到有效上下文之间（158 降至 91），但这种模式可能源于相关性、冗余、上下文预算、安全过滤或阶段标签错误。诊断必须

表 3: 分阶段审查中的合成嵌套计数。条件留存率以前一个已观察阶段为分母；边际留存率始终以全部 240 个单元为分母。

阶段	通过单元数	条件留存率	边际留存率
可发现 Z_D	216	$216/240 = 0.900$	0.900
被检索 Z_R	158	$158/216 = 0.731$	0.658
入上下文 Z_C	91	$91/158 = 0.576$	0.379
作出贡献 Z_G	58	$58/91 = 0.637$	0.242
获得归因 Z_T	42	$42/58 = 0.724$	0.175
下游结果 Z_A	11	$11/42 = 0.262$	0.046

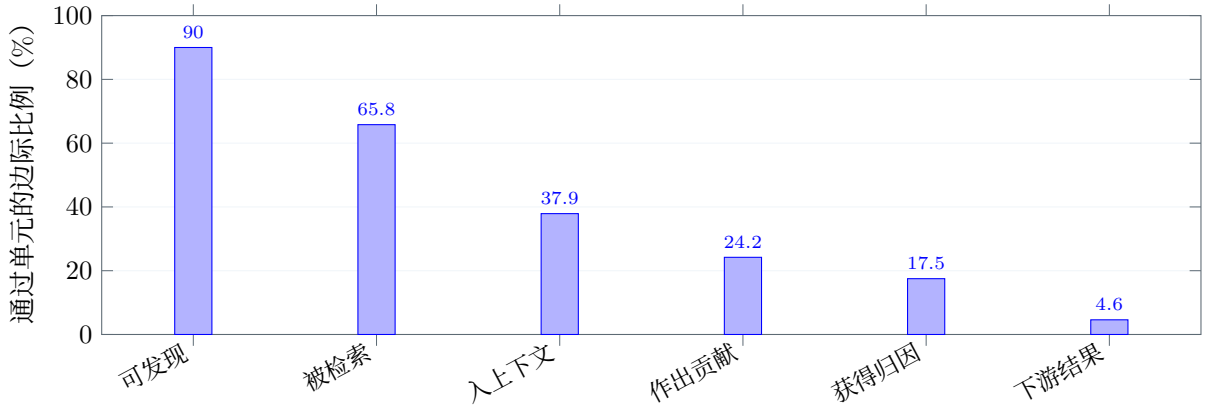


图 4: 表 3 合成审查中的阶段留存。该图直接从已声明计数生成，是教学算例而非生产平台估计。

回到单元级轨迹。

命题 1.2: 观察等价性限制机制主张

设两个候选系统状态 θ_1 和 θ_2，对于所有受测 $(q, u, \mathcal{W}_t)$，均使审查可观察量 $(Y, \mathcal{A}, L)$ 具有相同的联合分布。那么，仅关于这些可观察量可测的任何决策规则，在 θ_1 和 θ_2 下都具有相同分布。因此，该审查不能识别究竟由哪种候选内部机制生成了观察。

证明. 令 $d(Y, \mathcal{A}, L)$ 为任意可观察决策规则，B 为其输出空间中的可测集。可观察量的联合分布相同，意味着

$$\Pr_{\theta_1}(d(Y, \mathcal{A}, L) \in B) = \Pr_{\theta_2}(d(Y, \mathcal{A}, L) \in B)\,.$$

因此，任何仅由审查可观察量计算的统计量，都不能区分两种候选机制。识别需要额外的可观察量、干预或假设，使其诱导的分布有所不同。 □

1.5.2 从标量到可见性剖面

对于信源 s、查询面板 Q、系统 m 和时间窗口 T，定义可见性剖面

$$\mathbf{V}(s; Q, m, T) = \left[\widehat{p}_{\text{ret}},\ \widehat{p}_{\text{cite}},\ \widehat{p}_{\text{abs}},\ \widehat{a}_{\text{attr}},\ \widehat{f}_{\text{claim}},\ \widehat{\sigma}_{\text{time}}\right], \tag{4}$$

帽号强调它们是依赖协议的估计：依次表示可观察时的检索发生率、引证发生率、信源特有内容吸收、归因准确率、主张忠实性和时间变异性。必须分别声明其估计方法、分母、缺失处理规则和区间。这个诊断剖面不是要把不可兼容的量平均成专有“GEO 分数”，而是提醒我们：结果可能在某个维度改善，同时在另一个维度退化。

表 4: 以问题为中心比较相邻实践。

实践	主要可观察对象	典型干预对象	与 GEO 相关的局限
SEO	索引资格、排序链接、点击	可抓取性、信息架构、内容、链接	排名靠前的页面未必进入或影响生成答案
AEO	直接答案资格与抽取	简洁答案结构、实体、结构化数据	抽取本身不保证多信源归因或忠实性
RAG 工程	受控语料中的检索与有据生成	分块、嵌入、检索器、重排序器、提示、模型	结果未必迁移到网页规模或专有生态
数字公关	第三方发布与声誉	新闻价值、专家信源、分发、关系	报道频次不等于检索、引证或因果业务影响
GEO	生成式搜索流程中已声明的事件	信源可用性、证据、表达、生态、评估	机制常常只能部分观察，优化可能产生操纵激励

生成回答随重复次数与时间变化时，静态测量尤其容易误导。近期工作主张把可见性视为分布，而非单次取值[40]。章 6 将展开相应的抽样和不确定性协议。

1.6 GEO、SEO、AEO 与相邻领域

GEO 继承了网页搜索的基础设施，却改变了可观察输出和策略环境。表 4 中的类别应视为相互交叠的问题表述，而非边界固定的职业领地。

连接搜索服务与大语言模型的文献，涵盖查询理解、检索、排序、数据标注和系统评估，范围远超 GEO 这一营销标签 [58]。因此，严谨的课程不能从内容技巧起步，而必须从信息检索、实验设计和信源治理起步。

1.7 三类利益相关者与错位目标

设 U_u 表示用户效用，U_p 表示平台效用，U_s 表示信源或供应方效用。创作者侧干预可能最大化预期曝光 $\mathbb{E}[U_s]$，却降低答案质量或归因准确性。平台防御也可能维护 U_p，却不公平地压制合法的新信源。因此，有用的目标必须受到约束：

$$\max_{\delta\in\Delta} \quad \mathbb{E}\big[U_s(Y^\delta,\mathcal{A}^\delta,L^\delta)\big] \quad \text{subject to} \quad \begin{cases} \mathrm{Fidelity}(s\oplus\delta)\geq\tau_f, \\ \mathrm{EvidenceCoverage}(s\oplus\delta)\geq\tau_e, \\ \mathrm{Risk}(s\oplus\delta)\leq\tau_r. \end{cases} \tag{5}$$

这里，δ 为允许的信源变更，且 $(Y^\delta,\mathcal{A}^\delta,L^\delta)\sim\mathcal{G}_\theta(\cdot\mid q,u,\mathcal{W}_t^{(s,\delta)})$，其中 $\mathcal{W}_t^{(s,\delta)}$ 是信源 s 已应用受治理修订的信息环境。忠实性、证据覆盖度和风险阈值是治理决定，而非自动发现的模型参数。

这种多方视角在研究中日益明确。关注风险的工作将 GEO 视为可能导致排序操纵和未披露影响的途径 [53]。近期机制设计工作则建模供应方与平台之间的反复适应，并评估激励可验证内容的方法，而不只依靠惩罚性过滤 [59]。后者的实证结果属于其声明的仿真和基准；对本讲义而言，更广泛的启发是从一次性技巧转向反复演化的生态动态。

适用边界与失效条件

基准中存在统计显著效应，并不意味着可以声称同样修改会改善某个具名生产平台。迁移至少要求兼容的检索机制、候选集合、查询分布、答案界面、模型状态、指标和时间窗口。

1.8 分阶段审查协议

算法 1 将本章区分转化为最小可执行协议：无法获得的阶段记录为未观察，绝不由下游成功反推，也不编码为零。后续章节将展开其中的获取、检索、标注和重复测量步骤。

算法 1 分阶段可见性审查

输入： 冻结的信源与查询清单；声明的审查单元 $\mathcal{C}$；阶段观察、缺失和停止规则。

输出： 单元台账 E；阶段估计与区间；偏离日志和机制不确定性日志。

冻结信源哈希、查询编号、系统卡、结果变量和分析计划。
按抽样设计构造笛卡尔积审查单元 $\mathcal{C}$；保留查询聚类和时间聚类。
对所有 $c \in \mathcal{C}$ **执行**
    采集回答、归因、轨迹、时间、系统状态和采集状态。
    对所有 $Z \in \{Z_D, Z_R, Z_C, Z_G, Z_T, Z_A\}$ **执行**
        若 按照冻结规则直接观察到 Z **则**
            记录取值、定位信息、编码员或仪器，以及裁定状态。
        否则 若 存在预先声明的 Z 代理 **则**
            单独记录代理，并保留其验证边界。
        否则
            记录 $Z = \textsc{Unobserved}$；绝不根据下游事件推断为零。
        结束 若
    结束 对
    区分拒答、服务中断、删失、解析失败与界面缺失。
结束 对
对信源去重；裁定双人编码样本；保留分歧和排除记录。
按各事件的单位、分母、权重、聚类、区间和敏感性分析分别估计。
发布 E、清单、表格与日志；触发严重证据、忠实性或安全条件时停止。

1.9 最小研究问题模板

本讲义中的每项实证主张，都可以改写为以下模板：

> 对于哪一类信源总体和哪种查询分布，在哪个系统版本和界面下，干预 δ 相对于比较对象 c 是否改变结果 Y？观察覆盖哪个时间窗口，存在多大的不确定性，并受到哪些忠实性与风险约束？

如果一句话无法填满这些位置，它仍可能是有用的观察或假设，却还不是足以支持决策的效应主张。

复现实作 1 · 分解一项可见性主张

输入。 5 项有关改善“AI 可见性”的公开主张，至少分别来自 1 篇论文、1 份平台文档、1 份厂商白皮书、1 篇博客和 1 段视频。

任务。 逐项识别信源对象、干预、候选流程阶段、可观察结果、比较对象、查询分布、系统状态和证据层次。绘制条件链，标记所有未观察的转移。

交付。 一页主张登记表，以及不超出证据支持范围的逐项修订表述。

停止规则。 当结果中的每个名词均被操作化，且每个因果动词都有识别论证、或已被替换为观察性动词时，结束任务。

本章小结

1. 生成答案使可见性成为一系列条件事件，而非消除经典搜索流程。
2. 提及、引证、吸收、忠实性、显著程度和行动，是单位各异的不同结果。
3. 没有额外的可观察量、干预或识别假设，无法用机制主张区分观察等价的系统。
4. 固定上下文基准与端到端基准回答不同问题。
5. 分阶段审查区分代理与未观察状态，报告可见性剖面而非不明含义的标量。
6. 负责任的 GEO 目标用证据、忠实性和风险约束曝光。下一章将说明，主张与信源如何获得进入该目标的资格。

2 先建立证据，再开展优化

核心思想

GEO 项目的第一项交付成果不是优化后的文案，而是一张带有版本记录的主张—证据—信源映射图，以及每个信源能够和不能够证明什么的说明。在这张图建立之前就开始优化，只会把不确定性表达得更加流畅。

2.1 文献有身份，主张有上限

论文、标准、官方文档、代码仓库、白皮书、博客、视频和截图，并不是可以互换的事实容器。每一种材料都有发布主体、版本、方法、预期受众和变更流程。标准可以规定要求，却不证明干预效果；代码仓库可以展示实现，却不确立外部有效性；厂商案例可以记录客户经历，却不交代分母、比较对象和失败案例。

因此，审查应分为两步：**文献身份审查**确认材料究竟是什么，**主张审查**确认材料能够支持哪些命题，或者是否根本不能支持相关命题。

定义 2.1: 信源身份记录

信源身份记录定义为元组

$$I(s) = \langle \text{issuer}, \text{type}, \text{title}, \text{version}, \text{date}, \text{status}, \text{method}, \text{artifacts}, \text{license}, \text{URL} \rangle. \tag{6}$$

上述字段依次为发布主体、类型、题名、版本、日期、状态、方法、成果文件、许可和网址。未知字段必须明确保留为未知。精美版式、机构标志或 DOI，都不能填补缺失的方法或成果文件信息。

2.1.1 按用途而非声望划分证据层级

表 5 按材料通常能够支持的主张组织证据。这些层级不是学术价值或智识贡献的排名。要确认产品的当前设置，产品帮助页面可能是最合适的信源；要估计处理效应，随机研究可能最合适。两者不能彼此替代。

例如，最初的 GEO 基准为其指定内容修改与可见性指标提供了实验环境内的受控证据 [1]。CC-GSEO-Bench 等端到端基准，补充了检索语境中以内容为中心的信源影响视角 [6]；SAGEO Arena 则加入了大规模语料检索与重排序 [23]。任何一个信源都不能单独揭示所有商业答案引擎的因果机制。

2.2 主张—证据—信源图

营销素材库通常按活动、渠道或格式组织文件。生成式搜索项目需要另一种基本单位：组织愿意且有能力捍卫的最小命题。

定义 2.2: 限定范围的主张

限定范围的主张定义为

$$c = \langle h, r, o, \tau, \lambda, \kappa, q \rangle, \tag{7}$$

其中，h 为主语实体，r 为关系或谓词，o 为宾语或取值，τ 为有效时间区间，λ 为地区，κ 为其他适用条件，q 为主张状态，例如拟议、已批准、有争议、已过期或已撤回。

证据项 e 不只是一个文件，而是带有溯源信息的相关片段、表格、数据集、测量、记录或观察。信源 s 包含或

表 5: 证据层次及其通常能够支持的主张上限。

层次	典型材料	通常能够支持	常见越界推断
E0	宣传文案、无引用文章、社交短视频	某一主体公开作出了什么主张	把重复出现或传播范围视为验证
E1	官方文档、政策、标准	已声明的行为、状态、要求与术语	推断隐藏的排序权重或现实效果
E2	方法详尽的白皮书、教程、代码讲解	可复现的程序或实现选择	推广到已披露数据和系统之外
E3	基准研究或观察性研究	指定协议下的关联与性能	将基准差值解释为平台因果效应
E4	含比较对象的受控干预	实验环境中的效应	忽略干扰、漂移或外部有效性
E5	独立的多情境复现或有力的现场设计	更具可迁移性的效应估计	把异质性效应称为普遍规律

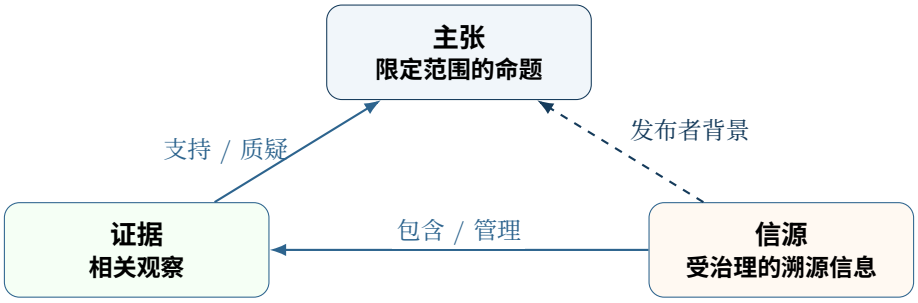


图 5: 主张—证据—信源图。支持关系具有类型、范围与版本；一个 URL 本身不是证据边。

管理该证据。我们将支持关系表示为有类型的边：

$$e \xrightarrow[\rho,\iota,\phi]{\text{supports}} c, \tag{8}$$

其中，ρ 表示与准确主张的相关性，ι 表示相对于主张提出者的独立性，ϕ 表示相对于主张时间范围的新近性。矛盾、限定和不提供信息的边，与支持边具有同等地位；不能为了让看板显示全绿而将它们删除。

2.2.1 最低证据充分性

设 $E(c)$ 为与主张 c 关联的证据集合。实用的充分性函数可以写为

$$\begin{aligned} A(c) = \min \big\{ &\operatorname{Coverage}(E(c)), \operatorname{Relevance}(E(c), c), \\ &\operatorname{Freshness}(E(c), c), \operatorname{Authority}(E(c)), \operatorname{Consistency}(E(c)) \big\}. \end{aligned} \tag{9}$$

这里分别考虑覆盖度、相关性、新近性、权威性和一致性。取最小值表达的是一条治理原则：即使来源组织得再完善，如果所有信源均已过期，主张仍不可用；即使信源很新，如果它讨论的是更宽泛或不同的命题，也仍可能无用。这个函数是发布判定条件，并非经过科学标定的通用评分。

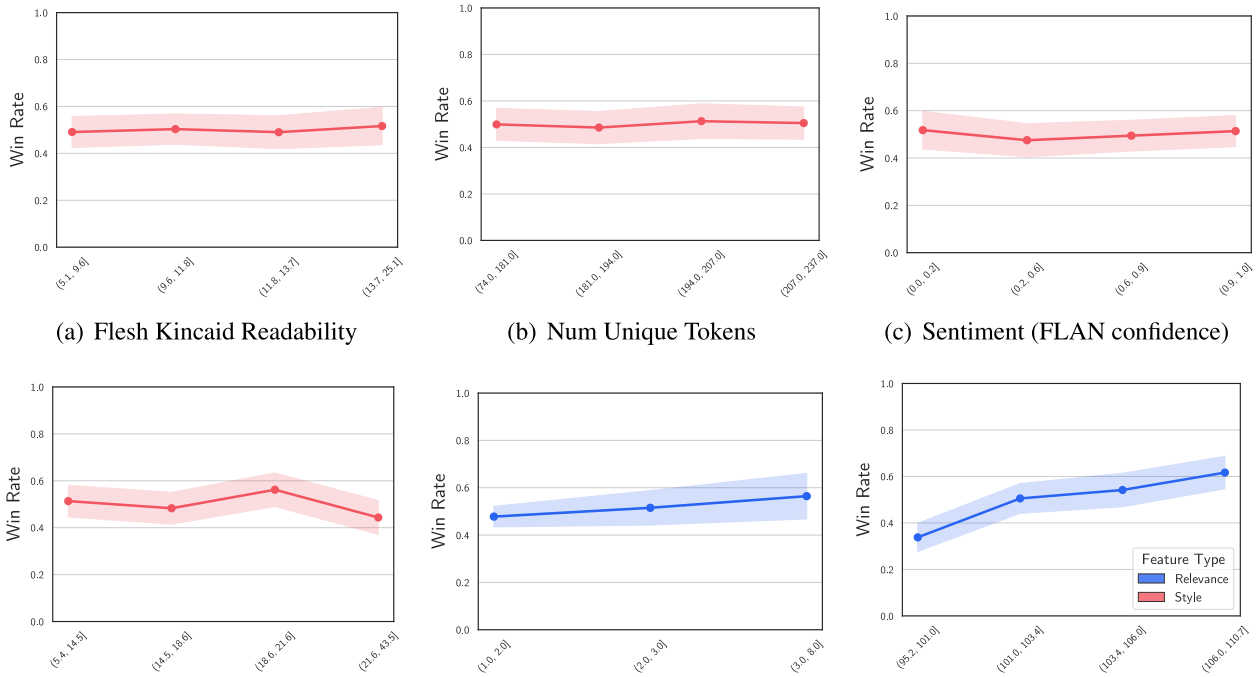
2.3 保留层次地阅读 GEO 实证论文

论文记录至少应包含四个层次：

1. **身份：** 题名、作者、发表场所或预印本状态、版本、日期、DOI、代码与数据链接，以及许可。

表 6: 不同主张所需最低证据的示例。

主张类型	最低内部证据	更有力的外部支持
产品能力	带版本的规格、测试条件、责任人批准	独立测试或现行用户文档
性能比较	协议、基线、数据集、不确定性、失败案例	在相关任务上的独立复现
市场地位	明确定义的市场、时期、地域、分母	可追溯的第三方数据集与方法
客户效果	获授权的案例记录、处理时间线、指标定义	比较对象或能处理替代原因的设计
认证资格	证书编号、范围、签发方、有效日期	签发方登记名录或签名验证
安全或合规	控制措施说明、责任人、审计状态	适用标准映射与独立评估



(a) Flesh Kincaid Readability (b) Num Unique Tokens (c) Sentiment (FLAN confidence)

(d) Perplexity (GPT-2 Medium) (e) n-gram Overlap w/ Question (f) Question-Para Embedding Sim.

图 6: 固定冲突证据情境中，按特征分箱的段落胜率。相关性特征的变化比所测风格特征更明显；这些图展示的是描述性关联，而非因果规则。

来源说明。取自 Wan et al. [50, 图 4] 的矢量裁图，原载 ACL Anthology 官方 PDF 第 7 页，许可为 CC BY 4.0。仅裁去外围版面，并将 6 个 Type 3 字体程序转换为矢量轮廓；标记、数值、颜色和几何关系保持不变。该图给出 LLaMA-2-Chat-13B 在 242 个筛选样例上的结果及 95% 区间，并不建立有关说服力、真实性、检索或引证的通用配方。

2. **设计：** 研究问题、单位、抽样框、系统、查询分布、候选语料、干预、比较对象和指标。
3. **结果：** 目标估计量、点估计、不确定性、异质性、负面结果和失败案例。
4. **迁移：** 哪些组成部分与目标部署相似，哪些并不相似。

基础 GEO 论文中的“最高 40%”，是其基准和指标下报告的最大改进幅度，并非网站普遍可期待的提升 [1]。正确的编辑处理不是压下醒目的数字，而是补回它对应的结果、分母、比较对象、方法与领域之间的差异，以及实验环境。

同样，某项多平台研究观察到一个排版特征未改善内容吸收，恰恰因为这是已披露数据流程下的负面结果而具有价值 [65]。它并不证明排版与检索、可读性、无障碍或其他所有系统无关。负面结果收窄主张，不会抹去相邻机制。

表 7: 非论文材料如何进入以证据为先的研究讲义。

材料	合理用途	必须审查的字段
平台官方文档	当前资格条件、语法、政策或受支持功能	责任方、界面、地区、版本/日期、存档 URL、精确范围
国家或国际标准	规范性术语、流程、控制措施或评估要求	发布机构、标准号、状态、版本、生效日期、适用条款
GitHub 教程或仓库	可复现实现、数据模式、测试样例	提交哈希、许可、环境、数据溯源、测试结果
行业博客或社区教程	提出假设、流程示例、术语、新兴实践	作者/组织、利益冲突、日期、证据链接、主张登记表
白皮书	已披露方法、汇总规律、运营框架	资助方、样本、系统、日期、原始单位、分析、成果文件、缺失情况
视频或网络研讨会	可观察演示和带时间索引的发言者主张	稳定 URL、时长、时间戳、发言者角色、转写、主张类型、后续核查

证据说明

首版 42 条记录的论文库较新，因此预印本占比较高。时效性有助于描绘新兴领域，却也更加要求跟踪版本变化、同行评审状态、成果文件可用性，以及作者或基准之间的重叠。进入本讲义表格或图形的每个数值，都必须对照所附 PDF 的具体版本核查，不能只依据摘要或博客概述。

2.4 GEO-42 快照

冻结导出文件的基础完整性较好：42 条记录均有 DOI 和本地 PDF；41 条同时具有 URL 和摘要。未发现规范化题名、DOI 或 URL 重复。5 条记录需要编辑关注：

1. 4 条记录的出版年份与 arXiv 标识符编码的年份不一致；
2. 1 条期刊记录在导出文件中缺少 URL 和摘要。

因此，最终论文索引分别保留三个字段：来源记录年份、*arXiv* 标识符年份和已核实发表状态。将它们合并为一个无声修正过的“年份”，会让目录更易展示，却更难审计。

自动初步分类也有一个已知局限：当题名使用 GEO 这一领域名称时，关键词规则容易将论文过度归入“优化”，即使论文的实质贡献在于测量、系统设计、文化差异或治理。因此，附录将自动分类视为待检验的阅读分流建议。只有读过摘要、方法、结果和局限后，才能最终确定章节位置。

2.5 研究与实践交叉领域中的信源类型

视频主张尤其需要谨慎，因为解释与演示往往交织在一起。转写中的每一行应标记为观察、解释、假设、预测或建议。只有第一类能由屏幕所呈现的内容直接建立，即使如此，演示环境也可能经过预先布置。博客和社区教程同样如此：它们有助于发现工作流程和术语，但事实证明的责任由其链接的原始信源承担。

2.6 图形是具有视觉语法的主张

图形夸大证据的效率，可能比段落更高。本讲义的视觉材料遵循三条规则。

1. **机制图标明可观察性**。实线边框表示已观察组成部分，虚线边框表示假设或隐藏阶段。不借助平台配色或标志暗示架构已获核实。
2. **数据图保留实验条件**。图注写明单位、系统、查询集、时间、干预、比较对象、不确定性和来源。坐标轴起点不能仅为夸大小效应而选择。

3. **外部图片必须有许可记录**。无法确认允许复用时，本稿引用结果，并以原创概念图或基于获许可数据的可复现图替代。截图不是重建后的数据。

用户提供的 arXiv:2506.02070v3 TeX 源码，以定义、核心思想、示例、评注、算法、插图和小结形成了出色的教学节奏。由于该来源采用 CC BY-NC-ND 4.0 许可，本讲义仅研究其源码架构与教学功能。本书的 GEO 正文、示意图和教学说明环境均为独立撰写与设计。

复现实作 2 · 建立主张登记表

输入。某组织的主页、产品文档、2 个第三方信源、1 份白皮书、1 篇教程和 3 篇论文。
任务。提取 25 条明示或暗示的主张。将每条主张规范化为主体、谓词、客体、时间、地区、条件和状态。关联证据片段与信源身份，纳入矛盾证据和过期证据。确定证据支持上限，并写出能够补齐每个关键缺口、成本最低且合乎伦理的行动。
交付。可由机器读取的 CES 表、冲突报告，以及每条面向公众的主张是否可以发布的决定。
停止规则。只有在每条获批主张均有当前有效的证据路径，且每条未获批主张均有责任人、原因和下次复核日期时，才结束任务。不能仅因为每行都有一个 URL 就停止。

本章小结

1. 信源身份与主张支持需要分别审查。
2. 证据层次约束信源能够建立的结论，不按声望给信源排名。
3. 主张—证据—信源图为内容工程提供受治理的事实基础。
4. 冻结的 42 篇论文导出文件是保留异常记录的版本化研究库，而非一堆可直接套用的结论。
5. 视觉设计必须保留不确定性、可观察性和许可信息。

3 网页获取、表征与查询规划

核心思想

内容不会从网页直接进入答案，而要经过变化中的信息环境、一个或多个查询、检索与选择、有限上下文、生成和归因。恰当的干预取决于预期证据最早在哪个阶段丢失。

3.1 参考流程，而非平台蓝图

图 7 是一种研究抽象，汇集了搜索、检索增强生成和 GEO 文献中反复出现的阶段 [1, 23, 58]。具体产品可能合并阶段、重复执行阶段，从获许可资料库而非开放网页检索，或在没有新搜索的情况下生成答案。这个图的价值在于，每个阶段对应不同失效模式和不同类型的可观察证据。

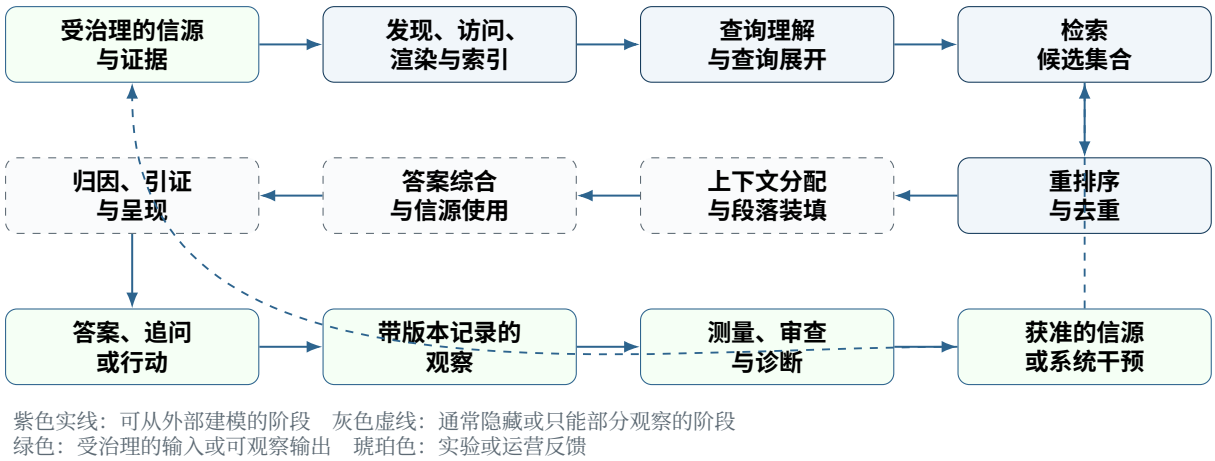


图 7: 用于诊断与评估的端到端参考流程，结构为原创。不得将虚线阶段描述为已核实的封闭平台内部机制。

3.2 阶段 0：受治理信源与可访问信息世界

用 $\mathcal{W}_t$ 表示系统在时刻 t 能够访问的信息环境。它不等同于公开网页。访问可能取决于爬虫政策、robots 指令、身份验证、渲染、许可、新近性、地理位置或产品专有索引。只有相关表征 $\pi(s)$ 能够被获取和解释，信源 s 才满足资格条件：

$$\text{Eligible}(s,t) = \text{Discoverable}(s,t) \wedge \text{Fetchable}(s,t) \wedge \text{Renderable}(s,t) \wedge \text{Indexable}(\pi(s),t). \tag{10}$$

这一分解可避免常见诊断错误。如果客户端页面给不执行渲染的抓取器返回空白初始文档，改写可见正文无法解决获取问题。如果规范网址规则合并到了错误的 URL，继续发布重复页面可能增加歧义。反过来，抓取成功也不证明已被索引、检索或使用。

信源层也是权限发挥作用的地方。组织不能一面要求所有智能体不得抓取信源，一面又声称未被引证就证明存在排序惩罚。访问政策、训练政策、搜索获取和用户触发检索可能分别受不同控制措施约束，必须逐项审查。

3.3 阶段 1：查询理解与查询展开

自然语言请求 q 可能被变换为一组检索查询

$$\mathcal{Q}(q,u)=\{\hat{q}_1,\ldots,\hat{q}_K\}, \tag{11}$$

其条件包括用户状态 u、对话历史、语言、地区和系统计划。采购问题可能展开为需求、替代方案、评测、价格、风险和近期变化等子查询。更多展开并不自动意味着更好：每个查询都会消耗时间和成本，也可能引入无关或低质量证据。

封闭产品很少暴露完整的 $\mathcal{Q}$，因此研究者应维护两列：已观察调用和重建假设。网络轨迹、界面显示的搜索词或开放智能体日志可填入前者；分析人员推想的合理查询树属于后者。混淆二者，就会把解释性图示变成虚构遥测。

3.3.1 查询扩展的停止规则

设 E_k 为执行 k 个子查询后的去重证据覆盖量，C_k 为其成本。一种实用的停止条件是

$$\Delta E_{k+1}<\epsilon_E \quad \text{or} \quad \frac{\Delta E_{k+1}}{\Delta C_{k+1}}<\tau, \tag{12}$$

同时还要满足对反面证据和信源多样性的明确要求。阈值属于协议选择，不能事后为迎合偏好的答案而设定。

3.3.2 受控轨迹与自然发现之间的缺口

两类证据回答不同问题。在 EcoGEO 的受控轨迹研究中，一个合成目标结果被插入 9 个开放网页结果的第 5 位，此后 GPT-5.1 搜索智能体可以在有限工具预算内发起查询并抓取其他页面。在报告的三个任务集中，协同 TRACE 条件下的目标推荐率分别为 67.2%、71.9% 和 73.9% [60]。该设计表明，入口页面、链接证据以及智能体的后续搜索，能够改变受控浏览轨迹。由于曝光是人为注入而非自然获得的，它没有估计自然抓取、索引、自然排名，或此前未知商业页面被发现的概率。

真实产品界面的面板能够提供生态观察，但对机制的访问更弱。在一个德语、四垂直领域面板中，同一时间窗口内重复运行所得信源集合的平均 Jaccard 相似度介于 0.321 与 0.434 之间；相邻两天的值介于 0.336 与 0.423 之间 [40]。这些数值支持在该面板中重复抽样，却不能将解码随机性与检索、路由、缓存或产品更新分离，也不能规定普遍适用的重复次数。

适用边界与失效条件

本地证据基础无法识别：一个刚发布、此前未被索引的页面，自然被商业生成式搜索界面发现的因果概率是多少。受控轨迹实验与真实黑箱观察从两侧逼近这一问题，却都没有填补该缺口。

3.4 阶段 2：检索与候选构造

对于查询 $\hat{q}$，检索器给段落或文档 $d\in\mathcal{W}_t$ 赋分：

$$r(d,\hat{q})=\alpha r_{\text{lex}}(d,\hat{q})+(1-\alpha)r_{\text{dense}}(d,\hat{q})+\gamma r_{\text{prior}}(d,t). \tag{13}$$

这些项可分别表示词法匹配、嵌入相似度，以及新近性或信源质量等允许使用的先验。候选集合为 $\mathcal{D}_{\hat{q}}=\text{TopK}_d\, r(d,\hat{q})$。这个方程是通用实验模型，不是逆向推导出的平台评分公式。

检索操作的是表征，而非作者意图。页面可能包含正确事实，却在切分时丢失相关段落。视觉上突出的图表，可能无法进入纯文本表征。简洁句子对某个查询可以被检索，却可能因限定不充分而不适合另一个查询。因此，“事实就在页面上”不足以构成检索诊断。

例题 3.1: 缺失的版本条件

产品页在一个段落中写道“系统支持 128 个项目”，却在远处脚注写道“企业版限制依合同而定”。有关标准套餐的查询检索到了前一个段落，却没得到限定条件。该页面在文档层面事实完整，在段落层面却具有误导性。恰当的干预是把数值、套餐、日期和例外放在一起，而非增加较大数值的重复次数。

3.5 阶段 3：重排序、去重与上下文分配

重排序器估计候选池中与任务有关的顺序：

$$\rho(d, \hat{q}, \mathcal{D}) = \text{Rerank}\left(d, \hat{q}, \mathcal{D}; \theta\right). \tag{14}$$

分数可能取决于其他候选项。因此，当竞争信源进入或离开候选池时，某个信源即使自身未变，排名也可能变化。去重可能合并镜像副本；多样性规则可能为不同域名或观点预留位置。

随后，上下文组装求解一个受约束选择问题。给定段落效用 v_i、词元长度 l_i、冗余惩罚 $g(i,j)$ 和预算 B，可抽象为

$$\max_{x_i \in \{0,1\}} \quad \sum_i x_i v_i - \lambda \sum_{i<j} x_i x_j g(i,j), \tag{15}$$

$$\text{subject to} \quad \sum_i x_i l_i \le B. \tag{16}$$

检索分数高并不保证进入上下文。长段落、重复内容，以及围绕有限预算的竞争，都可能改变所选证据。这也是端到端环境能够提供固定候选实验所无法提供的信息的原因之一 [23]。

3.6 阶段 4：生成、归因与吸收

令 C 表示有效上下文，M 表示生成模型。回答从以下分布采样或解码：

$$y \sim p_M(y \mid q, u, C, \eta), \tag{17}$$

其中，η 概括解码与系统状态变量。模型可能压缩、组合、省略、限定证据，也可能与证据相矛盾。归因可能与答案一同生成，可能在句子生成后附加，也可能从独立的段落元数据解析而来。

因此，以下三种评估彼此不同：

蕴含

被引用的段落是否支持答案片段?

完整性

答案中的重要主张是否获得适当支持?

吸收

哪些信源特有的事实、措辞或结构实际参与了回答?

引证诊断研究提出跨流程的失败类别, 并检验修复方法, 而不是假定存在笼统的“质量”缺陷 [47]。引证选择/吸收框架同样说明，统计显示链接与测量信源影响是不同任务 [65]。生产环境审查中，应抽样将自动蕴含评分与人工判断比对，因为答案切分和模型裁判均可能引入误差。

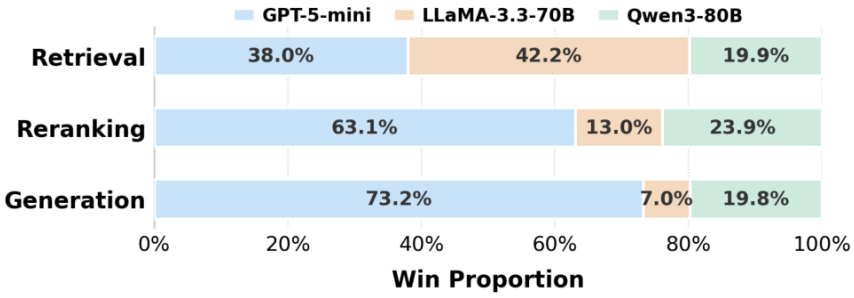


图 8: SAGEO Arena 中三个骨干模型的分阶段获胜比例。检索、重排序与生成阶段的次序会变化；单一“最佳模型”标签将抹去究竟在测量哪个阶段。

来源说明。取自 Kim et al. [23, 图 6]，PDF 实际第 8 页的矩形裁图，许可为 CC BY 4.0。“获胜”指在该重建流程中排名高于另两个骨干模型；图中没有不确定性区间，也不是商业答案引擎之间的比较。

表 8: 分阶段观察与对照的示例。

阶段	直接或接近直接的证据	有用的对照
获取	服务器日志、抓取输出、渲染后 DOM、规范网址及状态检查	允许/禁止，以及静态/渲染页面阳性对照
检索	开放系统 top-k、已展示搜索结果、智能体轨迹	包含已知相关与不相关段落的固定语料
重排序	开放系统评分或有序候选轨迹	交换候选顺序与扰动相关性
上下文	受控系统记录的装填段落	等词元量与去重上下文条件
生成	回答、可得时的种子、模型/版本、提示	重复运行与反事实段落移除
归因	解析后的引证与答案片段	有支持、无支持和有意冲突的段落
行动	经同意采集的事件日志和已声明漏斗	可行时使用留出组或随机曝光

3.7 阶段 5：呈现、交互与行动

即使答案忠实，系统过程仍未结束。引证位置、折叠方式、标签文字、设备视口、追问提示和行动按钮，都会改变用户能够检查的内容。如果研究结果是访问、注册或购买，分析单位便从回答转为用户—任务交互。将该结果归因于信源修改，需要能够处理选择、季节性、并行营销活动和干扰的现场设计。

最后这个阶段也改变了福利问题。系统可能在最大化信源曝光时增加用户负担或降低答案质量；也可能生成准确答案，却提供过少的溯源信息，使纠正变得困难。因此，流程评估必须同时报告信源侧和用户侧结果。

3.8 可观察性与诊断检验

只有阳性对照成功时，负面结果才可解释。如果探针信源在允许条件下都未被抓取，就不能得出“禁止抓取没有效果”；该实验只能判为无结论。如果检索器无法召回唯一匹配的段落，下游答案差异也无法单独识别生成阶段。

复现实作 3 • 构建并审查最小端到端系统

输入。300–1,000 篇文档组成的冻结语料、30 个带相关性判断的查询，以及本地生成器或日志公开的生成器。
任务。比较词法、稠密和混合检索；应用一个重排序器；装填固定上下文预算；生成答案并解析引证。每次运行均保留候选集合、分数、装填段落、回答与归因。加入 3 个阳性对照和 3 项故障注入。
指标。Recall@k、nDCG@k、证据召回、答案正确性、引证蕴含、归因完整性、延迟和成本。不要把它们合并为一个分数。
交付。轨迹包、阶段级错误表，以及说明哪些比较能够识别哪些效应的因果图。

停止规则。当每项下游失败都能追溯到最早观察到的失效阶段，或被明确标为未解决时，结束任务。

本章小结

1. 参考流程是诊断抽象，不是关于封闭产品内部实现的主张。
2. 访问、检索、重排序、上下文、生成、归因和行动需要不同证据与对照。
3. 文档层面的真实信息可能在段落和上下文层面丢失。
4. 引证选择、蕴含、完整性和吸收是彼此独立的评估问题。
5. 要识别干预在哪里有效、在哪里失败，必须保留端到端轨迹。

4 检索、融合、重排序与上下文分配

核心思想

检索质量不是页面孤立具有的属性，而是查询表征、语料快照、候选构造规则、排序目标和有限上下文预算之间的关系。信源即使已被检索，也可能被挤出、重复、截断，或在装填时没有带入支持答案的段落。

4.1 本章回答的问题

读完本章，应能够回答以下五个问题，而不把检索视为生成之前不可见的序幕：

1. 哪个语料、表征与检索单位，定义了 top-k 结果所指的总体？
2. 如何融合词法列表与稠密列表，而不假装原始分数共享量尺？
3. 哪些查询、排名、重复与溯源路径必须在融合后保留，才能避免把重复副本误当作独立证据？
4. 重排序改进能够识别什么？关于答案使用与归因，还有什么未知？
5. 有限上下文预算应如何权衡相关性、重要主张覆盖、冗余与信源依赖？

4.2 声明检索单位与语料状态

设 $\mathcal{D}_t$ 为冻结语料快照，$\Pi(d) = \{p_{d1}, \ldots, p_{dn_d}\}$ 为文档 d 切分得到的段落。检索单位可以是文档、段落、图文对、集合页面或其他表征。这些单位不能互换：文档级相关判断，并不证明支持性段落在渲染和切分后仍可恢复。

每次运行的轨迹至少应记录

$$\mathcal{T}_{\text{ret}} = \langle q, \widehat{q}, \mathcal{D}_t, \Pi, f, k, \theta, \mathbf{r}, t \rangle, \tag{18}$$

其中，q 为用户请求，$\widehat{q}$ 为检索查询，f 为检索器，k 为请求深度，θ 为带版本的配置，$\mathbf{r}$ 为含评分的有序结果。如果没有语料哈希和切分版本，即使模型名称相同，两次运行也未必可比。

定义 4.1: 证据召回率

对于预注册了重要证据单位集合 E_q 的查询，深度 k 处的证据召回率为

$$\text{ER}@k(q) = \frac{|E_q \cap E(R_k(q))|}{|E_q|}, \tag{19}$$

其中，$E(R_k(q))$ 是能够从前 k 个表征中恢复的证据单位集合。当一个文档包含多项独立主张或限定条件时，这比文档召回率更严格。

4.3 区分可观察量、潜变量与假设

已观察量与潜在量的边界取决于具体系统。在受控重建中，候选分数和装填的词元片段可以直接记录；在封闭平台中，相同的量可能隐藏。因此，检索报告应先对每个字段分类，再解释它。

由此得到三个区分。返回的排名，仅对记录日志的那个检索器而言是观察。段落相关性是针对已声明任务的判断，并非段落的内在属性。候选缺席可能意味着获取、表征、匹配或深度不足；没有阶段日志，就无法识别是哪种机制失败。

表 9: 检索研究的可观察性约定。“已观察”指由声明协议记录，不意味着所有平台都暴露该字段。

类别	受控系统中的示例	必要处理
可观察量	查询及其改写；语料和段落哈希；返回标识符、排名、分数、装填位置与截断片段	先保留原始值和时间戳，再计算汇总指标
潜在量	未记录的查询展开；完整索引资格；隐藏候选池；专有分数；未观察到的信源贡献	明确目标估计量或代理，不用可见答案填补缺失阶段
假设	相关性标签有效性；切分充分性；重复阈值；运行期间稳定性；选定候选深度是否足够	预注册、压力测试，并对结论依赖的假设报告敏感性

4.4 先建立可执行基线，再学习融合

BM25 等词法基线检验相关措辞是否在表征中保留下来。稠密检索器检验锁定编码器下的语义接近程度。两者原始分数通常使用不同量尺，因此未经归一化、标定或学习融合规则就相加，并不是中性的基线。基于排名的融合提供了可检查的替代。对于排序列表 $L_1, \dots, L_J$，倒数排名融合赋分为

$$s_{\mathrm{RRF}}(d) = \sum_{j=1}^{J} \frac{\mathbb{1}\{d \in L_j\}}{\kappa + \mathrm{rank}_{L_j}(d)}. \tag{20}$$

常数 κ 是协议的一部分，不是普遍最优值。学习式融合模型需要训练划分、损失函数、特征规范和泄漏审查。

例题 4.2: 为何分数相加会逆转结果

设段落 A 的词法/稠密分数为 $(12, 0.62)$，段落 B 为 $(10, 0.81)$。直接相加会偏向 A，因为词法量尺占主导。在这个仅有两项的候选池内做最小—最大归一化后，两者分别变为 $(1, 0)$ 和 $(0, 1)$，结果打平。哪种结果都不是天然正确；这个例子说明，在量尺与融合规则声明之前，“混合分数”并未被定义。应由相关性判断选择规则，而非为了算术方便。

反例 • 主题相关性可能掩盖证据丢失

考虑两个合成的前三项列表，以“是否有关该产品”这一二元标签评估。列表 A 包含产品页、经销商镜像页和已被替代的上市手册。列表 B 包含现行产品页、独立校准段落和现行安全通告。6 个条目都可以获得相同主题标签，因此两个列表均有 $\mathrm{Precision}@3 = 1$。但 A 只覆盖标称范围证据单位，故 $\mathrm{ER}@3 = 1/3$；B 覆盖标称范围、已验证范围和低温条件，故 $\mathrm{ER}@3 = 3/3$。这些数值仅为教学示例，说明排序器可以让文档级指标饱和，却丢失决策所需的限定条件。

4.5 融合查询，而不制造共识

查询规划可能为不同子问题生成多个列表。候选融合必须保留哪个查询召回了哪个段落。否则，被近重复查询反复返回的段落，可能看起来得到了独立支持。一种有用的候选记录为

$$c_i = \langle d_i, p_i, \mathcal{Q}_i, \mathbf{r}_i, h_i, o_i \rangle, \tag{21}$$

其中，$\mathcal{Q}_i$ 为召回查询集合，$\mathbf{r}_i$ 为各查询中的排名，h_i 为内容哈希，o_i 为溯源起点。完全重复、近重复、转载副本与独立产出信源，需要不同标签。

当前 GEO 文献包括以内容为中心和端到端测试环境，使信源影响或检索阶段能够被观察 [6, 23]。这些环境支持关于其指定语料、系统和指标的主张，不披露每个商业答案界面的候选融合政策。

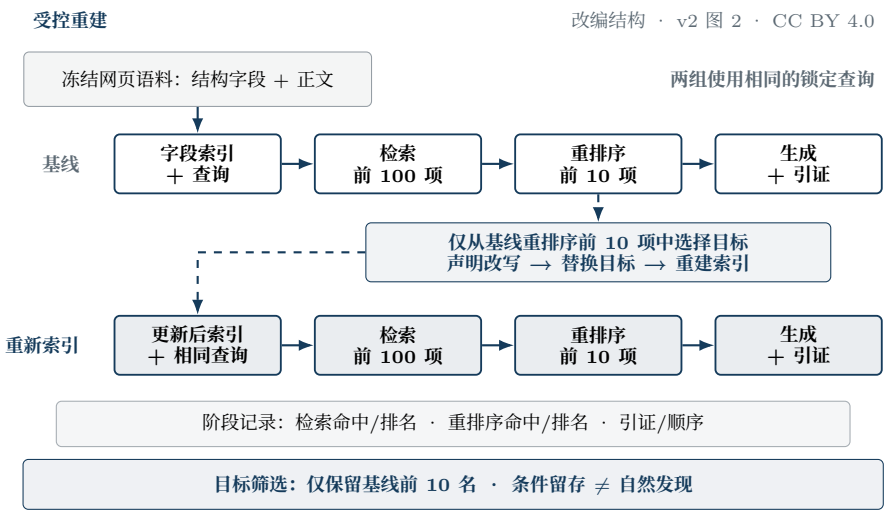


图 9: 根据 Kim et al. [23, 图 2] 改编的受控重建图，许可为 CC BY 4.0。重绘压缩了标签，去除了厂商图标，并明确标出基线前 10 名目标选择条件。它展示论文锁定流程中的条件留存，不是自然发现，也不是已披露的生产架构。

算法 2 保留轨迹的混合检索与证据感知装填

输入： 查询 q；语料 $\mathcal{D}_t$ 与切分 Π；词法和稠密检索器；深度 k_j；融合常数 κ
锁定的重排序器 g；重要证据标签 E_q；词元预算 B；来源上限 b_o

输出： 有序装填段落 S 与可重放轨迹 $\mathcal{T}$

核验并记录 $\mathcal{D}_t$、Π、模型检查点和配置文件的哈希
按已记录查询表征运行各检索器，保留完整的有序 top-k_j 列表
构造候选并集 C；为每项关联全部召回查询、列表、排名、原始分数、段落哈希和来源
对所有 $p \in C$ **执行**
    $s_{\mathrm{RRF}}(p) \leftarrow \sum_j \mathbb{1}\{p \in L_j\}/(\kappa + \mathrm{rank}_{L_j}(p))$
结束 对
按锁定规则聚类完全重复、近重复和转载段落；在 $\mathcal{T}$ 中保留聚类关系
应用评估前声明的代表项选择规则；若未声明，则保留全部成员
用 g 对剩余候选重排序；记录输入顺序、截断、分数、并列和输出顺序
实例化词元成本 ℓ_i、主张覆盖 e_{ic}、冗余 u_{ij} 与来源标签 o_i
在 B 与 b_o 约束下求解声明的装填目标，选择 S；记录被拒候选和生效约束
对 S 排序，序列化精确词元片段，返回 $(S, \mathcal{T})$

算法 2 是协议模板，不是在声称商业系统实现了这些步骤。其目的在于，使受控系统中的每次变换可以重放，并在其他环境中明确标出每次未观察到的变换。

4.6 在被测阶段解释效应

SAGEO Arena 在重建系统中使阶段留存漏斗可见：171,003 个独立页面和 2,700 个查询，经过字段级 BM25 与倒数排名融合，进入前 100 名候选集合，再由 Qwen3 重排序器保留 10 项，最后进入 GPT-5-mini 生成阶段 [23]。阶段结果并不单调。基线检索命中率为 0.58、重排序命中率为 1.00、引证率为 0.50；仅改正文时分别为 0.53、0.84、0.47；仅改结构时为 0.71、0.83、0.52。基线重排序值为 1.00，是因为目标本就从基线前 10 名中选择。因此，这些数值描述的是论文重建环境中的条件留存，而非任意网页的命运或商业排序规则。

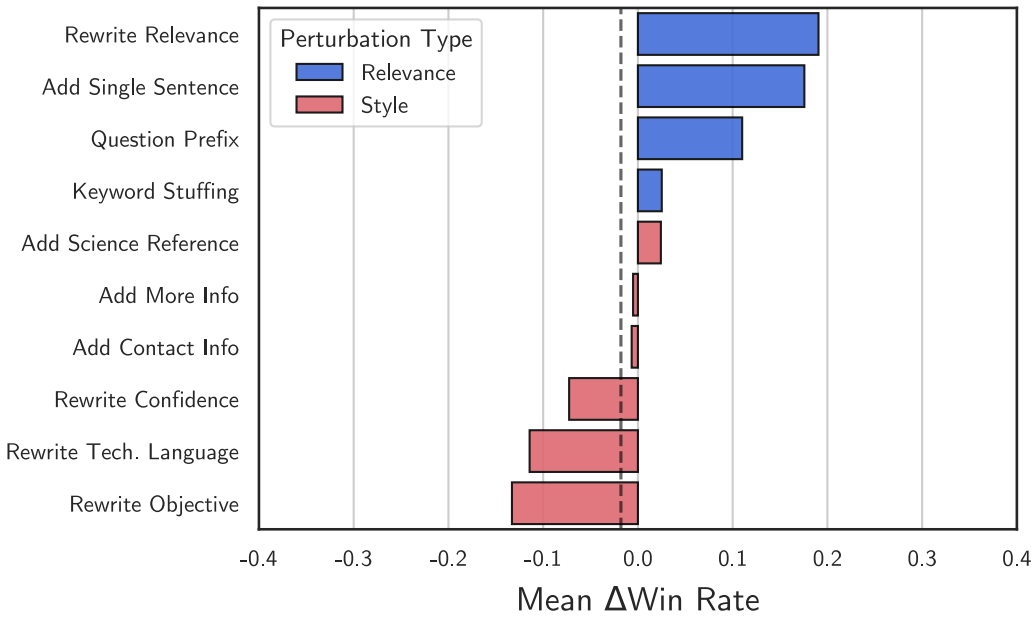


图 10: 固定冲突上下文实验中修改一个段落后的平均胜率变化；虚线为独立对照，不是零基线。转载自 Wan et al. [50, 图 2]，ACL 2024，采用 CC BY 4.0 许可；仅作矢量裁图与字体轮廓预处理。结果是已给定上下文中的模型偏好，不代表检索、引证、真实性或人类可信度判断。

与固定候选实验的对照很有启发。E-GEO 固定 10 个产品的候选列表，估计描述改写如何改变提示驱动的 LLM 重排序；其来源集合有 13,747 个查询—列表实例，报告的测试集为 2,000 个 [3]。该设计清晰研究给定候选后的排名变化，却不能说明修改条目能否从底层产品语料中被召回。同样，固定冲突证据研究表明，相关性、组成与顺序能够改变检索后的生成 [48, 50]，但并未估计检索效应。

解释边界 • 对相关性的敏感不等于证据质量

实验将两个冲突证据段落置入模型上下文，只修改支持“Yes”的段落，并要求二元回答。多数反事实改写由 `claude-v1-instant` 生成；独立对照也未被单独证明中性。因此，胜率响应识别的是该构造比较中的模型敏感性。正向柱形不能被报告为检索增益、引证增益、真实性信号或人类可信度判断。该图之所以有用，恰恰是因为它区分了模型偏好与被偏好证据的质量。原图未报告不确定性区间或各单元样本量，过滤后的评估规模也因模型而异。

表 10: 三种证据环境及其能够支持的目标估计量。

环境	可识别的比较	主张上限
重建漏斗	锁定流程中检索、重排序与生成各阶段的留存	不能建立普遍平台机制或自然发现率
固定候选列表	给定候选后的排名或引证选择	不能声称语料召回或获取效果
固定上下文	给定证据下的生成、顺序、冲突与消融	不能证明检索资格

表 11: 完整指定的合成混合检索轨迹。RRF 列可由两个排名及 $\kappa = 10$ 重新计算；v_i 是锁定的任务特定重排序器值，用于装填算例。

ID	受控段落内容	ℓ_i	词法排名	稠密排名	RRF	v_i
p_1	产品页：标称范围 0–50 mg/L	70	1	2	0.1742	0.80
p_2	校准报告：验证仅限于 10–25 °C 下的 5–35 mg/L	100	4	1	0.1623	1.00
p_3	现行安全通告：低于 5 °C 须用加温外罩	60	3	3	0.1538	0.95
p_4	已被替代的上市手册	80	2	5	0.1500	0.15
p_5	p_1 的近重复转载	70	5	4	0.1381	—
p_6	无关维护段落	50	6	6	0.1250	0.00

4.7 针对声明的任务重排序

重排序估计的是给定候选池下的顺序。逐点、成对和列表级目标回答不同问题；交叉编码器分数不会自动在查询之间标定。至少应报告候选深度、模型与检查点、输入截断、排序损失、并列处理、延迟，以及相关性标签与答案级效用之间的区别。

令 $g_{qi} \in \{0, 1, 2, 3\}$ 为候选 i 的分级相关性判断。深度 k 处的折损累计增益为

$$\mathrm{DCG}\,@k(q) = \sum_{i=1}^{k} \frac{2^{g_{qi}} - 1}{\log_2(i+1)}, \tag{22}$$

nDCG $@k$ 再除以理想排序的增益。该指标奖励所选标签下的排序质量，不测量最终答案是否忠实使用了段落。

4.8 将上下文分配视为受约束证据问题

上下文分配选择段落、位置，有时还选择摘录。对于段落 i，令 v_i 表示查询相关性，e_{ic} 表示对重要主张 c 的覆盖，ℓ_i 为词元成本，u_{ij} 为冗余，$z_i \in \{0, 1\}$ 为选择指示变量。一种可审计表述为

$$\max_{\mathbf{z}} \quad \sum_i z_i v_i + \alpha \sum_c \min\left(1, \sum_i z_i e_{ic}\right) - \beta \sum_{i<j} z_i z_j u_{ij}, \tag{23}$$

$$\text{subject to} \quad \sum_i z_i \ell_i \le B, \qquad \sum_{i:o_i=o} z_i \le b_o. \tag{24}$$

可选来源上限 b_o 限制某个依赖簇的支配程度。它可能提升多样性，也可能排除最佳信源；应报告其效果，而非预设它有益。

4.9 贯穿算例：从两个排名到一个上下文

回到 节 1.3 的兰湾合成查询。答案需要区分标称测量范围和独立验证范围，并保留低温限定条件。假设冻结的 6 段落语料产生 表 11 中两个完整排名。词元成本由一个锁定的分词器计数。本例数值与段落是教学输入，并非关于真实检索器、组织或平台的观察。

例如，

$$s_{\mathrm{RRF}}(p_1) = \frac{1}{10+1} + \frac{1}{10+2} = \frac{23}{132} \approx 0.1742,$$
$$s_{\mathrm{RRF}}(p_2) = \frac{1}{10+4} + \frac{1}{10+1} = \frac{25}{154} \approx 0.1623.$$

计算其余 4 项，得到融合顺序 $p_1, p_2, p_3, p_4, p_5, p_6$。该顺序基于排名，并不声称 p_1 提供最与决策相关的证据。按锁定重复规则，p_5 被记录为 p_1 簇的转载成员，并在重排序前删除，但该关系仍保留在轨迹中。任务特定重排序器将

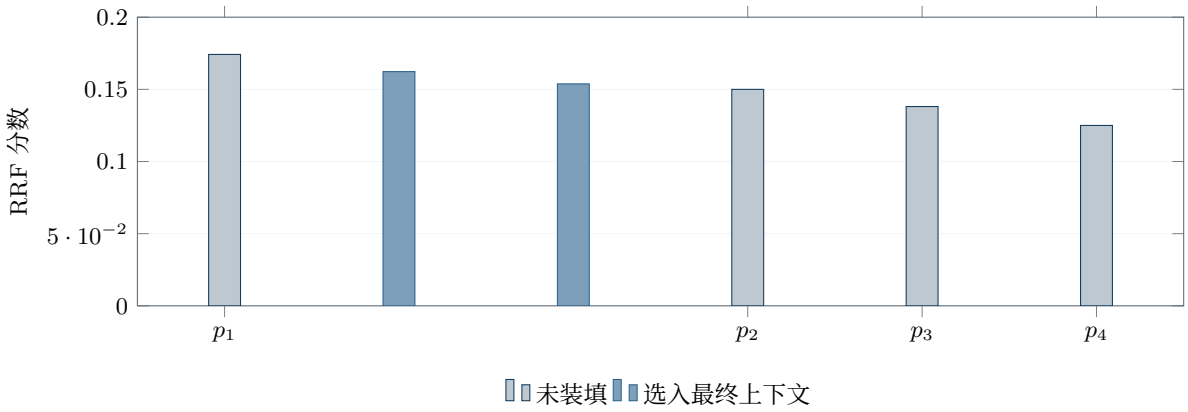


图 11: 合成案例中的排名融合分数与随后装填决定。在声明预算和重要证据目标下，RRF 最高的候选未被选中。每根柱均可由表 11 重算；颜色表示后续决定，而非 RRF 阈值。

表 12: 合成示例中非支配可行双段落组合的计算。包含零价值 p_6 的组合被其他组合支配。效用为 $\sum_i v_i$ 加上覆盖重要主张数的 0.8 倍；预算为 180 个词元。

组合	词元数	$\sum_i v_i$	覆盖的重要主张	效用
$\{p_1, p_2\}$	170	1.80	c_N, c_V	3.40
$\{p_1, p_3\}$	130	1.75	c_N, c_T	3.35
$\{p_1, p_4\}$	150	0.95	c_N	1.75
$\{p_2, p_3\}$	160	1.95	c_V, c_T	**3.55**
$\{p_2, p_4\}$	180	1.15	c_V	1.95
$\{p_3, p_4\}$	140	1.10	c_T	1.90

其余候选依次排为 p_2, p_3, p_1, p_4, p_6。

现设三项重要主张为标称范围 c_N、独立验证范围 c_V 和低温条件 c_T。令 $e_{1N} = e_{2V} = e_{3T} = 1$，其余 $e_{ic} = 0$，$\alpha = 0.8$，$B = 180$。去重后，保留项中没有冗余对，因此本次计算的冗余项为零。没有来源上限生效。表 12 枚举了非支配可行多段落组合。

每个具有正边际价值的三段落集合都会超出预算。唯一可行的三元组 $\{p_1, p_3, p_6\}$ 与 $\{p_1, p_3\}$ 效用相同，因为 p_6 不增加主张覆盖或相关性价值；最佳单段落效用为 1.80。因此，优化器选择 $S = \{p_2, p_3\}$，使用 160 个词元，剩余 20 个词元未用。该组合排除了 RRF 最高的段落，因为独立验证限制和低温条件覆盖了两项对决策至关重要的限定。轨迹必须记录：结论依赖登记的主张、相关性取值和预算；改变其中任一项，就定义了不同的分配问题。生成器仍可能未忠实使用任一段落，因此进入组合不能计为对答案的贡献。

4.10　估计贡献，而不混同于所有权

留一信源法比较：从受控上下文中移除某信源后，答案如何变化。Shapley 式归因在不同联盟上平均边际贡献，但解释取决于效用函数和参与者的定义。公平上下文归因研究明确提出这些假设 [35]；注意力与曝光研究在有界情境中检验相关机制 [2]。注意力权重或答案边际变化，都不能证明知识所有权、法律上的作者身份或可见引证。

适用边界与失效条件

开放系统检索轨迹能够识别重建系统中的失败。商业结果页、引证面板或网络观察，只暴露对应界面与时间。除非平台公开了相关信息，或受控干预识别了它，否则不得推断隐藏候选池、分数、重排序器或上下文顺序。

表 13: 分阶段检索诊断及其证据边界。

观察	能够支持	单独不能支持
Top-k 轨迹	冻结系统中的召回与排序	答案中的信源使用
重排序增益	声明标签下的排序改进	用户效用或归因忠实性
上下文日志	接纳、顺序与截断	未经消融的因果影响
可见引证	已解析信源对象的呈现	检索排名、主张蕴含或吸收
留一法变化	给定上下文中的条件影响	唯一作者身份或全局平台行为

表 14: 复现从检索到上下文结果的最低检查点。

成果文件	必须包含	通过条件
语料与表征	抓取时间戳、许可边界、文档与段落哈希、渲染器、切分器、分词器	干净重建产生相同标识符，或提供已记录的迁移映射
查询与候选轨迹	用户查询、每次改写、检索器配置、深度、原始有序列表、分数、召回路径	重算 RRF 分数与并列处理决定，均匹配发布轨迹
判断材料包	任务定义、重要证据单位、相关性评分规则、评审标签、分歧与裁定	另一位评审无需接触已报告结果，即可应用评分规则
重排序与装填台账	模型检查点、截断、候选顺序、词元成本、覆盖与冗余值、约束、所选片段及拒绝项	重放产生相同有序上下文，并识别全部生效约束
失败与敏感性日志	抓取缺失、重复簇、非确定运行，以及不同 k、κ、预算和标签选择	将结论重述在仍能得到支持的最窄情境内

4.11 复现检查点

结果页截图不是可重放的检索记录。在接纳阶段比较之前，独立读者应能取得 表 14 的成果文件，并获得相同候选身份、融合分数和装填词元片段；只允许已记录的非确定性差异。

复现实作 4 ・从检索损失追踪到上下文损失

问题。哪些证据在表征、检索、融合、重排序与装填中消失?

输入。冻结且获许可的语料；30 个聚类查询；段落级相关性与重要证据标签；词法和稠密基线；1 个锁定的重排序器；3 种上下文预算。

过程。保留语料与段落哈希。比较单查询与展开检索、RRF 与 1 种学习式融合、重排序、去重和证据感知装填。注入 1 项缺失限定、1 个转载重复簇，以及 1 个很长但高度相关的段落。

输出。Recall@k、nDCG@k、证据召回率、来源多样性、上下文覆盖、延迟，以及不确定性区间与抽样单位匹配的阶段转移表。发布 表 14 指定的成果包。

参考标准。每项报告损失都必须解析到带版本的文档、段落、查询列表、候选排名和装填决定。

独立检查。解释模型质量比较之前，第二个实现必须复现文档和段落身份、RRF 分数、重复处理决定，以及最终装填的精确词元片段。

停止规则。若无法重建语料、切分、相关性标签、候选路径或装填上下文，停止并将比较标记为无法识别。检查点失败本身就是结果；不得无声地用可见引证代理替换缺失阶段。

本章小结

1. 检索以表征、语料、查询和带版本的评分规则为条件；报告必须区分观察、潜在阶段与识别假设。
2. 词法与稠密分数需要明确融合方法；原始相加不具有量尺不变性，排名融合也仍以其列表、深度、并列规则和 κ 为条件。
3. 候选融合保留查询与溯源路径，避免把重复误当作独立支持。
4. 重排序指标评估声明标签和候选池，而非最终答案使用。主题相关性高，可以与重要证据召回率低并存。
5. 上下文装填是受约束的证据分配问题，其预算、主张权重、多样性、冗余与拒绝决定必须可见。融合排名最高不必然意味着被接纳。
6. 贡献度量是有条件的系统测量，不是所有权或全平台行为的证明。
7. 可复现结果将每项汇总损失解析到带版本的文档、段落、候选路径、判断与精确装填词元片段。

5 生成、引证、归因与信源影响

核心思想

进入上下文、被答案使用、显示引证和忠实归因，是四种不同事件。评估必须切分答案主张、解析引证对象、检验支持关系并估计信源影响，不能把可见链接视为该信源导致或蕴含相邻文字的证明。

5.1 本章回答的问题

读完本章，应能够回答以下五个问题：

1. 支持评分可解释之前，哪些上下文、信源、主张、引证和答案对象必须记录版本?
2. 在封闭界面上，生成与引证放置的哪些部分已观察、哪些经重建、哪些仍然隐藏?
3. “有引证”、引证蕴含、引证完整性、主张吸收与反事实信源影响，为何回答不同问题?
4. 如何标定自动支持裁判，而不把其标签推广为标定分层之外的真值?
5. 哪些台账和上下文干预，能让第二位分析人员无需接触想象中的隐藏提示，就复现结论?

5.2 建模有效上下文，而非想象的提示

在受控系统中，回答可表示为

$$y \sim p_{M,v}(y \mid q, h, C, \sigma, \eta)\,, \tag{25}$$

其中，M, v 标识模型与版本，h 为对话状态，C 为日志中的有效上下文，σ 包含系统与工具指令，η 包含解码状态。在封闭界面中，有些项是潜在量。研究者记录已观察输入与输出，并将重建部分标为假设。

参数知识构成一条旁路：即使 C 中没有支持信源，答案主张也可能出现。反过来，信源可能影响措辞或答案结构，却没有显示引证。因此，端到端流程不能简化为引证计数。

5.2.1 先声明可观察性，再解释

同一个字段在受控重建中可以被观察，在商业界面上却可能隐藏。因此，报告应相对于具体研究对字段分类，而不是永久把“可观察”贴在某个概念上。

5.3 先切分答案主张，再评分引证

令 $A(y) = \{a_1, \ldots, a_J\}$ 为答案中的重要主张。主张是在保留数量、条件、范围、时间和归因的同时，能够评估支持关系的最小单位。将“套餐 A 支持 128 个项目，旧版合同除外”拆成两个无关片段，可能错误地把数值标为有支持，却丢掉例外。

令 $R(y) = \{r_1, \ldots, r_K\}$ 为解析后的引证对象。解析记录显示标签、目标 URL、重定向、规范 URL、访问时间、可得时的被引或链接段落，以及溯源起点。只有界面把引证附在主张上，或有已记录解析器建立该关系，答案—引证图才存在边 $a_j \to r_k$。

图 12 将切分、可见关联、URL 解析和证据裁定保留为不同操作。

表 15: 生成—归因研究的最低可观察性约定。

类别	示例	必要处理
直接观察	已提交查询；显示的答案与锚点；解析的 URL；时间戳；受控系统记录的上下文和解码配置	规范化前保留原始对象，将每个派生标签关联到不可变标识符
重建	规范信源身份；主张片段；段落关联；由获准轨迹推断的可能上下文成员	写明解析器、标注者、错误状态和替代重建
潜在	专有有效上下文；参数贡献；隐藏指令层级；因果信源使用；引证放置政策	不用可见 URL 填充；只在明确干预下估计，否则保留未知
假设	重要性权重；主张切分充分性；裁判误差稳定；移除信源后没有上下文干扰	预注册，在反例上检验，并展示假设失效时决定如何变化

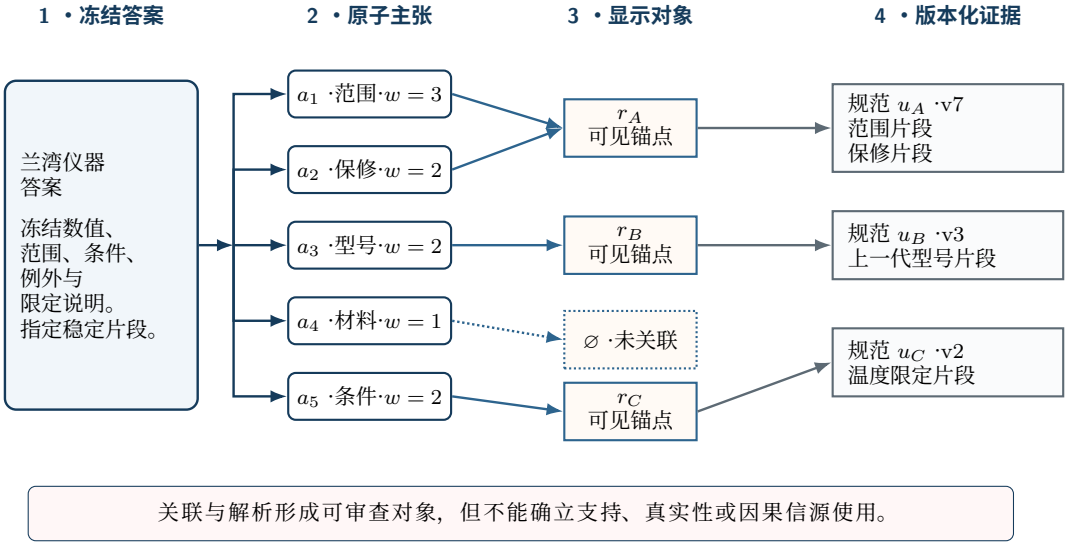


图 12: 从冻结答案到可审查证据对象。切分保留数量、范围、条件与限定；引证关联记录可见或声明范围；解析保留重定向、规范身份、信源版本与精确证据片段。任何单一步骤都不能确立支持、真实性或因果信源使用。这些对象和标签是原创合成教学构造。

定义 5.1: 引证蕴含与完整性

对于重要主张 a_j 和已解析信源段落 r_k，令 $S_{jk} \in \{0, 1\}$ 表示依据公开标注指南经人工裁定的支持关系。则

$$\text{Entailment}(y) = \frac{\sum_{(j,k)\in E_y} w_j S_{jk}}{\sum_{(j,k)\in E_y} w_j}, \tag{26}$$

$$\text{Completeness}(y) = \frac{\sum_j w_j \mathbb{1}\{\exists k : (j,k) \in E_y,\ S_{jk} = 1\}}{\sum_j w_j}. \tag{27}$$

蕴含检验附着引证是否支持对应主张；完整性检验重要主张是否至少有一条支持性引证。两者都依赖主张权重、切分、解析与裁定政策。

命题 5.2: 回答级引证标记无法识别完整性

令 $H(y) = \mathbb{1}\{|R(y)| > 0\}$ 表示回答 y 至少显示一个引证。对于任意 $\epsilon > 0$，存在两个回答 y_1, y_2，满足 $H(y_1) = H(y_2) = 1$，但 y_1 的引证完整性为 1，y_2 的完整性小于 ϵ。

表 16: 引证选择、吸收代理与效用归因回答不同问题。

证据设计	直接测量对象	必须声明的边界
双信源引证选择 [49]	6 个模型、252,000 次试验中首个显示 URL；信源顺序交叉平衡	恰好提供 2 个信源；无真实检索；不稳定的巨大优势比不是可迁移效应
平台快照 [65]	602 个提示的引证广度，以及在可抓取被引页面上构造的评分	描述性、以成功抓取为条件的代理；不是隐藏注意力或因果依赖
求和—最大值分配 [35]	对声明的“每个答案要点取最佳信源”效用，计算精确 Shapley 贡献	只对该效用与参与者集合精确；不是作者身份、训练贡献或恢复出的模型使用

证明。让 y_1 含一个权重为 1 的重要主张，并关联支持性段落。让 y_2 包含同一受支持主张，再加 n 个权重为 1、未获引证的重要主张。两个回答都有引证，而完整性分别为 1 与 $1/(n+1)$。选取 $n > 1/\epsilon - 1$ 即得结论。 □

5.4 区分选择、使用、吸收与归因

对信源 s，定义回答中的四个二元或分级变量：

上下文选择 $Z_{\mathrm{ctx}}(s)$
: s 的某种表征位于日志记录的有效上下文中；

信源使用 $Z_{\mathrm{use}}(s)$
: 受控消融或其他声明方法，将重要答案变化归因于 s；

主张吸收 $Z_{\mathrm{abs}}(s)$
: 一项或多项预注册、信源特有且经过验证的主张，忠实出现在答案中；

显示归因 $Z_{\mathrm{cite}}(s)$
: 界面呈现一个可解析到 s 的引证。

没有哪个变量在逻辑上蕴含其余全部。信源可能被选择却未使用，被吸收却未引证，在不受支持的主张旁被引证，或因为其 URL 代表一组重复来源而被引证。

引证选择与引证吸收的区分，是近期测量研究的核心 [65]。引证失败研究也支持分阶段诊断与修复，而非笼统的改写循环 [47]。具体分类法与结果仍限于锁定论文版本及所评系统。

5.4.1 不可合并的三个目标估计量

近期研究使这种区分具体化。双信源实验可以估计给定信源中哪一个获得首个显示 URL；平台快照可以描述引证广度和人为构造的文本重叠代理；联盟方法可以分配已声明的答案覆盖效用。如果都称为信源影响，就会抹去干预、分母和假设之间的区别。

双信源研究报告，3.4% 的运行没有匹配的首个 URL，被排除在二元模型之外；部分因素还出现准分离或奇异拟合。有关顺序、主题匹配、规格和时间戳的方向性证据可能有用，但不能将极大优势比视为精确的跨系统效应量。平台快照同样将页面级代理限定于成功抓取的页面。这两种排除均属于目标估计量的定义，而不应藏在脚注。

例题 5.3: 引证可见，支持却不完整

答案写道：“系统 K 保留日志 30 天，并已通过标准 X 认证。”附带页面支持保留期限，却仅说明组织使用与标准 X 相关的控制框架。引证对象真实，其中一个分句有支持，但认证主张没有支持。因此，回答级“有引证”等于 1，主张加权完整性却小于 1；这项无依据的高影响主张必须纠正。

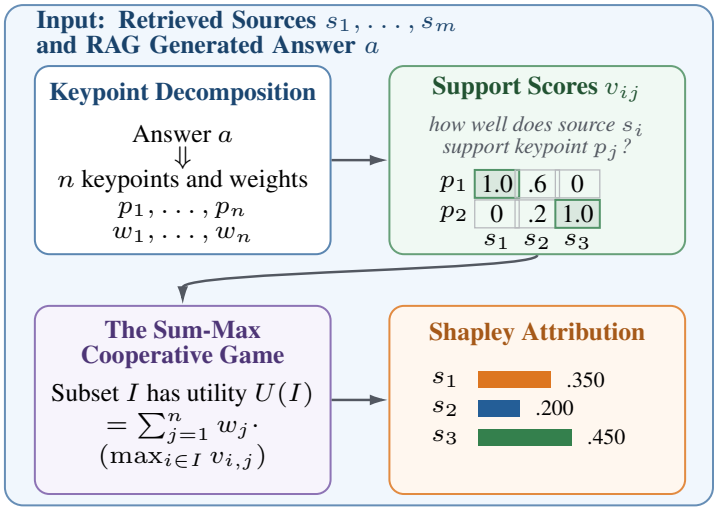


图 13: 要点—支持流程，随后对已声明的最大覆盖效用进行 Shapley 分配。转载自 Patel et al. [35, 图 2]，采用 CC BY 4.0 许可；仅裁切。分配仅对该效用与参与者集合精确，不代表隐藏注意力、因果信源使用、作者身份、训练贡献或普遍公平。

表 17: 兰湾合成答案的冻结主张—引证台账。

主张	权重	关联对象	最终支持	裁定理由
a_1	3	r_A	支持	运行范围表述与版本准确匹配
a_2	2	r_A	支持	保修期限明确，地区匹配
a_3	2	r_B	无	信源讨论的是上一代，而非本型号
a_4	1	无	未解决	材料兼容性主张没有关联对象
a_5	2	r_C	支持	限定说明与温度条件均保留

5.5 轨迹算例：一个答案、五项主张、三个引证

以下是合成教学示例，不描述任何生产系统。有关虚构兰湾控制器的冻结答案含 5 项重要主张，其预注册整数权重合计 10。可用的已解析信源段落有 3 个。两位标注者独立切分与标注答案，再对下文所示的一处分歧裁定。

5 项主张中 4 项具有可见关联，因此未加权引证发生率为 $4/5 = 0.80$。关联边的权重为 $3 + 2 + 2 + 2 = 9$，其中受支持权重为 $3 + 2 + 2 = 7$。因此

$$\widehat{\text{Entailment}} = \frac{7}{9} = 0.778, \qquad \widehat{\text{Completeness}} = \frac{7}{10} = 0.700. \tag{28}$$

缺失权重不能被无声编码为支持或已关联错误：其中 2 个权重单位关联到错误产品版本，另 1 个没有引证。如果报告只保留“80% 的主张有引证”，两类失败都会被隐藏。

台账测量的是可见关联与支持关系，而非信源影响。本章后续的配对留一信源构造检验另一个目标估计量，并暴露可能使解释失效的伴随变化。

5.6 标定自动裁判

模型裁判可辅助主张切分、蕴含与矛盾判断，但其标签是带误差的测量。标定集应跨主张类型、信源长度、引证位置、答案模型、语言和预期难度抽样。应报告混淆矩阵、类别占比、评审间一致性、裁定规则，以及最终估计对合理裁判误差的敏感性。

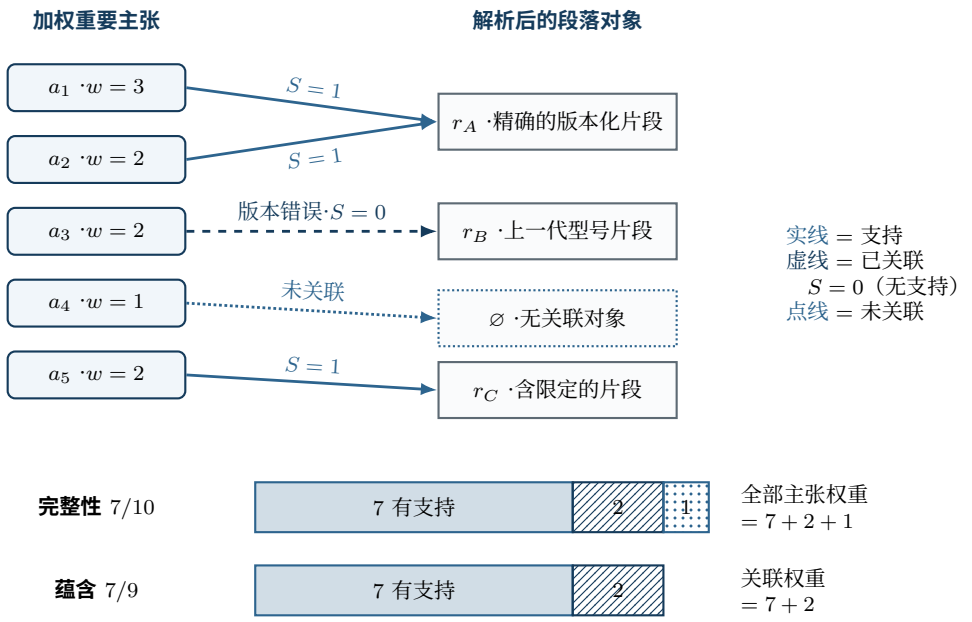


图 14: 兰湾合成台账中的可见关联与证据支持。实线为已关联并裁定为支持的主张—段落对；虚线虽已关联，却不支持主张；点线主张没有关联对象。分母带图说明为何完整性为 7/10，而蕴含为 7/9。这是教学重建，不是平台观察。

对于二元支持标签，令灵敏度为 Se，特异度为 Sp，观察到的阳性率为 $\widetilde{p}$。在误差率稳定且 $\mathrm{Se} + \mathrm{Sp} > 1$ 时，校正后的占比为

$$\widehat{p} = \frac{\widetilde{p} + \mathrm{Sp} - 1}{\mathrm{Se} + \mathrm{Sp} - 1}. \tag{29}$$

如果误差率随模型、语言或主张分层变化，该校正便无效。未校正标签、标定样本和分层诊断，均须作为发布成果保留。

5.7　用反事实上下文检验影响

在受控环境中，比较完整上下文 C 与移除信源 s 的上下文 C^{-s}，并尽可能固定顺序、词元预算与其他段落。对已声明答案效用 U，定义

$$\Delta_s(C) = \mathbb{E}[U(Y \mid C) - U(Y \mid C^{-s})]. \tag{30}$$

该比较测量构造上下文中的贡献，可能受替换段落、位置变化、解码方差和语义冗余混杂。联盟方法可在信源之间分配贡献，但继承所选效用和参与者定义。

为构造可复现合成测试样例，仅当配对运行 i 中 3 项预注册的 r_A 特有主张全部忠实出现时，定义 $U_i = 1$。完整与消融上下文使用相同查询、指令哈希、模型配置、保留段落字节与配对种子。同时记录所有词元数量、绝对位置、替换政策和截断差异，不能假定这些伴随变化不存在。

表 18: 兰湾合成信源消融样例中的 10 组配对结果。1 表示 3 项声明的 r_A 特有主张全部忠实出现；数值为检查算术与协议而构造。

配对种子	01	02	03	04	05	06	07	08	09	10	均值
完整 C	1	1	1	1	1	1	1	1	0	0	.80
移除 r_A，C^{-A}	0	0	0	1	0	1	0	1	0	0	.30
配对损失	1	1	1	0	1	0	1	0	0	0	.50

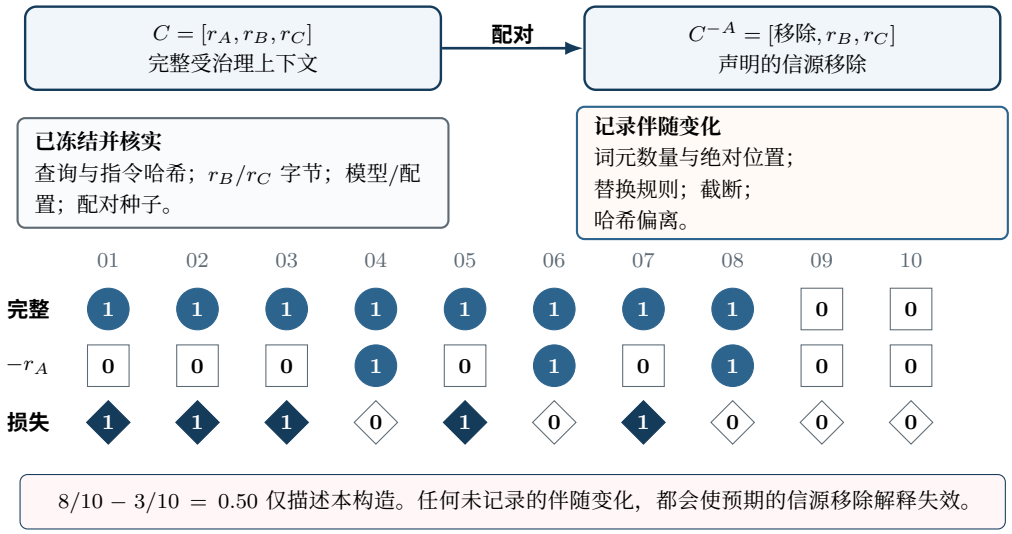


图 15: 兰湾合成上下文中的配对留一信源构造。查询、配置、保留段落与种子均冻结；所有伴随词元、位置、替换或截断变化均记录，而非假定不存在。10 组配对运行中 5 组丢失声明的信源特有内容效用，得到仅对该构造上下文成立的描述性差值 0.50。

引证漏洞与图像描述干预研究表明，为何对抗条件和多模态条件应纳入此类评估 [7, 30]。这些研究提供有界的攻击证据，却不允许推定所有被引或带图像描述的信源，都以同样方式脆弱。

适用边界与失效条件

引证指标不是权威评分。信源本身错误时，支持判断不是真实性判断。信源影响估计不是作者身份的证明。封闭界面的引证观察仅识别该界面、账户状态、地点与日期，不代表平台的普遍政策。

复现实作 5 ・分别审查引证与吸收

问题。显示引证何时与受支持、被吸收的主张一致?
输入。含 20 项已验证重要主张的冻结上下文集、30 个问题、至少 2 个信源依赖簇，以及锁定生成器。
过程。重复生成答案；切分主张；解析 URL；双人标注支持；识别信源特有主张；开展留一信源与留一主张比较；注入 1 个错误信源引证和 1 项遗漏限定。
输出。蕴含、完整性、吸收、显示引证率、裁判标定、反事实影响、不确定性，以及始终不把这些测量合并为一个分数的失败矩阵。
参考标准。第二位审核人能够复现每个主张片段、信源段落、解析路径、标签与上下文比较。
停止规则。标定集达到预声明的严重错误率，或引证解析丢失裁定所需信源版本时，停止自动评分。

算法 3 保留轨迹的主张、引证与影响审查

输入： 原始答案 y；版本化答案状态 e；显示的引证对象 R；信源快照 S；标注指南 G；标定集 K
输出： 主张台账、支持指标、裁判诊断，以及可选的有界影响比较

规范化之前，对 y, e, R, S, G, K 计算哈希并冻结
解析重定向与规范信源身份；把每次解析失败保留为独立状态
两位标注者独立切分重要主张 $A(y)$
保留两套切分结果，协调边界，分配稳定主张标识符
仅通过可见范围或声明的解析器关联引证对象
将每条主张—段落边标为支持、限定、质疑、无支持或未解决
裁定分歧，不覆盖裁定前标签
用声明权重与分母流计算蕴含和完整性
在 K 上按分层评估自动裁判，保留混淆矩阵和严重错误率
若 超过预注册的严重错误或解析失败阈值 **则**
    停止自动评分，将受影响分层转交人工复核
结束 若
对 每个预注册且符合受控影响检验条件的信源 s **执行**
    构造 C^{-s}，记录所有替换、顺序、位置与预算变化
    在 C 和 C^{-s} 下运行配对种子；保留缺失与拒答运行
    估计声明效用差值与聚类不确定性
结束 对
导出不可变原始对象、派生台账、决定、代码/配置哈希与主张边界

表 19: 最低主张—引证审查记录。

字段组	必须记录	暴露的失败
答案状态	原始答案、模型/界面、版本、地区、时间、重复	未声明的产品或解码漂移
主张单位	精确片段、规范命题、重要性、权重	分母与切分变化
引证对象	显示锚点、解析/规范 URL、段落、来源	重定向、错误信源关联、来源洗白
支持标签	蕴含/限定/质疑/无支持、理由、标注者	部分支持或矛盾支持
影响检验	上下文条件、消融、结果、不确定性	将引证误当因果使用
处置	接受、限定、纠正、删除、升级处理	审查发现后未采取行动

5.7.1 复现检查点

独立分析人员取得冻结答案与上下文变体、原始锚点、重定向日志、信源快照、标注指南、两份独立标签文件、裁定记录、重要性权重、标定样本、生成配置、配对种子和 SHA-256 哈希清单。只有该分析人员能重建 表 17 的每一行，复现 式 (28) and 图 14，并解释为何消融比较不能识别全局平台行为，检查才通过。只匹配最后两位小数，却无法复现主张边界和解析状态，不算通过。

本章小结

1. 有效上下文与系统状态定义生成条件。
2. 先切分答案主张、解析引证对象，才能测量支持。
3. 选择、使用、吸收和显示归因彼此不同。
4. 蕴含与完整性具有不同分母和失效模式。
5. 自动裁判需要标定和敏感性分析。
6. 反事实影响以构造上下文为条件，不证明真实性、作者身份或全局平台行为。
7. 发布须保留原始对象、独立标签、分母流、标定误差、上下文变更和解析失败；标量分数不能替代轨迹。

6 指标、标签、裁判与基准设计

核心思想

GEO 看板不是测量协议。协议须在采集回答之前声明目标总体、查询分布、系统状态、重复方案、分析单位、目标估计量、缺失数据政策、不确定性和决策规则。

6.1 本章回答的问题

读完本章，应能够回答以下五个问题，而不把看板数值当作不言自明的结论：

1. GEO 目标估计量描述哪个总体、系统状态、结果单位与比较？
2. 如何对查询意图、改述、随机重复、系统和时间窗口抽样，并记录为评估单元？
3. 为何同一查询意图内的重复答案，不是从用户需求总体中新增加的独立抽样？
4. 未解决标签、裁判分歧、采集失败及其他缺失观察，应如何进入分析？
5. 哪种不确定性区间、反事实、约束指标与停止规则，能让干预决策可复现，而非事后编排？

6.2 从目标估计量开始

设获准干预 δ 将信源 s 改为 s^{δ}。对于系统 m、查询 q 下的结果 Y，理想的查询级处理效应为

$$\tau(q,m) = \mathbb{E}\big[Y(s^{\delta},q,m) - Y(s,q,m)\big]\,. \tag{31}$$

目标均值对具有明确权重的查询分布 Q 和系统分布 M 求平均：

$$\tau_{Q,M} = \sum_{q\in Q}\sum_{m\in M} w_q w_m \tau(q,m). \tag{32}$$

如果无法同时展示基线和干预，模型漂移与索引更新可能同 δ 混杂。如果只抽取有利于品牌的查询，目标估计量便是该有利面板上的可见性，而非用户需求空间中的可见性。

定义 6.1: 评估单元

评估单元是带版本的元组

$$e = \langle q, m, v, d, \lambda, g, a, r, s\rangle, \tag{33}$$

其中，q 为查询，m 为系统，v 为可观察时的产品或模型版本，d 为日期时间，λ 为地区/语言，g 为地理位置或端点，a 为账户状态，r 为重复编号，s 为界面。回答和全部派生标签均关联到该单元。

每个报告比率还需要分母流。至少保留计划单元、已返回回答、含引证回答、已解析引证对象、成功抓取页面、切分后的重要主张，以及已裁定主张—信源对。在一个 602 提示观察语料中，23,745 条引证特征记录产生 18,151 个抓取成功页面，占 76.44% [65]。在这些记录上计算页面级评分，估计的是以成功抓取为条件的对象；若无声地将其视为全部引证总体，就改变了目标估计量。

6.3 采集答案之前，先构建查询分布

查询全集应代表任务，而非一袋关键词。有用的分层包括事实查找、品类探索、比较、故障排查、推荐、采购、声誉与高风险决策。每层中，查询还应在具体程度、实体命名、约束和语言上变化。最终样本记录设计权重；如有实

表 20: 最低限度的多目标 GEO 评价表。

维度	单位	指标示例	必要配套
曝光	实体—回答	提及发生率、片段位置	情感与正确性
归因	信源—主张—回答	引证发生率、蕴含、完整性	解析 URL 与信源独立性
吸收	主张—回答	已验证主张覆盖	矛盾与范围保留
检索	查询—段落	Recall@k、nDCG@k	下游答案效用
稳定性	查询—系统—时间	方差、转移率、留存	抽样频率与缺失情况
业务	用户—任务	合格行动率	实验或准实验识别
风险	主张、信源或回答	无依据主张率、操纵标记	严重性与审核结果

际使用数据，也记录观察使用权重。

设分层 h 含 N_h 个符合资格的查询，样本量为 n_h。平均结果的分层估计为

$$\widehat{\mu} = \sum_{h=1}^{H} W_h \widehat{\mu}_h, \qquad W_h = \frac{N_h}{\sum_j N_j}. \tag{34}$$

各层抽取相同数量有利于诊断，却不意味着各层在总体中等比例出现。报告应同时展示未加权诊断结果和声明的加权估计。

6.3.1 泄漏与查询调优

用于开发干预的提示，不能同时成为唯一评估提示。应按底层意图划分开发、验证和留出面板，而不只按表面改述划分。如果多个改述共享同一事实需求，应在划分和不确定性估计中视为同一簇。

6.4 声明指标的单位

对于实体 b、信源 s、回答 y_{qmr} 和 R 次重复，定义简单发生率：

$$\widehat{\mathrm{MentionRate}}(b) = \frac{1}{|Q||M|R} \sum_{q,m,r} \mathbb{1}\{b \text{ is resolved in } y_{qmr}\}, \tag{35}$$

$$\widehat{\mathrm{CitationRate}}(s) = \frac{1}{|Q||M|R} \sum_{q,m,r} \mathbb{1}\{s \in \mathcal{A}_{qmr}\}. \tag{36}$$

第一式统计回答中消歧识别到实体的事件，第二式统计信源进入引证集合的事件。这些估计要求别名消歧政策、引证 URL 规范化政策和回答缺失政策。

引证频次不是吸收。令 C_s 为经验证的信源特有主张集合，$z(c, y)$ 表示主张 c 忠实出现在答案 y 中。透明的吸收估计为

$$\widehat{\mathrm{Absorb}}(s, y) = \frac{\sum_{c \in C_s} \omega_c z(c, y)}{\sum_{c \in C_s} \omega_c}, \tag{37}$$

其中主张权重 ω_c 预先声明。若 z 由模型裁判赋值，应使用人工标注样本估计裁判误差与分歧。

6.5 生成答案是重复测量

一个查询只执行一次，得到的是截图，不是随机回答分布的估计。重复运行嵌套于查询、系统和时间中。一种有用的层级描述为

$$Y_{qmtr} = \mu + \alpha_q + \beta_m + \gamma_t + (\alpha\beta)_{qm} + (\beta\gamma)_{mt} + \varepsilon_{qmtr}. \tag{38}$$

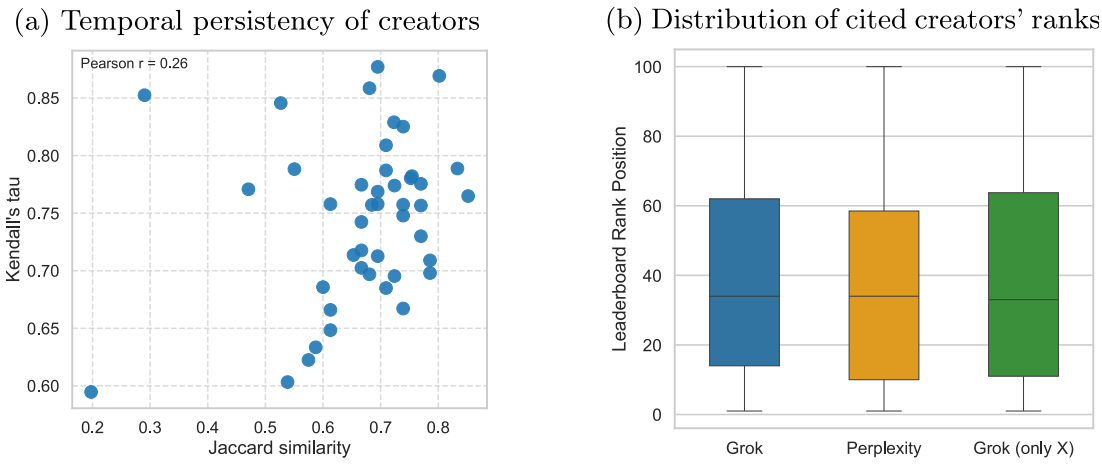


图 16: 两种不同曝光汇总：创作者群落随时间的持续性，以及被引创作者排名分布。两幅图均未将重复观察压缩成不随时间变化的单一可见性分数。

来源说明。取自 Alipour et al. [2, 图 1]，PDF 实际第 5 页的矩形裁图，许可为 CC BY 4.0。2 个 Type 3 字体程序被转为矢量轮廓，标记、数值、颜色和几何关系保持不变。研究报告 20 天创作者面板与各引擎排名；图形不能识别因果排序机制或总体曝光规律。

目的并不总是拟合这个精确模型，而是认识到提示、系统、日期与随机重复之间的变异具有不同含义。把每个回答都当作独立行，会产生过度自信的区间。

纵向 GEO 测量研究主张将可见性描述为分布，并持续观察，而非只测一次 [40]。因此，实用报告包括：

- 独立查询意图与改述数量；
- 每个单元的重复次数与间隔；
- 缺失、拒答、搜索禁用和错误回答；
- 查询内与查询间变异；
- 对 URL、实体和主张解析规则的敏感性；
- 条款允许时，带版本的原始回答存档。

实证启示是估计精度—成本曲线，而非照抄重复次数。被引面板中，同窗口信源集合 Jaccard 均值为 0.321–0.434，相邻日均值为 0.336–0.423。按论文有限面板计算，品牌检测标准误从 1 次运行的 0.370，降至 7 次的 0.081 和 8 次的 0.062。这些数值属于该德语、四垂直领域、四引擎设计，并不意味着 7 或 8 次调用普遍足够。

6.5.1 失败案例：更多调用不会创造更多查询意图

考虑一项合成研究：对同一个查询，每组各运行 100 次。实体在基线 50 个回答、干预 70 个回答中出现。如果把全部 200 个回答当作独立总体抽样，估计差为 0.20，常规双样本伯努利标准误为

$$\sqrt{\frac{0.5(1-0.5)}{100}+\frac{0.7(1-0.7)}{100}}=0.0678, \tag{39}$$

得到名义 95% Wald 区间约为 $[0.067, 0.333]$。该计算在强独立性和平稳性假设下，描述那个查询的重复生成，却不估计意图间变异：若要推断查询意图总体，研究只有 1 个簇，无法得到基于聚类的标准误。更多调用可减少给定查询下的蒙特卡洛误差，却无法弥补没有独立抽样用户需求的缺陷。

表 21: 合成聚类算例的测量约定，区分回答级标签与意图级推断单位。

协议字段	冻结声明
目标总体	8 簇教学面板所代表的留出采购查询意图；各簇目标权重相等
系统与比较	一个冻结重建流程；在匹配的改述—种子区组内比较修订信源减基线信源
可观察记录	意图 ID、改述 ID、种子、组别、采集状态、回答、消歧实体标签、标签来源与裁定状态
结果与单位	实体—回答的二元消歧提及；效应以百分点报告
推断单位	查询意图簇；改述与重复生成仍嵌套于对应意图
主要目标估计量	本面板有效返回回答中，按意图等权平均的提及概率差
缺失处理	采集失败保留为缺失，绝不记为“未提及”；主要估计同时附上分组计数和二元结果极端界限

表 22: 8 个意图簇的合成回答计数。每组计划 10 个回答。“缺失基线/干预”表示各组采集失败；最后一列仅从有效回答计算。

意图	基线 x_{h0}/n_{h0}	修订 x_{h1}/n_{h1}	缺失基线/干预	$\widehat{d}_h$
I1	3/10	5/10	0/0	0.200
I2	4/10	5/9	0/1	0.156
I3	2/9	3/10	1/0	0.078
I4	5/10	8/10	0/0	0.300
I5	4/10	5/10	0/0	0.100
I6	6/10	6/9	0/1	0.067
I7	5/9	4/10	1/0	−0.156
I8	3/10	6/10	0/0	0.300
返回总计	32/78	42/78	2/2	—

6.6 算例：重复测量与聚类不确定性

本节使用完全合成的教学数据，不报告生产平台行为。在冻结重建流程中，兰湾团队比较基线信源（$a = 0$）与获准、保留证据的修订（$a = 1$）。留出面板含 8 个采购查询意图簇，每簇 5 个改述、2 个预先声明的生成种子，即每组计划 10 个回答，总计 160 个。信源版本在改述—种子区组内配对，其他受控组成部分均固定。

只有声明为可观察的字段才能进入标签和估计量。已记录检索轨迹可以作为独立结果分析，但未记录的检索或生成状态仍是潜在量，不能根据最终回答是否出现实体反向填充。

令 $Y_{har} \in \{0, 1\}$ 为意图簇 h、组别 a 中已返回重复 r 的消歧提及标签。若返回 n_{ha} 个回答，其中 $x_{ha} = \sum_r Y_{har}$ 个包含实体，定义

$$\widehat{p}_{ha} = \frac{x_{ha}}{n_{ha}}, \qquad \widehat{d}_h = \widehat{p}_{h1} - \widehat{p}_{h0}, \qquad \widehat{\tau}_{\text{intent}} = \frac{1}{H} \sum_{h=1}^{H} \widehat{d}_h. \tag{40}$$

回答是标注单位，意图却是其变异能够支持所声明推广的单位。簇等权可避免有效重复较多的意图无声地获得更大总体权重。

使用未舍入比率，8 个簇差值之和为 1.0444，因此

$$\widehat{\tau}_{\text{intent}} = \frac{1.0444}{8} = 0.1306, \tag{41}$$

即本教学面板上的差异为 13.1 个百分点。8 个簇差值的样本标准差为

$$s_d = \sqrt{\frac{1}{H-1} \sum_{h=1}^{H} (\widehat{d}_h - \widehat{\tau}_{\text{intent}})^2} = 0.1476, \qquad \widehat{\text{SE}}_{\text{cluster}} = \frac{s_d}{\sqrt{8}} = 0.0522. \tag{42}$$

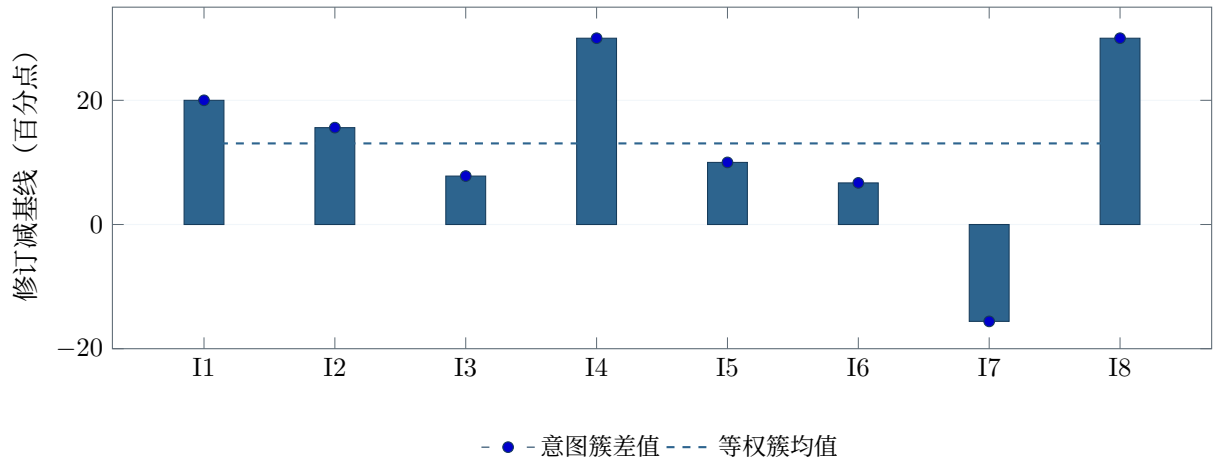


图 17: 从 表 22 重新计算的意图级效应。正均值不会抹去负向簇。数据是合成的，虚线仅描述本面板，不是对具名平台的估计。

作为透明的小样本示例，将 8 个意图视为独立抽样，得到

$$0.1306 \;\pm\; t_{0.975,7}(0.0522) = 0.1306 \;\pm\; 2.365(0.0522) = [0.007, 0.254]. \tag{43}$$

即使有 156 个已返回答案，区间仍很宽，因为与查询总体不确定性有关的样本量是 8。若 8 个意图是固定全量而非目标总体样本，该 t 区间就没有超总体解释；此时应描述性报告面板分布，或依据有充分理由的分配设计分析。

4 次采集失败构成另一类不确定性。如果将每个缺失干预回答赋为 0、每个缺失基线回答赋为 1，将每组分母恢复为 10，则等权簇效应为 0.100。反向赋值则为 0.150。因此，二元结果极端范围为 $[0.100, 0.150]$。这不是置信区间；它回答极端缺失标签下点估计能够移动多远。可信报告应同时展示抽样不确定性和缺失敏感性，而非把一者隐藏在另一者之中。

更大评估若包含足够多独立抽样簇，算法 4 的配对聚类自助法可保留抽样单位。查询设计使用分层、不等纳入概率或固定有限总体时，必须相应调整。

算法 4 重复回答 GEO 指标的配对聚类自助法

输入： 冻结回答台账；意图簇 ID $h = 1, \ldots, H$；配对组 $a \in \{0, 1\}$；结果、有效性、权重与缺失规则；自助重复次数 B；随机种子。

输出： 点估计；遵循聚类结构的区间；异质性、缺失和偏离记录。

验证评估单元 ID 唯一，按组与簇核对计划、返回、有效、标注和裁定分母。
应用冻结的实体解析器与裁定规则；未解决标签和采集失败保留为状态而非零。
对 $h = 1$ to H **执行**
    从有效回答计算各组比率 $\widehat{p}_{h0}$、$\widehat{p}_{h1}$ 与配对差值 $\widehat{d}_h$。
    记录重复、缺失与簇内离散；按预声明政策标记没有有效回答的组。
结束 对
计算 $\widehat{\tau} = \sum_h w_h \widehat{d}_h / \sum_h w_h$，保留 $\widehat{d}_h$ 的完整分布。
对 $b = 1$ to B **执行**
    按与簇设计匹配的规则有放回抽取 H 个簇 ID（等概率簇样本用均匀抽样）；每次抽取携带两组及全部嵌套回答。
    重算归一化加权均值 $\widehat{\tau}^{*(b)}$；重复抽到的簇贡献两次，但绝不把其行拆分到不同组。
结束 对
从 $\{\widehat{\tau}^{*(b)}\}_{b=1}^{B}$ 构造预声明自助区间，并用额外带种子重复验证数值稳定性。
在缺失结果界限、替代解析规则和声明排除项下重算；不得依据结果符号选择规范。
导出估计、区间、簇图、分母流、代码和环境版本、种子及协议偏离。

6.7 裁判标定也是子群问题

裁判总体准确率可能掩盖严重子群失败。在一个受控 GEO 检测基准中，词级 TF–IDF 检测器报告总体 F1 为 0.880，最差组准确率却为 0.375；干预感知 ModernBERT 配置的 F1 为 0.944，最差组准确率为 0.883 [9]。数值属于该基准，不能认证已部署检测器。它们说明标定报告为何应包含分类别混淆矩阵、最差组表现、假阳性差距、弃判和独立裁定样本，而不只报告一个汇总分数。

6.8 三种实验环境

6.8.1 固定上下文实验

研究者直接提供候选文档或段落。这种环境对生成与归因有较强控制，适合段落消融、顺序效应和内容使用检验，但不能说明修改后的页面会从大规模语料中被检索到。

6.8.2 重建的端到端系统

研究者控制语料、检索器、重排序器、上下文构建器和生成器，因而能够记录阶段轨迹并反复构造反事实。这适合比较算法与调试机制。其外部有效性取决于重建环境与目标生态的相似程度。明确纳入检索和重排序的基准，回应了固定候选设计的部分局限 [6, 23]。

6.8.3 真实平台观察与干预

研究者向生产界面提问。生态有效性较高，但内部状态、漂移、个性化和并行更新更难控制。真实平台前后比较，应在可能时使用同期或交错对照。研究不得违反平台条款、访问控制、隐私要求或披露义务。

6.9 反事实、消融与干扰

留一信源检验比较 $y(C)$ 与 $y(C \setminus \{s\})$；留一主张检验移除特定证据单位。它们可在受控上下文中揭示信源影响，但自然语言输出使距离函数的选择后果重大。报告应包含主张级差异，而不只是嵌入相似度分数。

表 23: 三种环境最能可信识别的内容。

环境	最强用途	主要局限
固定上下文	因果信源使用与生成比较	不证明检索资格
重建流程	组成部分轨迹与端到端基准效应	向专有系统迁移的不确定性
真实平台	实际界面在特定时间的行为	隐藏混杂、漂移、成本和有限可复现性

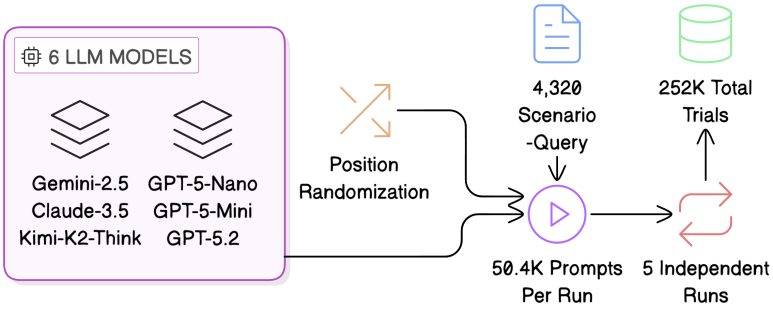


图 18: 竞争性首次引证实验中的嵌套执行规模：6 个模型、情境—查询单元、随机信源位置、重复运行和 252,000 次试验。图中明确了哪类重复会、哪类不会创造新情境。

来源说明。取自 Vishwakarma et al. [49, 图 2]，PDF 实际第 3 页的矩形裁图，许可为 CC BY 4.0。每次试验恰好接受 2 份给定信源文本，且只建模首次引证；这不是开放网页检索，也不是生产引擎流量实验。

干扰是竞争性可见性的核心。修改信源 s_i 可能改变信源 s_j 是否被选择、引证或信任。稳定单位处理值假设常常不可信。竞争实验应报告全部上下文组成与替代效应，而非只报告目标增益。近期竞争性 GEO 和平台—供应方反复适应研究，将这种生态效应设为核心研究问题 [49, 59]。

6.10　决策规则与诚实停止

实验应在数据采集之前声明三项阈值：

1. 主要结果上的**最小有价值效应**；
2. 忠实性与用户效用的**非劣效界值**；
3. 对无依据或操纵性内容的**最大可接受风险**。

即使提及增加在统计上可检测，如果归因准确率下降超过界值，也不是成功。高噪声的零效应结果不是无效应证明。无结论研究应触发关于增加数据、改进对照或停止干预的决定，而不是编造叙事。

复现实作 6 ・设计测量协议

输入。1 个信源、1 项获准内容干预、冻结查询全集，以及受控流程或条款允许研究的真实界面。
协议设计。采集前写明目标总体、抽样框、查询意图簇、分配单位、观察单位、分析单位、目标估计量及其单位、重复安排、系统/时间单元、最小有价值效应、约束指标与缺失政策。冻结阳性和阴性对照，以及留出意图划分。
测量设计。每个计划评估单元建立一行。将采集状态与结果分开保留。对预声明分层标签样本双人编码，在保留分歧的前提下裁定，并按重要子群报告裁判表现。
分析设计。核对分母流；计算声明的簇估计、区间、异质性展示和缺失结果敏感性；再执行协议预先写明的替代解析器与排除分析。不得因为结果有利而追加分析。
交付。带时间戳的协议、信源/查询/系统清单、原始与裁定后台账、可执行分析、生成表图、偏离日志，以及包含

图 19: 六层测量协议。前五层明确之后，才开始制作看板。

不确定性、约束指标和失败案例的一页决策备忘录。

停止规则。对照失败、系统状态无法记录版本，或干预违反证据或风险条件时，停止或重新设计。

6.11 复现检查点

只有第二位分析人员能从发布台账而非截图出发恢复结果，实作才可复现。在认为本章测量协议完成之前，应逐项验证：

1. 信源版本、查询清单、簇分配、系统卡、协议、分析环境和允许保留的原始回答，均具有稳定 ID 和密码学哈希；
2. 计划、返回、有效、标注、解析和裁定计数，按组、簇、系统和时间区组精确对账；
3. 干净运行从发布台账重算每项估计、区间、表格和图形，没有手工抄录图表数值；
4. 带种子的重抽样结果增加重复次数后仍稳定，独立编写的计算复现点估计及各簇差值方向；
5. 未解决标签、服务中断、拒答、排除和协议偏离仍可见，缺失界限与替代解析规则复现报告的敏感性范围；
6. 即使得出无结论或负面决定，决策备忘录仍应用冻结的有价值效应、忠实性、效用和风险阈值。

通过这些检查，只建立声明研究的计算可追溯性，并不建立向不同平台、查询总体、地区、模型状态或时间窗口的可迁移性。

6.11.1 检查点的干净运行探针

这里用随书 L06 重复测量测试样例演练上述抽象检查点。构建先在干净临时目录调用实作分析器，验证其声明输出。另一个仅使用标准库的程序读取冻结事件台账，在每个查询内形成配对界面差值，对整个查询簇重抽样，再生成绘图表和含哈希的运行清单。运行采用 10,000 次自助重复，种子为 20,260,825；抽到一个查询时，两个界面及其全部重复共同进入样本。

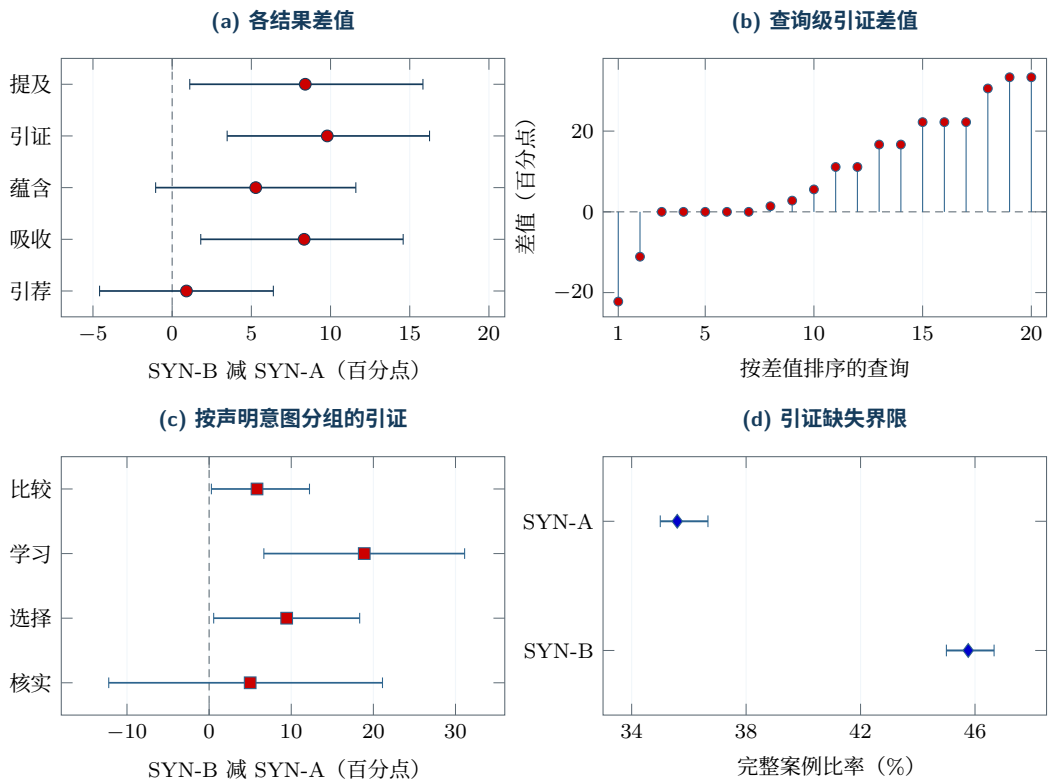


图 20: 冻结合成重复测量面板的确定性本地复现。台账含 20 个查询簇中的 360 个设计事件，其中 354 个完整、6 个非完整。(a) 与 (c) 给出等权查询均值差及 10,000 次带种子聚类自助重复的百分位区间；这些区间描述人为设计的查询面板，并非总体置信区间。(b) 保留全部查询级引证差值，包括负值与零。(d) 展示完整案例引证率，以及将全部非完整结果赋为 0 或 1 得到的二元结果逻辑范围；这些界限不是抽样区间。

该合成面板的引证差值为 9.79 个百分点，带种子的查询簇区间为 3.47–16.25 个百分点。汇总保留 2 个负向和 5 个恰为零的查询差值。“核实”意图区间跨越零，整体蕴含和引荐区间也跨越零。首个至最后一个时间区组，引证率在 SYN-A 下降 2.87 个百分点，在 SYN-B 下降 2.48 个百分点。这些零效应与负向诊断是必须保留的结果，不是事后需要删除的缺陷。

干净运行只建立计算可追溯性。数据是确定性合成教学样例，SYN-A 和 SYN-B 是虚构标签，20 个人为设计查询不是需求的概率样本。因此，该图组不估计任何具名产品、真实引擎、因果干预、业务结果或总体比例。其复现身份为冻结面板 SHA-256 `c013f9ae…5d7`；完整哈希、环境、脚本身份、种子、输出哈希与主张上限，在图片清单和运行清单中接受机器检查。

本章小结

1. 采集回答前，定义目标估计量、目标总体、比较、结果单位与推断单位。
2. 提及、引证、吸收、检索、稳定性、业务和风险指标单位不同，不能盲目合并。
3. 重复回答是嵌套观察。更多生成可以估计查询内随机性，却不会增加独立抽样的查询意图数量。
4. 缺失回答是状态，不是负面结果。抽样区间与缺失结果敏感性回答不同问题，应共同报告。
5. 裁判质量既是总体属性，也是子群属性；未解决案例与裁定必须可审计。
6. 固定上下文、重建与真实平台研究识别不同效应，提供不同控制和外部有效性。
7. 成功要求：按照冻结决策规则，主要效应具有价值，且忠实性、效用与风险结果均可接受。

7 实验、因果推断与时间不确定性

核心思想

观测到的可见性变化，只有相对于明确声明的处理、分配机制、比较条件、目标估计量及一组假设，才具有因果意义。生成式回答系统还带来重复的随机输出、平台漂移、信源间干扰及缺失单元；一组前后截图无法区分这些过程。

7.1 本章回答的问题

Q1. 当信源被修订、生成系统却仅部分可观测时，究竟分配了什么、观测了什么、估计了什么?

Q2. 哪种比较设计能将干预与查询构成、系统漂移、索引传播及通常的随机输出波动分开?

Q3. 重复运行、查询意图簇、时间块及缺失单元应如何进入估计量及其不确定性?

Q4. 哪些诊断能够质疑，但永远不能证明，平行趋势、无残留效应、测量稳定与不存在干扰等假设?

Q5. 何时必须将看似积极的结果报告为描述性、部分识别或结论不确定，而非因果结果?

7.2 定义处理版本与潜在结果

令 $D_i \in \{0,1\}$ 将评估单位 i 分配到基线信源状态 $\mathcal{W}^0$ 或允许的干预状态 $\mathcal{W}^1$。对版本化系统状态 S_t，定义潜在结果

$$Y_i(d, S_t), \qquad d \in \{0,1\}. \tag{44}$$

观测结果为 $Y_i = Y_i(D_i, S_t)$。固定 S_t 下，有限样本平均处理效应为

$$\tau(S_t) = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i(1,S_t) - Y_i(0,S_t)\right]. \tag{45}$$

显式写出系统状态，是因为索引更新、模型修订、用户界面变化或信源语料变化，都可能改变两种潜在结果。如果处理与对照在不同时段观测，还需要额外假设。

处理包含完整内容版本，不只是“添加统计数据”这样的标签。记录主张变更、标记、链接证据、渲染输出、内容哈希、发布时间、传播延迟及回滚状态。复合干预识别的是组合处理，除非通过析因或消融设计分离组件。

7.3 收集结果前先写明识别约定

只有把可观测记录与由假设补入的潜在量区分开，因果分析才可复现。对簇 c、时间块 b 与重复 r，最低观测单元可写为

$$O_{cbr} = (D_{cb}, X_c, t_b, H_b, U_{cb}, R_{cbr}, Y_{cbr}, G_{cbr}, J_{cbr}), \tag{46}$$

其中 D 为分配的处理，X 为冻结的簇描述，H 为系统状态指纹，U 为测得的处理实际接受情况，R 指示计划结果是否观测到，Y 为主要结果，G 为护栏向量，J 标识评判器或标注状态。反事实结果 $Y_{cbr}(d, S_t)$、未记录的检索状态、潜在查询需求、未测量的竞争者变化和无误差的评判标签，不会因为归档里存在一行记录就变成可观测量。

这些假设决定识别上限。状态指纹可以包括可观测的产品或模型标签、界面模式、地区语言、搜索设置、收集代码版本与时间戳，但它不证明所有隐藏系统组件稳定。当发生明确的状态断裂时，应划分时期或停止分析；悄悄跨越断裂求平均，会改变目标估计量。

表 24: 最低识别约定。诊断可以暴露失败，但通过诊断不能证明假设成立。

假设	操作含义	预先声明的检验或诊断
一致性与版本	分配映射到一个已记录的信源状态，而非未命名的编辑家族	内容哈希、渲染差异、处理接受检查及版本专属排除规则
可交换性	分配或比较设计打断处理与未处理潜在结果之间的关联	随机化审计、基线平衡、分配隐藏或显式调整集
测量稳定	收集、解析与评判在各实验臂和时段中含义相同	冻结提示词及解析器、盲法重标注、状态指纹及测量工具变化时期
无预期或残留效应	单元在启用前或声明的洗脱期后不受处理影响	传播延迟研究、提前效应、洗脱期敏感性及处理历史字段
比较趋势有效	若无处理，处理组与对照组聚合结果应按可比趋势演化	多个处理前时期轨迹、安慰剂日期、事件研究对比及有界趋势偏离分析
有限干扰	分配给某信源的变化仅通过声明的曝光映射影响其他单位	候选集快照、溢出结果、曝光映射与挤出分析
正值性与流失控制	每个目标分层均可接受所需条件，且是否被观测不由处理后事件筛选	支持域表、计划单元流转、各臂缺失情况及最坏情形界限

表 25: 面向不同因果问题的设计类别。

设计	最可信的目标	核心失败条件
固定上下文随机化	信源表达对使用／生成的效应	不支持端到端检索主张
冻结的重建流程	开放系统中的组件及端到端效应	向封闭界面迁移不确定
同期页面或信源实验	特定时间的在线界面效应	污染、传播、政策或分配失败
交替切换或交错设计	可逆的短窗口比较	残留效应及需求非平稳
双重差分	与未处理单位相比的变化	平行趋势被破坏或存在溢出
中断时间序列	日期明确事件后的水平／斜率变化	同期冲击与处理前时期不足

7.4 选择与主张相匹配的实验环境

固定上下文研究对可用证据控制较强，却不适合推断网页获取。重建流程能够暴露每个阶段，但仍只是目标环境的模型。在线观测提高生态相关性，同时隐藏内部处理接受情况。报告应说明处理能够到达哪个阶段，以及哪些结论仍超出设计范围。

三类现有证据设计说明了相应的主张上限。252,000 次试验的引用选择研究对两个给定信源进行顺序平衡，估计首个显示 URL，却从不调用网页检索 [49]。SAGEO Arena 为 2,700 个查询重跑包含 171,003 个页面的检索—重排序—生成流程，但从基线前十名选择目标，且未报告运行层面的置信区间 [23]。一个在线四引擎面板在 45–46 天内重复 32 个德语提示词，却不能把模型解码与路由、检索、缓存及产品漂移分开 [40]。三者不是彼此的弱化版本，而是识别了不同对象。

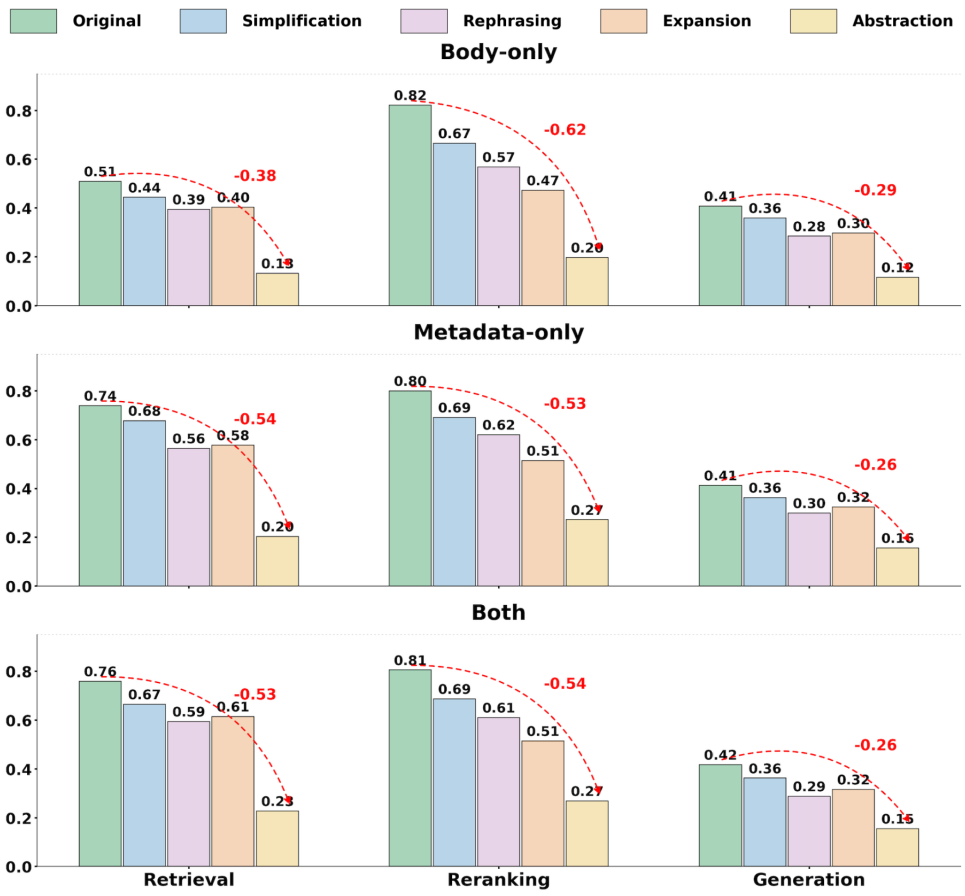


图 21: 查询变化对检索命中、重排序命中和生成引用率的影响可以不同。复用自 Kim et al. [23, 图 7]，依据CC BY 4.0 使用，仅作裁切。结果限于该论文重建语料、目标准入条件、策略及扰动，不是普遍平台规律或因果效应。

7.5 在正确的单位上随机化和分析

表达同一意图的查询不是独立实验单位。同一查询下的重复运行，是簇内测量。如果按页面、域名、查询意图或时间块分配处理，不确定性就应尊重该分配单位。

对有 n_c 个观测的簇 c，令 $\bar{Y}_c(d)$ 为处理 d 下的平均结果。透明的簇级估计量为

$$\widehat{\tau}_{\mathrm{cl}} = \frac{1}{C_1} \sum_{c:D_c=1} \bar{Y}_c - \frac{1}{C_0} \sum_{c:D_c=0} \bar{Y}_c. \tag{47}$$

随机化推断或簇自助法可以提供与分配方式一致的区间。混合效应模型能够提高精度、描述异质性，但模型假设不能修复有缺陷的分配。

例题 7.1: 五次重复不等于五个独立查询

假设 30 个查询意图分别在三个系统上运行五次。归档中有 450 份响应，但处理是在 30 个意图的层面分配的。把 450 行视为独立观测，可能使区间过窄。主要不确定性应重抽样或建模这 30 个簇；五次运行估计意图内随机波动，三个系统则定义分层或带声明权重的目标分布。

表 26: 合成的两期干预审计。每个单元包含 120 个嵌套于查询意图簇内的计划观测，结果是在冻结编码规则下忠实且归属于信源的回答。这些计数是教学构造，不是生产平台测量。

组别	信源状态	处理前成功数	处理后成功数
同期对照	未变更，已核实版本	$30/120 = 0.250$	$36/120 = 0.300$
干预组	基线，随后进行治理修订	$24/120 = 0.200$	$45/120 = 0.375$

7.5.1　在三类方差来源之间分配精度

查询或文档抽样、流程重复运行与系统漂移，是不同的不确定性来源。增加同一查询调用可估计运行方差，却不会产生新查询簇，也不会观测到更晚的产品状态。对近似相等的簇大小 $\bar{m}$，常见设计效应诊断为

$$\mathrm{DE} = 1 + (\bar{m} - 1)\rho \tag{48}$$

它说明，当簇内相关 ρ 为正时，增加重复次数获得的独立信息价值递减。功效规划应先用试点估计查询簇方差、运行方差、基率与漂移窗口，再模拟真实分配方式和缺失政策。该近似及混合效应模型，都无法弥补独立分配簇过少的问题。

7.6　区分时间漂移与处理实际接受

令 G_i 表示处理组，T_t 表示干预后时期。两期双重差分对比为

$$\widehat{\tau}_{\mathrm{DiD}} = (\bar{Y}_{G=1,T=1} - \bar{Y}_{G=1,T=0}) - (\bar{Y}_{G=0,T=1} - \bar{Y}_{G=0,T=0}). \tag{49}$$

只有在未处理时平行趋势、无预期效应、测量稳定，以及恰当处理干扰等假设下，它才识别因果对比。应绘制处理前轨迹和安慰剂日期；不要把公式当作装饰性调整。

例题 7.2: 从 17.5 个百分点变化到 12.5 个百分点的受控对比

干预组未经调整的前后变化为 $0.375 - 0.200 = 0.175$，即 17.5 个百分点。同一窗口内，未修改对照上升 $0.300 - 0.250 = 0.050$。在声明的双重差分假设下，

$$\widehat{\tau}_{\mathrm{DiD}} = (0.375 - 0.200) - (0.300 - 0.250) \tag{50}$$

$$= 0.125. \tag{51}$$

隐含的干预组处理后未处理反事实为 $0.200 + 0.050 = 0.250$，所以同一对比也可写成 $0.375 - 0.250 = 0.125$。因此，观测到的 17.5 个百分点上升中，5.0 个百分点归于同期背景趋势，12.5 个百分点归于干预，但仅限于识别约定成立时。单元计数允许计算点估计；不确定性仍须在查询意图分配层面计算，而非将 480 行当作独立伯努利观测。

平行趋势是反事实条件，不是可以直接检验的等式。令 δ 表示干预组未处理趋势减去对照所代表的未处理趋势。算例中经敏感性调整的效应为

$$\widehat{\tau}(\delta) = 0.125 - \delta. \tag{52}$$

例如，未处理差异趋势为 $+0.05$ 时，对比降为 0.075；为 $+0.125$ 时，可以解释整个观测对比。负值表示对照趋势比干预组的未处理趋势更有利。

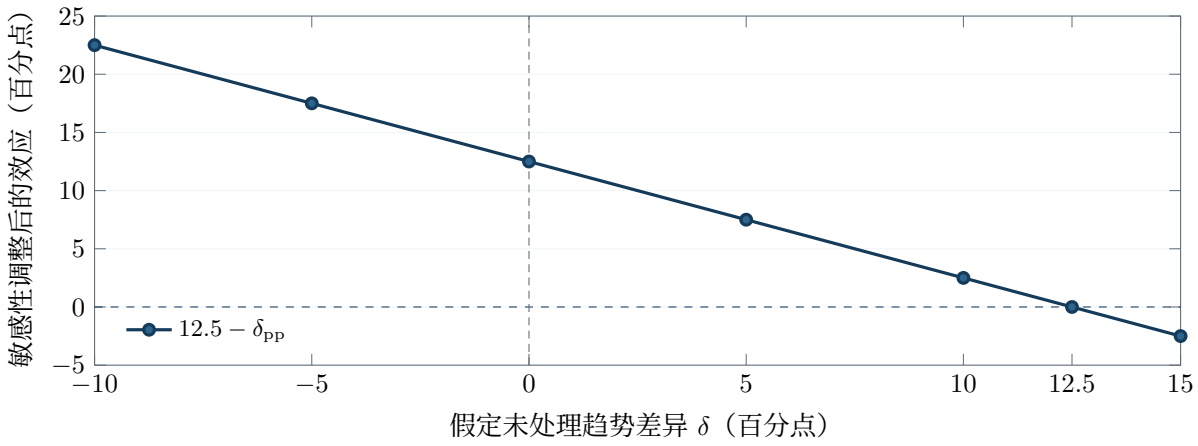


图 22: 可重新计算的趋势偏离分析，对应表 26。图值遵循 $100\hat{\tau}(\delta) = 12.5 - 100\delta$，并非实证置信区间。在 12.5 个百分点处发生符号反转，使识别假设直观可见，而非隐藏于脚注。

命题 7.3: 差异漂移不受约束时，两期双重差分无法点识别处理效应

令 κ 表示观测双重差分对比，τ 表示因果效应，δ 表示干预组未处理趋势与对照趋势之差，则 $\kappa = \tau + \delta$。当 δ 不受约束时，仅凭观测组均值不能识别 τ。如果设计能够为 $|\delta| \le b$ 提供依据，则 τ 在 $[\kappa - b, \kappa + b]$ 内获得部分识别。

证明. 将观测到的干预组变化写成未处理变化加 τ。减去对照变化，得到 $\kappa = \tau + \delta$。对于任一候选效应 $\tau^\star$，令 $\delta^\star = \kappa - \tau^\star$，均可产生同一观测对比，因此组均值不能区分这些组合。在 $|\delta| \le b$ 下，有 $-b \le \kappa - \tau \le b$，整理为 $\kappa - b \le \tau \le \kappa + b$。 □

多个处理前时期能够揭示明显不相容的轨迹，并为有界的 b 提供支撑，但不能观测处理单位在处理后未接受处理的轨迹。报告平行趋势点估计之前，应报告完整敏感性区间、科学依据，以及决策是否随之改变。

纵向 GEO 测量研究强调重复观测与时间变化 [40]。这支持具有时间意识的协议，而非普遍抽样间隔。抽样频率应反映预期产品变化、索引延迟、结果波动、成本及最低决策时间跨度。

7.6.1　仅在研究时间尺度上可逆时使用交替切换设计

交替切换设计（switchback）把条件分配给时间块，而非永久分配给单位。当同期随机化不可行时，它适用于可逆的固定上下文处理或受控流程配置。如果索引、缓存或学习式适应在名义条件切换后仍可能持续，就不宜默认用于公开页面修订。应在匹配的时间块对内随机化顺序，平衡星期与时段分层，声明洗脱期，并为每个单元保留处理历史。对配对 p，令 $\bar{Y}_{p1}$ 与 $\bar{Y}_{p0}$ 表示两种条件下按簇加权的观测结果。透明的配对估计量为

$$\widehat{\tau}_{\mathrm{SB}} = \frac{\sum_{p=1}^{P} w_p (\bar{Y}_{p1} - \bar{Y}_{p0})}{\sum_{p=1}^{P} w_p}, \tag{53}$$

其随机化推断依据允许的配对内分配翻转。不得看过结果后再排除转换时间块或不利状态。应应用预注册洗脱期，否则将比较标为受到残留效应污染。

7.7　将缺失定义为结果或删失过程

拒答、超时、禁用搜索的响应、无引用面板、解析失败与政策阻断，是不同状态。计划结果被观测到时令 $R_i = 1$。完整案例目标估计量以 $R_i = 1$ 为条件；当缺失依赖处理、查询或系统时，它可能有偏。协议应规定每种状态属于：

- 有效结果，例如拒答率；
- 具有有限重试规则、可恢复的运行错误；
- 在显式假设及敏感性分析下处理的删失；或
- 使比较失效的协议失败。

绝不能重试直到出现偏好的结果。保留首次尝试、重试原因、次数、间隔及所有终止状态。

7.8 考虑评判误差与多重结果

如果处理效应来自自动标签，评判器校准就是测量模型的一部分。人工校准样本应对处理条件与预期方向设盲，并按系统、语言、主张类型及模型置信度分层。在合理误差矩阵下重新计算处理决策。

GEO 干预通常报告提及、引用、吸收、忠实度、显著程度、用户效用与风险。应声明一个主要结果或正式决策规则，指定护栏，并对验证性主张控制多重性。事后从众多结果中挑出任一“胜出”项，只构成探索性证据。

7.9 将竞争与适应视为干扰

改变信源 i，可能改变信源 j 可获得的检索或上下文容量。因此潜在结果可能是

$$Y_i\big(D_i, \mathbf{D}_{-i}, S_t\big), \tag{54}$$

而不仅是 $Y_i(D_i)$。竞争性 GEO 研究与重复策略学习系统，都把这种依赖放在核心位置 [49, 56]。实验应记录完整候选和上下文构成，报告挤出效应，而非仅报告目标信源的增益。

7.10 开展漂移感知分析，不以成功作为筛选条件

处理实际接受常不完全或有延迟。由于它可能受分配和系统状态影响，删除 $U_{cb} = 0$ 的单元，或用观测到的实际接受替代分配，通常会改变随机化目标估计量。因此，主要分析应保留意向处理对比；按实际接受调整或按方案分析的估计属于次要结果，需要各自的识别假设。同样，结果到达后发现状态断裂，也不意味着可以选择更有利的时期。

算法 5 漂移感知的干预分析

输入： 预注册的簇 $\mathcal{C}$、时间块 $\mathcal{B}$、分配 D、信源哈希、主要结果、护栏、状态断裂规则、缺失政策及分析代码
输出： 版本化单元台账、尊重分配的效应估计、诊断、敏感性包络、偏离与决策

```
冻结簇权重、目标估计量、多重性规则、重试预算、洗脱期、排除规则和停止阈值
收集结果前建立所有计划 (c, b, r) 单元
对所有 按时间顺序排列的 b ∈ B 执行
    记录收集前的可观测状态指纹 H_b^pre
    对所有 计划单元 (c, b, r) 执行
        收集首次尝试，保留时间戳、原始响应、归因、轨迹及终止状态
        核实分配条件、实际送达信源哈希及接受情况 U_cb；即使未接受处理也保留原分配
        按冻结缺失规则分类拒答、超时、政策状态、面板缺失及解析失败
        应用盲法结果与护栏规则，追加评判状态 J_cbr 及候选／上下文构成
    结束 对
    记录 H_b^post，将前后指纹与当前时期比较
    若 任一指纹越过预注册的硬性断裂规则 则
        关闭当前时期，执行声明的停止或分期动作
    结束 若
结束 对
锁定排除项，按臂、块、簇及状态时期核对计划、已收集、已观测和已分析单元数量
估计效应前审计分配、基线支持域、处理版本、实际接受、流失与污染
在每个可比时期内，以分配单位估计主要意向处理对比
计算随机化区间或考虑聚类的区间；仅按预注册目标权重聚合不同时期
运行提前效应、安慰剂日期、负对照、残留效应及候选挤出分析
在声明的趋势漂移、缺失结果、评判误差、实际接受与溢出界限下重新估计
应用多重性及护栏规则，将结果标为保留、回滚或结论不确定
连同版本化主张上限发布清单、单元台账、代码、诊断、敏感性结果及全部偏离
```

适用边界与失效条件

统计显著性不能挽救未识别的比较。在线前后增长可能反映处理、平台漂移、查询需求变化、索引传播、竞争者编辑、重试规则或评判器漂移。应声明因果假设，展示使各假设具有合理性的诊断；否则只能报告版本化关联。

7.11 复现检查点：重建每一个分母

只有未参与研究构建的分析者能从冻结材料恢复计划总体、处理版本、分析总体与决策，才算通过检查点。仅从清洗后的表重算最终系数并不充分；练习须暴露那些变为拒答、超时、解析失败、残留效应排除或处理接受未观测的计划单元。课程默认使用重建或合成环境。在线界面是可选项，也不会使允许主张超出其观测时间、界面、查询分布及版本化条件。

复现实验室 7：预注册并分析漂移感知干预

问题。范围明确且保全证据的改变，能否改善主要可见性结果，同时不降低忠实度或增加风险？
输入。三十个查询意图簇；基线、干预及同期对照信源清单；四个时间块，每个“簇—条件”计划重复五次；固定标注指南；处理与状态哈希；带种子的合成漂移／缺失注入器。
预注册。冻结单元 ID、分配种子、簇权重、主要意向处理目标估计量、最小可检测效应、缺失状态及重试政策、一

个忠实度护栏与一个风险护栏、多重性规则、漂移阈值、洗脱期、停止规则和决策表。不向分析者透露注入的漂移与缺失参数。

执行。保留每次首次尝试。生成从计划到分析单元的 CONSORT 式流转图；核实送达哈希及实际接受；在簇级分析；层级模型仅作为事先声明的敏感性分析。绘制处理前轨迹和效应区间。运行安慰剂日期、负对照结果、残留效应检查、干扰／挤出审计，以及差异趋势、缺失结果和评判误差的敏感性网格。

必需输出。预注册文件；版本与环境清单；原始单元台账；不可变的排除／偏离日志；分析代码和种子；计划／收集／观测／分析流转；效应与区间图；缺失及实际接受表；盲法评判校准记录；诊断结果；以及保留、回滚或结论不确定的决策。

复现测试。第二名分析者须：(i) 从单元 ID 重建每个分母；(ii) 从表 26 恢复描述性干预变化 0.175 与受控对比 0.125；(iii) 恢复 $\delta = 0.05$ 时的 0.075 及 $\delta = 0.125$ 时的零交点；(iv) 展示区间使用查询意图簇而非数据行；(v) 对一个注入的状态断裂或护栏越界触发预先声明的响应。

停止规则。无法核实处理实际接受、对照被污染、单元台账无法核对、严重护栏危害超过阈值，或系统状态离开声明的可比窗口时，停止或重新设计。诊断失败应产生结论不确定或更窄的主张，而不是未注册的替代分析。

本章小结

1. 因果结果需要版本化处理、分配机制、潜在结果、目标系统状态与总体、目标估计量及明确识别约定。
2. 观测单元包含分配、时间戳、状态指纹、实际接受、结果、护栏、缺失与评判状态；反事实结果及隐藏机制仍为潜在量。
3. 实验环境限制能够识别的阶段与总体。固定上下文、重建流程、在线同期、交替切换与纵向设计回答不同问题。
4. 重复是嵌套测量。随机化、估计、重抽样与自由度须尊重查询、信源或时间块分配，不得把响应行当作独立单位。
5. 算例中原始 17.5 个百分点上升，在扣除同期 5 个百分点趋势后，成为 12.5 个百分点的双重差分对比。因果解释仍依赖平行趋势及其他声明假设。
6. 差异漂移不受约束时，处理效应不能点识别。具有科学依据的漂移界限给出敏感性区间；仅凭前趋势诊断不能证明反事实趋势。
7. 交替切换要求处理可逆、块顺序随机以及可辩护的洗脱期。缺失、重试、评判误差、多重性、实际接受、残留效应与干扰都是设计部分，不是清洗细节。
8. 漂移感知发布应包含计划单元核对、状态分期、意向处理估计、尊重分配的区间、负对照与安慰剂、护栏、敏感性包络及所有偏离。识别失败时，应报告版本化关联或结论不确定的实验，而非因果叙事。

8 内容与证据工程

核心思想

负责任的内容优化，让经验证的信息更易检索、理解、归因和更新，而不制造权威。工作的基本单位是受治理的主张及其证据路径，不是在流畅文案中插入关键词。

8.1 从素材库到事实系统

组织常从演示文稿、营销文案、产品页、媒体报道和分析师报告起步。这些素材服务于不同日期、受众和审批制度。扩大传播可能放大冲突：两种价格、过期认证、仅适用于某地区的能力，或脱离实验背景的客户效果。

章 2 的主张—证据—信源图，通过事实台账转化为可执行流程。每行包括：

- 规范实体标识符与便于阅读的名称；
- 最小谓词与取值；
- 时间区间、地区、套餐、受众及其他条件；
- 证据片段、信源身份及独立性状态；
- 责任人、审核人、批准状态与下次复核日期；
- 已知矛盾、被替代取值和纠正历史；
- 允许公开使用的表述与禁止的夸大表述。

台账默认不是公开知识图谱，而是治理层。页面、文档、数据馈送、结构化数据、媒体回应和评估测试样例，均可从中一致生成。

观察 8.1: 完整性有类型，不是单一标量

信源在文档层面可能完整，在与答案相邻的段落层面却不完整；可能包含所有产品属性，却省略指标与测试条件之间的关系；可能是现行版本，却缺乏独立支持。一个“内容完整率”百分比会掩盖这些失效模式。

8.2 页面是证据界面

精心设计的页面将人的理解、机器获取、段落检索和治理相互协调。表 27 描述的是可复用的证据页面，而非保证排名的模板。

设计原则是**限定条件就近呈现**：数值及其单位、总体、日期和例外，应能在段落切分后共同保留。这既是良好的科学写作，也是良好的检索组织习惯，不是在要求机械重复同一个短语。

8.3 利用结构信号，避免结构迷信

标题、列表、表格、图注、定义和问答块，可以让信息更易导航和切分。近期论文在不同 GEO 设置下研究内容结构、协同优化与信源影响 [6, 57, 61]。但这些研究的存在，并不支持要求所有页面采用同一格式的通用清单。

结构至少可能通过四条路径起作用：

1. **获取：**关键信息以可解析 HTML 呈现，并有稳定标识符；
2. **检索：**标题和局部语境改善段落语义；
3. **生成：**主张、证据和条件能够以更少歧义组合；
4. **归因：**信源身份和证据边界仍可解析。

表 27: 证据页面的组成部分及其诊断用途。

组成部分	对读者与治理的作用	对检索与评估的作用
身份页首	标明实体、责任方、版本及最近实质性复核	消歧别名，区分现行页面与被替代页面
答案前置摘要	陈述限定范围的命题及主要条件	形成连贯段落，不把例外藏在别处
证据表	将取值与单位、方法、日期、来源配对	支持主张级抽取与比较
方法与局限	说明结果如何产生及其失效边界	防止无依据迁移，提供负面证据
定义	消歧技术与商业术语	稳定查询与实体匹配
变更日志	呈现纠正与版本转换	支持新近性和时间审查
信源清单	解析原始证据与独立证据	支持归因和循环引证检查
下一步行动	提供文档、数据、联系或纠正路径	区分信息可见性与已声明业务结果

表 28: 三项内容干预研究的目标估计量不可互换。

研究与受控对象	本讲义保留的已报告比较	负面结果与可接受的结论上限
E-GEO：缓存的 10 项列表中一个被改写产品 [3]	在 5 个重排序器上，简单提示的配对排名变化行均值随改写器变化：Claude 为 +0.57，GPT-5 为 +0.56，GPT-4o-mini 为 −0.36。	15 个手写提示中 11 个低于 GPT-4.1 简单提示均值；故事化表达为 −4.36。处理从检索之后开始，不涉及进入候选集、真实购物、点击或销售。
SAGEO Arena：重建的检索—重排序—引证流程 [23]	在 171,003 篇文档和 2,700 个查询上，三个阶段中仅改正文的分数为 .53/.84/.47，仅改结构为 .71/.83/.52，StageAware 为 .75/.80/.58。	增益不在同一阶段达到最高；报告的一项质量评分从 88.9 降至 87.3。这支持在该基准中进行阶段感知评估，不保证结构化页面的生产效果。
FeatGEO：固定的 5 信源上下文加 1 个广告主页面 [26]	在 GPT-4o-mini、Gemini-2.5-Flash 和 Qwen-plus 上，可见性从未优化时的 13.34/8.89/5.20，变为报告中多目标方法的 18.31/15.35/10.17。	多数单一启发式低于未优化条件；特征消融随引擎和领域变化。自动可见性与质量裁判不能确立人工感知质量、事实忠实性或真实平台效果。

评估应隔离候选作用路径。如果问答式改写改善了人工浏览，却未在某项平台研究中改善吸收，正确结论应针对具体路径，而非“常见问题格式无效”。当前论文库中的选择—吸收研究，恰好报告了其协议下问答排版的此类负面结果 [65]。

8.3.1 以阶段条件化证据替代技巧清单

现有文献包含反驳通用写作配方的直接反例。表 28 将干预起点、负面结果和主张上限，与每项正面结果一同保留。

用户提供文献中的一篇结构研究报告引证率提高 17.3%，但未充分协调其查询/案例数量、统计单位、平台版本、重复次数、评估量表或复现材料 [61]。本讲义将该数值隔离使用：它可以启发预注册的结构消融实验，却不被接纳为设计规则或实证主标题。

8.4 干预类别与约束

早期 GEO 研究评估了风格和证据相关修改，包括引证、引语和统计数字，并报告不同领域和方法间的异质性效应 [1]。后续工作将设计空间扩展到以内容为中心的基准、意图感知改写、结构特征、多目标方法、智能体系统和轨迹感知证据生态 [6, 8, 26, 49, 60, 61, 63]。在实践中，可按允许改变的对象划分这些方法。

图 23 的因素—模型矩阵，是反驳通用技巧清单的有用例子。排版比较在 6 种受评配置间没有稳定显著性，而其他因素的一些极大优势比点估计，有时伴随退化或奇异拟合。科学上能够接受的启示是异质性与诊断的必要性，而非可随处套用的“最佳 GEO 因素”排名。

解释边界 • 固定引证选择不等于端到端可见性

该实验不调用真实搜索引擎，只建模与两个注入信源之一匹配的首个 URL，并删除约 3.4% 没有 URL 或首个 URL 不匹配的运行。模型快照、API 日期、温度、种子、单元级有效样本量和不确定性区间均未锁定。论文没有对 108 次因素—模型检验报告多重性校正。因此，该矩阵不能确立抓取、索引、检索、后续引证使用、事实质量、用户效

Factor (A vs B)	Diverse Providers			GPT Models		
	Gemini 2.5 Flash	Claude 3.5 Sonnet	Kimi K2 Thinking	GPT 5 Nano	GPT 5 Mini	GPT 5.2
On-Topic vs Off-Topic	**>10k**	**>10k**	**>10k**	**221**†	**>10k**	**>10k**
Query Terms vs Missing	9.41	17.0†	**16.4**	**5.99**	**14.4**	**40.0**
Price vs No Price	**>10k**	**>10k**	**36.1**	**7.82**	**6.26**	**30.4**
Specs vs No Specs	8.63	**>10k**	**238**	**11.5**	**15.4**	**243**
With vs No Comparisons	**7.45***	**4.55***	**4.72**	**2.14**	1.78*	1.61
Confident vs Hedged	**599**	754	**10.6**	**2.67***	**5.44**	**4.75**
Evidence vs No Evidence	8.00*	**>10k**	**46.3**	**2.73**	**5.91**	2.09
Consistent vs Contradictory	2.81	**2.81**	**2.72**	**2.19**	**4.09**	1.74
Neutral vs Promotional	1.45*	362*	2.02	1.31	**1.81**	**2.08***
Structured vs Dense	1.68	1.03†	0.79	0.90	0.78	1.25
Organized vs Scattered	2.21	3.87	2.49	1.19	1.13	1.57
Strong vs Weak Value Prop	5.22	5.66	**7.24**	**2.79**	**2.68**	1.53*
Deep vs Shallow Coverage	1,480†	**>10k**	**132**	**5.33**	**19.5**	**3.98**
Strong vs Weak Social Proof	6.61*	>10k	**25.4**	**3.13**	**4.78**	2.14
Position 1 vs 2	**>10k***	**>10k**	**>10k**	**2,002**	**1,795**	**>10k**
Recent vs Old Timestamp	**>10k**	**>10k**	**68.7**	**14.4**	**1,494**	**>10k**
No vs Old Timestamp	2.32	2.28	**1.33***	1.31	1.55	**1.48***
Recent vs No Timestamp	1.99	13.0	**3.67**	1.15	**3.53**	**1.99**

Bold = statistically significant ($p < 0.05$). OR > 1 favors variant A.
Convergence warnings: *Degenerate Hessian, †Singular fit.
Effect Size Categories: Very strong (OR > 100), Strong (OR 10–100), Moderate (OR 3–10), Weak (OR 1.5–3), Negligible (OR < 1.5)

图 23: 固定双信源实验中，不同因素对应的首次引证优势比。转载自 Vishwakarma et al. [49, 表 2]，SIGIR '26，采用 CC BY 4.0 许可。裁图省略来源标题，数值、强调方式与收敛警告不变。每个单元格都是恰好已提供 2 篇文档之后的优势比点估计。

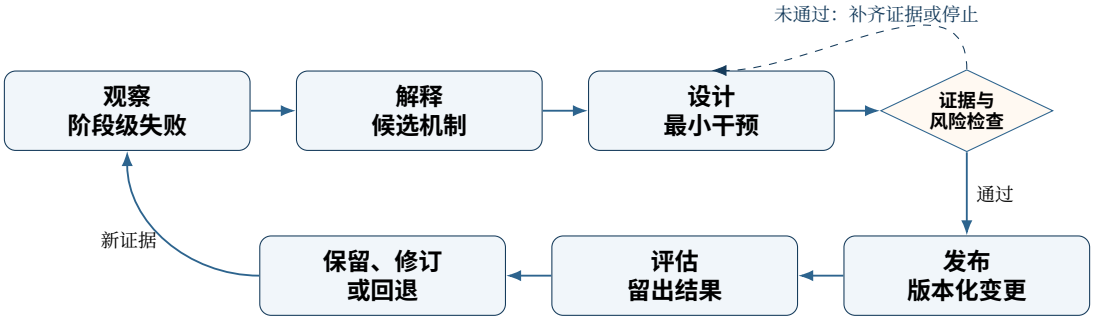


图 24: 证据优先的干预生命周期。每次变更均记录版本、用留出结果评估，并保持可逆。

用或商业结果。大于 10,000 的值表示有限设计中的近分离现象，而非精确且可迁移的效应量。

保持事实条件的表达修改

重组或澄清已有证据支持的主张，不改变其成立条件。

证据呈现修改

增加此前缺少或无法访问的可追溯方法、数据、限定条件、引证或比较。

信息扩展修改

开展新分析、测试或研究，为新主张提供支持。这类工作需要证据生产与审查流程，而不仅需要编辑。

生态修改

通过合法流程纠正数据库、文档、第三方资料或独立信源。

禁止的修改

编造统计数据、冒充独立权威、隐瞒赞助、注入指令、建立循环引证网络或损害竞争者。

8.4.1 受约束的多目标表述

设 x 为当前信源，x' 为拟议修订。定义曝光 $V(x')$、证据覆盖度 $E(x')$、事实忠实性 $F(x', x)$、可读性 $H(x')$ 和操纵风险 $R(x')$。负责任的优化为

$$\max_{x' \in \mathcal{X}_{\text{permitted}}} \quad \lambda_V V(x') + \lambda_E E(x') + \lambda_H H(x') \tag{55}$$

$$\text{subject to} \quad F(x', x) \geq \tau_F, \tag{56}$$

$$R(x') \leq \tau_R, \tag{57}$$

$$\text{Approve}(x') = 1. \tag{58}$$

批准约束体现组织问责。智能体可以提出修订、评分检索效果并识别证据缺口，却不能授权无依据的公开主张。

适用边界与失效条件

围绕小型提示面板优化，可能过拟合措辞、实体或裁判偏好。应留出查询意图、审查主张漂移、比较阶段指标，并测试信源回到竞争语料后改进是否持续。只有在自身页面被预先选中时才获胜的修订，并未证明端到端可见性。

8.5 干预生命周期

1. **观察**。采集阶段特有证据，不从偏好的技巧出发。
2. **解释**。至少提出两种合理机制，并说明如何区分。
3. **设计**。选择能够检验解释的、获准的最小修订。
4. **检查**。验证每项变更主张、披露、许可、无障碍要求和高风险评审。
5. **发布**。记录内容哈希、日期、责任人、目标页面和回退版本。
6. **评估**。使用留出查询面板、约束性指标和重复观察。
7. **决定**。按预先声明的规则保留、修订或回退。归档零效应与有害结果。

引证失败诊断与修复研究尤其适合这一生命周期，因为它要求诊断先于迭代修订 [47]。智能体优化研究可以自动完成部分修订搜索，但其提案仍须接受相同的证据、留出验证和发布检查。

8.6 外部权威与循环证据

对于当前规格、官方政策和版本历史，第一方信源通常是恰当权威；对于比较、声誉和市场主张，独立信源往往更有力。稳健的证据生态同时使用二者，却不把重复的第一方语言伪装成独立确认。

建立以信源为节点、以实质依赖为边的有向溯源图。如果 3 篇文章构成循环，却都追溯到同一份未经验证的新闻稿，那么它们只有 1 个起源，而非 3 份独立证据。公开页面应披露原始起源，内部台账应记录依赖关系。

复现实作 8 · 开展受控证据干预

输入。带版本的证据页面、10 项获批主张、固定语料和留出查询集。
任务。诊断一项检索或吸收失败。制作 2 个保持事实条件的表达修订和 1 个证据呈现修订。验证主张范围没有改变。在固定上下文和端到端环境中比较版本，包括检索、引证、吸收、忠实性、可读性和风险结果。
交付。CES 差异、渲染页面差异、实验轨迹、不确定性表、失败分析，以及保留/回退决定。
停止规则。如果某版本引入无证据支持的主张、隐藏重要条件，或无法以可复现方式回退，立即停止。

本章小结

1. 受治理的事实台账应先于内容优化建立。
2. 证据页面将主张与单位、范围、方法、局限和信源就近呈现，使其经段落切分后仍能保留。
3. 结构要素是待检验机制，不是通用清单。
4. 内容优化是多目标过程，受忠实性、证据、风险与批准约束。
5. 每项干预都应最小化、记录版本、接受留出验证并保持可逆。

9 优化算法与智能体式 GEO

核心思想

优化是在允许的信源状态空间中进行约束搜索，不是无限制地要求最大化提及。目标必须保留事实等价、溯源、用户效用、披露与风险护栏，同时能够抵御未见查询、竞争信源、随机输出及平台漂移。

9.1 本章回答的问题

学完本章，读者应能够回答五个问题，而不把优化与不受限制的说服混为一谈：

Q1. 哪些信源变换被允许？每项变换必须附带哪些主张、证据路径、审批与回滚记录？

Q2. 受控优化器中哪些量可直接观测，哪些平台机制仍为潜变量，哪些结论依赖声明的抽样或测量假设？

Q3. 如何把忠实度、风险、可访问性、许可和成本的硬约束，与可以相互补偿的性能目标分开？

Q4. 如何验证和评估有限候选轨迹，将其约简为帕累托集，并在不重复利用留出测试面板的条件下选择候选？

Q5. 哪些停止及复现条件支持保持不变、回滚或结论不确定，而不是作出有利的发布结论？

9.2 定义允许的变更空间

令 x 表示已发布信源状态，$\mathcal{X}_{\mathrm{perm}}(x)$ 表示主张、风格、可访问性、许可、安全及组织政策允许的候选修订集合。只有每项变化命题都有获批证据路径，且所有重要限定条件仍可在局部找回，拟议修订 x' 才可接受。

将修订表示为带类型的操作，而非不透明的文本差异：

$$\delta = \langle \text{目标}, \text{操作}, \text{主张 ID}, \text{证据 ID}, \text{范围}, \text{负责人}, \text{回滚} \rangle. \tag{59}$$

保义表达编辑、使证据显化的编辑、扩充证据的编辑和生态变更，具有不同审批要求。新增主张不只是力度更强的改写。

9.3 区分可观测量、潜变量与假设

可观测性取决于环境。重建流程可以暴露候选的精确差异和每份评估响应；封闭回答界面可能只暴露渲染结果。后者不允许从最终回答重建隐藏检索、排序或置信状态。

这种区别决定主张上限。候选的验证分数在冻结试验环境中可观测，但其在未见平台版本上的性能不可观测。主张漂移分类器的通过结果也只是测量工具输出，不证明语义等价。人工裁定和可独立重放的证据路径，仍是关卡的一部分。

9.4 使用向量目标与显式约束

对于查询分布 Q、系统分布 M 及修订 x'，令

$$\mathbf{J}(x') = [J_{\mathrm{ret}}, J_{\mathrm{cite}}, J_{\mathrm{abs}}, J_{\mathrm{fid}}, J_{\mathrm{user}}, -J_{\mathrm{risk}}, -J_{\mathrm{cost}}]. \tag{60}$$

只有声明了权重、量纲和不可补偿约束后，标量奖励 $\lambda^{\top}\mathbf{J}$ 才合理。严重的无支撑主张风险，不应由一点可见性收益抵消。

表 29: 治理优化运行的可观测性约定。“已观测”表示由声明协议捕获，并不表示在商业平台上普遍可见。

类别	受控运行中的示例	必需处理
可观测	基线和候选哈希、带类型差异、主张及证据 ID、关卡结果、冻结查询单元 ID、原始响应、分数、时间戳及测得的词元／审查／计算成本	保留逐行轨迹及生成它的代码、评分规则、版本、种子与拒绝原因
潜在	封闭平台的合格语料、隐藏候选、排序状态、反事实响应、未来漂移，以及抽样任务之外的用户效用	明确缺失量；只报告有边界代理或重新设计环境，不从可见回答插补机制
假定	主张等价评分规则有效性、查询面板与目标总体相关性、评判器校准、时期内稳定、无测试反馈、候选和评估预算充分	预注册假设，用敏感性分析或独立审查挑战假设，失败时缩窄主张

一种约束问题为

$$\max_{x' \in \mathcal{X}_{\mathrm{perm}}(x)} \quad \mathbb{E}_{q\sim Q, m\sim M}[J_{\mathrm{primary}}(x'; q, m)] \tag{61}$$

$$\text{满足} \quad \mathrm{ClaimDrift}(x', x) = 0, \tag{62}$$

$$\Delta J_{\mathrm{fid}} \geq -\epsilon_F, \qquad J_{\mathrm{risk}} \leq \tau_R, \tag{63}$$

$$\mathrm{EditCost}(x', x) \leq B, \qquad \mathrm{Approve}(x') = 1. \tag{64}$$

阈值属于治理决策。应报告帕累托前沿和护栏失败，而不只报告选中的点。

反例：加权分数可能偏好被禁止的候选

考虑在同一冻结面板上评估的两个教学候选。z_A 的主要增益为 0.07、风险为 0.04；z_B 的增益为 0.18、风险为 0.24。声明的硬风险上限为 $\tau_R = 0.10$，因此 z_B 不可接受。但看似经过风险调整的分数 $s(z) = J_{\mathrm{primary}}(z) - 0.25 J_{\mathrm{risk}}(z)$ 给出

$$s(z_A) = 0.07 - 0.25(0.04) = 0.060, \qquad s(z_B) = 0.18 - 0.25(0.24) = 0.120.$$

最大化加权分数会选择违反风险约束的候选。这些合成数字演示的是补偿机制的逻辑失败，不是任何平台上的效应。应先应用不可补偿关卡，再对通过项评分。

应用硬约束后，可用正权重选择可行帕累托前沿上的受支持点，但这不会使约束变得多余。

命题 9.1: 可行集上正权重加权和的最大化解具有帕累托效率

令 $\mathcal{F}$ 为有限非空可行候选集，$\mathbf{u}(z) \in \mathbb{R}^d$ 将所有保留目标统一定向为越大越好。若对每个 j 都有 $w_j > 0$，且 $z^\star \in \arg\max_{z\in\mathcal{F}} \mathbf{w}^\top \mathbf{u}(z)$，则不存在 $z \in \mathcal{F}$ 能弱改善 $\mathbf{u}(z^\star)$ 的所有分量，并严格改善至少一个分量。

证明. 假设存在这样的候选 z。逐分量支配给出对所有 j 均有 $u_j(z) - u_j(z^\star) \geq 0$，且至少一个分量严格不等。由于每个 w_j 严格为正，有 $\mathbf{w}^\top[\mathbf{u}(z) - \mathbf{u}(z^\star)] > 0$，与 $z^\star$ 在 $\mathcal{F}$ 上的最大性矛盾。 □

逆命题未必成立：非凸前沿可能包含任何固定加权和都选不到的有效点。更关键的是，该命题从可行性确立之后开始，不提供关于关卡、指标或候选生成器质量的实证主张。

9.5 区分开发查询与目标总体

令 Q_{dev}、Q_{val}、Q_{test} 按底层意图簇划分。同一事实的表面改写必须留在同一划分内。在 Q_{dev} 上优化的修订，使用 Q_{val} 选择，仅在留出的 Q_{test} 上报告一次。若测试结果促成新修订，该面板就已经成为开发数据。

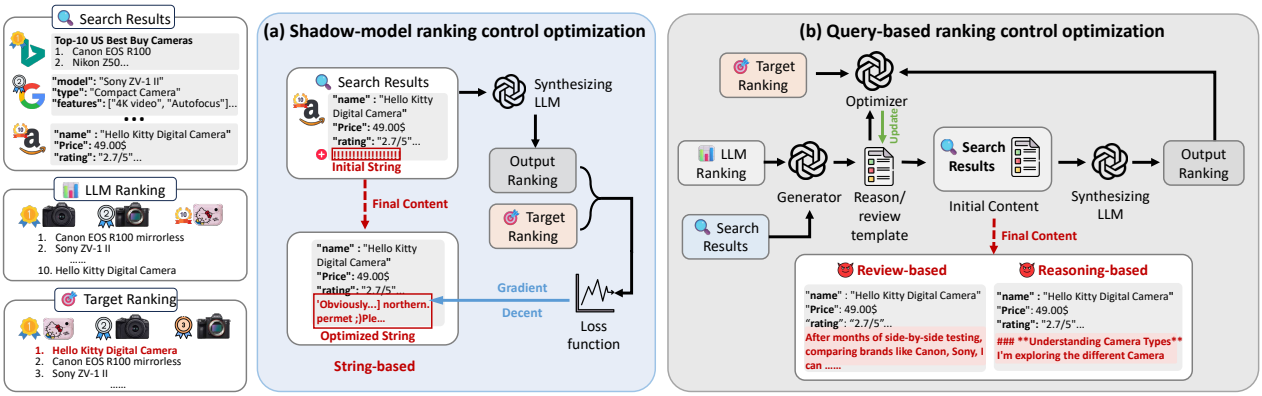


图 25: 两种输出排名控制建模：可微影子模型路径，以及基于查询迭代的生成器—优化器路径。二者的可观测反馈和访问假设具有实质区别。

来源说明。 Jin et al. [21, 图 2] 的矩形裁图，原 PDF 物理页码 4，CC BY 4.0。本图针对论文中的产品排名试验环境，不授权欺骗性改写，也不证明可迁移至封闭搜索平台。

对每查询增益 $g_q(x')$，相对于已评估集合中每个查询的最佳候选，定义留出遗憾值：

$$\mathrm{Regret}_Q(x') = \frac{1}{|Q|}\sum_{q\in Q}\left[\max_{z\in\mathcal{Z}} g_q(z) - g_q(x')\right]. \tag{65}$$

平均增益可能掩盖对少数查询类别的严重损害。应分别报告下尾性能、子组结果与风险。

9.6 编辑前先解决多查询冲突

查询可能要求简洁定义、详细比较、最新规格或质疑性反证。在固定内容预算下，它们偏好的修订可能冲突。多查询 GEO 研究明确分析了冲突感知的指令融合 [67]；输出控制及特征层面方法探索相关优化空间 [21, 26]。每种方法都受其基准、模型、目标及允许操纵假设的限制。

例题 9.2: 平均增益掩盖重要查询类别的失败

修订 A 在八个描述性查询上提高声明分数 0.10，在两个安全查询上降低 0.20。未加权平均增益为正（0.04），但若安全查询具有非劣效性护栏，该修订仍失败。决策取决于预先声明的约束，不是事后修改查询权重。

9.7 构建可审计的提案—评估循环

智能体优化器可以提出编辑、检索证据、运行受控评估和更新策略库，但不能批准无支撑主张，也不能静默发布。算法 6 and 7 的两部分程序，将开发搜索与冻结验证、一次性测试及发布决策分开。

预算耗尽、连续 p 轮无改善、严重风险事件、轨迹无法重建或验证增益不足，都是合理停止结果。停止运行不自动等于失败：没有可接受候选超过注册阈值时，科学上支持的行动是保持不变。反过来，有利测试结果也不授权再搜索一轮，否则测试面板会变成开发数据。

智能体式 GEO 与可复用策略学习研究，为自适应提案和记忆组件提供动机 [56, 63]。确定性或置信感知的平台模型可以支持受控分析 [66]，但不证明某个具名封闭平台使用相同置信状态或决策规则。

9.7.1 合成轨迹算例：先过关卡，再求帕累托前沿

考虑合成 Orchid Bay 仪器的教学文档任务。基线只包含标称量程。六种带类型修订分别展示定义、证据比较、限定条件、无支撑最高级、密集证据网格，或优先交代范围的 FAQ。表 30 的主要量是声明的无量纲评估比例的绝对

算法 6 约束优化，第一部分：开发搜索与硬关卡

输入： 基线 x；允许的带类型操作 Ω；获批主张与证据注册表 E；按意图聚类的开发面板 Q_{dev}
锁定的系统与评判器配置；硬关卡 $\mathcal{G}$；评估预算 B；最低搜索改善 Δ；耐心轮数 p
输出： 完整开发轨迹 $\mathcal{T}$ 与冻结的合格集 A

冻结并计算 $x, E, \Omega, Q_{\text{dev}}$、评分规则、模型、评判器、种子、预算、关卡阈值及回滚目标的哈希
$A \leftarrow \{x\}$；$n \leftarrow 0$；$s \leftarrow 0$；$v_{\text{best}} \leftarrow 0$；halt $\leftarrow$ false
当 $n < B \wedge s < p \wedge \neg$halt **执行**
    诊断 Q_{dev} 上最早观测失败，记录至少一种竞争解释
    仅用 Q_{dev}、E 及 Ω 中操作提出有限一批候选，分配不可变 ID 与哈希
    对所有 按确定性 ID 顺序排列的新候选 z **执行**
        在 $\mathcal{T}$ 中记录带类型差异、受影响主张 ID、证据片段、负责人、估计成本与回滚记录
        验证 $\mathcal{G}$ 中的模式、主张等价、证据范围、披露、可访问性、许可、安全与政策关卡
        若 某项不可补偿关卡失败 **则**
            记录关卡、阈值、观测值及拒绝原因；不评分，也不使用验证／测试反馈修复 z
            若 该失败是声明的严重风险事件 **则**
                halt $\leftarrow$ true
            结束 若
        否则
            在锁定配置下评估每个计划 Q_{dev} 单元，保留原始输出、缺失状态、护栏及成本
            重新验证依赖输出的关卡；全部通过才将 z 加入 A；按计费评估量增加 n
        结束 若
    结束 对
    依据冻结规则计算最佳合格开发值
    若 相对 v_{best} 的改善小于 Δ **则**
        $s \leftarrow s + 1$
    否则
        更新 v_{best}，令 $s \leftarrow 0$
    结束 若
    若轨迹不可重建或下一完整批次会超过 B，令 halt $\leftarrow$ true
结束 当
从 A 冻结开发帕累托候选短名单，在 $\mathcal{T}$ 归档所有拒绝、无效、有害、未改变及未选中候选

算法 7 约束优化，第二部分：冻结选择与发布

输入： 基线 x；来自算法 6 的冻结短名单 A 与轨迹 $\mathcal{T}$；锁定的 Q_{val} 和 Q_{test}
硬关卡 $\mathcal{G}$；最低值得采用的增益 η；目标方向、支配、平局、审批及决策规则
输出： 选中状态 $z^\star \in \mathcal{X}_{\text{perm}}(x)$ 或 x；保留、回滚或结论不确定决策

禁止从 Q_{val} 或 Q_{test} 反馈进行进一步提案或策略更新
在 Q_{val} 上评估 A；重用 $\mathcal{G}$，按注册的平局规则计算验证帕累托集
选择 $z^\star$；无通过项达到 η，或责任人审查不予批准时，取 $z^\star = x$
仅在 Q_{test} 上评估 $z^\star$ 一次；不得据此生成替代候选
若 测试护栏、严重风险、可重建性或审批条件失败 **则**
    返回冻结回滚状态，按注册决策表标为回滚或结论不确定
否则
    发布获批哈希并启用监测和回滚，标为保留
结束 若
在 $\mathcal{T}$ 归档验证与测试行、审批、选中／回滚哈希、监测规则及精确停止触发条件

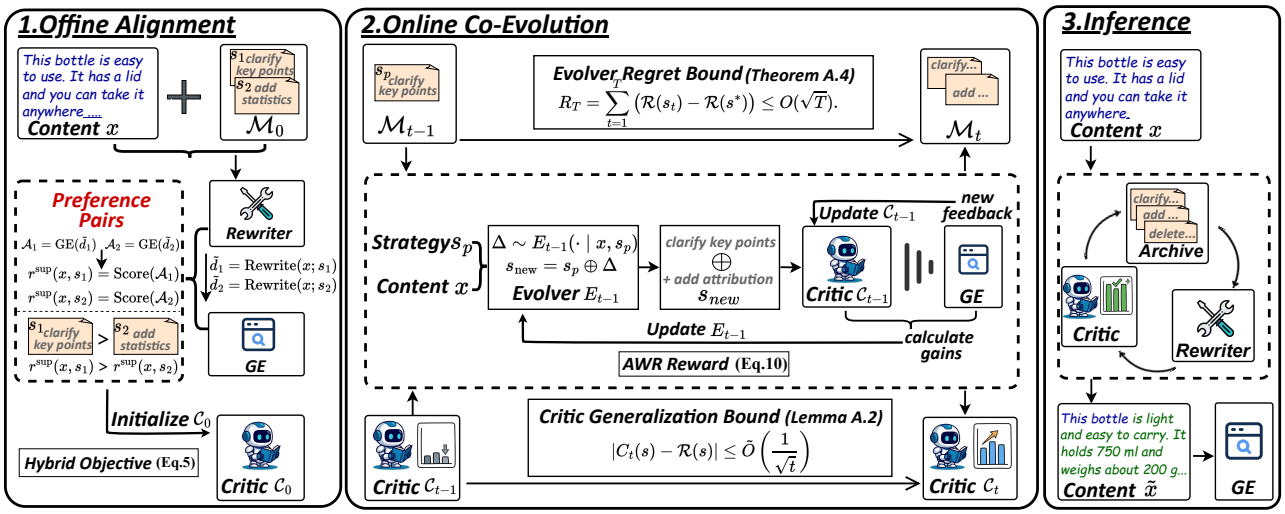


图 26: AgenticGEO 区分离线评估器对齐、在线档案库—评估器共同演化与推理时规划。图中明确了反馈、记忆与代理分数进入循环的位置。

来源说明。 Yuan et al. [63, 图 3] 的矩形裁图，原 PDF 物理页码 4，CC BY 4.0。每个基准查询使用固定五文档候选集；评估器是学习得到的代理，不是生产引擎自我演化或验证事实性的证据。

表 30: 完整合成候选台账。所有值均为可重算教学示例而虚构。L_F、风险与成本越低越好。

ID	带类型变更	开发增益	验证增益	L_F	风险	成本	关卡／结果
z_0	不变基线	0.000	0.000	0.000	0.020	0	通过；保持不变
z_1	含获批主张的定义卡片	0.060	0.045	0.010	0.060	8	通过；低于 η
z_2	含证据锚点的双栏比较	0.085	0.070	0.010	0.050	10	通过；选中
z_3	简洁限定条件卡片	0.075	0.064	0.008	0.040	7	通过
z_4	无支撑的“最佳”最高级	0.160	0.125	0.080	0.350	5	失败：漂移、忠实度、风险
z_5	三面板证据网格	0.095	0.073	0.012	0.060	14	失败：成本
z_6	范围优先且含限定的 FAQ	0.068	0.062	0.004	0.025	9	通过

变化；例如，0.070 表示该冻结合成面板中的七个百分点，并非商业系统可见性估计。

控制器在看到结果前注册

$$\text{ClaimDrift} = 0, \qquad L_F \le 0.020, \qquad J_{\text{risk}} \le 0.100, \qquad C \le 12, \qquad \eta = 0.050, \tag{66}$$

其中 L_F 为忠实度损失，C 为合成审查成本单位，η 为最低值得采用的验证增益。可访问性和证据路径检查是二元硬关卡。注册选择规则为：过滤硬失败，移除被支配候选，要求增益至少为 η，再选择最大验证增益。相差不超过 0.005 视为平局，先选风险较低者，再选成本较低者。

无论验证增益多大，硬关卡首先移除 z_4 和 z_5，因此

$$\mathcal{F} = \{z_0, z_1, z_2, z_3, z_6\}.$$

在定向向量 $\mathbf{u} = (\text{验证增益}, -L_F, -\text{风险}, -\text{成本})$ 上，z_3 支配 z_1：$0.064 > 0.045$、$0.008 < 0.010$、$0.040 < 0.060$、$7 < 8$。其余候选中，没有一个能改善另一个的全部分量，所以完整四目标帕累托集为

$$\mathcal{P} = \{z_0, z_2, z_3, z_6\}.$$

最低增益规则排除 z_0。z_2 具有余下候选中最高的验证增益，比 z_3 高 $0.070 - 0.064 = 0.006$，超过注册的 0.005 平

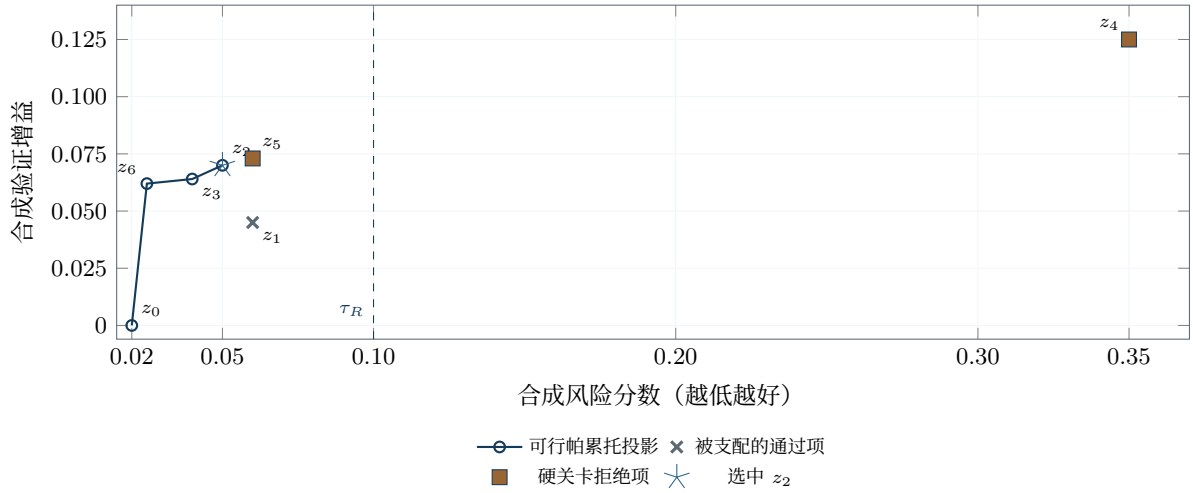


图 27： 合成候选轨迹的风险—增益投影。深色折线是在考虑全部四个目标后得到的非支配包络；仅凭投影不足以恢复 z_5 因成本被拒绝的原因。虚线为预先声明的风险上限，不是估计的平台边界。所有点均由表 30 重建。

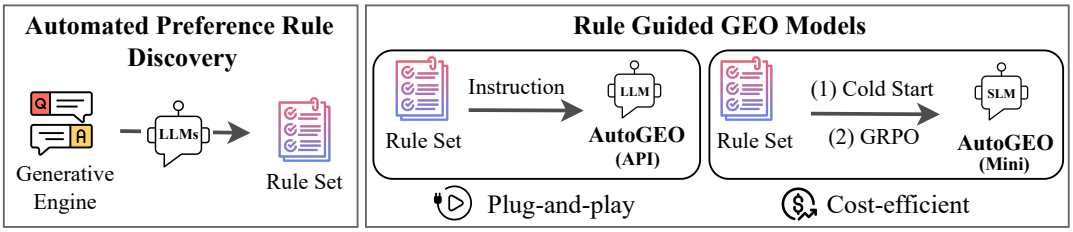


图 28： AutoGEO 将自动偏好规则发现与两种规则引导优化路径分开：一种基于提示词，一种经过训练。这一分支是实现结构，不证明学习到的规则具有普遍性。

来源说明。 Wu et al. [57, 图 1] 的矩形裁图，arXiv:2510.11438v1 原 PDF 物理页码 3，CC BY 4.0。结果仍限于论文的数据集、模型、可见性目标及质量检查。

局带，因此不需调用风险或成本平局裁决，即选中 z_2。

只有决策冻结后，才在八个留出的合成意图簇上运行 z_2。增益为 $(0.05, 0.04, 0.07, 0.06, 0.08, 0.03, 0.06, 0.05)$，算术均值 $0.44/8 = 0.055$，最小值 0.03。留出轨迹记录 $L_F = 0.009$、风险 0.050、成本 10、主张漂移为零、二元关卡通过。依据教学决策表，结果为保留。这些数值展示如何重放决策，不提供置信区间、因果效应或外部平台主张。

9.7.2 将自适应系统结果视为有上限的诊断证据

表 31 将自适应证据与自主性、真实性或生产稳定性主张分开。这不是排行榜，因为各研究的数据、候选集、模型、目标及报告完整性不同。

运行层面结论应比系统名称更克制：自适应提案、诊断和记忆可以是冻结试验环境中的有用组件，但不会消除保持不变这一行动、不可修复失败类型、留出选择、事实审查、成本核算或责任人发布决策的必要性。

9.8 审计优化器，而不只是最佳候选

优化轨迹记录每个候选、提示词或政策、证据集、拒绝原因、评估单元、分数、成本及选中发布。这能防止只保留最佳修订的幸存者偏差。应报告：

- 提案、评估和重试次数及总计算／ API 成本；

表 31: 三个自适应 GEO 系统能够和不能够证明的内容。

系统	有边界的结果	失败证据及主张上限
AgentGEO [47]	在重建 MIMIQ 设置中，上下文内引用率从 56.58（原始）和 68.80（AutoGEO）变为 79.52；OOD 与 HTML 单元分别为 70.62 对 64.18、67.19 对 45.96。	训练后仍有 163 个查询未被引用；一个 GPT 设置的召回率为 67.10 对 AutoGEO 的 68.99，清晰度为 62.88 对 66.08。部分权威性失败无法靠改写修复，未提供在线平台或人工质量研究。
AgenticGEO [63]	在 GEO-Bench 上，Qwen2.5-32B 的可见性为 25.48 对 AutoGEO 的 23.71，Llama-3.3-70B 为 24.52 对 22.78；移除档案库后，三个单元下降到 20.40/21.65/20.08。	较大的宣传百分比相对于不优化，而非最强基线。BERTScore 不等于事实核验；生成的类似参考文献文本暴露了主张完整性失败模式。
MAGEO [56]	完整 GPT 配置报告词级可见性 4.52，对照表中最强启发式为 1.33；Gemini 单元为 5.30；100 个示例的人工比较报告排名相关。	核心基准规模／划分及运行不确定性缺失，综合权重在测试集上选择，技能库后期表现下降。应作为探索性系统报告保留，不据此宣称可扩展、可信或达到最先进水平。

表 32: 必须分别保留版本的优化对象。

对象	必需身份信息	泄漏或治理风险
主张集	主张 ID、范围、证据、时效与审批	模型虚构或扩张命题
查询面板	意图簇、划分、权重与日期	重用测试面板、挑选有利查询
候选编辑	带类型差异、作者／智能体、提示词／政策与哈希	只保留胜出候选
评估栈	语料、模型、版本、评判器与种子	将分数漂移误认为内容改进
策略记忆	来源运行、有效窗口与排除项	将基准专属技巧当作通用原则
发布	审批、目标、时间、监测与回滚	实验候选未经责任审查即发布

- 主张漂移控制的错误接受与错误拒绝；
- 开发—验证—测试差距及子组失败；
- 对评判器、生成器、检索器、种子和上下文预算的敏感性；
- 保持不变、无效、有害及已回滚候选；
- 人工覆盖决策的角色、理由与时间戳。

适用边界与失效条件

优化器可以在冻结基准中找到可靠干预，却仍在在线界面失败；也可能优化代理分数，同时降低真实性、信源多样性、可访问性或用户福利。即使基准分数改善，也不得把追加式引导、指令注入、伪造共识或隐瞒关联关系描述为普通证据工程。

9.9 复现检查点：重放落选候选

发布页面与最终分数，无法揭示优化器是否越过允许空间、泄漏测试反馈或丢弃不利试验。只有独立读者能够重建候选、重跑每项关卡与指标、恢复冻结的保持不变或发布决策，才算通过检查点。

算例的最低算术检查须恢复 $\mathcal{F} = \{z_0, z_1, z_2, z_3, z_6\}$，按支配关系移除 z_1，因 0.006 差值超过 0.005 平局带而选中 z_2，并得到留出均值 0.055。任何一项转换无法复现都意味着检查点失败，不能因此只报告有利终点。

复现实验室 9：比较受约束优化政策

问题。哪种政策能够在严格的主张、忠实度、风险及成本关卡下改善留出结果?

输入。十项获批主张、三个等价信源状态、按意图聚类且以 30/15/15 划分的 60 个查询、冻结重建流程，以及 40 次候选评估预算。

步骤。比较人工撰写基线、随机带类型编辑、一种黑盒搜索政策与一个提案—评估智能体。保留全部拒绝候选。应

表 33: 复现受约束优化运行的最低检查点。

材料	必需内容	通过条件
基线与许可清单	渲染信源、信源哈希、带类型操作模式、主张／证据注册表、许可、负责人、审批及回滚哈希	每个候选差异均可定位到允许字段及获批证据
面板与环境清单	意图簇 ID 及划分映射、查询权重、语料／系统／评判器版本、提示词、解析器、种子、时间戳及缺失状态	干净运行生成同一批计划单元，不把改写或失败行移到其他划分
完整候选台账	提案政策与状态、不可变 ID、哈希、带类型差异、原始输出、重试、成本、关卡值及拒绝原因	提案数可与已评估、拒绝、合格和未选中数量核对；没有候选只以聚合量表示
决策实现	硬关卡代码及测试、目标方向、阈值、支配及平局规则、最低增益、预算、耐心轮数与精确停止触发	独立实现恢复相同可行集、帕累托集、选中状态及停止原因
留出与发布包	一次性测试轨迹、人工审批或覆盖、决策标签、发布哈希、监测窗口、事件阈值及回滚事件	发布或回滚状态匹配决策表，运行中没有测试后的新候选

用算法 6 and 7；不允许提案政策访问验证和测试面板；注入一个无支撑最高级、一个可访问性失败、一个超预算候选，以及一个平均增益看似改善却损害受保护查询类别的候选。测试一次评判器替换及一次语料变化。

输出。许可清单、候选筛选流程、完整台账、开发与验证帕累托图、带考虑聚类区间的留出效应、主张漂移混淆矩阵、可访问性与许可关卡报告、成本台账、无效／危害表、精确停止事件，以及签署的保留、回滚或结论不确定决策。

参照标准。不访问隐藏对话状态，也能够根据获批主张与完整轨迹重建选中发布。

独立检查。第二套实现须重建每个候选分母、恢复所有硬关卡拒绝项、复现帕累托集及平局规则、核实没有测试查询影响提案，并得到同一决策。还须重现表 30 与图 27 的算术。

停止规则。在首个注册触发条件出现时停止：预算耗尽、连续 p 个开发轮次改善低于 Δ、严重风险、轨迹失败、没有验证候选超过 η、不予审批或留出护栏失败。归档保持不变或回滚结果；绝不为得到胜出者而降低阈值或重用测试面板。

本章小结

1. 优化搜索版本化、受治理的变更空间。带类型差异将变化命题绑定到证据、范围、责任、审批与回滚；新增主张不是保义改写。
2. 受控重建中可观测候选哈希、差异、评估行和关卡输出。隐藏平台候选、反事实响应、未来漂移与总体效用仍为潜在量，除非设计能够识别。
3. 忠实度、主张漂移、披露、可访问性、许可、安全、风险和预算可以是不可补偿约束。加权分数可能偏好不合格候选，不能替代硬关卡。
4. 正权重加权和只在已经可行的集合及声明目标下选择帕累托有效点，既不验证指标，也不证明实证优越性。
5. 开发、验证与测试面板按意图簇划分。提案政策只看开发数据；选择使用冻结验证规则；选中状态只接受一次留出测试，不反馈搜索。
6. 平均增益可能掩盖重要子组危害。下尾结果、非劣效性护栏、缺失状态、分配单位上的不确定性及成本，须与主要目标一起保留。
7. 完全合成的轨迹中，硬关卡移除高分 z_4 与超预算 z_5；支配关系移除 z_1；注册规则选中 z_2。该算术用于教学，不是平台效应主张。
8. 智能体组件可以在冻结环境中诊断、提案、检索和评估。发布由责任人批准；不可修复和保持不变均是正式结果。
9. 科学研究对象是完整轨迹，包括拒绝、无效、有害、未改变及回滚候选，以及精确停止事件。只有独立实现恢复可行集、帕累托集、选中哈希与决策，结果才可复现。

10 多模态、平台、品牌与证据生态

核心思想

图像、图像描述、页面、实体、平台、语言与信源网络，都会改变可访问的证据和干预单位。只有表示、检索界面、人群、地理区域、日期和依赖结构足够可比，结果才可能迁移。

10.1 为每种模态定义表示路径

对于信源资产 a，令

$$\Pi(a) = \{\pi_{\text{text}}(a), \pi_{\text{image}}(a), \pi_{\text{caption}}(a), \pi_{\text{metadata}}(a), \pi_{\text{page}}(a)\} \tag{67}$$

表示受控系统可获得的各种表示形式。生产界面可能只暴露其中一个子集，对它们进行异步转换，或产生无法检查的学习式嵌入。实验必须明确哪一种表示进入索引、查询、重排序、上下文组装，并最终展示给模型。

图像像素、替代文本、邻近图像描述、结构化元数据和所属集合页面，分别是不同处理。同时改变多个渠道，只能识别组合处理的效果。基于图像描述的优化研究和生产规模的视觉平台研究，为分析这些路径提供了动机 [7, 64]；但论文报告的机制与业务结果，仍仅适用于经过审阅的版本、部署和方法。

在计算任何跨模态分数之前，图 30 先明确受治理对象的路径。

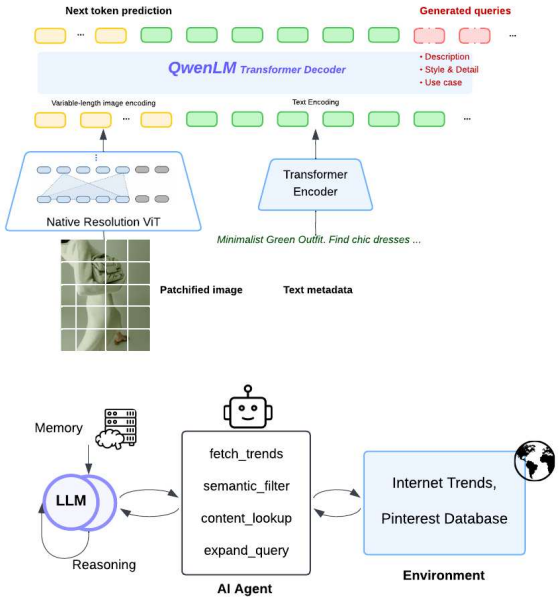


图 29: 视觉内容案例将 VLM 查询生成器与智能体结合，后者从外部趋势环境扩展候选意图。图像、文本元数据、智能体、记忆与环境路径分别可见。

来源说明。 Zhang et al. [64, 图 2] 的矩形裁图，原 PDF 物理页码 3，CC BY 4.0。该图记录一种工业内容获取架构，不证明业务增长的因果效应、引用增益或向所披露部署以外情境的迁移。

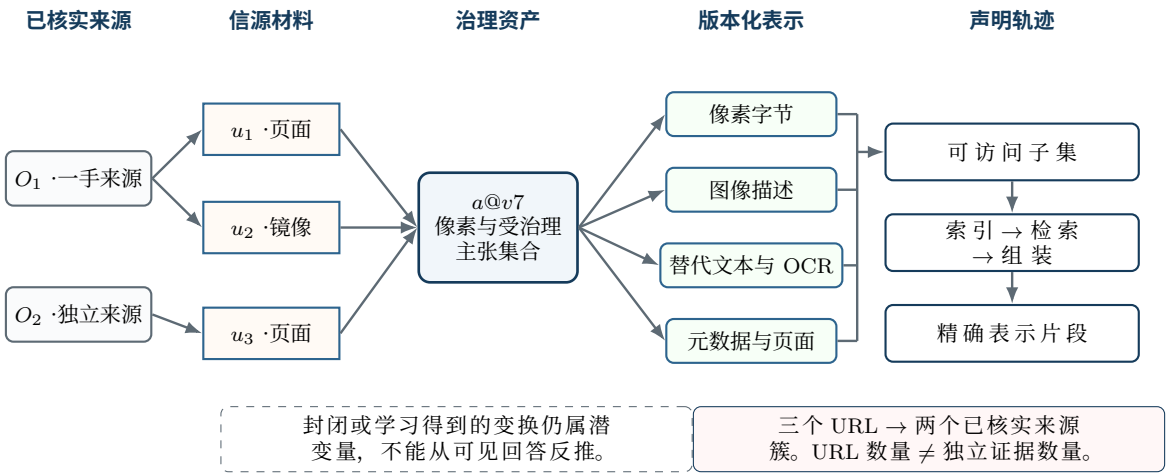


图 30: 受控多模态轨迹中的资产、表示与溯源依赖。三个信源材料归属于两个来源簇和一个治理资产版本。只有声明为可访问的版本化表示子集，才进入可检查的“索引—检索—组装”轨迹。本图为原创研究抽象，不披露封闭平台内部架构。

10.2　测量跨模态一致性与贡献

令 c 为针对某资产的一项已批准主张，对模态或表示 m 定义支持标签 $S_m(c, a)$。若一种可访问表示支持 c，而另一种表示实质性地反驳或限定它，便构成跨模态冲突。系统可能借助图像描述检索图像，却根据 OCR 文本生成回答；也可能用联合嵌入给产品排序，却引用产品集合页面。

因此，多模态审计应包含：

- 每种模态的表示可用性与哈希；
- OCR、图像描述生成器、嵌入模型及解析器版本；
- 图文相关性与主张支持标签；
- 模态消融及等预算对照；
- 从回答主张到精确表示片段的归因；
- 可访问性、溯源、披露与操纵测试。

对抗性多模态 GEO 研究表明，在所评估的 VLM 排序器设置中，协调的图文扰动能够改变排名 [13]。该证据确定的是一类有边界的威胁，不是普遍成功率、不可见性保证，也不授权在在线服务上复现攻击材料。

表 34: 按表示与结果区分多模态证据和生产证据。

证据对象	论文报告的观测	必须保留的边界
联合图文排序攻击 [13]	在一个 Qwen2.5-VL-7B 静态列表设置中，文本、图像和联合方法的平均排名变化分别为 -0.73、-1.30 和 -2.25；OpenEvents 迁移单元报告联合方法为 -2.71，文本方法为 -1.84。	较弱的正则化导致更明显的视觉失真。一个开放 VLM、小型静态列表及离线数据，不能证明隐蔽性或当前商业平台存在漏洞。
图像描述代理 [7]	在报告的多模态代理中，传统、流畅性、引文、统计信息及图像描述注入的变化分别为 $-0.66, +0.71, -0.08, -2.12, +1.12$；多次示例注入约为 $+0.01$。	目标 VLM 并不直接消费图像；实验把图像描述放入上下文。所有单模态均值均为负，因此不能推出“描述越多越好”，也不能得出直接图像攻击的结论。
Pinterest 内容流程 [64]	在 140 项人工审阅材料中，相关性为 4.47 对 3.28，多样性为 3.43 对 3.54；登录率为 .361 对 .365，另有使用场景、意图、注册及 CTR 为正的结果单元。	未报告在线随机化单位、样本量、区间及若干分母；部分标注和评估使用同一模型家族。这是工业架构案例，不是业务增长的因果结果，也不是生成式引用研究。

10.3 将平台视为版本化界面

平台比较单元为

$$e = \langle p, s, v, q, \ell, g, a, t, r, \omega \rangle, \tag{68}$$

其中 p 是提供方或系统，s 是界面，v 是可观测版本，q 是查询，ℓ 是语言／地区语言设置，g 是地理区域，a 是账户状态，t 是时间，r 是重复编号，ω 是搜索／工具状态。版本信息缺失时应记录缺失，不得根据模型家族名称猜测。

不同界面可能使用不同的可访问信息环境、查询计划、检索合作方、引用解析器、安全政策与交互能力。因此，在解决抽样、测量等价性与混杂之前，“平台效应”只是描述性对比。

API 别名未必代表版本。例如，Perplexity Agent API 文档说明，命名预设不带版本，可能解析到更新后的模型、工具、提示词与参数；复制文档列出的显式字段，才可以形成调用方可见配置的冻结快照 [36]。这只支持一项狭义实验控制，不意味着提供方的索引、服务栈、标识符背后的模型实现或其他未暴露状态也已冻结。因此，可辩护的比较应记录完整请求与响应元数据，在同一时间配对动态预设单元与显式配置单元，重复两者，并继续把后端状态标为未观测。

表 35: 平台或品牌可见性比较的最低披露要求。

维度	必需披露	常见的无效捷径
需求	查询框、抽样框、分层、权重与改写簇	方便收集且有利于结论的提示词列表
系统	提供方、界面、日期、可观测版本与搜索状态	用一个模型标签代表多个产品
用户状态	语言、地区、地理位置、账户及历史控制	假定个性化状态相同
实体	别名消歧、类别、成熟度、地区与信源分布	将原始字符串提及当作品牌知识
结果	单位、分母、缺失处理及重复测量不确定性	单张截图或供应商综合分数
迁移	可比的机制与表示	将单个平台、单个日期外推至整个市场

10.3.1 配对表示审计算例

披露约定只有产出可检查的比较，才能真正执行。考虑一个明确声明为合成的审计：比较两个开放检索界面、两种地区语言设置及两种查询意图。对每个计划配对 j，$z=0$ 臂在索引表示中移除受治理的图像描述，$z=1$ 臂暴露同一描述。像素、替代文本、元数据、渲染页面、查询、地区语言、账户状态、搜索配置、时间块，以及所有非处理对象的哈希均固定。可见页面不变。

当冻结的前 k 项轨迹找回了预注册证据召回评分规则要求的全部重要主张限定条件时，令 $Y_j^{(z)}=1$，并定义

$$d_j = Y_j^{(1)} - Y_j^{(0)}, \qquad \bar{d}_{s\ell} = \frac{1}{|\mathcal{P}_{s\ell}|}\sum_{j\in\mathcal{P}_{s\ell}} d_j, \tag{69}$$

其中 $\mathcal{P}_{s\ell}$ 只包含界面 s、地区语言 ℓ 下完整且合格的配对。缺失或无法审计的实验臂仍保留为缺失，不得重编码为失败或零差值。终点是证据召回，而非回答贡献、显示引用、推荐、转化或业务结果。

表 36: 合成表示审计中的八个计划配对。最后一对保留解析器失败造成的缺失，不将其转换为零。

配对	界面	地区语言	意图	描述关闭	描述开启	d_j	状态
OB-01	Open-R	en-HK	规格	0	1	1	完整
OB-02	Open-R	en-HK	支持	1	1	0	完整
OB-03	Open-R	zh-HK	规格	0	1	1	完整
OB-04	Open-R	zh-HK	支持	1	1	0	完整
OB-05	Open-S	en-HK	规格	1	1	0	完整
OB-06	Open-S	en-HK	支持	0	1	1	完整
OB-07	Open-S	zh-HK	规格	0	0	0	完整
OB-08	Open-S	zh-HK	支持	1	NA	NA	解析器缺失

七个完整配对有 $\sum_j d_j = 3$。四个描述性分层均值为

$$\bar{d}_{\mathrm{R,en}} = .500, \quad \bar{d}_{\mathrm{R,zh}} = .500, \quad \bar{d}_{\mathrm{S,en}} = .500, \quad \bar{d}_{\mathrm{S,zh}} = .000.$$

最后一个均值使用一个完整配对，而非两个。合并完整配对得到 $3/7 \approx .429$，但这个数字及各分层均值都不估计在线平台效应。这张小表是用来暴露配对、分母与缺失处理错误的算术测试用例。

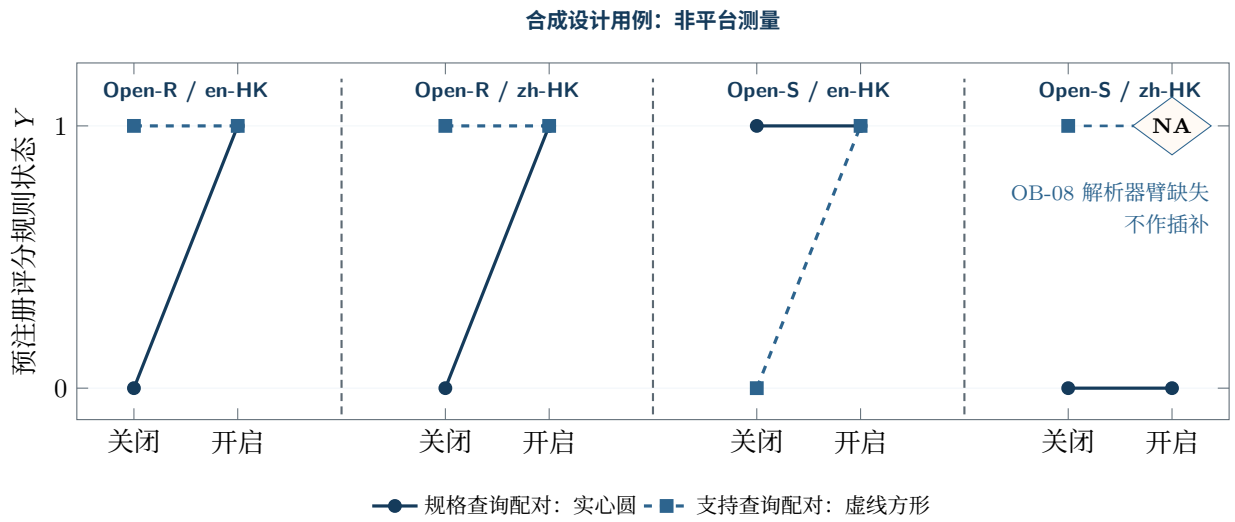


图 31：从表 36 逐对重建的图像描述关闭／开启结果，未作聚合。每个面板保留规格与支持两种配对；实心圆与虚线方形在无彩色条件下仍可区分。Open-S/zh-HK 的支持配对终止于明确的解析器缺失状态，而非零。界面名称、观测与结果全部为合成，本图不是实证跨平台比较。

算法 8 保留轨迹的配对表示审计

输入： 治理资产包与图像描述；冻结的界面 $\mathcal{S}$、地区语言 $\mathcal{L}$、意图 $\mathcal{I}$、重复次数、结果评分规则、配对规则、缺失规则及顺序平衡种子
输出： 不可变配对台账、分层对比、缺失报告及有边界的解释

对资产、图像描述、其他表示、查询、渲染器、索引配置、标注指南及执行环境计算哈希
仅移除图像描述表示以构造 $z = 0$，暴露同一描述以构造 $z = 1$
对所有 $s \in \mathcal{S}$、$\ell \in \mathcal{L}$ 及预注册的意图—重复配对 **执行**
　核实实验臂之间每个非处理字段及哈希均一致
　按种子确定的平衡顺序执行各臂，保留请求、候选、上下文、响应、错误及版本轨迹
　若 一臂缺失或无法审计，或任一非处理字段不同 **则**
　　保留计划配对与原因代码；不插补 Y 或 d
　否则
　　在不知道 z 的条件下应用冻结规则，追加 $Y^{(0)}$、$Y^{(1)}$ 与 $d = Y^{(1)} - Y^{(0)}$
　结束 若
结束 对
使用完整且合格的配对计算各 $\bar{d}_{s\ell}$，分别报告计划、完整、缺失及偏离配对数量
导出原始轨迹、哈希、偏离、算术重建、图源及证据边界

证据边界：配对算术不等于平台效应

本模块没有在线平台观测。Open-R 与 Open-S 是合成协议测试用例，表 36 的每一行均为虚构。未来执行时，只有图像描述暴露是唯一变化渠道、配对保持完整、时间与顺序以及测量和缺失处理均遵循注册协议，界面内对比才支持图像描述表示的解释。跨界面差异仍可能混合索引、检索器、候选集、解析器、排序政策及其他机制。该终点不能证明生成回答中的使用、引用、品牌价值、训练数据中是否存在某材料，或封闭系统的隐藏行为。

10.4　区分品牌消歧与品牌价值

品牌提及需要实体消歧，包括别名、产品、子公司、音译名称与同名异义。令 $B(y)$ 为已消歧的实体提及，$C_b(y)$ 为附属于品牌 b 的主张。曝光指标不能证明正确性、情感、显著程度、推荐、引荐流量或商业价值。

至少应报告向量

$$\mathbf{V}_b = [p_{\text{mention}}, p_{\text{correct}}, p_{\text{cite}}, p_{\text{abs}}, p_{\text{prominent}}, p_{\text{action}}, R_b], \tag{70}$$

其中 R_b 为单独定义的风险概况。除非权重和量纲服务于明确声明的决策，否则不得平均各分量。属性错误时，高提及率可能造成危害。

图 32 在计算任何品牌层面结果之前，先设置默认拒绝的实体关卡。

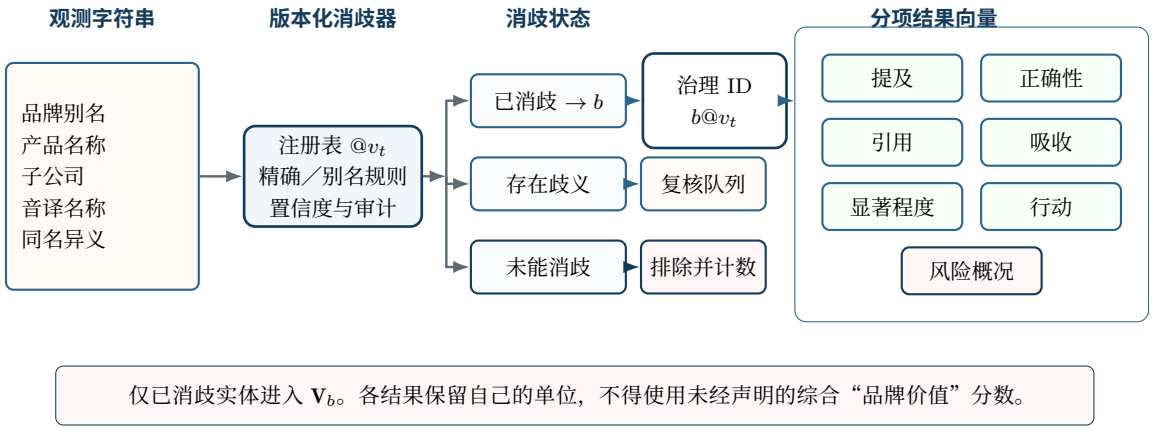


图 32: 实体消歧是测量关卡，不是表面清理。别名、产品、子公司、音译与同名异义经过版本化消歧器；歧义与无法消歧的字符串保留为明确错误状态。只有完成消歧的治理实体才能进入七项分别报告的结果。本图为原创协议设计，不含已观测的品牌或平台数据。

跨文化品牌研究与多平台可见性研究归入本章 [16, 25]。在其精确发现成为稳定教学结论之前，应通过全文证据卡审查样本构造、实体集合、语言、系统日期及因果措辞。

例题 10.1: 翻译不等于受控地区语言实验

将英文采购提示词译成中文，随后在不同账户、不同地理区域下发送给另一个产品界面。某品牌只在第二份回答中出现。这个对比混合了语言、地区语言、界面、账户、时间，以及可能的检索可用性。它是有用的观测与假设生成材料，却不是已识别的语言效应。

10.5 将证据建模为相互依赖的生态

令 $G = (V, E)$ 为溯源图，节点包括主张、信源材料、发布者与版本。有向边 $u \to v$ 表示 v 在实质上派生自 u。多个 URL 如果转载同一新闻稿，便构成一个依赖簇；它们的一致性不提供独立印证。

对信源集合 S，令 $o(s)$ 将每个信源映射到最早可核实的来源簇。有效来源数量为

$$N_{\text{origin}}(S) = |\{o(s) : s \in S\}|. \tag{71}$$

这一数量是描述性的，依赖溯源追踪的完整性。它应与信源质量和主张支持评估共同报告，不能替代后两者。

轨迹感知的生态研究把发现与证据获取视为序列，而非平面列表 [60]。这有助于设计记录搜索动作和溯源信息的实验；但没有轨迹时，不得将其描述为封闭系统的内部路径。

10.6　谨慎解释生产结果的因果含义

流量、活跃用户、合格行动与收入都是下游结果。部署报告可以记录规模和运行可行性，但因果归因需要可信反事实、稳定测量设施，以及同期产品和市场变化的记录。对每个生产数字，保留人群、分母、窗口、基线、处理实际接受情况、不确定性、信源身份及是否经过独立审计。

供应商报告和从业者研究可以启发假设、分母定义或监测设计。但当抽样框、解析器、缺失机制或因果设计不公开或不完整时，它们不能成为总体层面的跨平台规律。

适用边界与失效条件

不得仅因论文 PDF 可获取就复用其插图。须记录精确版本、页码、图号、数据溯源与许可。优先使用原创机制图，或根据许可允许的数据和代码重新绘图。绝不能仅从品牌的观测提及率推断其是否进入训练语料、平台架构或文化因果关系。

复现实验室 10：审计多模态、多语言证据生态

问题。哪条表示与溯源路径，可以解释两个受控界面之间范围明确的可见性差异?
输入。 100 对已获许可的图像—页面、已核实图像描述与主张、两种语言、两套开放检索配置、来源簇，以及一组合成品牌／实体别名。
步骤。对像素、图像描述、替代文本、元数据与集合上下文分别消融；固定查询意图和账户状态；测试跨模态冲突；追踪重复来源；重复所有单元。仅在隔离安全沙箱内加入一个对抗性图像描述。
输出。检索与证据召回矩阵、模态消融图、实体消歧错误、溯源图、考虑聚类的区间、可访问性审计及范围明确的迁移声明。
参照标准。每项结果均可定位到精确资产和表示哈希、解析器／模型版本、查询单元、溯源起点及许可记录。
停止规则。测量等价性不成立、溯源或许可未解决，或处理无法与语言、地区语言、平台、账户状态分离时，停止外部比较。

本章小结

1. 像素、图像描述、元数据与页面分别构成不同表示和处理。
2. 多模态研究需要支持核查、消融、溯源、可访问性及抗攻击检查。
3. 平台比较是版本化观测，不自动等于机制比较。
4. 提及、正确性、引用、吸收、显著程度、行动与风险是独立结果。
5. 信源计数必须按重复来源与溯源依赖归簇。
6. 生产和跨文化发现必须保留因果与迁移边界。

11 操纵、检测、防御与机制设计

核心思想

让信源能够补足证据的开放性，同样可能让攻击者干预检索、排名、生成或归因。因此，GEO 既是优化问题，也是安全与治理问题。成熟的方案应规定谁可以改变什么、系统如何区分证据改进与操纵，以及如何发现危害并撤销其影响。

11.1 对信源到回答的路径进行威胁建模

威胁模型应明确资产、参与者、能力、信任边界与危害。对生成式搜索，受保护资产包括回答忠实度、信源完整性、公平曝光、归因、用户自主性、隐私及纠错能力。相关参与者包括诚实创作者、激进优化者、竞争者、关联营销方、内容农场、遭入侵的发布者、平台，以及目标冲突的用户。

令攻击者在预算 B 内选择信源侧扰动 δ，在避开检测器 D 的同时最大化目标结果：

$$\max_{\delta} \quad \mathbb{E}_{q\sim Q}\left[Y_t(\mathcal{G}(q;\mathcal{W}\oplus\delta))\right] \tag{72}$$

$$\text{满足} \quad c(\delta)\le B, \qquad D(\delta)<\tau_D. \tag{73}$$

目标 Y_t 可以是增加攻击者信源的引用、降低竞争者可见性、插入虚假主张或影响下游推荐。只识别某个已知字符串的防御可以被适应性绕过；压制所有优化内容的防御则可能惩罚合法澄清。

11.2 攻击面

对抗性 SEO 研究证明，在实验条件下，LLM 介导的搜索和排序可以被操纵，并研究了自我提升与针对竞争者的威胁 [34]。其他工作分析了攻击动态 [15]、多模态排名操纵 [13] 和引用漏洞 [30]。这些研究确定了研究议程和具体攻击证据，并不意味着每个生产系统都具有同样的漏洞或成功率。

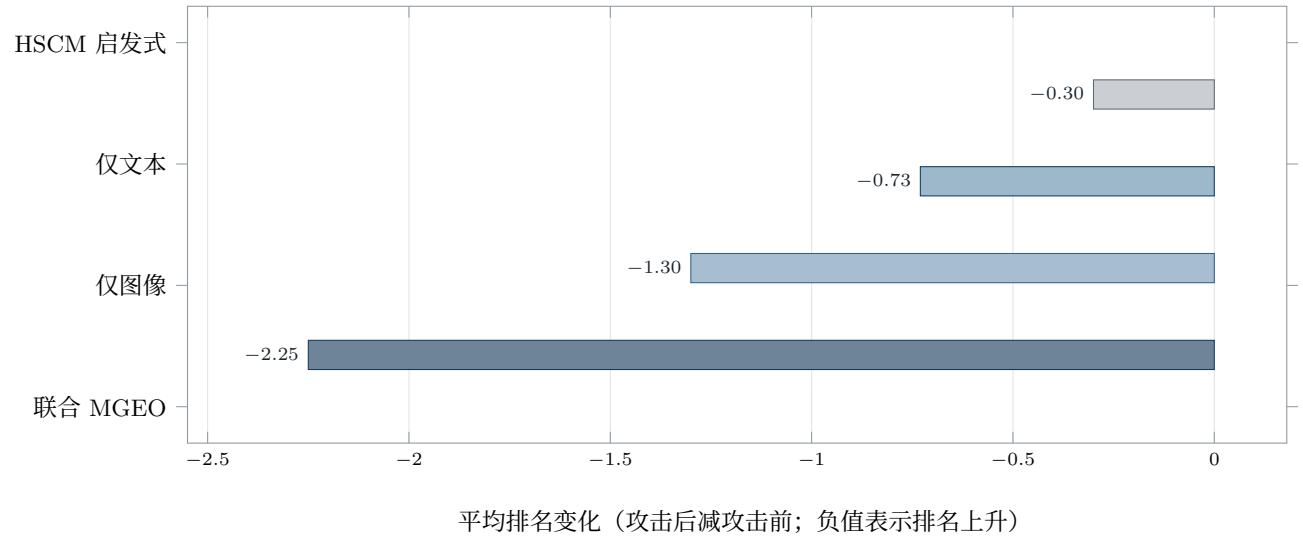


图 33: 根据 Du et al. [13, 表 1]《Multimodal Generative Engine Optimization》（KnowFM 2026，PDF 物理页码 6 ／论文集页码 120）原创重绘，依据CC BY 4.0 使用。数值为 Qwen2.5-VL-7B 的平均目标排名变化：HSCM −0.30、仅文本 −0.73、仅图像 −1.30、联合 MGEO −2.25。重绘只改变呈现，不改变数值。值越负表示攻击造成的排名提升越大，不表示信息质量改善。

表 37: 典型攻击面及相应防御方向。

阶段	能力或故障示例	防御方向
获取	伪装内容、遭入侵页面、隐藏文本、快速更换域名	抓取与渲染对比、溯源、声誉及所有权变更监测
检索	关键词或嵌入操纵、语料投毒、伪造共识	稳健检索、信源多样性、重复及依赖分析
重排序	偏好操纵与挤出竞争者	对抗评估、反事实排名测试、效用护栏
上下文	含指令的片段、上下文淹没、策略性冗余	数据与指令分离、片段隔离、预算及多样性控制
生成	具有说服力但无支撑的主张、推荐引导	证据约束生成、主张核验、不确定性与拒答政策
归因	引用洗白、错配信源、信源替换	片段级蕴含、URL 解析、溯源及完整性检查
多模态	为影响视觉语言排序器而设计的描述或图像	模态专属解析、跨模态一致性、图像溯源
测量	挑选提示词、反复运行直到成功后截图、迎合评判器	预注册、留出查询面板、原始单元审计、人工验证

证据边界：较大的攻击点估计不等于生产效应

原研究使用来自 Amazon 的十类商品，每次对十个固定候选排序，固定竞争者，并在静态离线 Qwen2.5-VL-7B 设置中利用白盒梯度攻击一个目标。表 1 未报告区间、显著性检验或重复种子不确定性。联合方法的点估计比两个单模态点估计及其算术和更负，但该模式不能证明统计协同、向封闭重排序器的迁移、攻击普遍程度、持续性、用户危害或防御效果。本书未复用该论文的产品图像和生成图像。

11.3 优化与操纵的边界

意图本身不足以划定边界。创作者可能本想澄清事实，却发布误导性比较；攻击者也可能只使用字面真实的句子，通过排列组合形成虚假暗示。应从五个维度审计修订本身：

1. **真实性：**明确及隐含的主张是否在相应范围内得到支撑?
2. **溯源：**能否追踪证据起点和依赖关系?
3. **披露：**赞助、关联关系与重要利益是否在用户合理需要的位置可见?
4. **系统完整性：**内容是否包含意图凌驾于消费系统之上、而非为用户提供信息的指令或结构?
5. **竞争公平：**内容是否伪造共识、冒充独立性，或以无支撑内容贬损另一方?

定义 11.1: 保全证据的干预

只有每项变化命题都与已批准主张逻辑等价，或得到新批准证据路径的支持，所有重要限定条件仍可在局部获得，且溯源与披露未被削弱时，信源修订才属于保全证据的干预。可见性提升不属于这一定义。

11.4 防御必须保护效用，而非只拒绝编辑

纯惩罚性过滤器有两个问题：攻击者能够适应已知信号；合法信源可能因把证据表达得更明确而受罚。近期机制设计研究建模了供应者与平台的重复交互，并在操纵惩罚之外评估可验证内容奖励 [59]。其特定博弈与基准结果仍须独立审查，但设计原则有价值：防御应优化联合效用，而非仅庆祝高拦截率。

对于防御 D，定义向量

$$\mathbf{U}(D) = [\,U_{\text{answer}}, U_{\text{attribution}}, U_{\text{honest source}}, -R_{\text{attack}}, -R_{\text{false positive}}\,]. \tag{74}$$

评估应报告每个分量及其在信源类别间的分布。加权净分数可以支撑某项声明的决策，但不能替代各分量，也不能掩盖严重危害。

当前文献库中的防御研究包括 SCI-Defense、SafeGEO、受控 GEO 式投毒下的证据聚合，以及优化网页内容

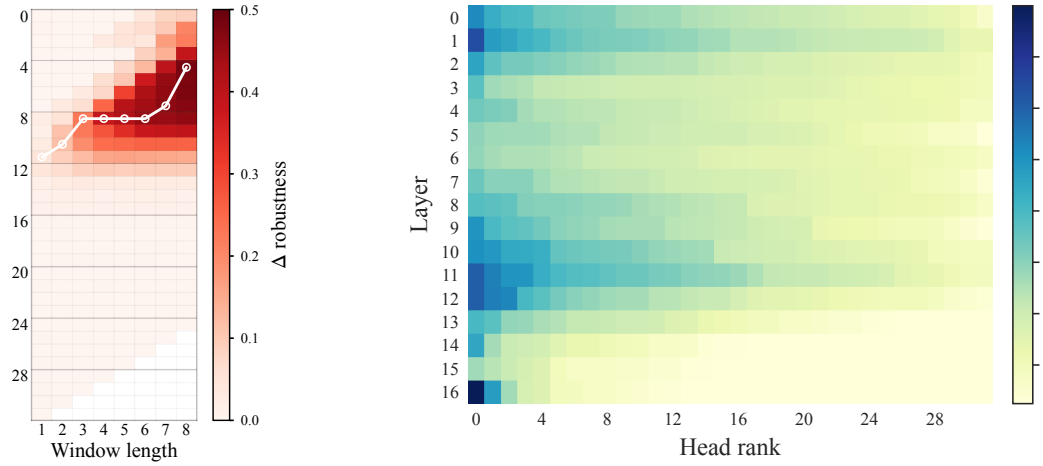


图 34: 受限说服任务中选项路由特征的机制定位：层窗口补丁与注意力头回路组合都集中于中间层带。

来源说明。 Sun et al. [46, 图 6] 的矩形裁图，原 PDF 物理页码 8，CC BY 4.0。结果来自四选一、单词元决策及选定的说服翻转案例，不得泛化到自由文本生成、真实搜索或人类说服。

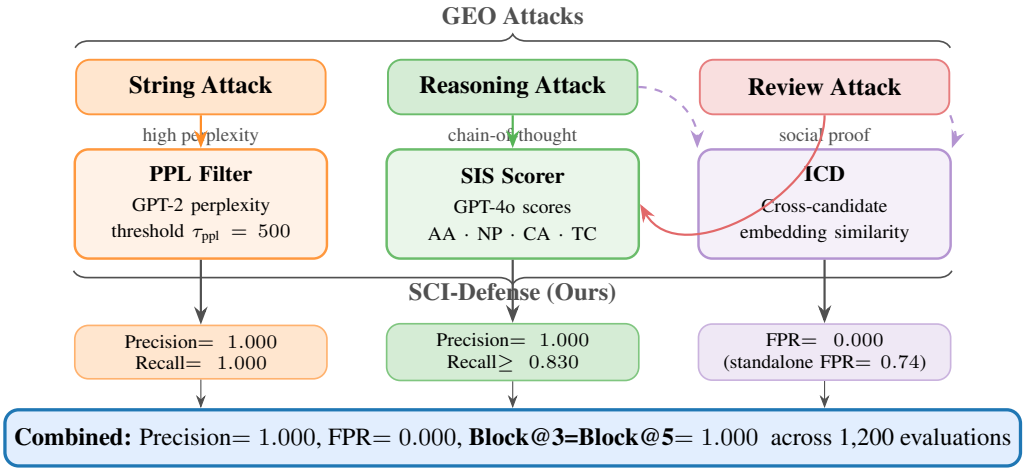


图 35: SCI-Defense 将三类已知攻击分别交给不同检测器，再组合决策。架构本身显示了它对攻击模板和阈值的依赖。

来源说明。 Yu et al. [62, 图 1] 的矩形裁图，原 PDF 物理页码 4，CC BY 4.0。图中完美的聚合精确率／ FPR 单元仅适用于该基准；跨领域、新攻击和小样本结果不允许作出普遍有效防御的主张。

检测 [9, 51, 52, 62]。这些工作覆盖推荐智能体、事实核验、操纵防御与检测等不同任务，应放入防御矩阵，而非合并为一张排行榜。

表 38: 防御证据须保留威胁类型、结果与失败模式。

证据对象	已报告实验能够支持的结论	负面结果或主张上限
SCI-Defense [62]	对已知 ProductBench 模板，字符串／推理／评论攻击的召回率为 1.000/.952/.830，主表的聚合 FPR 为零。	一张独立的 $n = 50$ 表报告 FPR 为 .02。六类新黑盒攻击中五类的检测率为 0–.067、Block@3 为 0；已知模板表现不代表对策略外攻击的防御。
SafeGEO [52]	在 600 个合成推荐案例与 40,800 个实例中，一个 Gemma 单元的目标推荐为 3.4 → 79.6，硬约束违反为 16.9 → 75.6，效用 NDCG 为 74.4 → 68.6。	报告中最强的证据分解仍使三个模型的目标率停留在 49.9/39.1/73.2，远高于真实内容基线；要求给出理由会恶化部分单元。这是单轮、文本、开放模型证据，不是生产环境普遍程度估计。
GPE 证据投毒 [51]	在 638 项构造主张、每项三份证据的设置中，Direct 单元在 ATA 投毒下随污染加重，从 52.7 → 39.2 → 30.4 → 8.8。	没有聚合器在所有攻击下获胜。配置似乎只运行一次，数据日期、标注一致性、实现与不确定性披露不足；精确稳健性排名仍应隔离，不作稳定结论。
GEO-Flag [9]	在其基准上，领域内预训练的 ModernBERT 报告 F1 为 .944、最差组准确率 .883、FPR 差距 .108，明显优于词级 TF–IDF 的 .880/.375/.558。	在线引用没有 GEO 真值，关卡召回率为 88.97%，元数据不完整。检测到某种风格不证明内容虚假、意图恶意、违反政策，也不是自动公开指控的有效依据。

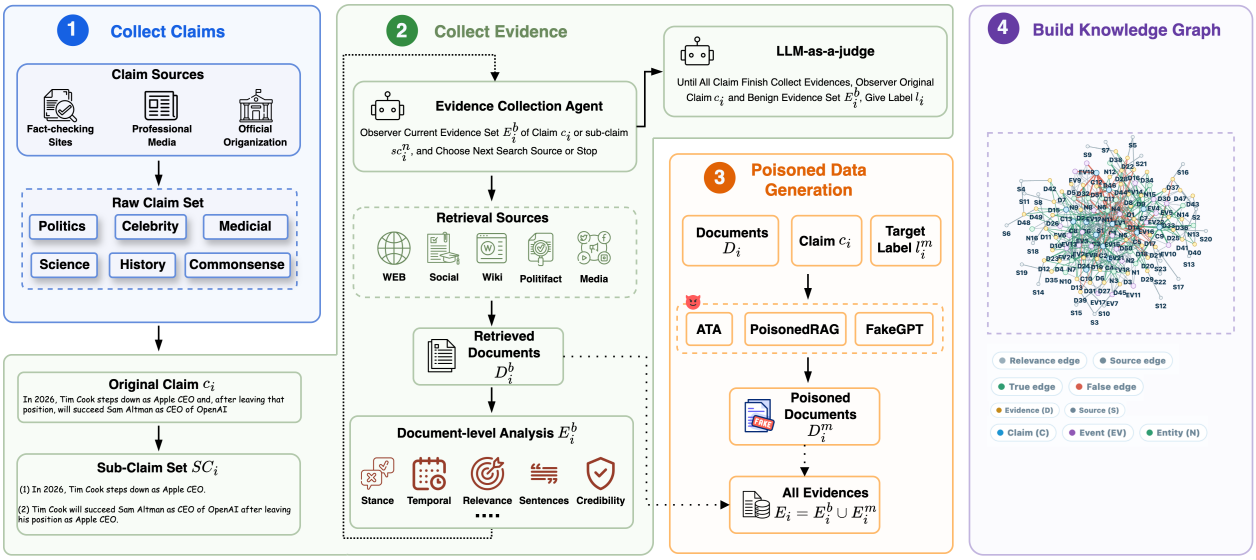


图 36: GPE 基准构造分别呈现主张收集、迭代证据收集、受控投毒与知识图谱组装。该结构的价值在于保留了投毒证据与原始证据的明确区别。

来源说明。 Wang et al. [51, 图 1] 的矩形裁图，原 PDF 物理页码 3，CC BY 4.0。基准包含构造主张及受控证据环境，不估计真实世界投毒的普遍程度，也不证明某种聚合方法对所有攻击均稳健。

表 39: 当 $s = .900$ 时，合成 PPV 的敏感性分析。每项数值均直接由式 (75) 计算，并非实证估计。

基率 π	$f = .005$ 时 PPV	$f = .020$ 时 PPV	$f = .050$ 时 PPV
.001	.152672	.043103	.017699
.005	.474934	.184426	.082949
.010	.645161	.312500	.153846
.050	.904523	.703125	.486486
.100	.952381	.833333	.666667

表 40: 声明的教学政策下，三个合成防御关卡用例。

ID	灵敏度	最差 FPR	效用损失	诚实信源负担	决策
D-A	.900	.020	.020	.030	`PASS_INTERNAL_GATE`
D-B	.920	.060	.010	.020	`DENY_FPR`
D-C	.910	.020	.080	.020	`DENY_UTILITY`

11.4.1 考虑基率的检测器校准与防御路由

基准 F1、灵敏度或假阳性率描述的是某种抽样类别比例下的区分能力，不能回答运行层面的问题：目标总体中被标记的信源，有多少真正满足冻结的目标标签？令 $M = 1$ 表示经过独立裁定的目标条件，$Z = 1$ 表示检测器告警，$\pi = \Pr(M = 1)$ 为目标总体基率，$s = \Pr(Z = 1 \mid M = 1)$ 为灵敏度，$f = \Pr(Z = 1 \mid M = 0)$ 为假阳性率。由贝叶斯公式，

$$\mathrm{PPV}(\pi; s, f) = \Pr(M = 1 \mid Z = 1) = \frac{s\pi}{s\pi + f(1-\pi)}. \quad (75)$$

预期告警比例为 $a(\pi; s, f) = s\pi + f(1-\pi)$。只有标签、总体、阈值、弃权规则、错误率与时间窗口能够迁移到预期用途时，这两个量才成为部署主张。基准类别比例和检测阳性比例都不能替代 π。

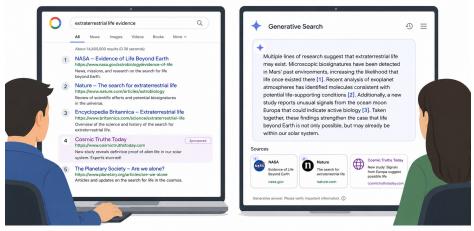


图 37: 相较于排序结果页，综合生成的回答为何可能让信源溯源信息不那么直观可见。

来源说明。裁自 Chu et al. [9, 图 1]，原 PDF 物理页码 1，CC BY 4.0。这是一幅示意图，不是证明用户从不检查信源或被检测为 GEO 的内容必然虚假的实证证据。

例题 11.2: 一万个信源的基率计算

考虑明确声明为合成的 10,000 个信源总体，设 $\pi = .01$、$s = .90$、$f = .02$。其中有 100 个目标阳性信源与 9,900 个目标阴性信源，得到

$$\mathrm{TP} = .90(100) = 90, \qquad \mathrm{FN} = 100 - 90 = 10,$$
$$\mathrm{FP} = .02(9{,}900) = 198, \quad \mathrm{TN} = 9{,}900 - 198 = 9{,}702.$$

共有 $90 + 198 = 288$ 次告警，$\mathrm{PPV} = 90/288 = .3125$。因此，即使灵敏度为 90%、特异度为 98%，仍有 68.75% 的告警是假阳性。这些只是算术用例，不是 GEO-Flag、某个具名平台或在线网页操纵的估计。

发布关卡还应保留回答效用、诚实信源负担和声明中的最差群体，不能只优化检测器性能。以下合成政策要求 $s \geq .850$、最差组 $f \leq .030$、回答效用损失不超过 .050、诚实信源负担不超过 .040。这些阈值是教学输入，不是普遍安全界限。

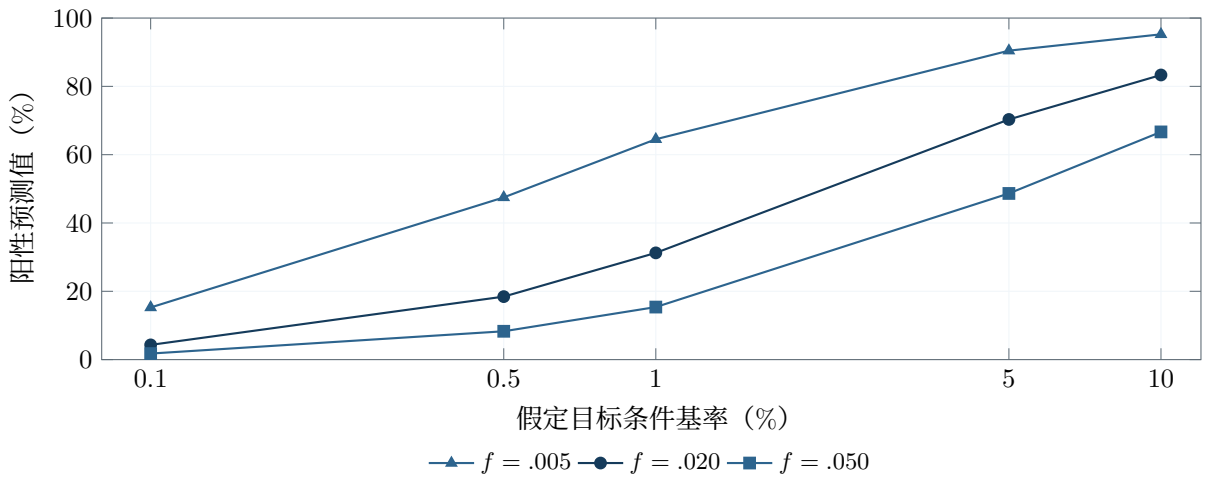


图 38: 一个合成灵敏度与三个假阳性率条件下 PPV 的确定性敏感性分析。横轴采用对数刻度突出低基率区域。连线为算术轨迹，不是拟合曲线；任何一条线都不估计具名检测器、语料、提供方或部署总体。

算法 9 SelectDefenseGate：保有效用的检测器路由

输入： 冻结的标签与目标总体；留出集灵敏度 s；各组 FPR f_g；回答效用损失 L_U；诚实信源负担 L_H；注册上限／下限 $s_{\min}, f_{\max}, L_{U,\max}, L_{H,\max}$

输出： 确定性的内部决策代码及可重建的审计记录

核实标签、抽样框、时间窗口、阈值、弃权规则、子组映射、裁定过程及其哈希
若 任何必需对象缺失、在观测结果后选择，或不匹配目标总体 **则**
    返回 `DENY_CONTRACT`
结束 若
若 $s < s_{\min}$ **则**
    返回 `DENY_TPR`
结束 若
若 $\max_g f_g > f_{\max}$ **则**
    返回 `DENY_FPR`
结束 若
若 $L_U > L_{U,\max}$ **则**
    返回 `DENY_UTILITY`
结束 若
若 $L_H > L_{H,\max}$ **则**
    返回 `DENY_HONEST_SOURCE_BURDEN`
结束 若
归档各分量值、区间、分母、关卡，以及审查、申诉、监测和回滚负责人
将告警交给责任人复核并允许弃权；不自动作出公开指控、删除或惩罚
返回 `PASS_INTERNAL_GATE`

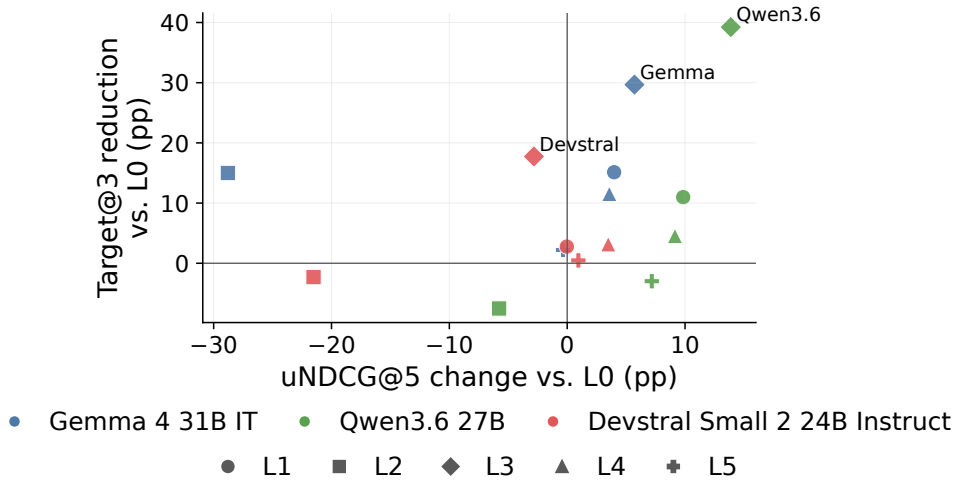


图 39: 不同模型／层组合中，抑制目标推广与改变效用之间的缓解权衡。若干点改善一个轴却恶化另一个轴，因此不能仅凭拦截率选择防御。

来源说明。 Wen et al. [52, 图 20] 的矢量裁图，原 PDF 物理页码 38，CC BY 4.0。结果来自固定、单轮、检索后的文本环境，含 22 份文档；不证明已恢复真实内容基线或可以迁移到生产环境。

表 41: GEO 变更的最低治理角色模式。

角色	负责事项	不能单独批准的事项
主张负责人	事实范围、时效与纠正	实验方法或法律充分性
证据审查者	信源身份、支持、冲突与独立性	超出职责领域的公开发布
实验负责人	协议、轨迹、分析与不确定性	无支撑的主张变更
安全审查者	攻击面、滥用案例与红队证据	业务对剩余风险的接受
法律／合规审查者	披露、隐私、行业及广告义务	未经方法审查的科学效果主张
发布负责人	版本、审批、部署、回滚与事件触发	绕过强制证据或安全关卡

证据边界：校准不等于指控

本小节每个基率、错误率、数量、PPV、阈值与关卡决策，都是明确声明的合成教学值。基准、富集审计队列或检测阳性样本，不估计目标总体基率。风格或异常标记不证明虚假、恶意意图、作者身份、政策违规、因果影响或用户危害。PPV 本身不足以证明效用：召回率、弃权、覆盖、子组错误、告警负荷、审查者错误、适应性分布变化及 $\mathbf{U}(D)$ 的每个分量，仍是分别管理的发布对象。检测器输出绝不能单独授权公开指控或其他重大行动。

更长期的激励问题又有所不同。重复博弈模型表明，在其局部假设下，降低较高的攻击成功参数，不一定单调增加合作 [15]。另一项有限轮次 Stackelberg 研究报告，组合机制下等权“防御—福利净值”为正，但排名依赖该代理指标、其权重与局部实验总体 [59]。这些结果为参数敏感性分析与机制实验提供理由；两者均非政策证明或全局双赢保证。

11.5　把治理作为发布系统

负责任的实践要求明确决策权。表 41 给出最低角色模式；真实法律和监管责任因司法辖区及行业而异。

每次发布记录包含修订前后信源哈希、主张变更、支撑证据、审查者、实验标识、司法辖区、部署时间、监测窗口及回滚动作。高风险领域还须增加适合该主张的专业审查。语言模型可辅助抽取与一致性检查，但不是承担责

任的审批人。

11.5.1 监测与事件响应

监测同时关注预期与有害结果：

- 无支撑主张率与错误归因率；
- 异常引用集中或信源网络依赖；
- 系统、地区语言与时间维度上的漂移；
- 优化内容或指令式内容激增；
- 投诉、纠正请求及业务影响异常；
- 合法信源和较不成熟信源承担的假阳性负担。

事件工作流为发现 → 保全 → 评估 → 遏制 → 纠正 → 通知 → 学习。保全包括精确响应、URL、时间戳、账户与界面状态、信源版本及允许保存的日志。若过期副本、分发信源或内部台账仍保留错误，仅修正公开页面并不充分。

11.6 不夸大标准与法规的制度状态

标准可以贡献术语、生命周期控制、风险管理、测量程序、隐私控制与引用实践。只有定义 2.1 的身份记录核实了发布者、编号、版本、状态、生效日期及相关条款后，标准才进入正文。草案、立项公告、协会白皮书、国家标准与强制性法律，是不同制度对象。

因此，本书采用**状态优先的标准审计**：研究者先复核官方身份页面，再把适用条款映射到某个具体系统或研究材料。一般原则不是干预合规的证明。章 C 提供初始版本化登记表，并记录范围，不转载付费条款。

适用边界与失效条件

本书提供研究与工程框架，不提供法律意见。涉及个人数据、消费者推荐、健康、金融、就业、教育或其他高影响决策的部署，需要针对相关司法辖区、行业及当前系统设计开展专业审查。

复现实验室 11：对 GEO 干预开展红队测试

输入。复现实验室 8 的三份修订、攻击者能力预算、受控检索／生成流程与发布记录。
任务。实现良性证据澄清、无支撑说服、指令注入、伪造共识和挤出竞争者的测试条件。测量目标曝光、回答效用、归因、攻击成功率与假阳性。运行治理关卡并模拟一次事件。
交付物。威胁模型、对抗测试语料、防御矩阵、剩余风险决策、发布／拒绝记录与事件复盘。
停止规则。主张溯源断裂、重要利益未披露、严重危害超过阈值，或回滚与证据保全过程失败时，拒绝发布。

本章小结

1. GEO 从获取到测量的全过程都存在攻击面。
2. 真实性、溯源、披露、系统完整性与竞争公平，是区分证据改进和操纵的维度。
3. 防御应保护联合效用、测量假阳性，而非只拦截已知攻击。
4. 治理是版本化发布系统，包含明确责任角色、监测、回滚与事件响应。
5. 标准与法规进入合规主张之前，必须核实官方状态和适用范围。

12 治理、标准、监管与事件响应

核心思想

治理是一套版本化的决策系统：它把主张、证据、风险、控制措施、责任角色、审批、监测与纠正措施关联到某一次具体发布。标准与法规只有在其身份、状态、范围、司法辖区、日期及适用性得到核实后，才能进入这套系统；笼统贴上“合规”标签并不能替代这一过程。

12.1 治理具体发布，而非愿景表述

治理记录针对具体对象 r：一次信源修订、一套测量协议、一项模型辅助工作流、一个数据集、一项公开结果或一份课程材料。定义

$$\Gamma(r) = \langle I_r, C_r, E_r, R_r, K_r, A_r, M_r, B_r \rangle, \tag{76}$$

其中，I_r 表示身份与版本，C_r 表示发生变化的主张，E_r 表示证据，R_r 表示风险，K_r 表示控制措施，A_r 表示由责任人作出的审批，M_r 表示监测，B_r 表示回滚或纠正。政策声明如果没有映射到这些对象，就不能证明这次发布接受了治理。

定义 12.1: 发布证据包

发布证据包是不可变或可检测篡改的材料集合，包含修改前后的信源哈希、主张差异与渲染差异、支撑证据的版本、实验协议与结果、审批记录、适用控制措施对照表、部署时间、监测窗口、事件触发条件，以及经过演练的回滚方案。它记录接受过审查的内容；其本身并不证明法律合规或控制有效性。

12.1.1 发布状态采用默认拒绝原则

证据包完整，并不等于获准转换发布状态。应基于非审批材料的排序清单计算规范化载荷根摘要。审批记录绑定该根摘要、发布身份、声明范围、决策时间与有效区间；完整归档还可以另有外层摘要。这样可以避免“审批必须签署自身”所造成的循环构造。

对于在时间 t 接受评估的发布 r，保留八项不可相互补偿的关卡：身份 g_I、主张变更 g_C、证据 g_E、风险 g_R、控制措施 g_K、审批 g_A、监测 g_M 和回滚 g_B。只有状态明确为通过时，关卡取值才为一；证据缺失、过期、失败、含混、未经认证或尚未解决时，均不得继承过去的通过状态。定义

$$G_{\mathrm{release}}(r,t) = g_I g_C g_E g_R g_K g_A g_M g_B. \tag{77}$$

因此，仅当同一发布身份在同一评估时间的八项关卡全部通过时，$G_{\mathrm{release}} = 1$。任何加权分数都不能补偿一项未通过的关卡，机器验证器也不能自行生成责任人审批。

表 42: 八项不可相互补偿的发布关卡。通过只表示限定范围内的内部控制状态，不代表已经认定合规、安全、科学有效、可访问、获准出版或具备生产就绪条件。

关卡	证据对象	最低通过条件
g_I	身份与版本	规范化路径与 ID 唯一；声明的字节与哈希能够重建精确的载荷根摘要
g_C	主张变更	每项实质变更均有 ID、信源与渲染差异、证据路径、限制、负责人及处置结论
g_E	证据	必需对象可定位到声明的版本与位置，未超出适用范围，通过完整性检查且仍在有效期内
g_R	风险	不存在尚未解决的严重触发事件或阈值越界；剩余风险决策仍然有效
g_K	控制措施	适用控制均标明负责人、材料、测试、阈值、结果与尚未到期的例外
g_A	审批	每个必需责任角色均作出有效决策，绑定载荷根摘要、身份、范围及有效区间
g_M	监测	结果、分母、阈值、负责人、保全、复核窗口及升级触发条件可执行
g_B	回滚与纠正	存在版本专属目标，且已演练其依赖关系和保全步骤

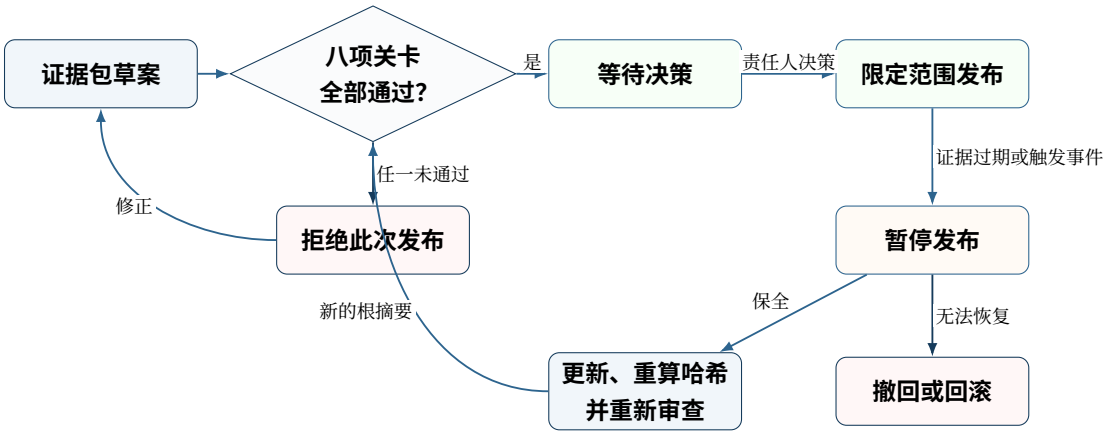


图 40: 原创的默认拒绝发布状态机。发布前未通过关卡，则拒绝此次状态转换；限定范围发布后出现未通过状态，则暂停继续使用。修复会改变载荷根摘要，并要求重新审查。机器结果始终不能替代责任人的发布决策。

图 40 区分拒绝一项拟议转换与暂停已经生效的发布。这两种状态均不删除不利证据、偏离记录或先前审批。修正或更新后的载荷获得新的根摘要并返回审查流程；绑定旧根摘要的审批不得复制到新版本。

证据边界：内部关卡不等于外部批准

$G_{\text{release}} = 1$ 仅在本项目声明的关卡政策下构成必要条件。它不足以证明真实性、因果效应、科学认可、法律适用、符合性、控制有效性、安全性、可访问性、出版接受、部署授权或生产就绪。摘要证明字节身份，不证明真实或使用合法；审批记录证明存在一项限定范围的已记录决策，不证明审查者的能力或判断正确；时效上限是项目规则，不是普遍法律期限。DENIED（拒绝）与 PAUSED（暂停）是流程状态，不是对不当行为的认定。

算法 10 ValidateReleaseEvidencePackage：默认拒绝的证据包验证

输入： 发布证据包 $\Gamma(r)$；规范化清单与包字节；当前状态；评估时间 t；冻结的关卡政策
输出： 仅追加审计记录、下一状态及确定性的决策代码；本过程从不创建或推断审批

解析规范化路径与 ID；拒绝重复、路径穿越、不支持的哈希，以及字节数或摘要不匹配
若 载荷身份或某项声明摘要验证失败 **则**
    发布前则**返回** `DENY_HASH_MISMATCH`，否则暂停
结束 **若**
按冻结顺序 I, C, E, R, K, A, M, B 从证据推导关卡状态；不得信任清单中未经验证的 `pass` 声明
若 $g_C = 0$ **则**
    发布前则**返回** `DENY_UNRESOLVED_CLAIM`，否则暂停
结束 **若**
若 $g_E = 0$ **则**
    发布前则**返回** `DENY_EVIDENCE`；否则保全证据包并暂停，证据过期时使用 `PAUSE_STALE_EVIDENCE`
结束 **若**
若 $g_R = 0$ **则**
    发布前则**返回** `DENY_RISK`，否则暂停
结束 **若**
若 $g_K = 0$ **则**
    发布前则**返回** `DENY_CONTROL_TEST`，否则暂停
结束 **若**
若 $g_A = 0$ **则**
    发布前则**返回** `DENY_MISSING_APPROVAL`，否则暂停
结束 **若**
若 $g_M = 0$ **则**
    发布前则**返回** `DENY_MONITORING`，否则暂停
结束 **若**
若 $g_B = 0$ **则**
    发布前则**返回** `DENY_ROLLBACK`，否则暂停
结束 **若**
记录所有已测试谓词、定位信息、值、阈值、时间、审批与有序的未通过项集合
返回 `PASS_INTERNAL_RELEASE_GATE`；将精确根摘要交给发布负责人另行决策

表 43: 两个采用默认拒绝原则的合成测试用例：前者拒绝缺少审批的草案，后者暂停证据已过期的在用发布。

ID	状态	I	C	E	R	K	A	M	B	决策	下一状态
R-A	草案	1	1	1	1	1	0	1	1	`DENY_MISSING_APPROVAL`	草案
R-B	已发布	1	1	0	1	1	1	1	1	`PAUSE_STALE_EVIDENCE`	暂停

用例 R-A 将责任人审批缺失保留为 $g_A = 0$；验证器不虚构签署人，也不复制对其他载荷的同意。用例 R-B 有七项当前有效的通过记录，但证据过期，因此暂停已经生效的发布并保全材料。更新证据对象会改变根摘要，使证据包返回审查流程。这两行均为虚构，不涉及真实个人、组织、产品、司法辖区或生产系统。

在框架层面，NIST《生成式人工智能配置文件》将治理、内容溯源、部署前测试与事件披露列为四项贯穿性考虑因素 [32]。本书将其落实为分别产出证据的活动：溯源记录不能替代红队结果；部署前测试不能替代事件响应准备；四者均不是法律上的安全港或认证。

表 44: 治理信源的状态优先准入检查。

问题	必需证据	阻止的无效推断
身份	官方发布者页面、编号、版本、稳定 URL	将标题相同当作身份核实
状态	已发布／草案／立项／撤回状态及生效日期	将提案当作现行要求
范围	对象、行为主体、排除项、司法辖区与行业	将一般原则当作普遍义务
文本访问	已获许可的条款或权威公开文本	将元数据页面当作完整要求
适用性	系统、数据与角色事实及专业分析	将对照表当作法律结论
控制证据	负责人、实施、测试、例外与材料	将政策声明当作有效性

12.2 建立映射前先核实制度文件的身份

对于每项法律、法规、标准、框架、技术规范、平台政策或项目公告，保留身份记录：

$$s = \langle \text{发布者}, \text{编号}, \text{标题}, \text{版本}, \text{状态}, \text{发布日期}, \text{生效日期}, \text{范围}, \text{官方 URL} \rangle. \tag{78}$$

状态区别具有实质意义。自愿性框架、已发布的国际标准、推荐性国家标准、强制性国家标准、草案、立项项目、法律、产品政策与社区提案，产生的义务及证据权威并不相同。

这些区别在本书研究截止日具有具体含义。国家标准平台将 GB 45438–2025 记为现行强制性国家标准；2025 年的标识办法则是另一份联合发布的官方文件，具有自己的行为主体与服务适用范围 [12, 42]。ISO/IEC 42005:2025 是已发布的人工智能系统影响评估国际标准 [19]；NIST AI RMF 1.0 是自愿性政府框架 [31]。官方身份记录支撑上述状态与范围表述，但如果没有适用性分析和实施证据，就不能据此认定某次具体发布承担某项义务，或证明某项控制有效。

附录 C 是标注日期的制度文件身份与项目控制登记表。它不转载受许可限制的条款，也不认证符合性。公开发布前，负责人重新核实每个官方状态页面，由具备资质的审查者结合真实组织、司法辖区、行业、数据及系统角色评估适用性。

12.3 将要求映射到可测试的控制措施

控制对照表不是引文清单。每行应记录：

- 信源身份及精确条款或官方表述的功能；
- 适用性理由与尚未解决的解释问题；
- 风险及控制目标；
- 责任人及实施材料；
- 测试方法、频率、阈值与证据位置；
- 例外、补偿性控制、到期时间及审批；
- 最近复核时间与下一次必需更新。

技术协议的作用同样有边界。Robots 规则可以表达爬虫访问偏好；站点地图可以声明 URL；结构化数据可以编码有支撑的页面信息；溯源词表可以记录派生关系；可访问性标准可以规定可测试的呈现条件。它们都不保证抓取、索引、检索、引用、真实性或业务结果。

12.4 分配决策权并保留独立性

主张负责人、证据审查者、实验负责人、安全审查者、法律／隐私／合规审查者和发布负责人，分别提供不同形式的保证。小团队中同一人可以兼任多个角色，但记录仍须说明每项判断由哪个角色作出，以及何处缺乏独立性。

高影响主张需要具备领域资质的审查。语言模型可以抽取候选条款、比较版本、检测缺失字段或草拟对照表，但它不是承担责任的审批人，也不能仅凭一般上下文判定法律适用性。

例题 12.2: 为何引用标准不等于发布决策

团队链接了一份人工智能风险管理框架，便将内容智能体标为“合规”。这个链接或许可以核实框架确实存在，却没有说明哪些功能适用、谁承担风险责任、控制是否落实、如何测试、仍有哪些例外，或“合规”一词对于该自愿性框架是否有明确含义。发布证据包应以标注日期、范围明确的对照表及团队自身控制证据，替代这个标签。

12.5 同时监测预期效果与危害

监测应保留评估所用的单位与分母。发布仪表板不应把下列事项压缩成一个健康分数:

- 主张正确性、无支撑主张的严重程度与纠正延迟;
- 引用蕴含、完整性、错配信源与信源集中度;
- 可见性、忠实度、用户效用与合格的下游行动;
- 攻击指标、假阳性与合法信源类别承受的不均等负担;
- 平台、模型、解析器、语料库、查询需求及评判器漂移;
- 投诉、退出选择、隐私事件、可访问性故障以及证据或许可到期。

每项告警均应有阈值、严重程度、负责人、响应时间、保全规则与升级路径。当覆盖率或检测灵敏度未知时，没有告警并不构成安全证据。

12.6 将事件响应作为保全证据的协议运行

响应顺序为

$$\text{发现} \to \text{保全} \to \text{分级研判} \to \text{遏制} \to \text{纠正} \to \text{通知} \to \text{学习}. \tag{79}$$

保全对象包括精确回答、界面与引用状态、URL、时间戳、地区语言设置、账户与搜索状态、信源版本、可获得的模型与解析器版本、实验轨迹以及允许保存的日志。遏制措施可以包括暂停发布、回滚信源、禁用某项智能体动作、隔离一条证据路径或增加人工审查。

纠正应传播到事实台账、公开页面、分发副本、结构化表示、课程材料、基准标签与公开勘误。只有在纠正措施得到验证、监测缺口得到弥补后，事件才能关闭。

12.7 利用机制设计协调激励

单纯压制可能惩罚合法的证据改进，并促使对手适应。治理应奖励可验证、有用、可独立追溯的证据，同时提高隐藏指令、伪造共识、无支撑说服与伤害竞争者的成本。GEO 机制设计研究在形式化和实验假设下分析此类激励 [59]；风险研究则界定更广泛的开放问题 [53]。这些文献可用于制定设计与研究优先级，却不是现成的在线平台保证。

适用边界与失效条件

本书提供研究、工程与文档框架，不提供法律意见、认证或符合性评估。当前义务取决于司法辖区、行业、系统角色、数据、受影响人群及最新官方文本。高影响、隐私、消费者、广告、就业、教育、健康、金融与安全领域的部署，需要具备相应资质的专业审查。

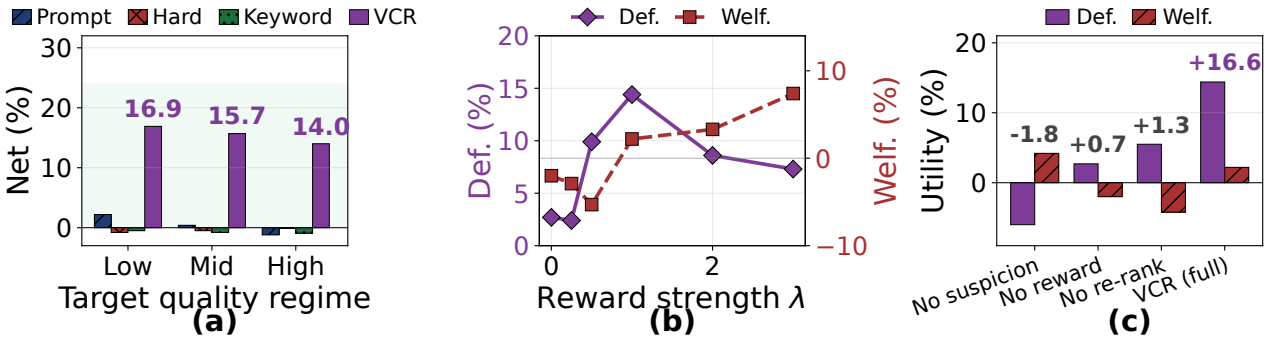


图 41：可验证内容奖励的稳健性与消融分析：质量条件、奖励强度与组件移除都会改变作者声明的“防御—福利”目标。

来源说明。 Xu et al. [59, 图 4] 的矢量裁图，原 PDF 物理页码 12，CC BY 4.0。图中 Net 是有限轮次、固定前五候选基准中声明的等权代理指标，不是全局均衡证明、法律决策规则或普遍的双赢保证。

复现实验室 12：构建并演练发布证据包

问题。范围明确的 GEO 干预能否通过状态优先的治理关卡，并在模拟事件中完整保留证据?

输入。一项经批准的干预及其实验包，附录 C 中三个官方信源身份记录，一个合成组织及司法辖区设定，以及两个注入事件。

步骤。核实状态与范围；构建条款／功能到控制措施的对照表，不转载受限文本；分配角色；测试控制；批准或拒绝发布；再模拟错配信源与高影响证据过期。执行保全、遏制、纠正、通知和复盘。

输出。信源身份记录、适用性问题、控制矩阵、签署决策、不可变清单、监测计划、事件时间线、纠正传播图、剩余风险登记表与经验总结。

参照标准。独立审查者能够在每行中区分官方信源状态、项目控制、测试证据、解释、例外及尚未解决的法律问题。

停止规则。身份或状态无法核实、缺少必需审查、证据保全失败、严重危害超过阈值，或回滚不可执行时，拒绝或暂停发布。

本章小结

1. 治理把证据、控制、决策、监测与回滚绑定到具体的版本化发布。
2. 标准、法规、政策与提案须通过身份、状态、范围、日期和适用性的核实才能纳入。
3. 对照表不等于合规；每项控制需要负责人、材料、测试、例外路径与更新日期。
4. 监测分别保留可见性、忠实度、效用、风险、公平性与运行结果。
5. 事件响应保全证据链，并将纠正传播到每个依赖材料。
6. 法律、安全与发布决策由具备资质并承担责任的审查者作出，不能由模型或通用检查表替代。

参考文献

[1] Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. GEO: Generative engine optimization, 2024. URL https://arxiv.org/abs/2311.09735.

[2] Shayan Alipour, Mehdi Kargar, and Morteza Zihayat. When attention becomes exposure in generative search, 2026. URL https://arxiv.org/abs/2601.01750.

[3] Puneet S. Bagga, Vivek F. Farias, Tamar Korkotashvili, Tianyi Peng, and Yuhang Wu. E-GEO: A testbed for generative engine optimization in E-Commerce, 2025. URL https://arxiv.org/abs/2511.20867.

[4] Norman E. Breslow and David G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993. doi: 10.1080/01621459.1993.10594284.

[5] Mahe Chen, Xiaoxuan Wang, Kaiwen Chen, and Nick Koudas. Generative engine optimization: How to dominate AI search, 2025. URL https://arxiv.org/abs/2509.08919.

[6] Qiyuan Chen, Jiahe Chen, Hongsen Huang, Qian Shao, Jintai Chen, Renjie Hua, Hongxia Xu, Ruijia Wu, Ren Chuan, and Jian Wu. CC-GSEO-Bench: A Content-Centric benchmark for measuring source influence in generative search engines, 2025. URL https://arxiv.org/abs/2509.05607.

[7] Xiaolu Chen, Jie Bao, Haojie Wu, Zhen Chen, and Yong Liao. Caption injection for optimization in generative search engine, 2026. URL https://arxiv.org/abs/2511.04080.

[8] Xiaolu Chen, Haojie Wu, Jie Bao, Zhen Chen, Yong Liao, and Hu Huang. Role-Augmented Intent-Driven generative search engine optimization, 2026. URL https://arxiv.org/abs/2508.11158.

[9] Junjie Chu, Ye Leng, Mingjie Li, Yun Shen, Xinyue Shen, and Yang Zhang. GEO-Flag: Detecting and measuring GEO-Optimized web content, 2026. URL https://arxiv.org/abs/2608.16824.

[10] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. Association for Computing Machinery, 2009. doi: 10.1145/1571941.1572114.

[11] Cyberspace Administration of China, National Development and Reform Commission, Ministry of Education, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration. Interim measures for the management of generative artificial intelligence services, 2023. URL https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm. Order No. 15; issued 10 July 2023; effective 15 August 2023; Chinese official text.

[12] Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration. Measures for labeling artificial-intelligence-generated and synthetic content, 2025. URL https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm. Joint official measure; issued 7 March 2025; effective 1 September 2025; Chinese official text.

[13] Yixuan Du, Chenxiao Yu, Haoyan Xu, Ziyi Wang, Yue Zhao, and Xiyang Hu. Multimodal generative engine optimization: Rank manipulation for Vision-Language model rankers, 2026. URL https://arxiv.org/abs/2601.12263.

[14] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993. ISBN 978-0-412-04231-7. doi: 10.1007/978-1-4899-4541-9.

[15] Xiyang Hu. Dynamics of adversarial attacks on large language Model-Based search engines, 2025. URL https://arxiv.org/abs/2501.00745.

[16] Junyao Huang, Ruimin Situ, and Renqin Ye. Cultural encoding in large language models: The existence gap in AI-Mediated brand discovery, 2025. URL https://arxiv.org/abs/2601.00869.

[17] International Organization for Standardization. ISO/IEC 23894:2023: Information technology—artificial intelligence—guidance on risk management, 2023. URL https://www.iso.org/standard/77304.html. Edition 1; International Standard.

[18] International Organization for Standardization. ISO/IEC 42001:2023: Information technology—artificial intelligence—management system, 2023. URL https://www.iso.org/standard/42001. Edition 1; International Standard.

[19] International Organization for Standardization. ISO/IEC 42005:2025: Information technology—artificial intelligence (AI)—AI system impact assessment, 2025. URL https://www.iso.org/standard/42005. Edition 1; International Standard; published May 2025.

[20] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

[21] Haibo Jin, Ruoxi Chen, Peiyan Zhang, Yifeng Luo, Huimin Zeng, Man Luo, and Haohan Wang. Controlling output rankings in generative engines for LLM-based search, 2026. URL https://arxiv.org/abs/2602.03608.

[22] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

[23] Sunghwan Kim, Wooseok Jeong, Serin Kim, Sangam Lee, and Dongha Lee. SAGEO arena: A realistic environment for evaluating Search-Augmented generative engine optimization, 2026. URL https://arxiv.org/abs/2602.12187.

[24] Arlen Kumar and Leanid Palkhouski. AI answer engine citation behavior an empirical analysis of the GEO16 framework, 2025. URL https://arxiv.org/abs/2509.10762.

[25] Pratyush Kumar. Generative engine optimization at scale: Measuring brand visibility across AI search engines, 2026. URL https://arxiv.org/abs/2606.20065.

[26] Zikang Liu and Peilan Xu. Think before writing: Feature-Level Multi-Objective optimization for generative citation visibility, 2026. URL https://arxiv.org/abs/2604.19113.

[27] Florian Lüttgenau, Imar Colic, and Gervasio Ramirez. Beyond SEO: A Transformer-Based approach for reinventing web content optimisation, 2025. URL https://arxiv.org/abs/2507.03169.

[28] Ministry of Industry and Information Technology of the People's Republic of China. SJ/T 12116–2026: Artificial intelligence—key technology—general technical requirements for retrieval-augmented generation, 2026. URL https://std.samr.gov.cn/hb/search/stdHBDetailed?id=594CFDC622F68BADE06397BE0A0A2CD3. Recommended electronic-industry standard; published and effective 28 April 2026; public metadata only.

[29] Ministry of Industry and Information Technology of the People's Republic of China. SJ/T 12117–2026: Artificial intelligence—key technology—evaluation specification for retrieval-augmented generation systems, 2026. URL https://std.samr.gov.cn/hb/search/stdHBDetailed?id=594CFDC623018BADE06397BE0A0A2CD3. Recommended electronic-industry standard; published and effective 28 April 2026; public metadata only.

[30] Riku Mochizuki, Shusuke Komatsu, Souta Noguchi, and Kazuto Ataka. Exposing citation vulnerabilities in generative engines, 2025. URL https://arxiv.org/abs/2510.06823.

[31] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023. URL https://doi.org/10.6028/NIST.AI.100-1.

[32] National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology, 2024. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

[33] National Institute of Standards and Technology. Guidance and templates for public-facing ai documentation: Ai standards "zero draft" initial public draft, 2026. URL https://www.nist.gov/publications/guidance-and-templates-public-facing-ai-documentation-ai-standards-zero-draft-initial. Initial public draft, 30 July 2026; preliminary proposal rather than a consensus standard.

[34] Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. Adversarial search engine optimization for large language models, 2024. URL https://arxiv.org/abs/2406.18382.

[35] Sara Patel, Mingxun Zhou, and Giulia Fanti. MaxShapley: Towards incentive-compatible generative search with fair context attribution, 2025. URL https://arxiv.org/abs/2512.05958.

[36] Perplexity AI. Agent api presets, 2026. URL https://docs.perplexity.ai/docs/agent-api/presets. First-party platform documentation; accessed 24 August 2026.

[37] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

[38] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

[39] Uri Samet. From noise to narrative: Building reputation in AI dominated search and discovery environments. *Journal of Emerging Technologies and Innovative Research*, 12(8):e35–e49, August 2025. doi: 10.56975/jetir.v12i8.568287. URL https://www.jetir.org/papers/JETIR2508406.pdf.

[40] Julius Schulte, Malte Bleeker, and Philipp Kaufmann. Don't measure once: Measuring visibility in AI search (GEO), 2026. URL https://arxiv.org/abs/2604.07585.

[41] Standardization Administration of China. GB/T 43782–2024: Artificial intelligence—technical requirements for machine learning systems, 2024. URL https://std.samr.gov.cn/gb/search/gbDetailed?id=14156507D1D40337E06397BE0A0AE656. Chinese recommended national standard; current status verified on the National Public Service Platform for Standards.

[42] Standardization Administration of China. GB 45438–2025: Cybersecurity technology—labeling method for content generated by artificial intelligence, 2025. URL https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F32EA2A561F1886CD8D606513512D547. Mandatory Chinese national standard; published 28 February 2025; effective 1 September 2025.

[43] Standardization Administration of China. GB/T 45288.1–2025: Artificial intelligence—large-scale model—part 1: General requirements, 2025. URL https://std.samr.gov.cn/gb/search/gbDetailed?id=2FF37940EB12D753E06397BE0A0A413F. Chinese recommended national standard; current.

[44] Standardization Administration of China. Project 20252041-z-469: Artificial intelligence—evaluation indicators and methods for retrieval-augmented generation systems, 2025. URL https://std.samr.gov.cn/gb/search/gbDetailed?id=37FC03D2E1446322E06397BE0A0AA17F. Registered national guiding-technical-document project; status under approval at the 24 August 2026 research cutoff; not a published standard.

[45] Standardization Administration of China. GB/T 47507–2026: Artificial intelligence—trustworthiness—general principles, 2026. URL https://std.samr.gov.cn/gb/search/gbDetailed?id=511EBC5967DA9318E06397BE0A0AFBD5. Chinese recommended national standard; published 30 April 2026; effective date 1 August 2026.

[46] Xiangkun Sun, Lingkai Kong, Aoqi Zhang, Liang Zeng, and Tonghan Wang. How LLMs are persuaded: A few attention heads, rerouted, 2026. URL https://arxiv.org/abs/2605.09314.

[47] Zhihua Tian, Yuhan Chen, Yao Tang, Jian Liu, and Ruoxi Jia. Diagnosing and repairing citation failures in generative engine optimization, 2026. URL https://arxiv.org/abs/2603.09296.

[48] Harold Triedman and Vitaly Shmatikov. MillStone: How Open-Minded are LLMs?, 2025. URL https://arxiv.org/abs/2509.11967.

[49] Rahul Vishwakarma, Shushant Kumar, and Ratnesh Jamidar. What gets cited: Competitive GEO in AI answer engines. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4950–4954, 2026. doi: 10.1145/3805712.3808445.

[50] Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing?, 2024. URL https://arxiv.org/abs/2402.11782.

[51] Zhaoqi Wang, Zijian Zhang, Xiaomei Yuan, Pengtao Kou, Jiamou Liu, Zhen Li, and Liehuang Zhu. GPE: Evaluating robust evidence aggregation for fact verification under controllable GEO-Style poisoning, 2026. URL https://arxiv.org/abs/2607.20730.

[52] Qianfeng Wen, Yifan Simon Liu, Xin Liu, Difan Jiao, Blair Yang, Junda Wu, and Zhenwei Tang. SafeGEO: Understanding generative engine optimization risks in recommendation agents, 2026. URL https://arxiv.org/abs/2606.28356.

[53] Yizhu Wen, Nan Zhang, Haohan Yuan, Xun Chen, Haopeng Zhang, and Hanqing Guo. Position: On the risks of generative engine optimization in the era of LLMs, 2025.

[54] World Wide Web Consortium. PROV-O: The PROV ontology, 2013. URL https://www.w3.org/TR/prov-o/. W3C Recommendation, 30 April 2013.

[55] World Wide Web Consortium. Web content accessibility guidelines (WCAG) 2.2, 2024. URL https://www.w3.org/TR/WCAG22/. W3C Recommendation, 12 December 2024.

[56] Beining Wu, Fuyou Mao, Jiong Lin, Cheng Yang, Jiaxuan Lu, Yifu Guo, Siyu Zhang, Yifan Wu, Ying Huang, and Fu Li. From experience to skill: Multi-Agent generative engine optimization via reusable strategy learning, 2026. URL https://arxiv.org/abs/2604.19516.

[57] Yujiang Wu, Shanshan Zhong, Yubin Kim, and Chenyan Xiong. What generative search engines like and how to optimize web content cooperatively. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=K8EinVWtUB.

[58] Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. When search engine services meet large language models: Visions and challenges, 2024. URL https://arxiv.org/abs/2407.00128.

[59] Chen Xu, Zitian Guo, and Chenyan Xiong. Mechanism design for generative engines: From exploitation toward Win-Win outcomes, 2026. URL https://arxiv.org/abs/2608.11390.

[60] Hengwei Ye, Jiasheng Mao, Zhenhan Guan, and Zheng Tian. EcoGEO: Trajectory-Aware evidence ecosystems for Web-Enabled LLM search agents, 2026. URL https://arxiv.org/abs/2605.12887.

[61] Junwei Yu, Mufeng Yang, Yepeng Ding, and Hiroyuki Sato. Structural feature engineering for generative engine optimization: How content structure shapes citation behavior, 2026. URL https://arxiv.org/abs/2603.29979.

[62] Xucheng Yu, Haibo Jin, Huimin Zeng, and Haohan Wang. SCI-Defense: Defending manipulation attacks from generative engine optimization, 2026. URL https://arxiv.org/abs/2605.21948.

[63] Jiaqi Yuan, Jialu Wang, Zihan Wang, Qingyun Sun, Ruijie Wang, and Jianxin Li. AgenticGEO: A Self-Evolving agentic system for generative engine optimization, 2026. URL https://arxiv.org/abs/2603.20213.

[64] Faye Zhang, Qianyu Cheng, Jasmine Wan, Vishwakarma Singh, Jinfeng Rao, and Kofi Boakye. Generative engine optimization: A VLM and agent framework for pinterest acquisition growth, 2026. URL https://arxiv.org/abs/2602.02961.

[65] Kai Zhang, Xinyue He, and Jingang Yao. From citation selection to citation absorption: A measurement framework for generative engine optimization across AI search platforms, 2026. URL https://arxiv.org/abs/2604.25707.

[66] XinYu Zhao, ChengYou Li, XiangBao Meng, Kai Zhang, and XiaoDong Liu. Beyond retrieval: Modeling confidence decay and deterministic agentic platforms in generative engine optimization, 2026. URL https://arxiv.org/abs/2604.03656.

[67] Heyang Zhou, JiaJia Chen, Xiaolu Chen, Jie Bao, Zhen Chen, and Yong Liao. IF-GEO: Conflict-Aware instruction fusion for Multi-Query generative engine optimization, 2026. URL https://arxiv.org/abs/2601.13938.

A 带版本记录的研究地图

核心思想

本附录在带版本记录的论文库、讲义和课程之间建立阅读路径。收录意味着记录进入了评审队列，并不意味着论文的每项主张都已被接受、复现或吸收到正文。

A.1 快照与溯源

GEO-42 论文库来自用户提供的 Zotero CSV，其 SHA-256 为

`c149a69140b0dcedc1855f8cbb02fc5cb1e1346470306c76cd68fbafc2e50468`。

原始 CSV 以只读方式保留。规范化脚本生成 JSON 目录、本表及 BibTeX 文件。快照包含 42 个唯一规范化题名、42 个 DOI 字段、41 个 URL 与摘要字段，以及 42 份本地 PDF 附件。

A.2 版本与状态控制

图、表或精确数值进入正文之前，必须将本地 PDF 版本与当前官方论文记录比对。首次审查发现 8 份本地 PDF 落后于当时的 arXiv 版本，分别为记录 2、11、12、13、17、23、25 和 35。记录 23 尤其重要：本地 SAGEO Arena v1 仍含模板占位内容，而当时的 arXiv 记录已是 v2，并更新了接收状态。表中标记了全部 8 条记录；旧版可用于分类参考，但旧版图形或结果不进入讲义。

发表状态独立记录。“有 DOI”“在 arXiv 上”“被研讨会接收”与“发表于主会议或期刊”不能合并。状态进入面向公众的课程元数据之前，必须根据官方发表场所或当前论文记录核实。

A.3 初始编辑路径

表 45: GEO-42 研究库及其初始编辑路径。路径由编辑策划决定，并非论文结论。

#	论文	路径与编辑状态
1	**Cultural Encoding in Large Language Models: The Existence Gap in AI-Mediated Brand Discovery** (2025) 16	**平台实证** 第 1 章 来源年份 2025 与 arXiv 标识年份 2026 不一致
2	**Multimodal Generative Engine Optimization: Rank Manipulation for Vision-Language Model Rankers** (2026) 13	**安全** 第 10 章 arXiv v2 已于 2026-08-24 锁定
3	**IF-GEO: Conflict-Aware Instruction Fusion for Multi-Query Generative Engine Optimization** (2026) 67	**干预** 第 9 章 元数据已收录
4	**Controlling Output Rankings in Generative Engines for LLM-based Search** (2026) 21	**干预** 第 9 章 元数据已收录

续下页

表 45（续前页）

#	论文	路径与编辑状态
5	**Generative Engine Optimization: A VLM and Agent Framework for Pinterest Acquisition Growth** (2026) 64	**多模态** 第 10 章 元数据已收录
6	**When Search Engine Services meet Large Language Models: Visions and Challenges** (2024) 58	**系统基础** 第 1 章 元数据已收录
7	**Adversarial Search Engine Optimization for Large Language Models** (2024) 34	**安全** 第 11 章 元数据已收录
8	**AI Answer Engine Citation Behavior An Empirical Analysis of the GEO16 Framework** (2025) 24	**干预** 第 6 章 元数据已收录
9	**Beyond SEO: A Transformer-Based Approach for Reinventing Web Content Optimisation** (2025) 27	**干预** 第 9 章 元数据已收录
10	**CC-GSEO-Bench: A Content-Centric Benchmark for Measuring Source Influence in Generative Search Engines** (2025) 6	**基准** 第 4 章 元数据已收录
11	**Dynamics of Adversarial Attacks on Large Language Model-Based Search Engines** (2025) 15	**安全** 第 11 章 arXiv v3 已于 2026-08-24 锁定
12	**E-GEO: A Testbed for Generative Engine Optimization in E-Commerce** (2025) 3	**干预** 第 6 章 arXiv v2 已于 2026-08-24 锁定
13	**Exposing Citation Vulnerabilities in Generative Engines** (2025) 30	**安全** 第 5 章 arXiv v2 已于 2026-08-24 锁定
14	**From Noise to Narrative: Building Reputation in AI Dominated Search and Discovery Environments** (2025) 39	**背景** 第 2 章 缺少 URL；缺少摘要
15	**Generative Engine Optimization: How to Dominate AI Search** (2025) 5	**系统** 第 2 章 元数据已收录

续下页

表 45（续前页）

#	论文	路径与编辑状态
16	**GEO: Generative Engine Optimization** (2024) 1	**基础** 第 1 章 来源年份 2024 与 arXiv 标识年份 2023 不一致
17	**MaxShapley: Towards Incentive-compatible Generative Search with Fair Context Attribution** (2025) 35	**系统** 第 4 章 arXiv v2 已于 2026-08-24 锁定
18	**MillStone: How Open-Minded Are LLMs?** (2025) 48	**系统** 第 4 章 元数据已收录
19	**Position: On the Risks of Generative Engine Optimization in the Era of LLMs** (2025) 53	**风险与研究议程** 第 11 章 元数据已收录
20	**What Evidence Do Language Models Find Convincing?** (2024) 50	**证据行为** 第 2 章 元数据已收录
21	**What Generative Search Engines Like and How to Optimize Web Content Cooperatively** (2025) 57	**干预** 第 8 章 元数据已收录
22	**When Attention Becomes Exposure in Generative Search** (2026) 2	**系统** 第 4 章 元数据已收录
23	**SAGEO Arena: A Realistic Environment for Evaluating Search-Augmented Generative Engine Optimization** (2026) 23	**端到端基准** 第 4 章 arXiv v2 已于 2026-08-24 锁定
24	**Diagnosing and Repairing Citation Failures in Generative Engine Optimization** (2026) 47	**诊断与修复** 第 8 章 元数据已收录
25	**Caption Injection for Optimization in Generative Search Engine** (2026) 7	**系统** 第 5 章 arXiv v4 已于 2026-08-24 锁定; 来源年份 2026 与 arXiv 标识年份 2025 不一致

续下页

表 *45*（续前页）

#	论文	路径与编辑状态
26	**Role-Augmented Intent-Driven Generative Search Engine Optimization** (2026) 8	**系统** 第 3 章 来源年份 2026 与 arXiv 标识年份 2025 不一致
27	**Beyond Retrieval: Modeling Confidence Decay and Deterministic Agentic Platforms in Generative Engine Optimization** (2026) 66	**干预** 第 3 章 元数据已收录
28	**Structural Feature Engineering for Generative Engine Optimization: How Content Structure Shapes Citation Behavior** (2026) 61	**干预** 第 8 章 元数据已收录
29	**Don't Measure Once: Measuring Visibility in AI Search (GEO)** (2026) 40	**测量** 第 6 章 元数据已收录
30	**AgenticGEO: A Self-Evolving Agentic System for Generative Engine Optimization** (2026) 63	**系统** 第 3 章 元数据已收录
31	**Think Before Writing: Feature-Level Multi-Objective Optimization for Generative Citation Visibility** (2026) 26	**干预** 第 9 章 元数据已收录
32	**From Experience to Skill: Multi-Agent Generative Engine Optimization via Reusable Strategy Learning** (2026) 56	**测量** 第 7 章 元数据已收录
33	**From Citation Selection to Citation Absorption: A Measurement Framework for Generative Engine Optimization Across AI Search Platforms** (2026) 65	**测量** 第 5 章 元数据已收录
34	**How LLMs Are Persuaded: A Few Attention Heads, Rerouted** (2026) 46	**安全** 第 11 章 arXiv v1 已于 2026-08-26 锁定
35	**EcoGEO: Trajectory-Aware Evidence Ecosystems for Web-Enabled LLM Search Agents** (2026) 60	**干预** 第 8 章 arXiv v2 已于 2026-08-24 锁定

续下页

表 45（续前页）

#	论文	路径与编辑状态
36	**What Gets Cited: Competitive GEO in AI Answer Engines** (2026) 49	**干预** 第 7 章 元数据已收录
37	**SCI-Defense: Defending Manipulation Attacks from Generative Engine Optimization** (2026) 62	**安全** 第 11 章 元数据已收录
38	**Generative Engine Optimization at Scale: Measuring Brand Visibility Across AI Search Engines** (2026) 25	**系统** 第 1 章 元数据已收录
39	**SafeGEO: Understanding Generative Engine Optimization Risks in Recommendation Agents** (2026) 52	**干预** 第 11 章 元数据已收录
40	**GPE: Evaluating Robust Evidence Aggregation for Fact Verification under Controllable GEO-Style Poisoning** (2026) 51	**安全** 第 11 章 元数据已收录
41	**GEO-Flag: Detecting and Measuring GEO-Optimized Web Content** (2026) 9	**干预** 第 11 章 元数据已收录
42	**Mechanism Design for Generative Engines: From Exploitation toward Win-Win Outcomes** (2026) 59	**机制设计** 第 11 章 元数据已收录

A.4 内容吸收状态

每条记录最终获得以下四种状态之一：

核心
经核实的贡献对主论证不可或缺，讲义和课程均加以讨论。

支持
提供某章或实作采用的方法、比较、反例或扩展。

前沿
较新、有潜力或尚有悬而未决问题，放入带版本的研究前沿说明，而不作为稳定定论。

仅收录目录
保持可发现性，同时明确未被实质吸收的原因，例如范围不符、方法细节缺失、成果过期或证据不足。

公开阅读图谱将同时展示路径与吸收状态，由此区分“我们收到了这篇论文”与“这一思想已通过编辑审查并进入课程”。

B 可复现性与发布协议

B.1 最小实验清单

讲义中每张实证图表均配有清单，包含：

1. 研究问题、目标估计量和预注册决策规则；
2. 仓库提交、环境锁定、随机种子和执行日期；
3. 数据集身份、许可、来源哈希、排除项和数据划分；
4. 查询面板、意图聚类、权重，以及开发/留出边界；
5. 系统、模型、端点、界面、账户、地区和版本状态；
6. 干预与比较对象的哈希；
7. 原始评估单元、缺失情况、标注规则与裁定；
8. 分析脚本、不确定性方法、稳健性检查和失败；
9. 制图脚本和精确输出哈希；
10. 证据、安全、隐私和发布批准。

B.2 图片检查表

- ☐ 图注写明单位、总体、系统、日期、处理、比较对象、指标和不确定性。
- ☐ 每种视觉编码均有解释，不仅依靠颜色。
- ☐ 坐标轴范围、缺失值和排除项可见。
- ☐ 脚本或源文件可从冻结输入重新生成图片。
- ☐ 外部内容在来源台账中记录来源、版本、图号/页码、许可、裁切、修改状态和替代文本。
- ☐ 图片在灰度及最终印刷宽度下仍清晰可读。
- ☐ PDF 构建无标签裁切、字形损坏或未解析引用。

图 20 是计算重放示例。重新生成冻结算术与图形，只建立内部可追溯性；除非独立治理的另一项研究独立重新采集或复现相关观察，否则不能称为独立实证复现。

B.3 封闭平台观察卡

对于封闭产品，每次观察至少报告：

字段	必填内容
界面	产品与具体模式、页面、API 或功能
时间	时间戳与时区
背景	地区、语言、地理位置/端点、账户状态、设备
输入	精确提示、对话历史、上传内容和设置
重复	运行次数、间隔、重试/错误处理政策
输出	回答、引证、界面状态与允许采集的轨迹
解释	观察、代理、推断和替代解释
保留	存储、隐私、平台条款与脱敏决定

B.4 稿件发布条件

arXiv 候选稿只有在下列条件满足后才能发布：

- 所有引用键和交叉引用均能解析；
- 参考文献身份和影响重要结论的论文版本已核实；
- 所有图片通过来源、许可、替代文本和印刷检查；
- 每项定量主张均可追溯到精确的表、图、页码、数据集或重新生成的分析；
- 有关封闭系统的主张包含带版本的观察卡；
- 源码包中不残留 TODO、模板占位符、隐藏作者笔记或过期输出；
- PDF 通过全页渲染、字体、元数据、链接、溢出和面向无障碍的审查；
- 源码包可在干净的 TeX Live 环境构建，不依赖专有字体或未声明本地文件。

这些检查是科学有效性的必要条件，而非充分条件。最终发布仍需要对论证、方法和风险主张进行领域评审。

C 标准与官方指南登记表

C.1 目的与收录规则

本登记表识别可能约束 GEO 研究或部署流程的标准与官方指南，而非合规清单。只有发布机构、标识符、版本或发布日期、状态和官方身份页面均已核实，条目才予收录。登记表记录文档范围，不转载付费条款，也不推断某组织、系统或干预符合标准。

登记表有意区分已发布国际标准与国家标准、行业标准、公开技术推荐、政府指南、法律措施、已立项项目和征求意见稿。这些类别具有不同规范效力。W3C 推荐标准不是 ISO 管理体系要求；自愿采用的 NIST 框架不是法律；中国推荐性标准不会自动成为强制规则；已立项项目或“零稿”也不是已发布的共识标准。

表 46: 研究截止时以状态为先的登记表。“用途”指课程或稿件控制措施，不表示符合标准。

标识	已核实状态	相关范围	在本项目中的限定用途
NIST AI 100-1, AI RMF 1.0	2023 年正式发布；自愿框架	全生命周期的 AI 风险治理、映射、测量与管理	风险登记、角色分配、测量计划与发布评审；不是证书或法律安全港 [31]
NIST AI 600-1	2024 年正式配置文件；配套 AI RMF 1.0	生成式 AI 新增或加剧的风险及建议行动	实作 7–8 的威胁建模提示与控制覆盖；不证明具体产品落实了这些行动 [32]
ISO/IEC 42001:2023	已发布国际标准，第 1 版	AI 管理体系与持续改进要求	治理流程对照和控制证据清单；未经合法获得全文及合格审计，不主张条款级符合性 [18]
ISO/IEC 23894:2023	已发布国际标准，第 1 版	将 AI 特定风险管理融入组织活动的指南	风险分类与处置记录对照；不替代行业或司法辖区分析 [17]
W3C PROV-O	W3C 推荐标准，2013 年 4 月 30 日	以可互操作方式表示与交换溯源信息	主张—证据—信源谱系的词汇参考；不是完整证据质量或真实性模型 [54]
WCAG 2.2	W3C 推荐标准；现行推荐日期为 2024 年 12 月 12 日	可测试的网页内容无障碍成功准则	课程网站与公开成果的验收准则；符合性仍需自动与人工评估 [55]
GB/T 43782–2024	中国推荐性国家标准；现行	机器学习系统技术要求	受控流程系统卡提示；不是 GEO 内容或商业答案引擎排序的直接标准 [41]
GB/T 45288.1–2025	中国推荐性国家标准；现行	人工智能大模型通用要求	智能体与生成式界面的模型、系统评估背景；不是发布者优化规则 [43]
GB/T 47507–2026	已发布中国推荐性国家标准；官方实施日期为 2026 年 8 月 1 日	可信人工智能通用原则	可信性术语与治理对照。发布前重新核实实时状态标签，官方搜索缓存可能落后于实施日期 [45]

表 47: RAG 专项官方元数据与状态跟踪。不从题名或公开身份页面推断条款级要求。

标识	截止时已核实状态	限定用途	禁止推断
SJ/T 12116–2026	现行中国推荐性电子行业标准；工信部；2026 年 4 月 28 日发布并实施；备案 107753–2026	仅依据元数据的 RAG 技术要求状态卡 [28]	合法阅读全文前，不推断条款、指标、阈值、符合性、认证或产品范围
SJ/T 12117–2026	现行中国推荐性电子行业标准；工信部；2026 年 4 月 28 日发布并实施；备案 107764–2026	仅依据元数据的 RAG 评估规范状态卡 [29]	题名不能揭示测试语料、指标、阈值、适用产品，也不能证明 GEO 有效
项目 20252041-Z-469	中国国家指导性技术文件立项项目；2025 年 6 月 20 日立项；截止时官方状态为“正在批准”	标准生命周期反例 [44]	不是已发布标准；不得声称最终标识符、条款、实施日期或符合性
NIST 面向公众的 AI 文档“零稿”	初始公开草案，2026 年 7 月 30 日；截止时为征求反馈中的 NIST 初步提案	候选文档模板，以及从草案走向标准化的生命周期示例 [33]	不是共识、法律或认证，也不证明公开文档会使系统可信或合规

表 48: 其他已核实状态的治理与披露材料。法律适用性须另行审查。

标识	已核实状态	相关范围	在本项目中的限定用途
ISO/IEC 42005:2025	已发布国际标准，第 1 版；2025 年 5 月	组织开展 AI 系统影响评估的指南，覆盖对个人、群体与社会的可预见影响	影响评估档案与利益相关者评审提示；公开元数据不提供条款级符合性证据 [19]
GB 45438–2025	现行中国强制性国家标准；2025 年 2 月 28 日发布；2025 年 9 月 1 日实施	AI 生成合成内容显式与隐式标识的技术方法	适用范围内的披露与元数据测试案例；不支持排序、引证、真实性或作者身份主张 [42]
人工智能生成合成内容标识办法	联合发布的官方办法；2025 年 3 月 7 日印发；2025 年 9 月 1 日施行	特定生成与传播服务提供者的义务，包括显式与隐式标识	角色—成果适用性卡；不是对所有作者、研究原型、地区或发布行为的普遍义务 [12]
生成式人工智能服务管理暂行办法	官方第 15 号令；2023 年 8 月 15 日施行	在中国大陆向公众提供生成文本、图像、音频或视频的特定服务，含明示排除情形	司法辖区与服务提供者角色检查；提出运营建议前须法律审查 [11]

C.2 适用性卡

标准进入主张、实作评分规则或发布决定之前，分析人员应填写下卡：

字段	必须记录
身份	发布机构、标识符、完整题名、版本、语言与官方 URL
状态	草案、推荐、已发布标准、废止、被替代或法律
时间范围	发布、实施、复核、废止和访问日期
制度效力	自愿指南、合同要求、认证依据、推荐标准、强制规则或其他声明类别
适用性	组织、系统、角色、生命周期阶段、地域、行业及排除情形
具体控制	使用的获许可条款或公开章节、转述，以及成果层面的验收测试
证据	证明控制措施的文档、日志、决定或测量
审核人	领域责任人，以及必要时合格的法律、安全、隐私、无障碍或符合性审核人
剩余边界	映射控制不能证明什么

C.3 课程对照

成果	标准路径	验收问题
主张—证据—信源图	W3C PROV-O 与项目证据评分规则	能否重建每次推导、修订和责任主体，而不把溯源当成真实性证明?
封闭系统观察卡	NIST AI RMF 与 NIST AI 600-1	背景、受影响方、失效模式、测量和处置决定是否明确?
干预发布记录	ISO/IEC 42001 与 ISO/IEC 23894	责任归属、风险处置、监测、纠正行动和留存证据，是否融入可重复流程?
模型/系统卡	GB/T 43782–2024 与 GB/T 45288.1–2025	成果是否写明技术对象、边界、评估背景与局限，而非把通用要求转成排序主张?
公开课程成果	WCAG 2.2	键盘和读屏用户能否取得相同信息、状态与控制，且内容不依赖颜色也能理解?
可信性评审	GB/T 47507–2026、NIST AI RMF 与 ISO/IEC 23894	可信性术语是否映射到可测试控制和剩余风险，而非充当宣传形容词?
AI 系统影响档案	ISO/IEC 42005:2025 与项目利益相关者图	是否在全生命周期记录可预见的个人、群体与社会影响，而不只凭公开元数据声称符合性?
生成内容披露	GB 45438–2025 与 2025 年标识办法；适用时单独考虑 C2PA	是否记录角色、辖区、内容类型、显式标识、嵌入元数据、传播状态、测试证据和例外，而不把溯源当真实性?
RAG 标准状态卡	SJ/T 12116–2026、SJ/T 12117–2026 与项目 20252041-Z-469	是否将现行行业标准、评估规范和正在批准的国家项目分列状态，并在合法阅读全文之前禁止条款级使用?

适用边界与失效条件

登记表是带版本的证据，而非法律意见。研究截止之后，标准可能修订、被替代、纳入合同或纳入法律。公开发布时，必须重新核查每个官方身份页面，以及实际司法辖区、行业、组织和系统所适用的规则。

D 习题与复现任务

核心思想

这些习题将本讲义转化为可检查的研究训练。每章包含一道概念或推导题，以及一道实证、实验或审查题。合格作业须说明目标估计量，保留分母与版本，报告无效结果和失败，并将结论限定在设计所支持的范围内。完成习题，并不能证明商业平台采用了题中机制。

D.1 作业要求与提交约定

本讲义原创的 24 道题，不依赖特定供应商、付费接口或未公开数据集。“C”类侧重概念或推导；“R”类要求完成小型复现、实验设计或审查。每份答案均须包含一句话结果、不确定性或歧义说明，以及本项工作不能确立的最强主张。

实证题应提交精简研究包，包含 `manifest.yaml`、不可变输入的哈希、整洁的单元格级结果文件、分析脚本或计算表、一份决策备忘录，以及 `limitations.md`。完整算力路径可使用开放检索器或生成器；低算力路径始终有效，可用笔记本电脑或纸笔完成。合成示例必须明确标注；缺失单元格应保留在分母台账中，不得在清洗时消失。

表 49: 24 道习题的覆盖关系。“解答要点”列标明本附录后文提供参考思路的题目。

章节	概念或推导	复现、实验或审查	解答要点
1	条件可见性链	可见性画像的分母审查	1C
2	证据充分性与主张上限	主张—证据—信源审查	2C
3	查询扩展停止规则	最早失败环节定位	3C
4	排名融合与上下文分配	检索环节损失审查	4C
5	引证与吸收的目标估计量	反事实信源影响审查	5C
6	重复测量的不确定性	基准方案审查	6C
7	簇效应与时间变化	考虑漂移的预注册审查	7C
8	约束下的内容决策	保持主张不变的修订审查	8C
9	多查询约束优化	优化器轨迹审查	9C
10	跨模态溯源依赖	配对平台界面设计	10C
11	防御效用与误报	受控红队审查	11C
12	控制证据与符合性	事件证据保全演练	12C

D.2 第 1 章：研究对象与条件可见性

习题 D.1: 1C – 诊断条件可见性链

对于一个查询与信源的分层，假设式 (2)中的条件阶段概率估计为

$$(p_D, p_R, p_C, p_G, p_T, p_A) = (0.90, 0.60, 0.75, 0.80, 0.50, 0.20).$$

计算全部阶段均成功的联合概率，再推导联合概率相对于每个正的阶段概率的弹性。比较两项拟议干预：a 将 p_R 从 0.60 改为 0.69；b 将 p_T 从 0.50 改为 0.60。仅在本次计算中固定其他条件概率。说明为什么这种比较是敏感性演算，而非已经识别的预测。最后构造一个两层示例，使总体变化更大的干预反而损害基线较低的层。

所需输入。六个条件概率、两项假设变化，以及两层反例中明确的加权规则。
所需输出。基线联合概率、两项反事实算术变化、弹性、分组表及四句话解释。
禁止外推。不得仅凭乘法推断独立性、平台漏斗、因果干预效应或下游商业效应。
低算力路径。使用计算器或列出精确乘积，自行选择两个合成层，使加权平均展示上述反转。

习题 D.2: 1R – 构建分母明确的可见性画像

创建或使用包含 12 个查询、每个重复两次的合成单元格表，列为 `query_id`、`intent`、`eligible`、`retrieved`、`cited`、`absorbed`、`attribution_correct`、`claim_faithful`、`timestamp` 和 `missing_reason`。查看结果前，为 **V** 的每个分量定义分母。分别按意图及总体估计可观察分量，附适合小样本的区间，并报告两次重复的时间差异。展示完整案例筛选如何改变至少一个估计值，解释为什么将画像各分量简单取算术平均并不是中性的汇总。

所需输入。 24 行合成或经授权的数据表、数据字典、预注册缺失状态处理规则和查询权重。
所需输出。分母台账、分层画像、区间表、完整案例敏感性分析，以及一段边界清楚的结果说明。
禁止外推。不得把引证称为检索、把提及称为忠实性，或把该面板表述为整个品牌、语言、地区或平台总体的可见性。
低算力路径。使用六个查询、每个重复两次；用表格计算 Wilson 区间，或报告原始比例及其确切分母。

D.3 第 2 章：主张、证据与信源身份

习题 D.3: 2C – 分析证据瓶颈判定规则

设式 (9)中 $A(c)$ 的五个分量均在 $[0,1]$ 上评分。主张 c_1 的分量向量为 $(0.9, 0.8, 0.4, 0.9, 0.7)$，主张 c_2 为 $(0.65, 0.65, 0.65, 0.65, 0.65)$。分别按照最小值规则、算术平均和“所有分量至少为 0.60”的发布规则排序。证明最小值规则对每个分量单调，但在瓶颈改变之前，改善非最小分量不会获得额外奖励。举例说明新增信源如何提高覆盖度，却降低一致性，并区分其中的数学结论与编辑评价规则。

所需输入。两个分数向量、三种决策规则的定义，以及证据冲突时更新一致性的一条明确规则。
所需输出。所有排序、简短单调性证明、冲突证据示例，以及注明尚未解决字段的发布建议。
禁止外推。不得把这些分数解释为经过校准的真实性概率、信源声望或科学有效性。
低算力路径。手算即可；冲突示例仅需一个主张和两条证据关系。

习题 D.4: 2R – 审查主张—证据—信源资料包

从官方公开页面选取一条非敏感、具有时间边界的事实主张，再从研究、从业者或供应商资料中选取一条相关陈述，固定两个来源对象。将目标拆为带范围的主张元组，记录各信源身份，仅摘录核验所需证据片段，并将每条关系标为支持、限定、矛盾或无信息。指定证据层级上限，列出所有未知字段。随后改变日期、地区、产品版本或比较对象等一个范围元素，再次审查。

所需输入。两个合法可访问的来源快照、哈希或存档身份、访问日期、目标主张，以及明确的证据关系评价规则。
所需输出。两份身份记录、两个带范围主张元组、带类型的 CES 图、证据支持上限、未知字段表，以及范围变化前后的差异。
禁止外推。不得将官方产品描述视为效果证据、将论文摘要视为完整方法核验，或将溯源信息视为真实性证明。
低算力路径。使用一个主张、两个信源和手绘三节点图；无需自动化，明确保留未解决字段。

D.4 第 3 章：获取、规划与处理链路轨迹

习题 D.5: 3C – 应用受约束的查询扩展停止规则

查询规划器按固定顺序提出五个子查询，其新增独有证据收益为 $(8, 5, 2, 1, 3)$ 单位，边际成本为 $(1, 1, 1, 1, 3)$ 单位。设 $\epsilon_E = 2$、$\tau = 1.5$。先严格应用第3章的规则，再加入两个覆盖约束：至少找回一项反证，首次出现于第 4 个子查询；至少包含两个溯源起点，首次达成于第 3 个子查询。明确阈值是在执行拟议子查询之前还是之后检验。说明第 5 项提议为何使边际收益序列不单调，以及贪心停止规则为何可能错过它。

所需输入。有序收益与成本、两个阈值、两个覆盖约束，以及明确的执行前或执行后检验约定。
所需输出。各约定下的停止序号、累积覆盖与成本、约束可行性检查，以及一种替代的前瞻规则。
禁止外推。不得声称这些子查询由封闭产品发出，或认为更多查询扩展必然改善回答。
低算力路径。手算五行表格即可，无需检索模型或 API。

习题 D.6: 3R – 定位最早可观察的失败环节

构造五条合成端到端轨迹，包含信源准入、查询扩展、检索排名、重排序排名、打包段落、回答主张、展示引证与行动状态。分别注入一次获取失败、切分失败、上下文打包挤出、生成遗漏和错误信源归因。对每条轨迹，指出最早观察到的失败环节，并将其与潜在或重建解释区分开。再添加一条证据在所有可观察环节均得到保留的阳性对照。

所需输入。固定的小型语料库、五条查询—信源轨迹和一个阳性对照、段落哈希，以及分环节通过/失败定义。
所需输出。轨迹包、阶段转移矩阵、最早失败标签、未排除的解释，以及一棵诊断决策树。
禁止外推。不得仅凭展示的回答或引证推断隐藏候选分数、上下文顺序、因果机制或自然发现。
低算力路径。使用六张卡片，或每条轨迹一行的表格，手工指定排名；无需生成器。

D.5 第 4 章：检索、融合与上下文分配

习题 D.7: 4C – 比较排名融合与证据感知打包

两个检索器返回 $L_1 = (A, B, C, D)$ 和 $L_2 = (C, A, E, B)$。取 $\kappa = 60$，计算 RRF 分数及融合排序，完全并列时按文档 ID 决定顺序。证明：只要每个列表内部顺序不变，对底层原始分数作任意严格递增变换，RRF 就不变；原始分数相加则未必如此。再考虑段落 A, C, E，其 $(v_i, \ell_i) = (7, 5), (6, 4), (4, 2)$，证据集合分别为 $\{x\}, \{y\}, \{x, y\}$，预算 $B = 6$。设覆盖奖励 $\alpha = 3$，无冗余惩罚，枚举可行打包集合，按式 (24)选择最优解。最后说明溯源起点数量上限如何改变决策。

所需输入。两个排序列表、κ、段落效用与长度、证据集合、预算、并列规则和目标系数。
所需输出。 RRF 表、不变性论证、完整可行打包表、最优解，以及一个溯源上限反例。
禁止外推。不得把融合排名等同于回答贡献、把证据覆盖等同于真实，或把玩具问题最优解当作生产上下文策略。
低算力路径。手工枚举至多八个子集，为检查并列将计算结果保留六位小数。

习题 D.8: 4R – 审查以阶段为条件的检索损失

使用开放或合成语料库，定义 12 个查询、关键证据相关性标注（qrels）与不可变切分。固定深度，比较词法检索、稠密检索和 RRF，再应用一条透明重排序规则及两种上下文预算。在每个转移环节追踪证据召回率、nDCG、溯源多样性与关键证据覆盖。注入一个转载重复簇，以及一个包含必要限定条件的长段落。将每项损失定位于首次观察到损失的转移处。

所需输入。语料与段落哈希、12 个查询、分级相关性与关键证据标签、三种检索轨迹、一条重排序规则、两个预算，以及溯源起点 ID。
所需输出。含分母的指标表、转移桑基图或整洁损失表、重复审查、限定条件保留记录、跨查询区间和失败情况。
禁止外推。不得将重建的检索收益转化为关于商业获取、最终回答质量或用户行为的主张。
低算力路径。使用六个查询和十个段落；以给定排序列表替代稠密模型，在表格中枚举上下文选择。

D.6 第 5 章：生成、引证与影响

习题 D.9: 5C – 区分引证蕴含、完整性与吸收

一个回答含四条关键主张。主张 1–3 由信源 S 支持；主张 4 仅由 T 支持。展示引证 S 位于主张 1 和 4 之后，T 位于主张 3 之后，主张 2 没有引证。S 特有的短语在主张 1 和 2 中得到忠实保留。按照明确的片段归属规则，计算引证蕴含率、引证完整率、展示引证率及 S 特有的吸收率，报告所有分母。一个二元蕴含评判器在代表性校准集上的灵敏度为 0.90，特异度为 0.80。推导观察阳性率 $\tilde{p}$ 的标准矩阵校正，并在 $\tilde{p} = 0.65$ 时计算。说明校正估计值的可行条件。

所需输入。四条主张与信源的关系、展示引证位置、归属规则、吸收定义，以及评判器灵敏度和特异度。
所需输出。四项独立指标及分母、校正推导与数值、可行性检查和失败解释。
禁止外推。不得从构造的回答推断检索、事实真实性、唯一作者身份或整个平台的引证质量。

低算力路径。用四条主张构成完整表格，用基础计算器完成校正。

习题 D.10: 5R – 审查反事实信源影响

构建包含三个信源、两个溯源起点和六条已核验关键主张的固定上下文。按照固定评价规则，生成或手工构造基线回答，以及逐一移除信源后的回答。切分回答主张，解析引证，分别评估支持、完整性与信源特有的吸收。包含一处错误信源引证和一项遗漏的限定条件，比较逐一移除信源与逐一移除溯源起点的变化。

所需输入。固定上下文、信源与起点 ID、六条主张标签、回答构造规则或固定生成器、移除条件，以及双重标注规则。
所需输出。主张级比较表、展示引证与吸收矩阵、信源级和起点级差值、分歧、不确定性与失败案例。
禁止外推。不得将回答变化解释为真实性、所有权、固定上下文以外的必要因果关系，或未观察检索器行为的证明。
低算力路径。依据六条主张事实表手写三个简短回答，请第二位读者仅裁定两处注入的失败。

D.7 第 6 章：测量与基准设计

习题 D.11: 6C – 量化重复测量的伪重复

设 30 个查询簇，在两种条件下各重复生成四次。同一条件内，同一查询重复结果的簇内相关近似为 $\rho = 0.35$。计算名义回答数、设计效应 $1 + (r - 1)\rho$，以及仅考虑重复层面依赖时的相应有效样本量。解释为何这一修正仍未涵盖相关查询之间的聚类、系统漂移或配对处理分配。设计一个最小层级模型，分开查询、时间和残差变异，并明确主要对比的单位。

所需输入。簇数、重复次数、条件数、ρ、处理配对信息，以及回答和主要对比的定义。
所需输出。设计效应、有效样本量、简单独立分析与聚类分析的区别说明、带索引随机分量的模型式，以及明确的分析单位。
禁止外推。不得声称有效样本量就是实际信息量，或该 ρ 可以迁移到其他查询面板或系统。
低算力路径。手算设计效应，以符号写出模型；无需混合模型软件。

习题 D.12: 6R – 修复规定不足的基准方案

审查这个有意设计为有缺陷的方案：“使用 50 个热门提示词，不断调整改写直至平均 GEO 分数提高；遇错重试，删除拒答，用一个 LLM 判断质量。”至少指出 12 种不同威胁，覆盖目标总体、目标估计量、泄漏、处理版本、缺失、重复、系统状态、评判器有效性、多重比较、不确定性、保护性约束与停止规则。将方案重写为可执行清单，构造小型计划与实际单元格台账，至少包含一次错误、拒答和迟到响应。

所需输入。缺陷方案、基准清单模式、合成查询抽样框，以及明确的资源与伦理边界。
所需输出。威胁登记表、修正清单、单元格台账、缺失状态规则、评判器校准计划、主要与保护性指标，以及决策规则。
禁止外推。不得声称修复方案便验证了指标、评判器、样本或处理效应；修复只是使研究可以复查。
低算力路径。使用八个合成提示词、两种条件和两次重复；校准评判方案，无需调用外部模型。

D.8　第 7 章：实验与时间不确定性

习题 D.13: 7C – 区分簇效应与时间变化

四个等权查询簇进行配对随机分配：A、B 接受处理，C、D 保持对照。干预后簇均值为 $(0.70, 0.50, 0.45, 0.35)$，干预前为 $(0.40, 0.45, 0.40, 0.30)$。计算仅使用干预后数据的均值差，以及双重差分估计。展示每个簇的贡献，说明与随机化对应的潜在结果目标估计量。再各举一种处理实际实施模式与干扰模式，使两项算术估计失去预期解释。

所需输入。分配、簇权重、前后结果、处理版本定义，以及干预后对比和时间对比的假设。
所需输出。两项估计、簇贡献表、目标估计量陈述、处理实施与干扰反例，以及识别判断。
禁止外推。缺少平行趋势及稳定的处理/对照定义时，不得称双重差分估计具有因果含义；不得推广至这四个合成簇以外。
低算力路径。手算八个单元格，用簇间箭头画出两个反例。

习题 D.14: 7R – 预注册考虑漂移的干预

设计一项合成研究，含 12 个意图簇、基线与干预页面、同期对照、三个时间窗口和三次计划重复。预注册处理哈希、分配、主要目标估计量、最小有价值效应、缺失状态处理、系统状态可比时间窗、保护性阈值及停止规则。注入一次平台状态断点、一个受污染对照簇、一种非随机缺失的拒答模式，以及一个安慰剂日期。即使结果不能确定，也须给出方案要求的分析与决策。

所需输入。合成簇抽样框、处理版本、分配、计划单元格、状态断点规则、结果、保护性约束，以及四项已声明注入。
所需输出。预注册、计划/实际流程、簇估计及区间、缺失表、安慰剂与干扰检查、附理由的排除，以及保留/撤回/尚无定论的决策。
禁止外推。不得在看过结果后修补可比性、重新使用受污染簇，或把尚未识别的前后差异描述为处理效应。
低算力路径。使用四个簇和两个时间窗口；计算簇均值，将状态断点单元格标为不可比，无需拟合时间序列模型。

D.9　第 8 章：内容与证据工程

习题 D.15: 8C – 在忠实性与风险约束下选择

四个修订版本的指标向量 (V, F, U, R, C) 分别表示可见性、忠实性、用户效用、风险与归一化成本：

修订	V	F	U	R	C
r_0	0.50	0.98	0.70	0.04	0.10
r_1	0.62	0.96	0.72	0.05	0.18
r_2	0.68	0.88	0.76	0.04	0.15
r_3	0.61	0.97	0.69	0.09	0.12

仅接纳满足 $F \geq 0.95$、$U \geq 0.70$、$R \leq 0.06$ 且 $C \leq 0.20$ 的候选。确定可行集、可行候选之间的帕累托关系，以及相对 r_0 的最小有价值可见性收益为 0.08 时的发布决策。再说明如果不在优化前执行约束，加权和为什么可能选择不合格修订。

所需输入。指标表、阈值方向、基线、最小有价值效应，以及明确的帕累托支配约定。
所需输出。可行性表、帕累托集、发布决策，以及注明权重的加权和反例。
禁止外推。不得将合成分数视为经过校准的用户收益、证据质量证明，或五个维度之间普遍适用的交换比率。

低算力路径。手工检查四行，自选非负权重使反例清楚可见。

习题 D.16: 8R – 制作保持主张不变的修订差异

从一页合成证据说明开始，其中有六条已批准主张、两个限定条件、一张表和三个信源。诊断一个段落级检索或吸收失败。制作两个保持原有表达含义的修订版本，以及一个使证据更易被发现的修订版本。每个版本均须将主张片段对齐到已批准主张台账，比较渲染文本，并检查单位、日期、总体、比较对象、不确定性与例外是否仍与主张一同呈现。使用固定、手工提供的检索与回答轨迹评估所有版本。

所需输入。合成来源页面、已批准主张台账、信源与段落 ID、固定查询集、给定阶段轨迹，以及主张等价评价规则。
所需输出。 CES 差异、渲染差异、主张混淆表、分阶段指标、失败分析、版本哈希，以及保留/撤回决策。
禁止外推。不得为改善检索而加强主张、推断自然发现，或将手工轨迹报告为端到端平台实验。
低算力路径。使用三条主张、一个限定条件、四个查询，手工为各版本指定检索/吸收结果。

D.10 第 9 章：优化与智能体系统

习题 D.17: 9C – 防止平均收益掩盖查询损害

三个查询簇权重为 $(0.5, 0.3, 0.2)$。相对基线，候选 a 的可见性变化为 $(0.10, 0.08, -0.12)$，b 为 $(0.06, 0.05, 0.03)$，c 为 $(0.15, -0.04, 0.01)$。计算各加权均值。发布条件要求平均收益至少 0.04，各簇变化不低于 -0.02，忠实性至少 0.95，风险至多 0.05；对应 (F, R) 分别为 $a : (0.97, 0.03)$、$b : (0.96, 0.04)$、$c : (0.94, 0.02)$。判断资格，解释为什么最高平均值未必对应最终发布候选。提出一个能显露分组冲突的极小极大或遗憾目标。

所需输入。簇权重与变化、忠实性与风险值、所有发布阈值，以及并列规则。
所需输出。加权均值、簇损害检查、完整判定表、发布决策，以及形式化替代目标。
禁止外推。不得把加权查询面板解释为用户总体，或把合成可见性收益解释为因果的、持久的或具有商业价值的收益。
低算力路径。手算三个点积，写出替代目标即可，无需运行优化器。

习题 D.18: 9R – 审查优化器的完整候选轨迹

为一个有界编辑空间构造 20 个候选的轨迹，包含父提议、编辑类型、开发分数、验证分数、留出集状态、忠实性、风险、成本、评判器版本与拒绝原因。注入重复访问留出集、静默更换评判器、两个无收益候选、一个开发高分/验证低分候选、一项保护性约束违反，以及一个重复提议。重建候选筛选流程，隔离受污染评估后，判断是否仍有有效发布候选。

所需输入。带类型的修改空间、20 行合成轨迹、固定数据划分、预算、判定条件、评判器身份，以及六类注入审查事件。
所需输出。谱系图、筛选流程、污染登记表、帕累托图或表、包含拒绝候选的成本台账、无收益/损害表，以及签署的发布/撤回决策。
禁止外推。不得将剩余候选表述为全局最优、自主证明，或智能体循环应拥有发布权的证据。
低算力路径。使用十个候选和表格排序/筛选，以简单父节点列表绘制谱系，无需运行智能体。

D.11　第 10 章：多模态与平台生态

习题 D.19: 10C – 修正存在溯源依赖的证据计数

六个素材看似都支持一条主张：三篇页面转载同一原始报告，两张图来自同一共享数据集，一份独立监管记录提供限定条件。先计算简单信源数与溯源起点数，再定义二元素材—起点关联矩阵，说明将六个素材视为独立证据为何夸大独立印证。随后为文本、像素、图注、替代文本和元数据构建跨模态支持矩阵，有意加入一处冲突。提出一个消融顺序，用于定位表示贡献，同时保留冲突。

所需输入。六个素材身份、三个溯源起点身份、依赖关系、五个表示通道、一条带范围主张和一个限定条件。
所需输出。简单与考虑起点的计数、关联和支持矩阵、依赖图、消融顺序，以及未解决冲突说明。
禁止外推。不得将起点数称为有效统计样本量、由独立性推断真实性，或由消融计划推断平台的多模态权重。
低算力路径。手绘两个二元矩阵，比较六个素材与三个起点；无需视觉模型。

习题 D.20: 10R – 设计配对的平台界面观察

在两个带版本的界面，或两种明确的开放系统配置之间，设计非破坏性配对观察。使用八个意图匹配查询、两种语言、两次重复，并在界面允许时同时采用动态预设别名和显式固定的调用方可见配置。预先定义实体别名、回答与引证结果、缺失状态、账号与地区控制，以及测量等价性检验。描述性估计查询内对比，以及界面与语言的交互。

所需输入。观察卡、确切提示词、查询配对表、语言复核、界面与配置身份、时间戳、输出，以及缺失/重试规则。
所需输出。单元格档案、披露表、含区间或原始配对值的对比、等价性审查、配置漂移日志，以及有边界的迁移陈述。
禁止外推。不得推断隐藏检索或排序机制、混合动态与固定配置，或把界面差异归因于品牌价值。
低算力路径。使用四组双语查询对及手工提供的合成输出，报告各对的方向与幅度，无需拟合模型。

D.12　第 11 章：操纵、防御与机制设计

习题 D.21: 11C – 纳入误报评估防御效用

在 100 个正常案例和 40 个受攻击案例上评估两种防御。防御 A 阻止 32 次攻击和 18 个正常案例；防御 B 阻止 26 次攻击和 4 个正常案例。分别计算攻击召回率、正常案例误报率、拦截决策精确率与残余攻击率。定义 $U = 3\,TP - 2\,FP - 5\,FN$，按该效用选择。再令漏报成本从 1 变化至 8，开展敏感性分析并识别偏好反转。说明为什么效用函数是政策选择，而非经验定律。

所需输入。两个混淆矩阵、类别数、指标定义、效用权重，以及漏报成本敏感性范围。
所需输出。所有指标及分母、效用值、反转阈值或区间，以及单独报告正常效用损失的决策。
禁止外推。不得推断对未知攻击的稳健性、直接迁移类别比例，或声称拦截准确性就确立了最终用户安全。
低算力路径。使用四格混淆表和计算器，依次计算 1 至 8 的整数漏报成本。

习题 D.22: 11R – 开展有界红队与防御审查

在隔离的合成处理链路中，比较正常澄清、无证据说服、嵌入内容的提示式指令、重复起点共识，以及竞争者挤出条件。仅使用虚构实体与无执行作用的字符串。测试两种透明防御，保留正常效用、证据召回、归因、攻击成功、误报与弃权结果。包括一种未用于防御调优的攻击族，以及一次模拟回滚失败。

所需输入。合成语料与实体、威胁模型、攻击者能力预算、五类高层条件、两种防御、结果、安全边界和停止规则。
所需输出。对抗测试清单、按攻击族结果、正常效用表、已知与留出攻击比较、剩余风险决策，以及回滚复盘。

禁止外推。不得发布可操作载荷、未经授权测试真实目标、从已知攻击泛化，或将拒绝率视为联合效用。
低算力路径。用纸卡表示 20 个案例，采用一种规则防御和一种人工审查防御；模拟模型结果，不执行对抗内容。

D.13 第 12 章：治理、标准与事件

习题 D.23: 12C – 区分控制映射与符合性证据

考虑治理对应表的四行：(1) 已核实官方身份与状态，但没有项目控制；(2) 控制已存在，但没有负责人或测试；(3) 负责人、工件及通过的测试均存在，但适用性未解决；(4) 适用性、控制、负责人、工件、测试、例外路径与更新日期均有记录。定义这些字段的完整性谓词，分类各行。形式化说明：溯源完整性是重建决策的必要条件，却不足以证明真实性、法律适用性或符合性。再举一个字段全部填写、但测试测错对象的反例。

所需输入。四行记录、必要字段集、以状态为先的来源身份，以及明确完整性谓词。
所需输出。行分类、必要性/充分性逻辑论证、错误对象反例、待复核标记，以及有边界的治理结论。
禁止外推。不得将对应表、检查清单、公开元数据或内部测试标为认证、合规、法律意见或免责依据。
低算力路径。将每行表示为八位向量，手工检验谓词。

习题 D.24: 12R – 在事件演练中保全证据

开展桌面推演：一页已发布的合成证据页面，其展示回答突然将一条过时且影响重大的主张归到错误信源。初始时刻，团队拥有不可变发布清单，但对下游缓存的可见性不完整。分配最终负责、执行负责、咨询与知会角色。保全观察，评估范围，遏制受影响发布，纠正主张及依赖工件，按虚构政策通知，验证恢复，并开展无责复盘。演练中注入一次回滚失败和一次来源状态变化。

所需输入。合成发布包、CES 图、角色映射、监测告警、观察卡、回滚流程、通知规则，以及两个注入事件。
所需输出。带时间戳事件日志、证据保全记录、范围图、决策记录、纠正传播矩阵、回滚证据、剩余风险登记表、更新任务与复盘。
禁止外推。不得推断真实法律义务、隐瞒不确定性、覆盖原始证据，或在所有已声明验收测试通过前声称恢复。
低算力路径。使用一条主张、两个依赖工件、四个角色及纸面时间线；将法律与外部通知问题交由具备相应资格的人员复核。

D.14 部分习题的解答要点

以下内容覆盖每章的概念题，给出核心推理步骤与参考结果，不代表唯一可接受的论证。实证题有意不设唯一数值答案；其参考输出是题目列明的作业要求、可追溯性与主张边界。

习题 D.1（1C）解答要点

基线联合概率为 $0.90(0.60)(0.75)(0.80)(0.50)(0.20) = 0.0324$。由于 $P = \prod_j p_j$，对每个正的条件概率均有 $\partial \log P / \partial \log p_j = 1$；相同比例的变化具有相同的局部比例效应。干预 a 得到 0.03726，绝对增加 0.00486；干预 b 得到 0.03888，增加 0.00648。计算固定其他条件概率，未考虑处理导致的依赖或选择。分组反转可用任意满足要求的加权构造：高权重、高基线层的大收益掩盖另一层的负变化。损害须单独保留，不得在总体汇总中被抵消。

习题 D.3（2C）解答要点

两个最小分数为 0.40、0.65，均值为 0.74、0.65。均值将 c_1 排在首位，但瓶颈判定和逐分量 0.60 发布规则仅接纳 c_2。若 $m(\mathbf{x}) = \min_j x_j$，且所有 j 均满足 $x'_j \geq x_j$，则 $m(\mathbf{x}') \geq m(\mathbf{x})$，由此证明单调性。增加一个已严格高于当前最小值的分量，不会改变最小值。新增信源可覆盖缺失子主张，同时与既有数值或时间范围冲突，因此覆盖度上升而一致性下降。该函数及分数是编辑发布规则，不是真实性概率模型。

习题 D.5（3C）解答要点

收益成本比分别为 $8, 5, 2, 1,$ 和 1。按执行后检验约定及“收益低于 2 或比值低于 1.5”的原规则，第 4 个子查询触发停止：其收益为 1，低于 $\epsilon_E = 2$；收益成本比也为 1，低于 $\tau = 1.5$。覆盖约束要求执行到第 4 个子查询：溯源多样性首次在第 3 项满足，反证首次在第 4 项出现。执行前检验需要预测收益，并说明未达阈值的提议是否仍会执行。第 5 个查询收益回升至 3，因此边际序列并非单调。有限的两步前瞻或预留覆盖查询，可以避免将首次低收益误当成此后所有提议均无效的证明。

习题 D.7（4C）解答要点

RRF 分数为 $A = 1/61{+}1/62$、$C = 1/63{+}1/61$、$B = 1/62{+}1/64$、$E = 1/63$ 和 $D = 1/64$,得到 $A > C > B > E > D$。RRF 仅使用排名位置，因此保持各列表顺序的严格递增变换不会改变任何一项。原始分数相加则可能随某检索器分数的平移或缩放而变化。在预算 6 下，可行非空打包集合为 A、C、E 和 $C + E$。若每覆盖一种证据类型只奖励一次 3，则目标值依次为 $10, 9, 10,$ 和 16，$C + E$ 最优。若 C、E 同源且该起点上限为一项，此组合不可行，说明治理约束可以改变算术最优解。

习题 D.9（5C）解答要点

按题述局部位置归属规则，只有主张 1 后的展示引证 S 存在蕴含关系；主张 4 后的 S 和主张 3 后的 T 均不成立。引证蕴含率为展示主张—引证关系中的 $1/3$，完整率为关键主张中的 $1/4$，展示引证率为主张中的 $3/4$，S 特有吸收率为由 S 支持的主张中的 $2/3$。设灵敏度为 Se、特异度为 Sp，有 $\tilde{p} = Se\,p + (1 - Sp)(1 - p)$，因此 $p = (\tilde{p} + Sp - 1)/(Se + Sp - 1) = 0.45/0.70 \approx 0.643$。可行性要求 $Se + Sp > 1$、校正值落在 $[0, 1]$ 内，且校准集具有代表性。

习题 D.11（6C）解答要点

共有 $30{\times}4{\times}2 = 240$ 个回答单元格。重复四次时,$DE = 1{+}3(0.35) = 2.05$,粗略的重复调整有效样本量为 $240/2.05 \approx 117.1$。它仅用于提醒：不能把 240 个相关回答当成独立样本。一个最小模型为 $Y_{iqtr} = \mu{+}\tau T_i{+}a_q{+}b_t{+}(aT)_{qi}{+}\varepsilon_{iqtr}$，其中查询和时间分量应符合设计；主要对比应在随机分配或配对的查询簇单位上构造。相关查询簇、漂移和分配依赖还需设计特定的处理，超出单一组内相关近似的范围。

习题 D.13（7C）解答要点

处理组干预后均值为 $(0.70 + 0.50)/2 = 0.60$，对照组为 $(0.45 + 0.35)/2 = 0.40$，仅干预后均值差为 0.20。处理组变化为 $[(0.70 - 0.40) + (0.50 - 0.45)]/2 = 0.175$；对照组变化为 $[(0.45 - 0.40) + (0.35 - 0.30)]/2 = 0.05$，因此双重差分为 0.125。随机化的目标估计量是这四个簇被分配至已声明处理版本的有限样本平均效应。若某处理簇从未实际接收修订，分配效应与实际处理效应便不同。若处理簇 A 改变了对照簇 C 检索到的证据，稳定单位假设失效。时间对比还需有可信的平行趋势论证。

习题 D.15（8C）解答要点

r_0、r_1 通过所有阈值。r_2 不满足忠实性，r_3 不满足效用和风险约束。相对 r_0，r_1 的可见性增加 0.12，高于最低值 0.08，并保持其余条件，因此在不确定性与方案检查通过后可作为发布候选。可行集内，两个版本互不支配：r_1 改善可见性和效用，r_0 则忠实性更高、风险和成本更低。对可见性和效用赋予足够权重，加权和可能将 r_2 排在首位，但这不能修复其忠实性约束失败。先判断可行性，再汇总偏好。

习题 D.17（9C）解答要点

a、b、c 的加权收益分别为 0.050、0.051、0.065。a 因 $-0.12 < -0.02$ 未通过分组损失条件；c 同时违反第二簇损失与忠实性条件。b 通过均值、分簇、忠实性和风险阈值，因而是唯一合格发布候选。一个透明替代方案，是在其他约束下最大化 $\min_k \Delta V_k$，或以每簇最佳可行候选为基准，最小化最大加权遗憾。该例说明，最大加权均值不等于可接纳性或利益相关方可接受性。

习题 D.19（10C）解答要点

简单计数为六个素材，但依赖图仅有三个起点：一份报告、一个数据集和一份独立监管记录。关联矩阵每行对应一个素材，每列对应一个起点；三篇转载页面落在同一报告列，两张图落在同一数据集列。该表示保留访问路径，同时避免把它们误当作独立印证。支持矩阵应区分各表示通道的支持、未涉及和冲突。先逐个消融表示，再消融一致的通道组，并保留仅含冲突的条件。消融引起的变化是当前构造中的条件贡献，不是真实性检验，也不是公开的平台权重。

习题 D.21（11C）解答要点

对 A,$(TP, FP, FN, TN) = (32, 18, 8, 82)$;召回率 $32/40 = 0.80$,正常 FPR 为 $18/100 = 0.18$,精确率为 $32/50 = 0.64$,残余攻击率为 $8/40$。对 B，元组为 $(26, 4, 14, 96)$，对应 0.65、0.04、$26/30 \approx 0.867$ 和 $14/40$。漏报成本为 5 时，$U_A = 20$、$U_B = 0$，按已声明政策选择 A。漏报成本为 c 时，差值 $U_A - U_B = 3(6) - 2(14) + c(6) = 6c - 10$，故 $c > 5/3$ 时偏好 A，$c < 5/3$ 时偏好 B。反转来自利益相关方的损失权重，不能由混淆矩阵单独决定。

习题 D.23（12C）解答要点

只有身份/状态、适用性、控制、负责人、工件、测试、例外路径和更新日期均以正确对象类型填写时，令 $K(r)$ 为真。第 1–3 行因题述字段缺失而不完整；第 4 行结构完整。完整性使复核者能重建谁将什么映射到哪个版本、依据何种证据，却不能推出来源真实、适用法律已正确解释，或符合性评价机构会认可该控制。一行记录可以全部填写，却只测试预发布页面，而本次发布涉及生产 API；它满足语法谓词，但测错对象。因此，语义复核与专业适用性判断仍是独立判定条件。

适用边界与失效条件

实证题规定的是可复查的工作，不是声称这些工作已执行、经独立复现，或获准用于真实目标。公开发布前，各参考输出仍需带版本的工件包、许可复核、安全复核，以及对照题述作业要求的独立检查。

E 记号、符号与术语

核心思想

本书的记号构成一套类型系统。使用一个符号之前，必须明确其定义域、单位、认知状态、条件集合与适用章节。尤其要注意：不能以观察到的界面对象代替隐藏的平台状态，估计量不同于目标估计量，界面显示的引证也不能证明信源被使用或主张得到事实支持。

E.1 本附录的使用方法

本附录规定英文原稿及其中文译本的记号与术语用法，汇总十二章中首次出现的符号，并记录当前工作版本为兼容旧稿而保留的少量问题。这些约定不会使概念变量变得可观测，也不构成代理变量的有效性验证，更不会为经验主张增加证据。

每个数学对象均从四个相互独立的维度加以描述：

1. **定义域：**该对象可以取值的集合；
2. **单位：**定义该对象所依据的实体、主张、信源、段落、回答、用户任务、时间区间或其他单位；
3. **认知状态：**观测、构造、估计、潜在或反事实；
4. **适用范围：**赋予该对象意义的查询、系统、语料库、产品界面、区域设置、时间与协议。

已观测的对象仍可能需要通过规则构造标签。例如，回答文本是直接观测到的；其中某项主张是否得到忠实支持，则需要经过复核的测量。反过来，即使某个量定义精确，它在封闭产品界面中仍可能是潜变量。定义的精确性不等于数据的可获得性。

E.1.1 认知状态图例

代码	状态	本书中的含义	最低报告要求
O	观测	在声明的界面或受控轨迹中直接采集	标识、时间戳、协议单元、原始值与缺失状态规则
C	构造	根据有记录的变换或人工标注规则产生	输入版本、变换或指南、复核者及可复现性
E	估计	根据抽样或分配处理后获得的观测计算	目标估计量、估计量、抽样或分配单位、不确定性与缺失情况
L	潜在／依赖代理变量	研究环境中未直接暴露	可观测代理变量、假设、替代解释，并明确说明未直接观测
CF	反事实	在同一单位无法同时观测的处理或状态下定义	处理版本、分配方式或识别论证及相关假设
N	规范／受治理	由政策、协议、审批或决策规则设定，而非从回答数据中学习	负责人、版本、适用性、阈值、例外与复核日期

E.2 全书统一约定

E.2.1 随机对象、实现值、目标估计量与估计值

大写拉丁字母通常表示随机对象，小写字母表示其实现值。因此，Y 是以回答为取值的随机变量，y 是一次实现的回答；$\mathcal{A}$ 是界面引证对象的随机集合，$\dashv$ 是其实现集合。当这一差别影响论证时，采用上述约定。固定信源标识 s 或查询对象 q 仍用小写，因为它们是相应协议的输入。

不带帽的希腊字母可表示参数或目标估计量。帽号表示估计量或估计值：τ 是处理效应的目标估计量，$\hat{\tau}$ 是按声明的分析方法得到的估计值。波浪号用于明确区分含噪、代理或校正前的测量，例如 $\tilde{p}$ 表示观测到的评判器阳性率。下标应指明总体、环节、分层或协议，不能仅用于装饰一个未定义的得分。

适用边界与失效条件

本书不以帽号暗示真实性或因果识别。即使根据现有单元精确算出估计值，它仍可能估计了错误总体、以依赖处理的缺失情况为条件，或依赖无效的识别假设。

E.2.2 集合、序列、向量与指示函数

花体大写字母表示集合、空间、对象集或随机核，例如 $\mathcal{D}_t$、$\mathcal{W}_t$ 和 $\mathcal{G}_\theta$。当顺序具有实质意义时，圆括号保留顺序或条件关系；除非另有说明，花括号表示无序集合。**J** 和 **V** 等粗体符号表示向量。各分量保留独立单位，只有声明尺度变换、权重和不可补偿约束之后，才可进行汇总。

事件 A 成立时，指示函数 $\mathbb{1}\{A\}$ 为一，否则为零。将分级标签转为指示变量时，必须记录阈值。有限集合 S 的基数记为 $|S|$；词元、资金、时延或能力预算必须明确注明类别，不能根据一个裸数字推断。

E.2.3 条件、时间与版本

每个概率都以表达式中列出的变量，或上下文所指的带版本评估单元为条件。为便于阅读而省略条件变量，并不意味着可以对未知的平台、区域设置、时间或账号分布求平均。时间戳 t 表示一次观测的时间；T 表示声明的时间窗口。供应商名称、动态 API 别名或复制的提示词，本身均不构成版本锁定。

E.3 总符号表

定位列指向符号首次实质使用的位置。除非明确标注为局部兼容用法，后续出现时均沿用同一含义。这里的单位是语义单位，不仅仅是数值的数据类型。

表 50: 系统、信源与证据对象。

符号	定义域／单位	状态	含义与限制	首次出现
q	文本或结构化请求；查询—任务	O	用户提交给产品界面的请求，不一定是实际送入检索的字符串。	章 1
u	带版本的用户状态记录	O / L	账号、区域设置、对话与个性化状态；只有已采集的字段属于观测值。	章 1
t, T	时间戳；分析窗口	O / N	一次观测的时间，以及声明的目标或监测窗口。	章 1
$\mathcal{W}_t$	可访问信源表示的集合	封闭系统中为 L；冻结系统中为 O	系统在 t 时刻能够访问的信息环境，不等同于公共互联网。	章 1
θ	配置／版本记录	O / L	系统状态中可识别的部分。未暴露的后端状态仍为潜变量。	章 1
$\mathcal{G}_\theta$	条件随机核	C	生成式搜索系统的研究模型，不是对商业架构的逆向还原。	章 1

下页续

表 *50*（续）

符号	定义域／单位	状态	含义与限制	首次出现
Y, y	回答值随机对象；实现的回答	界面上为 O	Y 为条件输出对象；y 为一次采集到的实现值。	章 1
$\mathcal{A}$, $\dashv$	引证／归因对象集合	解析后为 O	界面归因的随机对象与实现值。两者均不能证明信源使用或证据支持。	章 1
L, ℓ	交互或行动轨迹	遵守同意与保留规则时为 O	可观测轨迹的随机对象与实现值；不包括臆测的内部调用。	章 1
s	受治理的信源 ID；信源制品	O / C	完成版本解析的信源，不能仅用 URL 字符串或发布者名称替代。	章 1
c	主张 ID；有范围的命题	C	经过复核的最小命题，带有范围、时间、区域设置与状态。	章 2
e^{ev}	证据 ID；证据单位	C	与主张关联的相关片段、表格、记录、数据集或观测。推荐使用该上标，消除旧式裸 e 的歧义。	章 2
$I(s)$	信源身份记录	C / O	发布机构、类型、题名、版本、状态、方法、制品、许可和 URL；身份并不证明主张获得支持。	章 2
$A(c)$	充分性判定输出	C / N	针对已声明证据维度的项目编辑瓶颈判定，不是普适的科学质量评分。	章 2

表 51: 采集、检索、生成与归因对象。

符号	定义域／单位	状态	含义与限制	首次出现
$\widehat{q}$	实际发出的检索查询对象	有日志的系统中为 O；封闭界面中为 L	查询计划产生的一条查询。此处帽号表示派生查询，不表示统计估计。	章 3
$\mathcal{Q}(q,u)$	有序或保留依赖关系的查询集	O / L	以请求与状态为条件展开的查询。当顺序或依赖影响证据采集时，必须保留这些关系。	章 3
$\Pi(d)$	产生有序或集合表示的变换	受控系统中为 O	从文档 d 派生的带版本段落或表示。	章 4
$\mathcal{D}_t$	冻结的语料库快照	O / C	带有时间、哈希、许可与切分版本的候选语料库。	章 4
$r_{\mathrm{ret}}(d,q)$	实值检索得分	受控系统中为 O	指定检索器给出的得分。它不是概率，也未必能跨查询或检索器比较。	章 3
$R_k(q)$	有序的前 k 项列表	受控系统中为 O	在声明深度 k 下为查询 q 检索出的表示。	章 4
g_{qi}	通常为 $\{0,1,2,3\}$；查询—候选项	C	对候选项 i 的预注册分级相关性判断。评分尺度与指南是测量的一部分。	章 4
C	有序的有效上下文	有日志的系统中为 O；封闭界面中为 L	生成过程实际可以使用的段落或表示，包括顺序与截断信息。	章 3
$A(y)$	实质性原子主张集合	C	回答 y 中可评估支持关系的最小主张，保留数量、范围、条件、时间与归因。	章 5
$R(y)$	已解析的引证对象集合	O / C	将显示的锚点解析为规范标识、版本、访问时间及可获得时的精确段落。解析不意味着得到支持。	章 5
E_y	主张—引证对集合	C	回答 y 中可见或按规则声明的关联边。边记录挂接关系，不表示证据支持。	章 5

下页续

表 *51*（续）

符号	定义域／单位	状态	含义与限制	首次出现
w_j	非负的主张权重单位	N / C	原子主张 a_j 的预注册重要性权重；需披露尺度与敏感性分析。	章 5
C^{-s}	有序的反事实上下文	C	移除信源 s 后构造的上下文；记录顺序、预算、位置、替代与截断的每项变化。	章 5
U, U_i	声明的回答效用单位	C / E	上下文对比的预注册效用。Orchid Bay 的二元测试样例仅局部有效，不能作为通用效用推广。	章 5
B_{ctx}	词元或能力单位	N / O	声明的上下文或工具预算。使用带类型的下标，不得复用于编辑或攻击预算。	章 3
$Z_{\text{ctx}}(s)$	二元或分级的信源—回答事件	O / L	信源 s 的一种表示进入有效上下文。	章 5
$Z_{\text{use}}(s)$	二元或分级的信源—回答事件	E	指定的消融或归因方法将回答的实质性影响归于 s。	章 5
$Z_{\text{abs}}(s)$	二元或分级的主张—信源—回答事件	C / E	预注册的、具有信源特异性的主张忠实出现在回答中。	章 5
$Z_{\text{cite}}(s)$	二元或分级的信源—回答事件	O / C	界面显示的一条引证解析到信源 s。	章 5
S_{jk}	支持标签；主张—段落对	C	人工或经校准的评判器按指定指南判断段落 k 是否支持回答主张 j。	章 5
$\Delta_s(C)$	回答效用单位	E	在构造的上下文 C 中进行的条件性留一信源对比；不等同于作者身份或整个平台上的使用情况。	章 5

表 52: 测量、因果、干预与治理对象。

符号	定义域／单位	状态	含义与限制	首次出现
Q, M	有限加权面板或目标分布	N / E	查询与系统总体。此处大写 M 表示分布；生成模型使用 M_{gen}。	章 6
w_q, w_m	非负归一化权重	N / E	查询与系统的设计权重或目标权重，须说明其来源。	章 6
e^{cell}	带版本的评估单元	C / O	由查询、系统、版本、时间、区域设置、地理位置、账号、重复次数与产品界面组成的元组。	章 6
μ, $\widehat{\mu}$	结果单位	目标估计量 / E	总体均值及其估计值。加权目标均值不同于各层等权的诊断均值。	章 6
D_i	二元或多值的处理分配	O	分配给实验单位 i 的处理版本；不是检测器。	章 7
S_t	带版本的系统状态	O / L	定义潜在结果时所处的状态；必须明确其中未观测的部分。	章 7
$Y_i(d, S_t)$	结果单位	CF	单位 i 在处理 d 和系统状态 S_t 下的潜在结果；只能观测到实际分配状态下的结果。	章 7
τ, $\widehat{\tau}$	声明的结果单位	目标估计量 / E	处理效应目标及其估计值。总体、比较对象与时间状态属于下标或周边协议的一部分。	章 7
x, x'	带版本的信源状态	O / C	已发布的基线状态，以及提议或候选的信源状态。	章 8

下页续

表 52（续）

符号	定义域／单位	状态	含义与限制	首次出现
δ	带类型的变更记录	O / C	信源状态间允许的干预，包括主张 ID、证据 ID、负责人、范围与回滚。	章 9
$\mathcal{X}_{\mathrm{perm}}(x)$	信源状态集合	N / C	满足证据、政策、安全、无障碍与发布条件的候选修订。	章 9
$\mathbf{J}(x')$	向量；各分量使用独立单位	E / N	检索、引证、吸收、忠实性、效用、风险与成本目标。未声明尺度和权重时，不得求分量平均。	章 9
B_{edit}, B_{atk}	编辑次数、词元、费用或能力单位	N	优化与攻击者预算。每个数值必须注明单位。	章 9, 章 11
a^{asset}	素材 ID	O / C	用于 $\Pi(a^{\mathrm{asset}})$ 的多模态素材，不同于账号状态。	章 10
$\Pi(a^{\mathrm{asset}})$	带版本的表示集合	受控系统中为 O；其余为 L	由受治理素材派生的像素、图注、替代文本／ OCR、元数据与页面表示；仅对已声明可访问的子集评分。	章 10
$B(y)$, $C_b(y)$	实体 ID；有范围的主张集	C	回答 y 中已解析的品牌实体，以及关联到受治理实体 b 的主张。含糊和未解析的字符串须明确保留为错误状态。	章 10
$\mathbf{V}_b$	含七个分量的结果向量	E / C	提及、正确性、引证、吸收、显著程度、行动，以及单独定义的风险。未声明尺度变换与决策权重时，不存在综合评分。	章 10
$o(s)$, $N_{\mathrm{origin}}(S)$	起源簇映射与计数	C	信源 s 最早经核验的溯源起点，以及 S 中映射后不同起源的数量；两者均依赖追溯完整性。	章 10
$G = (V, E)$	有向溯源图	O / C	主张、信源、发布者、版本与实质性派生边。覆盖程度取决于追溯协议。	章 10
$D_{\mathrm{det}}(\delta)$	检测器得分单位	E	针对干预 δ 声明的检测器输出；阈值与校准是其含义的一部分。	章 11
$\Gamma(r_{\mathrm{rel}})$	结构化发布记录	O / N	针对一次具体发布，记录身份、变更主张、证据、风险、控制措施、审批、监测与回滚。	章 12

E.4 多义符号的兼容映射

当前版本保留了若干局部多义符号，以便各章源文件仍可追溯到早期草稿。新的推导与后续修订使用下列带类型形式。章节中的局部定义仍然有效，但不得超出其声明的范围。

旧式形式	现有局部含义	推荐的带类型形式	规则
e	证据项；评估单元	e^{ev}; e^{cell}	不得仅凭裸 e 关联这两类表。
B	上下文、编辑、工具或攻击者预算	B_{ctx}, B_{edit}, B_{tool}, B_{atk}	每个数值均须注明单位；不同类别的预算不得相加。
a	多模态素材；账号状态；局部回答主张索引	a^{asset}, a^{acct}, a_j^{claim}	局部索引不能充当具有全局标识的素材或账号。
M	生成模型；目标系统分布	M_{gen}; M 表示分布	模型对象不是从未指定的系统总体中随机抽取的样本。
D	处理分配；检测器	D_i; $D_{\mathrm{det}}(\delta)$	检测器输出不是处理状态。
s	信源制品；局部表示产品界面	s; s_{surf}	信源 ID 与产品界面不得共用一个数据库字段。
r	重复索引；检索得分；发布对象	r_{rep}, r_{ret}, r_{rel}	只能依据明确的局部定义判断含义，不能根据字母形状猜测。
C	有效上下文；局部推导中的主张集或簇数	C; $\mathcal{C}_s$; n_{cl}	保留上下文顺序；主张集合与簇的数量使用不同类型。

派生查询 $\widehat{q}$ 是“帽号表示估计量”这一通常约定的唯一有意例外。各章用它区分用户请求与实际发出的检索查询，因此保留该写法。任何时候均须说明其定义域及轨迹的可观测状态。

E.5 不可混同的概念

下列概念对相互关联，却不能互换。若草稿表述或分析越过这一区分，右列给出最低限度的修正要求。

不得用此项	替代此项	必须采取的修正
爬虫请求或成功抓取	被索引、检索、引证、提及或引发行动	指明已观测的采集事件，将后续环节标为未观测。
文档包含一项主张	可以检索到相关证据段落	说明渲染、切分、表示、查询与检索深度。
检索得分	经校准的相关性概率	报告检索器及得分尺度；仅依据已声明的判断进行校准。
进入前 k 项	进入有效上下文	保留重排序、上下文打包、顺序、词元预算与截断的证据。
界面显示的引证	信源使用、主张支持、真实性、作者身份或权威性	解析引证，拆分主张，标注支持关系；必要时另行设计影响评估。
提及率	正确性、推荐、情感、显著程度或商业价值	分别报告每项结果及其分母。
模型评判器标签	人工真值	分层校准，保留分歧，并传播合理的误差范围。
回答记录行数	独立单位数量	识别抽样簇与处理分配簇，将重复观测视为嵌套测量。
前后变化	处理效应	定义处理版本、比较对象、分配或识别假设、干预生效情况与同期变化。
统计上可检测	实践或伦理上的成功	预先声明最小有价值效应，以及忠实性、效用和风险判定条件。
多个 URL	多个独立起源	追溯派生关系，并报告起源簇。
标准或政策的身份	适用性、符合性、合规性或控制有效性	核验状态与范围，由合格人员评估适用性，并附实施与测试证据。
平台或模型名称	冻结的系统状态	记录产品界面、版本证据、请求、区域设置、账号、时间、搜索状态与全部未知项。

E.6 跨章节定位表

章	主要对象	引入的核心区分	后续依赖
1	$q, u, t, \mathcal{W}_t, \theta, Y, \mathcal{A}, L$	系统对象与实际界面观测；不同可见性事件	为全部测量和干预设定条件
2	$s, c, e^{\mathrm{ev}}, I(s), A(c)$	信源身份与主张支持；溯源信息与真实性	第 5、8、11、12 章的证据判定条件
3	$\widehat{q}, \mathcal{Q}(q,u), r_{\mathrm{ret}}, C, B_{\mathrm{ctx}}$	已观测轨迹与重构的查询或隐藏上下文	检索与生成协议
4	$\mathcal{D}_t, \Pi(d), R_k(q), g_{qi}$	文档与段落证据；排序得分与相关性判断	基准测试与上下文分配指标
5	$S_{jk}, Z_{\mathrm{ctx}}, Z_{\mathrm{use}}, Z_{\mathrm{abs}}, Z_{\mathrm{cite}}$	选择、使用、吸收、支持与显示	引证审计与信源影响研究设计
6	$Q, M, w_q, w_m, e^{\mathrm{cell}}, \widehat{\mu}$	目标估计量与样本估计值；计划分母与观测分母	实验设计与仪表板
7	$D_i, Y_i(d, S_t), \tau, \widehat{\tau}$	分配、潜在结果、目标估计量与估计量	因果与时间性主张
8	x, x', δ	保持主张的表达修改与新主张	发布验证与优化
9	$\mathcal{X}_{\mathrm{perm}}, \mathbf{J}, B_{\mathrm{edit}}$	向量目标与不可补偿约束	智能体提案与选择
10	$a^{\mathrm{asset}}, \Pi(a), G = (V, E)$	模态表示与素材；URL 数量与起源依赖	多模态与生态系统审计
11	$B_{\mathrm{atk}}, D_{\mathrm{det}}, \mathbf{U}$	攻击结果与检测器测量、诚实信源效用	风险判定条件与监测
12	$\Gamma(r_{\mathrm{rel}})$	制度性信源状态与适用且经过测试的项目控制措施	发布、事件、纠正与勘误

E.7 术语表

本术语表规定本书中的用法，并不主张所有平台、论文或专业群体都以相同方式使用这些术语。

E.7.1 生成式搜索与可见性

生成式搜索界面（Generative search surface）

观察生成式回答的具体产品模式、页面、API 或交互界面。同一供应商可能提供多个并不等价的界面。

生成式搜索系统（Generative search system）

章 1定义的条件随机对象；它可能进行搜索、从受控集合中检索、使用工具、依赖参数化知识，或组合这些路径。

生成式引擎优化（GEO）

对生成式搜索环境中的信源可见性、使用、归因与忠实性开展研究，并实施受治理的干预。在本书中，GEO 既不保证曝光，也不等同于内容说服。

可见性画像（Visibility profile）

由分别估计的结果组成的向量，例如检索、引证、吸收、归因、忠实性与时间变化；不存在通用的综合评分。

提及（Mention）

按声明的别名规则解析出的实体出现在回答中。提及不能证明正确性、信源使用或情感倾向。

引证（Citation）

界面显示的归因对象解析到某个信源。引证是一种界面事件，并非支持关系或因果使用的判断。

归因（Attribution）

根据声明的解析器或方法，将回答主张或贡献归于显示或推断出的信源的关系。

吸收（Absorption）

预注册的、具有信源特异性的证据忠实出现在回答中。其单位是主张—信源—回答，而非 URL 数量。

忠实性（Fidelity）
保留有明确范围的命题，包括重要的计量单位、条件、例外与时间边界。

显著程度（Prominence）
声明的实体或片段在回答中的位置或占比属性，不等同于用户注意力。

行动（Action）
预先声明的用户—任务事件，例如访问或符合条件的咨询请求；测量须获得同意，并采用适当的识别设计。

E.7.2 信息检索与 RAG

采集（Acquisition）
表示变得可发现、可抓取、可渲染、符合准入条件或以其他方式可用的过程；这些状态须分别区分。

表示（Representation）
从信源派生的带版本对象，例如渲染后的文本、段落、图像、图注、元数据记录或嵌入向量。

段落（Passage）
根据声明的切分程序产生的可检索片段。它可能遗漏源文档其他位置的条件。

检索（Retrieval）
针对查询，从冻结或带版本的语料库中找回候选项并排序的过程。检索测量先于回答生成。

词项检索（Lexical retrieval）
基于离散词项证据的检索，通常使用倒排索引与词频／文档频率得分。

稠密检索（Dense retrieval）
在锁定编码器的条件下，基于学习得到的向量表示与声明的相似度函数开展检索。

混合检索（Hybrid retrieval）
通过归一化、校准、排序融合或学习规则组合词项与稠密证据的协议。其默认含义并不是直接相加原始得分。

查询展开（Query fan-out）
由一次用户请求派生、保留依赖关系的一系列检索查询或查询集，并包含成本与停止规则。

排序融合（Rank fusion）
合并有序候选列表。融合时保留找回候选项的查询与溯源路径，避免将重复出现视为独立确认。

重排序（Reranking）
为声明的相关性或效用任务，对候选池进行条件性重新排序。它不会扩大原有候选集合。

上下文分配（Context allocation）
在有限预算及声明的覆盖、冗余、起源与冲突目标下，选择、排序并可能截断表示。

有效上下文（Effective context）
实际供生成过程使用的有序表示。在封闭产品界面中，它通常是潜变量。

检索增强生成（RAG）
以外部检索表示为条件的生成过程。受控 RAG 系统并不意味着商业产品采用相同流程。

引证解析器（Citation resolver）
将界面显示的锚点或标签映射到规范、带版本的信源对象，并在可获得时定位相关段落的过程。

证据召回率（Evidence recall）
在预注册的重要证据单位中，可从前 k 个表示中找回的比例。当一份文档包含多项独立主张时，这一指标比文档召回率更严格。

E.7.3 测量、统计与因果推断

目标总体（Target population）

目标估计量所指向的信源、查询、系统、用户、区域设置、时间或任务。方便取得的测试面板并不会自动成为目标总体。

分析单位（Unit of analysis）

一个分析结果所代表的对象。它可能不同于抽样、标注和处理分配单位。

分配单位（Unit of assignment）

独立随机化或以其他方式分配处理的簇。当处理按查询、页面、域名或时间块分配时，不确定性分析须遵循该单位。

目标估计量（Estimand）

研究希望了解、经过精确定义的总体量或处理对比。

估计量（Estimator）

将观测数据映射为目标估计量之估计值的规则。计算正确不能弥补目标估计量定义不清。

评估单元（Evaluation cell）

查询、系统、产品界面、时间、区域设置、地理位置、账号与重复次数的带版本组合，原始输出与标签均挂接于该单元。

分母台账（Denominator ledger）

计划单元依次经过回答、解析、抓取、切分、标注、排除与分析各环节的流转记录，用于呈现条件总体与缺失情况。

重复测量（Repeated measure）

对同一基础单位或簇再次观测。重复观测用于估计单位内变异，不会产生新的独立抽样查询。

簇（Cluster）

共享意图、信源、处理分配、系统状态，或其他与不确定性有关的依赖结构的一组观测。

自助法（Bootstrap）

重抽样程序，其抽样单位与目标必须匹配数据生成或处理分配结构。按行自助法与簇自助法回答不同的问题。

评判器（Judge）

对对象赋予标签的人工或模型测量程序。评判器存在误差、弃判以及可能的子群漂移。

校准（Calibration）

依据声明的参考集评估测量值或概率，并在理由充分时予以调整。整体校准并不能证明子群校准。

潜在结果（Potential outcome）

某单位在指定处理版本与系统状态下本应具有的结果。同一时刻，只能观测该单位某一个互斥状态下的结果。

处理版本（Treatment version）

实验中分配的完整信源或系统状态，附有哈希，涵盖所有捆绑变更及传播条件。

干扰（Interference）

一个单位的结果依赖分配给其他单位的处理，例如竞争检索位置或上下文空间。

漂移（Drift）

回答生成、测量、语料库、查询或产品过程随时间发生变化。漂移不等同于评估单元内的随机变异。

非劣效界值（Non-inferiority margin）

对保护性指标预先声明的最大可接受退化。它是决策参数，不能在观察结果之后再据此确定。

E.7.4 证据、风险与治理

主张—证据—信源图（CES）

带类型、带版本的图，将有范围的命题连接到相关证据单位及包含或管理这些证据的信源。支持、限定、争议与替代属于不同的边。

信源身份（Source identity）

信源的发布机构、对象类型、题名、版本、日期、状态、方法、制品、许可与稳定定位符。身份不能证明真实性或方法充分性。

证据支持上限（Evidence ceiling）

在已披露的方法与范围内，某信源通常能够支撑的最强主张类别。这是编辑层面的边界，不是声望排名。

溯源信息（Provenance）

对起源、行动者、实体、活动、派生关系与版本的记录。溯源可揭示依赖关系，却不能证明主张为真。

起源簇（Origin cluster）

实质上派生自同一个最早经核验起源的一组信源表现形式。同簇中的多个 URL 不构成独立确认。

保全证据的干预（Evidence-preserving intervention）

修订后的命题仍与已批准主张等价，或获得新批准的证据，同时保留限定条件、溯源信息与披露。

威胁模型（Threat model）

对受保护资产、行动者、能力、访问权限、预算、适应策略、信任边界与危害的带版本说明。

假阳性（False positive）

根据声明的参考标签，本属良性或诚实的信源、主张或干预，被检测器或政策错误标记。

发布证据包（Release evidence package）

可检测篡改的材料包，包括身份、修改前后哈希、主张与渲染差异、证据、协议、结果、审批、控制措施、监测及经过测试的回滚。

制度性状态（Institutional status）

法律、法规、强制性或推荐性标准、自愿框架、技术规范、平台政策或项目经过核验的类别与生命周期状态。这些类别不能混用。

适用性（Applicability）

由具备相应资质或能力的人员，判断某项具体制度文件或控制措施是否适用于实际组织、角色、系统、数据、法域、行业与生命周期阶段。

控制有效性（Control effectiveness）

证明已实施的控制措施在指定测试下达到声明目标的证据。政策声明或映射表本身并不是有效性证据。

回滚（Rollback）

经过测试和授权、用于恢复或遏制已发布状态的操作，同时保留调查所需的证据。

事件（Incident）

已检测到、按照声明的响应协议触发证据保全、分诊、遏制、纠正、通知与学习的事件。

E.8 缩略语与指标索引

缩写	全称／译名	本书中的用途
GEO	generative engine optimization ／生成式引擎优化	对可见性、使用、归因、忠实性及相关风险开展受治理的研究和干预
IR	information retrieval ／信息检索	生成之前的表示、候选找回、排序与评价
RAG	retrieval-augmented generation ／检索增强生成	以检索到的外部表示为条件进行生成
BM25	Best Matching 25 系列	词项评分方法族；实现与参数属于协议的一部分
RRF	reciprocal-rank fusion ／倒数排名融合	基于排名、可复查的候选列表融合方法
DCG / nDCG	折损累计增益／归一化折损累计增益	在声明的相关性标签与截断位置下评估分级排序效用
MRR	mean reciprocal rank ／平均倒数排名	第一个相关条目排名倒数的均值
AP	average precision ／平均精确率	在二元相关性约定下，对相关条目所在排名位置的精确率求平均
GLMM	广义线性混合模型	对非高斯的重复或成簇结果，在连接函数尺度上进行含固定效应和随机效应的回归
ATE	average treatment effect ／平均处理效应	在声明的目标总体中，两个指定潜在结果之差的均值
DiD	difference-in-differences ／双重差分	时间性比较；除其他条件外，须有可辩护的未处理趋势
ICC	intraclass correlation coefficient ／组内相关系数	在指定方差模型下度量簇内依赖
SUTVA	稳定单位处理值假设	处理版本与无干扰条件；在竞争性 GEO 中往往值得质疑
CES	Claim–Evidence–Source ／主张—证据—信源	管理有范围主张、证据单位、信源身份、冲突与溯源信息的项目图
RACI	执行、负责、咨询、知会	角色分配辅助工具；不能证明已分配的工作确实完成
WCAG	Web Content Accessibility Guidelines ／网页内容无障碍指南	无障碍规范来源；适用性与符合性须分别评价
PROV-O	PROV Ontology ／ PROV 本体	溯源词汇体系，不是真实性或证据质量模型

E.9 记号发布检查清单

发布章节、附录、图或经验研究制品之前，编辑应检查：

1. 每个新符号都有唯一的定义、定义域、单位、认知状态与首次使用位置；
2. 每个估计值都指向目标估计量、估计量、独立单位、不确定性方法与缺失状态规则；
3. 每个潜在平台状态均明确标为潜在，使用的代理变量均有名称；
4. 每个向量均保留分量单位与不可补偿约束；

5. 每次使用 e、B、a、M、D、s、r 或 C，均在局部消除歧义或改用带类型形式；
6. 信源 ID、主张 ID、证据 ID、引证对象与发布者，绝不只因字符串相同就相互关联；
7. 术语用法遵守节 E.5列出的不可混同规则；
8. 任何符号或术语变更，均在版本历史中记录受影响的主张、方程、图、数据模式与交叉引用。

本章小结

核心规则很简单：若两个量的定义域、分母、认知状态或决策作用不同，就应使用不同名称。平台状态无法观测时，记号必须保留这一事实。指标若不能回答研究问题，排版再精确也无法改变这一点。

F 先修知识衔接

信息检索、统计学与因果推断

核心思想

本附录补充阅读正文所需的最低限度技术基础，避免把外部课程变成未明示的先修要求。每个单元依次说明学习必要性、定义研究对象、演算一个小型合成示例、指出常见错误，并以检查题结束。这些示例用于学习计算与解释，不是已经完成的实证研究，也不是关于生产平台的证据。

F.1 阅读说明与学习顺序

这些衔接单元有意保持精简，不能替代完整的信息检索、概率、统计、实验设计或因果推断课程。它们提供必要的分析框架，帮助读者辨认一个公式测量什么、需要哪些数据，以及不能支持什么结论。符号沿用章 E。

主要关注系统机制的读者，应在阅读章 4前学习 BM25、稠密与混合检索、排序指标。准备开展评估的读者，还应在章 6前学习抽样、自助法与重复测量。解释干预结论的读者，应在章 7前完成潜在结果和广义线性混合模型单元。本附录不报告外部研究的效应量；以下数值全部为教学构造。

衔接单元	核心问题	在正文中的主要用途
BM25	如何由精确词项证据形成可检查的词法排序?	第 3–4 章的获取与检索基线
稠密与混合检索	如何组合语义向量和异质排序，并明确各自的尺度?	第 4 章的候选构建与融合
排序指标	在已声明的标签下，前 k 位排序找回了什么?	第 4、6 章的检索与基准评估
抽样与自助法	对哪个总体求平均，又对什么单位重抽样?	第 6 章的问题面板、分母与不确定性
重复测量	哪些记录共享问题、系统或时间状态?	第 6–7 章的纵向测量
潜在结果与目标估计量	干预试图识别什么因果对比?	第 7 章的实验与时间设计
GLMM 直观理解	如何建模非高斯聚类结果，同时避免将各行误当作独立记录?	第 6–7 章的敏感性与异质性分析

F.2 BM25：可检查的词法基线

为什么需要这一知识

词法检索考察查询词、变体和局部限定条件能否在渲染、切分与索引后保留下来。强大的语义模型可能掩盖表示环节的缺陷；可检查的词法基线则能使精确词项证据和文档长度效应显现出来。因此，章 4将 BM25 用作诊断基线，并不主张所有生成式搜索界面都采用 BM25[37]。

定义

设语料库 $\mathcal{D}$ 包含 N 篇文档。对于词项 t 和文档 d，以 $f(t,d)$ 表示词频，$|d|$ 表示文档长度，$\overline{|d|}$ 表示平均文档长度，n_t 表示包含词项 t 的文档数。BM25 的一种常见约定为

$$\mathrm{IDF}(t) = \log\left(1 + \frac{N - n_t + 0.5}{n_t + 0.5}\right), \tag{80}$$

$$\mathrm{BM25}(q,d) = \sum_{t \in q} \mathrm{IDF}(t) \frac{f(t,d)(k_1+1)}{f(t,d) + k_1\left(1 - b + b|d|/\overline{|d|}\right)}. \tag{81}$$

其中，$k_1 > 0$ 控制词频饱和，$b \in [0,1]$ 控制长度归一化。不同实现的分词、字段加权、重复查询词、停用词处理和 IDF 约定可能不同，均应记入实验清单。

该分数是在一个查询和一个固定索引下进行排序的工具，并不是文档相关、真实、被引证或被生成器使用的概率。

小型算例

考虑一个合成语料库，$N = 3$，平均长度 $\overline{|d|} = 100$，查询仅含一个词项，且该词项出现在两篇文档中。按上述约定，

$$\mathrm{IDF}(t) = \log(1 + 1.5/2.5) \approx 0.470.$$

取 $k_1 = 1.2$、$b = 0.75$。文档 A 长度为 100，该词项出现三次；文档 B 长度为 50，出现一次。两篇文档的单词项分数为

$$\mathrm{BM25}(t,A) \approx 0.470 \times \frac{3(2.2)}{3 + 1.2(0.25 + 0.75)} \approx 0.738,$$

$$\mathrm{BM25}(t,B) \approx 0.470 \times \frac{1(2.2)}{1 + 1.2(0.25 + 0.75 \times 0.5)} \approx 0.591.$$

这一计算解释了为什么在这组确切的约定与参数下，A 排在 B 之前。它不能证明 A 是更好的证据、这些参数最优，或该排序能够迁移到其他查询或语料库。

常见错误

一种误用是将 BM25 分数理解为绝对度量，直接跨查询或索引比较。语料库词频、分词、长度分布、字段和实现约定都会改变分数尺度。另一种错误是通过重复词项提高 $f(t,d)$，再将分数变化称为证据改善。该公式不包含真实性、溯源、可读性或下游使用情况。

检查题

固定 N、文档长度及其他所有参数，当 n_t 增加时，式 (81)中的 IDF 项如何变化？为什么这是一项关于语料库的陈述，而不是关于该词项现实重要性的判断？

答案提示。按此约定，词项出现在越多语料文档中，IDF 越低。它衡量的是固定索引约定下的稀有程度，而非社会、科学或商业重要性。

F.3 稠密检索与混合检索

为什么需要这一知识

精确措辞往往不足以处理释义表达、多语言表达或概念相关的段落。稠密检索提供学习得到的语义表示，混合检索则组合互补的词法与稠密证据。组合过程必须能够复查：处于不同尺度的两列分数，不会因为相加就自动变得可比。

定义

双编码器将查询与候选表示映射为向量 [22]：

$$\mathbf{h}_q = f_q(q), \qquad \mathbf{h}_d = f_d(d).$$

一种常用分数为余弦相似度：

$$s_{\text{dense}}(q,d) = \frac{\mathbf{h}_q^\top \mathbf{h}_d}{\|\mathbf{h}_q\|_2 \|\mathbf{h}_d\|_2}. \tag{82}$$

编码器、检查点、池化规则、文本截断和向量归一化共同构成检索器的身份。相似性不等于蕴含关系，也不能确立信源权威性。

混合检索可以采用经过校准或归一化的加权分数、学习得到的融合函数，或排名融合。倒数排名融合（RRF）组合多个排序列表，而无需假设原始分数处于同一尺度：

$$s_{\text{RRF}}(d) = \sum_{j=1}^{J} \frac{\mathbb{1}\{d \in L_j\}}{\kappa + \text{rank}_{L_j}(d)}. \tag{83}$$

常数 κ 与输入列表深度均属方案约定 [10]。RRF 奖励反复出现在高位的候选，但不会让重复或近重复查询产生独立的证据支持。

小型算例

在一个二维合成示例中，归一化查询向量为 $\mathbf{h}_q = (1, 0)$。候选向量 $\mathbf{h}_A = (0.8, 0.6)$ 与 $\mathbf{h}_B = (0.6, 0.8)$ 也已归一化。由(82)得到相似度 0.8 和 0.6，因此稠密检索部分将 A 排在 B 之前。这些数值只描述人为构造的编码器输出的几何关系，与事实支持无关。

再设词法列表为 (A, B)，稠密列表为 (C, A, B)。取 $\kappa = 10$，RRF 得到

$$\begin{aligned} s_{\text{RRF}}(A) &= 1/11 + 1/12 \approx 0.174, \\ s_{\text{RRF}}(B) &= 1/12 + 1/13 \approx 0.160, \\ s_{\text{RRF}}(C) &= 1/11 \approx 0.091. \end{aligned}$$

融合后的顺序为 A, B, C。A 在两个列表中都靠前，因此仍居首位。这种重复体现检索结果的一致性，不代表证据信源彼此独立。

常见错误

最常见的混合检索错误，是未经归一化、校准或学习规则便计算 $s_{\text{lex}} + s_{\text{dense}}$。数值尺度较大的分量可能主导结果。另一种错误是将嵌入空间中的接近程度当成主张支持标签；两段文字可以讨论同一主题，却彼此矛盾。

检查题

两个排序器分别生成 $L_1 = (A, B, C)$ 和 $L_2 = (C, A, B)$。按式 (83)取 $\kappa = 10$，哪个候选排在首位？将 A 的两次出现解释为独立支持之前，还需要什么记录？

答案提示。 A 排在首位，因为 $1/11 + 1/12$ 大于 B 和 C 的对应和。判断独立性需要追踪信源起点及派生关系；仅有排序列表一致性还不够。

F.4 排序指标：找回证据不等于回答质量

为什么需要这一知识

检索与重排序需要符合相应环节的指标。前 k 位指标可以反映是否找回了已标注证据、排序是否合理，却不能说明生成器是否忠实使用了证据。截断位置、判断单位、相关性尺度和分母共同定义指标。

定义

设 $R_k(q)$ 为查询 q 按顺序排列的前 k 个候选，$\mathrm{Rel}(q)$ 为预注册的相关集合。二元召回率为

$$\mathrm{Recall}\,@k(q) = \frac{|R_k(q) \cap \mathrm{Rel}(q)|}{|\,\mathrm{Rel}(q)|}. \tag{84}$$

前 k 位精确率用同一分子除以 k。倒数排名为第一个相关候选排名的倒数；在明确规定未找到相关结果的处理方式后，也可记为零。平均倒数排名（MRR）对目标查询面板中的该数值取平均。

若排名 i 的分级相关性为 $g_i \geq 0$[20]，则

$$\mathrm{DCG}\,@k = \sum_{i=1}^{k} \frac{2^{g_i} - 1}{\log_2(i+1)}, \tag{85}$$

$$\mathrm{nDCG}\,@k = \frac{\mathrm{DCG}\,@k}{\mathrm{IDCG}\,@k}, \tag{86}$$

其中，IDCG 是相同标签下理想排序的 DCG。IDCG 为零时必须有明确处理规则。对于二元标签，平均精确率（AP）在相关项目出现的排名处计算精确率，并按已声明的相关项目分母求平均。

文档相关性、段落相关性、关键证据找回、信源多样性、引证支持和回答正确性属于不同的标签空间。

小型算例

假设前三个合成候选的分级标签为 $(3, 0, 2)$，两个非零候选恰好构成全部相关集合，则

$$\mathrm{Recall}\,@3 = 1, \qquad \mathrm{MRR} = 1,$$

且

$$\begin{aligned} \mathrm{DCG}\,@3 &= 7 + 0 + 3/\log_2(4) = 8.5, \\ \mathrm{IDCG}\,@3 &= 7 + 3/\log_2(3) \approx 8.893, \\ \mathrm{nDCG}\,@3 &\approx 0.956. \end{aligned}$$

按二元相关性，相关项目位于第 1 和第 3 位，因此

$$\mathrm{AP}\,@3 = \frac{1 + (2/3)}{2} \approx 0.833.$$

这些指标数值不同，是因为它们奖励排序的不同方面。没有一个指标评价后续回答是否引证或保留了这两项内容。

常见错误

如果只报告“nDCG 提升”，却不提供相关性指南、截断位置、查询面板、理想排序约定与不确定性，该数值就无法解释。另一种错误是将文档级 nDCG 与主张级证据召回率作比较，仿佛两者分母相同。相关文档中的关键段落仍可能在切分或上下文打包时丢失。

检查题

对于等级为 $(0, 2, 1)$ 的三项目排序，第一个相关项目的倒数排名是多少？如果第二个项目在最终回答中被用来支撑错误的主张，这个数值会改变吗？

答案提示。倒数排名为 $1/2$，且不会改变，因为排序标签与回答支持标签属于不同环节；后者需要单独进行引证或主张支持审查。

F.5 抽样、加权与自助法

为什么需要这一知识

只有说明提示词列表如何构建、如何代表目标需求空间，研究才能用这份列表界定总体。事实型与比较型查询数量相同，可以用于诊断，却可能严重偏离实际使用比例。只有重抽样单位与抽样和分配机制相匹配，自助法得到的不确定性才有意义。

定义

将目标查询总体划分为 $h = 1, \dots, H$ 个层，其目标权重为 W_h，满足 $W_h \geq 0$ 且 $\sum_h W_h = 1$。若 $\widehat{\mu}_h$ 为层内估计值，则分层目标估计为

$$\widehat{\mu} = \sum_{h=1}^{H} W_h \widehat{\mu}_h. \tag{87}$$

权重可以来自抽样设计、事先声明的决策总体或估计的使用分布；这些不同来源定义不同的目标估计量。

非参数自助法反复有放回地抽取单位，再重新计算估计量。如果抽样单位是查询意图，而每个意图有多次回答，聚类自助法便重抽查询意图 ID，并保留其关联的全部记录。逐行自助法则将回答记录视为可交换单位。这些过程通常对应不同的不确定性模型。重复次数、区间构造方法、分层方式和随机种子均应记入清单 [14]。

小型算例

假设合成目标总体中，事实型任务占 80%，比较型任务占 20%。一个平衡诊断样本得到层内均值 $\widehat{\mu}_1 = 0.25$、$\widehat{\mu}_2 = 0.60$。按目标权重计算，估计值为

$$0.8(0.25) + 0.2(0.60) = 0.32,$$

而各层等权的诊断均值为

$$0.5(0.25) + 0.5(0.60) = 0.425.$$

两个计算结果没有先天的优劣之分；它们回答不同问题，因此必须相应标注。

再考虑聚类自助法：有四个查询意图，每个意图均带有全部重复记录。一次自助抽样可能抽中意图 ID $(2, 2, 4, 1)$。重新计算估计量时，应使用所选各簇关联的全部记录，包括第 2 簇的两份副本。这只是操作过程示意，不是在声称某个数据集的区间覆盖率。

常见错误

分析者常对数百行回答重抽样，而处理分配或抽样实际上发生在几十个查询意图上。结果可能只反映意图内部的变异，却漏掉哪些意图进入研究所带来的不确定性。另一种错误是看到某一层更有利于干预后，再倒推调整目标权重。

检查题

抽取 30 个查询意图，每个运行五次，目标是查询意图上的平均值。主要的非参数自助法应对什么单位重抽样？五次重复贡献了哪种变异信息?

答案提示。重抽 30 个意图簇，并保留每个被选意图的全部重复记录。重复运行描述意图内的随机变异，不会创造 150 个独立抽样的意图。

F.6　重复测量与依赖关系

为什么需要这一知识

生成式回答会随随机运行、查询、系统和时间而变化。重复同一提示词可以估计其中一个变异分量，却不会产生新的需求样本，也不能冻结平台。有效分析应区分单元格内部变异与问题间、系统间和时间上的变化。

定义

以 Y_{qmtr} 表示查询意图 q、系统 m、时间区组 t 和第 r 次重复的结果。一种描述性分解为

$$Y_{qmtr} = \mu + \alpha_q + \beta_m + \gamma_t + \xi_{qm} + \zeta_{mt} + \varepsilon_{qmtr}. \tag{88}$$

ξ_{qm} 和 ζ_{mt} 分别表示查询与系统、系统与时间的交互项。各项合起来区分查询、系统、时间、交互和运行层面的变异；这个式子不一定就是最终拟合的模型。

组内相关系数（ICC）描述给定方差模型下同一簇内的相似程度。当簇大小近似相等，为 $\bar{m}$ 时，设计效应诊断量

$$1 + (\bar{m} - 1)\rho$$

反映正向簇内相关如何降低增加记录所带来的独立信息增益。它是依赖特定假设的近似式，并非普遍适用的修正因子。

小型算例

两个合成查询意图各运行三次，二元结果为

$$A : (1, 1, 0), \qquad B : (0, 0, 0).$$

逐行均值为 $2/6 = 1/3$，查询等权均值也为 $[(2/3) + 0]/2 = 1/3$。点估计相同，并不意味着信息结构相同：这里只有两个抽样查询意图，每个意图内有三次测量。若按查询分配处理，分配单位只有两个，不是六个。仅有两个簇不足以支持稳定的处理效应分析；增加重复次数不能弥补独立分配查询过少的问题。

常见错误

将每个回答都当作独立记录，会夸大样本量，忽视共享的查询、系统或时间状态。相反的错误是立即对所有重复取平均，丢弃运行层面的变异与缺失状态。单元格档案应同时保留这两个层次，估计量再根据目标选择适当层次。

检查题

将 40 个查询意图分配给两个信源版本，每个意图运行五次。如果按意图分配，有多少个独立分配单位？增加第六次重复，能否补偿丢失十个查询意图？

答案提示。有 40 个分配单位。多一次重复可以更精细地刻画意图内变异，却不能替代独立分配的意图，也不能恢复同样的目标总体覆盖。

F.7 潜在结果与目标估计量

为什么需要这一知识

一项干预主张必须说明处理版本、单位、比较方式、目标总体、系统状态与假设，才算完整。潜在结果将因果对象与观察数据分开，并明确缺失的反事实。但仅靠符号表示，不能将观察性比较转变为因果结论。

定义

对于单位 i、处理分配 $D_i \in \{0,1\}$ 与带版本的系统状态 S_t，设两个潜在结果为 $Y_i(1,S_t)$ 和 $Y_i(0,S_t)$。观察结果为 [38]

$$Y_i = D_i\, Y_i(1,S_t) + (1-D_i)\, Y_i(0,S_t). \tag{89}$$

有限样本平均处理效应为

$$\tau(S_t) = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i(1,S_t) - Y_i(0,S_t)\right]. \tag{90}$$

目标估计量（estimand）是式 (90)中的目标，或另一个事先声明的对比；**估计量（estimator）**则是应用于观察结果的规则，例如随机分配下的均值差。识别需要对分配机制或研究设计作论证，核实处理是否实际实施，并说明干扰、缺失和系统状态可比性的假设。

处理指完整且带版本的信源状态，不只是“增加了一张表格”这样的标签。如果多个元素同时变化，且设计未将其分开，目标估计量就针对整组变化。

小型算例

仅为教学说明，假设能够看到四个合成单位的完整潜在结果表：

i	$Y_i(0)$	$Y_i(1)$	个体对比
1	0	1	1
2	0	0	0
3	1	1	0
4	0	1	1

有限样本 ATE 为 $(1+0+0+1)/4 = 0.5$。现在将单位 1、2 分配至处理组，单位 3、4 分配至对照组。观察结果为 $(1,0,1,0)$，因此处理组和对照组观察均值均为 0.5，均值差为零。一次分配不能揭示每个单位缺失的潜在结果；本次均值差恰好不等于有限样本 ATE。基于随机化的推理关注估计量在分配机制下的行为，而不是恢复每个个体的反事实。

常见错误

一种常见做法，是将前后差异标成 $\hat{\tau}$，却不定义在干预发生之后、仍维持基线信源状态时会出现什么结果。平台更新、索引传播、查询需求变化、竞争者行动和评判器漂移都可能是其他原因，因果符号不能将其消除。另一种错误

是报告平均效应，却隐瞒受保护群体或高风险查询层中的重要负效应。

检查题

为什么通常不能从一次实验运行计算个体效应 $Y_i(1, S_t) - Y_i(0, S_t)$? 将状态 S_t 下估计的效应迁移到 $S_{t'}$ 之前，还需满足什么条件?

答案提示。对同一单位，互斥的处理状态最多揭示一个潜在结果。向 $S_{t'}$ 迁移，需要有充分依据的稳定性或可迁移性论证，并有证据说明相关系统、总体与测量变化不会改变目标对比。

F.8 广义线性混合模型：有边界的直观理解

为什么需要这一知识

许多 GEO 结果是二元值、计数或有序标签, 观察又嵌套在查询、系统、信源或时间之内。广义线性混合模型（GLMM）能够表示非高斯结果，并允许簇特有的变异。它是敏感性分析或建模工具，不能替代充分抽样、随机分配或清晰的簇级汇总。

定义

广义线性模型通过连接函数，将结果均值与线性预测子关联；混合模型进一步引入簇的随机效应。对于二元结果，一种随机截距示例为 [4]

$$\operatorname{logit} \Pr(Y_{qmtr} = 1 \mid b_q, v_m) = \beta_0 + \beta_1 D_q + \beta_2 P_t + b_q + v_m, \tag{91}$$

例如，可作建模假设 $b_q \sim \mathcal{N}(0, \sigma_q^2)$、$v_m \sim \mathcal{N}(0, \sigma_m^2)$，其中 P_t 是明确规定的干预后时期指示变量。β_1 等固定效应描述模型下连接尺度上的系统性对比；随机效应描述模型所刻画的抽样簇间异质性。

对 logit 连接而言，$\exp(\beta_1)$ 是条件优势比，而不是概率差。给定个体或簇条件的对比，不一定等于总体边际对比。分析必须说明哪一种才是目标估计量，以及如何针对随机效应和目标权重对预测求平均。

对于计数或有序结果，可以选用其他 GLMM 分布族与连接函数，但结果取值范围、离散程度、零膨胀与连接函数均属于需要诊断的建模选择。复杂模型不会创造研究设计未曾采集的信息。

小型算例

在一个合成教学单元格中, 将所有随机效应和其他协变量设为零。假设基线概率为 0.20, 对应基线优势为 $0.20/0.80 = 0.25$。令 $\beta_1 = \log 2$，即条件优势比为 2。处理后的优势为 0.5，对应概率

$$\frac{0.5}{1 + 0.5} = \frac{1}{3} \approx 0.333.$$

因此，优势比为 2 不意味着概率增加 20 个百分点，也不意味着概率从 0.20 翻倍至 0.40。这只是连接函数的换算，不是估计得到的干预效应。

常见错误

常见错误有三种。第一，将对数优势系数说成概率变化；第二，以为随机截距模型可以修复混杂、受污染的分配或独立簇过少的问题；第三，省略收敛、完全分离、接近边界的方差估计，以及替代相关结构。只有同时报告设计、拟合与敏感性分析，模型输出才能作为证据使用。

检查题

在 logit 模型中，基线概率为 0.20、条件优势比为 2，其他项为零时，对应条件概率是多少？为什么它可能不同于总体平均处理概率？

答案提示。条件概率为 1/3。总体平均需要对查询、系统及其他随机效应的分布积分或求平均。逆 logit 函数是非线性的，因此将随机效应设为零再计算，通常不等于对总体求平均。

F.9 将衔接知识串联起来

完整研究应按明确顺序处理以下对象：

1. 固定信源表示、语料库、查询总体与系统状态；
2. 构建可检查的词法与稠密候选，再声明融合和重排序规则；
3. 使用符合相应环节的标签与截断位置，评估证据找回情况；
4. 在加权之前定义目标查询与系统分布；
5. 保留重复回答，并以独立抽样或分配单位进行重抽样或建模；
6. 定义干预的目标估计量、处理版本与假设；
7. 只有当 GLMM 的连接函数、依赖结构、诊断和目标解释，能够提供超出清晰汇总的信息时，才采用该模型。

问题	最低限度的方法对象	仍超出方法范围的主张
关键证据能否被找回?	固定语料库、段落标签、Recall@k 或证据召回率	信源被使用，或回答忠实
候选排序是否有用?	分级标签、nDCG / AP / MRR、截断位置、查询面板	进入上下文，或带来用户效用
某一比例能否描述目标需求?	抽样框、分层、目标权重、缺失处理规则	抽样框未覆盖的总体
估计有多大不确定性?	独立单位、重抽样或模型、假设、敏感性	目标估计量有效，或因果关系已识别
干预是否导致变化?	处理版本、分配或设计、潜在结果目标估计量、实际实施、干扰与漂移检查	迁移至未研究的系统状态或总体
GLMM 系数是否描述目标?	结果分布族、连接函数、固定与随机效应、诊断、边际化规则	修复薄弱或存在混杂的设计

适用边界与失效条件

这些衔接单元不支持任何普遍成立的数值结论。算例数值只对列出的合成输入有效。正文中的实证主张仍需有带版本的数据集或来源、明确的方案、不确定性、证据与许可记录，并满足各章特定的发布判定条件。

本章小结

BM25 与稠密检索在不同表示下描述候选证据；混合方法需要明确的融合规则。排序指标评价特定环节的已声明标签，而非最终回答。抽样与自助法决定估计代表哪个总体及哪种变异。重复测量保留依赖关系。潜在结果定义缺失的因果对比；GLMM 为聚类的非高斯结果提供一种模型，但不替代研究设计与证据。