九章格物 NineGrids

GENERATIVE ENGINE OPTIMIZATION · 信源研究

星阙实验室GEO研究中心 · 第 02 册

豆包生成式引擎优化（GEO）信源策略偏好分析

四组可观测的信号如何依次决定一次引用

出品：九章格物 · 星阙实验室

作者：李现龙

版本：V2.0

时间：2026 年 11 月

权威信源：基于火山引擎豆包搜索官方文档、豆包用户协议、ACL/EMNLP 学术论文、CNNIC 报告等公开资料编制

微信扫我加好友

扫码添加微信 · 获取 GEO 方案

© 2026 九章格物 · 星阙实验室 · ninegrids.cn

目录

核心摘要 3

偏好不是榜单，是四组信号各管一段 4

唯一写明的一层：站点权威度分四级 5

分级打给站点，不打给文章 7

配额决定竞争强度：二十条与五十条 9

内容形态偏好：能被切走的那一段 10

相关度评分：唯一公开的名次刻度 12

时效字段：时间是一道硬门槛 13

学术侧证据：可被核对的单句优先 14

协议划出的底线：引用不是担保 16

三种失效形态：内容对，却没被用上 17

同题竞争：相关度是相对排序 18

资源有限时的取舍：先改哪一段 19

四组信号的冲突与优先级 20

提问类型决定取源结构 21

垂类之间的偏好差异从何而来 22

偏好为什么会漂移 23

五种反驳与逐条回应 24

一次推演：把四组信号走一遍 26

哪些偏好还不能断言 27

附录 A　自查清单 28

附录 B　术语表 29

参考来源 29

摘 核心摘要

这份报告只回答一个问题：豆包在检索与生成环节，**更倾向于挑哪一类信源**。市面上的答案通常是一张「权重榜」，但凡是公开可查的部分里，没有任何一张这样的榜。能查到的是一组分散的、彼此独立的信号，它们分别在不同环节生效：站点权威度分级决定「站在第几档」，配额上限决定「要和谁抢名额」，内容切片决定「哪一段能被取走」，相关度评分决定「排在第几位」。

因此本报告不承诺「哪类站点一定被引」，而是把偏好拆成四组可观测信号，逐组给出公开口径、机制推论与可验证判据。全篇严格区分三层证据：**官方写明的事实**、**由事实推出的机制结论**、**没有公开依据因而不可断言的推断**。第 19 章把三层证据做成对照表，凡属第三层的，文中一律标注，不让它冒充事实。

三条主要结论

第一条，**唯一被官方写明的偏好，是站点级的权威度分级**。火山引擎把站点权威度分为非常权威、正常权威、一般权威、一般不权威四级，用途写明为主动搜索控制与搜索结果标注[1]。同一档内部不再细分——这意味着「正常权威」这一档里，同花顺、懂车帝、小荷健康、豆瓣、B站、什么值得买、澎湃与中小型企业官网被放在同一层[1]，档内谁胜出，取决于分级之外的其他信号。

第二条，**配额与形态是隐藏的门槛，比等级更早生效**。豆包搜索单次返回上限为 Global 版 20 条、Custom 版 50 条，平均时延分别为 1053 毫秒与 700 毫秒[2]；返回内容以「网页相关内容摘要」为主，摘要长度可配置[2]。一个站点哪怕在最高档，只要它的内容在摘要里无法自成一枚证据，也不会被取用——等级决定资格，形态决定取用。

第三条，**「被引用」与「被信任」是两件事**。豆包《用户协议》明确，搜索来源由第三方提供，平台对其真实性、合法性不作保证[3]；学术侧，ALCE 把引用质量拆成引用召回与引用精确两项，并测出当时最好的模型在 ELI5 上仍有约 50% 的陈述缺乏完整支撑[4]。这说明挂链是生成动作的结果，不是平台给出的信用背书。把被引当成被信任，会直接导致内容侧放松证据要求，反而削弱长期稳定性。

四组信号的生效顺序



图 0-1 四组信号的生效顺序（据第 02—07 章公开口径绘制，非实测）

本报告的写法

每一章保持同一节奏：先录公开口径并标编号，再给出由口径推出的机制结论，最后落到一条可以自己动手验证的判据。这样做的代价是行文偏慢，好处是任何一个判断都能被同行复算。凡遇到只能推断的地方，直接写「推断」并说明缺哪一项公开信息，不用形容词糊过去。

之所以坚持这个写法，是因为信源偏好这个题目上虚假的确定性太多。有人给出「官媒权重 3 倍于门户」的具体倍数，却指不到任何文档；有人断言「七天内权重最高」，却给不出观察样本。这些数字一旦进入决策，后面所有推断都会失真。研究如果建立在传说上，投入再多也只是在为传说付费。

01 偏好不是榜单，是四组信号各管一段

「豆包偏好什么信源」这个问题，最常见的答法是给一个排序：官媒优先，门户次之，企业官网靠后。这类排序在经验上并非全错，但它混淆了三样东西——**偏好**是引擎在同等条件下更倾向选谁；**权重**是这种倾向在打分公式里的具体系数；**结果**是某一次回答里实际挂了谁。三者可以完全脱钩：一个权重很高的站点，可能因为这一次没有对应内容而缺席；一个权重一般的站点，可能因为那句摘要恰好自证而被取用。

三件事必须分开，否则会得出反向结论

把三者混为一谈，最典型的误判是「我的站没被引，所以平台不信任我」。这个推论跳过了中间环节：没被引可能是不在候选池、可能是排在第 21 位之外、可能是摘要里没有一句能支撑答案。三种原因的补救动作完全不同——补准入、补相关性、补写法。用一个统一解释去覆盖三种病因，最后只能在所有方向上都试一遍，而每种都只试了一点。

概念	定义	可观测信号	是否公开
偏好	同等条件下更倾向选谁	分级档位、档内案例分布	部分公开（第 02 章）
权重	偏好折算成的打分系数	无对外接口	未公开
结果	某一次回答实际挂了谁	来源列表、答案脚注	可见，但带概率

表 1-1 偏好、权重、结果的分野（权重一栏依据「无公开接口」这一事实）

还有一层区别值得强调：偏好是**可以跨站点比较的定性判断**——「官方机构文档比转载稿更可能被选」，这句话在任何一次回答里都成立，不需要数字支撑。权重是**跨提问不变的定量系数**——这种说法必须有公式或阈值才能成立，而公开材料里没有。把定性判断包装成定量系数，是这一领域最常见的伪精确。

四组信号，各在一个环节生效

把公开文档里能查到的参数按环节归位，会得到一张比「权重榜」更可用的图：准入环节看站点段位，配额环节看单次条数上限，名次环节看相关度评分与发布时间，取用环节看切片与摘要能否自证。四组信号并不在同一个平面上竞争，它们像四道先后开启的门，前一道不通，后面的优势一律归零。



图 1-1 两个版本在配额与时延上的差别（数值取自官方产品简介[2]）

这张图里有一个容易被忽略的对应关系：**配额更大的版本，时延反而更低。**Custom 版返回 50 条，平均 700 毫秒；Global 版返回 20 条，平均 1053 毫秒。若按「条数越多越慢」的直觉，这组数字是反的。可核的解释只能停在文档写明的范围：两个版本是不同的调用形态，返回字段也不同——Custom 版额外返回网页正文与相关度评分[2]。至于内部为何这样配置，属未公开部分，本报告不作推测。

这组数字的用处不是记住它，而是理解它揭示的结构：**配额、时延、字段三者是被一起设计的。**调用方要更多字段，就要接受一种与默认形态不同的服务形态；反过来，默认形态的 20 条与 1053 毫秒，说明它更看重响应速度而非取源广度。对内容方，这意味着「被取用的机会」在两种形态下分布不同——面向公开产品的内容，实际竞争的是 20 条这一档的窄门。

四道门如何串联

把四组信号画成一条链，会看到一个对内容方有利也有害的结构。有利的一面是：后三道门都握在自己手里——写不写成可切句、给不给准确时点、留不留可核数据，全部可控。有害的一面是：第一道门（站点段位）由平台按站点整体判定，单篇文章无法临时改判，这意味着**站点层面的积累是慢变量，内容层面的写法是快变量**。快变量救不了慢变量的欠账，慢变量也替代不了快变量的当场表现。



图 1-2 四道门的串联关系与可控性（推断，依据各章公开字段）

串联结构还解释了一个常见的「努力无效」现象：有人在写法上投入很多，结果毫无变化。原因往往卡在第一道门——讲法再好，若站点处于「一般不权威」档，或在严格配置下被过滤，内容根本没进入评分环节。这时写法优化不是无效，而是**没有被执行到**。判断卡在哪一道门，比继续优化更重要。

为什么不存在一张完整的偏好榜

一个可核的事实是：官方公开的只有「分级」这一层，且分级只到站点，不到文章；公开的相关性信号只有 rerank 相关度评分一个字段[2]；公开的时效信号只有「网页内容发布时间」一个字段[2]。除此之外，排序公式、各信号的相对权重、配额如何在不同意图间分配，均无公开口径。任何声称掌握完整权重表的说法，都无法被第三方复算——这正是本报告把它归入「不可断言」的原因（详见第 19 章）。

反过来说，这份「不完整」恰恰给了内容方一个稳定的操作面：既然能改的是形态、时效、口径与可核性，就应当把资源集中在这四项上，而不是追逐每年都在变的所谓权重传闻。**可核的部分做完，剩下的才交给概率。**

一个可以立刻用的判断顺序

遇到「为什么没被引」时，按门的顺序自查，比按传闻自查有效得多：先看站点是否在可抓取、可分级的状态；再看同题下候选是否已经排满、自己是否在 20 或 50 条之外；再看发布时间是否满足问题的时效要求；最后看那段最可能被取用的摘要，能否脱离上下文独立成立。四步走完，多数「玄学问题」会变成一个具体的工程问题。

这个顺序还有副作用：它会主动排除掉一部分无效怀疑。例如当自查发现站点已进候选、时点也新，问题就只能落在写法上，再看同题竞争者的摘要形态即可定位。把自我诊断收敛到一个环节，比在四个环节同时加码更容易看到效果。

02 唯一写明的一层：站点权威度分四级

在所有公开材料里，唯一以「权威」为名、明文给出分级的，是火山引擎的站点权威度说明。这套分级把站点分成四档：非常权威、正常权威、一般权威、一般不权威[1]。它的用途也写得很清楚：一是在主动搜索中用于控制（Custom 版可指定权威等级），二是在搜索结果里作为标注字段返回[1]。这两条用途决定了它的性质——它是**可被调用方使用的参数**，而不是一份面向公众的排行榜。

四档的边界与用途

四档名称本身带方向性：前两档是「权威」的正向表述，第三档「一般权威」已经转向中性，第四档「一般不权威」则是明确的负向。这种命名方式暗示分级不是为了给站点排名次，而是为了在检索时做**准入切分**：调用方可以只要前两档，也可以全收。理解这一点很重要——分级的第一用途是过滤，第二用途才是排序。

档位（官方口径）	性质	对内容方的实际含义
非常权威	最高档	在只取高权威的调用配置里仍然入选；通常对应机构型、官方型来源
正常权威	正向档	主流可用档位；商业媒体、垂直平台、企业官网多落在此档
一般权威	中性档	可被返回，但在严格配置下可能被过滤；内容质量对结果影响更大
一般不权威	负向档	默认配置下也可能被降权或排除；属需要规避的站点形态

表 2-1 四档分级的性质与含义（档位名称取自官方说明[1]，含义栏为机制推论）

需要注意表 2-1 的一处边界：官方说明给出的是档位名称与用途，并未公布每一档的判定细则——哪些指标、多少阈值、多久复评一次，都没有公开。所以「我的站为什么在这一档」在公开层面无法回答，只能通过档内同类站点的分布去反推（这也是下一节要做的事）。

「正常权威」这一档里藏着什么

分级说明给出的示例站点，是理解这套分级最有价值的部分。被列为「正常权威」的包括同花顺、懂车帝、小荷健康、豆瓣、B站、什么值得买、澎湃，以及中小型企业官网[1]。把这八个名字放在一起看，会得到一个反直觉的结论：**它们的共同点不是「大」，而是「在一个具体领域里有稳定的结构化供给」。**

同花顺提供的是行情与财务数据，懂车帝提供车型库与参数，小荷健康提供疾病与药品条目，豆瓣提供条目化的书籍影视评分，B站提供创作者内容与分区结构，什么值得买提供商品与价格信息，澎湃提供原创新闻，中小型企业官网提供产品与服务的第一手说明。八类站点的形态完全不同，但都满足同一件事：**在某个具体问题上，它们能给出一个可以单独摘走的、带明确归属的事实。**

示例站点	供给形态	为何能被摘走
同花顺	行情、财务数据	数值带代码与时间戳，单条即可核
懂车帝	车型库与参数	参数按版本与工况分列，归属明确
小荷健康	疾病、药品条目	条目化定义，术语与适应症固定
豆瓣	条目评分与简介	评分与条目一一对应，不易误配
B站	分区化创作者内容	标题与分区提供强主题约束
什么值得买	商品与价格信息	型号、价格、渠道同现
澎湃	原创新闻	署名、时间、地点三要素齐全
中小型企业官网	产品与服务说明	第一手口径，可回溯到主体

表 2-2 正常权威档内八类站点的供给形态（站点名单取自官方示例[1]）

这张表还解释了一件常被误解的事：**中小型企业官网能进「正常权威」，不是因为它们体量大，而是因为它们是一手信息的持有者。**同一条产品参数，官网写的是原始口径，转载站写的是二次转述。分级把两者放在同一档，意味着在「谁说的一手信

息」这件事上，规模不构成门槛。这对内容方的启示很直接：官网不是「因为小所以吃亏」，而是「因为一手所以有位置」，前提是它真的把参数写清楚了。

顺着这条逻辑可以推出一个可操作判据：判断一个站点是否「有资格站在正向档」，不必看它的注册年限或流量，而看它**是否在某个具体问题上拥有独家的一手口径**。一家只做资讯搬运的企业站，即便版面精致、更新频繁，也在这套逻辑里没有独特位置；一家只有几十个产品页的制造企业，只要参数、型号、工况写得清楚，就具备一手属性。

两种常见误读

第一种误读是把分级当成「豆包 App 的排序权重」。可核的事实是，分级说明的用途写的是主动搜索控制与结果标注[1]，这两条都发生在检索侧；而 App 端的答案生成是否使用同一个分级、如何使用，没有公开说明。把检索侧的参数直接等同于生成侧的权重，是把两段链路压成一段。

第二种误读是把分级当成「一次判定、永久有效」。分级按站点判定，而站点会变：栏目改版、内容方向调整、原创比例下降、转载增多，都可能改变判定结果。公开材料没有给出复评周期，但「有判定就有复评」是任何分级体系的前提。**因此档位是状态，不是身份**——它是当前的结果，不是永久标签。这个理解差别很实际：前者会让人持续维护站点的供给结构，后者会让人拿到档位后就不再管理。

还有一处容易被误读：分级说明的示例站点是举例，不是穷举。名单里出现中小型企业官网，不等于「所有企业官网都在正向档」；名单里没有某个品类，也不等于该类站点被排除。把示例当分类全集，会得出「某类内容一定进不去」的错误结论——这类结论在公开材料里找不到任何依据。

档位与调用配置：两条过滤路径

分级的第二层含义来自用途。官方写法是「主动搜索控制」与「搜索结果标注」[1]，这两件事对应两条不同的过滤路径：一条是调用方主动指定只取某档，此时档位是硬阈值，档外内容连进入候选的机会都没有；另一条是档位作为标注随结果返回，由调用方自行决定如何使用，此时档位只影响后续判断，不影响召回本身。

两条路径对内容方的含义不同。硬阈值路径下，档位就是资格线，跨不过去谈什么都没有意义；标注路径下，档位只是众多信号之一，可切性、时点、相关度仍能起作用。由于无法判断具体调用走哪条路径，保守的做法是把档位当作资格线来对待——这样无论对方走哪条路径，至少不会因为这一项被直接排除。

这个保守做法还有一个副作用：它把「是否升档」从「优化选项」变成了「前置条件」。在预算有限时，这种判断会显著改变资源排期——先确认资格，再谈细节，而不是先做细节、最后才发现资格不够。

03 分级打给站点，不打给文章

上一章留下的问题是：既然分级只到站点，那么同一站点内部，谁在决定哪一页被取用？答案只能从公开字段里找。豆包搜索返回的字段里，除权威度之外，还有网页标题、相关内容摘要、内容发布时间，Custom 版另给网页正文与 rerank 相关度评分[2]。这四项都以「页」为单位，且都与站点段位无关。换句话说，**站点负责准入，页面负责胜负**。

一个站点内部的四种落差

把同一站点内常见的页面分成四类，可以看到段位相同、结果截然不同的原因。以下四类不是实测统计，而是把「字段能读到什么」翻译成页面形态的差异：

页面类型	在字段里呈现的样子	为什么结果不同
参数页 / 条目页	标题即对象名，摘要含型号与数值，发布时间明确	三个字段互相印证，单条即可支撑一个事实陈述
活动页 / 专题页	标题含营销词，摘要多为口号，时间常为活动期	摘要无实质数据，且时效窗口一过即失效
聚合列表页	标题泛化，摘要由多条摘要拼接，时间随列表刷新	拼接后主体模糊，难以判断某项信息属于谁
深度长文	标题可读，摘要截在铺垫段，结论不在摘要内	结论落在文中后段，摘要里看不到，等于没写

表 3-1 同站内四类页面的字段表现差异（机制推演，非实测统计）

四类里最值得单独说的是「聚合列表页」。它在人类阅读中是效率工具，把同类信息排成一屏；但在字段层面，它把多个主体的信息压进一段摘要，主体与数值的对应关系随之消失。模型拿到的是一段「谁都在、谁都不明确」的文本，既无法支撑具体陈述，也无法被判定为可靠。这就是为什么大量列表页在来源列表里长期缺席，而它们的站点段位并不低。

活动页与专题页的问题则不在信息量，而在信息结构：它们的标题往往承载营销语义（「限时」「焕新」「重磅」），摘要承载口号，而价格、型号、时间这些可核要素被压在图片或表单里。转成纯文本后，可核要素全部消失，剩下的只有不可核的情绪词。这类页面即使被取到，也很难支撑一个事实陈述。

站点段位是慢变量，页面写法是快变量

把两个变量分开计时，会得到一个清晰的操作节奏。站点段位按域名判定，复评周期不由内容方控制，改一次要等一个周期；页面写法可以在下一次发布时立即生效，且效果直接体现在摘要与标题里。这意味着任何一次优化，**先做能立刻生效的那一半**，再等慢变量兑现，顺序不能反。

反过来，也有一种常见的错误投入：把资源全部押在「抬段位」上，指望档次上去之后内容自然被引。这个链条在公开层面就断了一环——档次只解决准入，不解决取用。一个从「一般权威」升到「正常权威」的站点，如果页面依旧是活动页与聚合页的写法，结果不会有实质变化，只是从「不被考虑」变成「被考虑但不被取用」。

更细一层：快变量本身也有先后。标题改动成本最低、生效最快；摘要形态（首段写法）次之；正文结构重构成本最高，但影响最持久。实际排期应当是标题与首段先动，结构与栏目后动，而不是反过来先做一次大改版——大改版期间旧内容的字段表现会短暂混乱，反而放大不确定性。

域名结构在这里起了什么作用

一个常被问到的问题是：把内容放在主域名下，还是放在子域名、栏目路径下，是否影响判定？公开材料只给出「站点」这一级的分级单位，并未定义「站点」是主域还是子域。这一层属于未公开，因此只能给出保守判断：**在主域内以稳定栏目路径组织，是风险最低的做法**——它与「站点」这一单位最贴近，不会因拆分而把已有的段位稀释到多个子域。

这个判断的依据不是传闻，而是成本结构：分级按域名做，域名越多，需要单独积累判定结果的单元就越多。把内容集中在少数路径上，等于把有限的判定结果集中使用；把内容散在大量子域上，等于把同一份积累摊薄。前者是可推的策略，后者是可推的风险，两者都不涉及任何未公开的权重细节。

实践中的另一个常见动作是「多域名分发」，把同一篇内容同步发到自有站群。这在旧的分发逻辑里是扩大覆盖，在新结构里却可能反向：同一份内容出现在多个域名下，每个域名都要独立积累判定结果，而内容本身没有增量。真正有效的做法是把增量写在一个主域名上，其他渠道做引流与引用回指，而不是复制正文。

栏目层级与「可归属」

再往下一层，是栏目与页面命名。字段里最先被读到的是标题，标题若不携带对象名，后面的摘要再完整也无从归属。一个可操作的判据是：**把标题遮住站点名之后，它是否仍然指向唯一对象。**例如「2026 款某车型 625 Max 科技版 参数配置」这样的标题，脱离站点依然成立；「最新车型参数汇总」则必须依赖站点与栏目才能解释。前者在字段层面自带归属，后者把归属外包给了上下文，而上下文在切片时会消失。

这一条与段位无关，属于纯写法问题，也是本报告认为最值得优先处理的一项：它不需要等复评周期，不需要外部资源，只需要在发布前改一次标题与首段。

同站内容的分层策略

把上面几条合起来，一个站点内的内容其实应当分层。第一层是**可核层**：参数页、条目页、说明页，目标是每一页至少有一句能被独立取用；第二层是**解释层**：方法、机制、对比，目标是把可核层的事实讲透，为答案提供上下文；第三层是**流量层**：活动、专题、聚合，服务人类阅读与转化，但不承担被引任务。三层的写法要求完全不同，混在一层里做，必然两头不讨好。

分层的好处是资源可分配：可核层投入写法与更新，解释层投入结构，流量层不必为「被引」做额外改造。这样既不会为了被引而放弃人类可读性，也不会为了排版美观而牺牲可切性。这个判断不依赖任何未公开信息，只依赖一个公开事实：字段按页返回[2]，页面的角色由内容方自己定义。

04 配额决定竞争强度：二十条与五十条

第三个信号是配额。官方文档写明，豆包搜索单次返回上限为 Global 版 20 条、Custom 版 50 条[2]。这不是一个技术参数那么简单，它直接决定了同场竞争的对手数量：**上限越小，名次的价值越陡峭。**第 1 名与第 20 名在 20 条里都算入选，但在下游生成阶段，靠前的条目获得的使用机会更多；而在 50 条的配置里，第 20 名之后仍有位置，边际价值下降得更慢。

配额如何改变「谁值得争」

把配额与相关度评分放在一起看，会得到一个与直觉不同的结论。在 20 条的配置里，进入候选本身就是稀缺资源，争的是「进不进得去」；在 50 条的配置里，进入候选不再稀缺，争的是「排得够不够前」。两种配置下，内容方的投入重点不同：前者要保证在任何相关问法下都能被召回，后者要保证在同题竞争里名次靠前。

配置	上限	稀缺资源	投入重点
Global	20 条	候选名额	扩大可命中的问法面，先求进池
Custom	50 条	名次与字段完整度	提高单条相关度与可核性，争前排

表 4-1 两种配额下的稀缺资源差异（上限取自官方文档[2]，投入重点为推论）

需要标注的是，这组推论的前提是「下游生成阶段对前排更敏感」。这一条在公开材料里没有直接给出，属于合理推断而非事实。能确证的部分只有两个：一是上限数值[2]；二是 Custom 版额外返回 rerank 相关度评分与网页正文[2]。也就是说，配额与字段是公开的，下游如何加权不是。

即便只按公开部分推，也能得到一条稳固的判断：**配额越小，候选的平均质量越高，被选中的门槛就越高。**20 条里挤满了同题最强的内容，一条中等水平的内容几乎没有机会；50 条里长尾内容尚有位置，但与问题无关的内容同样多。两种配置下的策略因此不同：窄配额靠「极致相关」，宽配额靠「稳定相关 + 可核」。

时延预算与取源广度

官方给出的平均时延是 Global 1053 毫秒、Custom 700 毫秒[2]。这两个数字单看只是性能指标，放进整条链路里，它们约束的是「一次回答能给检索留多少时间」。时延预算越紧，单次能处理的候选就越受限；这与配额上限形成一对相互牵制的参数——配额是「最多取多少」，时延是「来不来得及取」。

对内容方的含义很实际：**能被快速判定为相关的内容，比需要多跳才能判定的内容更容易被取用。**所谓「多跳」，是指摘要里看不出与问题的对应关系，需要打开正文、跨段拼合才能确认。这类内容在时延预算紧张时，天然处于劣势。它不是被判为不可靠，而是被判为「不划算」——这恰恰是最难通过申诉改变的一种落选。

配额之外的隐性名额

还有一个常被忽略的层面：20 条与 50 条是单次请求的上限，而一次对话里可能发生多次检索。问题被拆成数个子问、或者用户在追问中补充了条件，都会触发新一轮取源。这意味着**总曝光不等于单次配额**，一个在首轮排到 30 名的站点，可能在追问轮次里进入前 20。

这解释了一种看似矛盾的现象：某些内容在长对话里偶尔出现，在单轮问答里几乎不见。原因不在内容质量波动，而在轮次结构——首轮竞争最激烈，追问轮次的候选面更窄，边缘内容反而有机会。理解这一点，可以避免用单次抽查的结果去否定整体判断。

轮次结构还有一层含义：**能承接追问的内容，比只能回答首问的内容更耐用。**首问通常是「某某是什么」「某某哪个好」，追问则是「为什么」「有没有例外」「在哪买」。前者靠条目与定义，后者靠细节与条件。把两类内容都准备好的站点，在整段对话里被反复取用的概率更高；只写首问答案的站点，往往在第二轮就消失。

配额视角下的长尾判断

长尾问法在窄配额下最难生存，因为它们对应的候选往往数量不多但质量参差，且相关度评分本身偏低。反过来，长尾问法在宽配额下反而有机会，因为竞争者少。这带来一个与常识相反的策略判断：**长尾内容不是「必争」，而是「看配置」。**在只取 20 条的形态里，把资源押在长尾上，产出比很低；在可扩展到 50 条、且调用方按评分取用的形态里，长尾内容才具备边际价值。

因此「要不要做长尾」这个问题，在没有配置信息的前提下无法回答。能给出的只是条件式判断：先确认自身内容主要面向哪种取源形态，再决定长尾投入的比重。这一条同样属于推断，其前提（下游对前排更敏感、宽配额更依赖评分）已在文中标注。

05 内容形态偏好：能被切走的那一段

第四组信号与内容形态有关，也是唯一完全由内容方掌控的一组。官方文档把返回内容描述为「网页相关内容摘要」，并说明摘要长度可配置；Custom 版另返回网页正文（txt / markdown）[2]。这两句话合起来，定义了下游能看到什么：**一段被截取的文本，或一份丢掉版式的纯文本。**加粗、底色、字号、图表、表格结构，在这个环节全部不保留。

摘要不是预览，是候选文本

把摘要当预览，是内容方最普遍的误判。预览服务于「要不要点开」，摘要服务于「能不能直接支撑一句话」。二者的长度逻辑相反：预览可以短、可以悬念；摘要必须能自成判断。**判断一段摘要是否合格，标准只有一条：把它单独贴出来，是否仍然成立。**

由此可以推出一条直白的写作规则：把最需要被引用的那句话，放在段落最前面，并把主体、版本、数值、单位、时点五要素压进同一句。这样即便摘要只截取开头，那句话也完整在内；反之，若结论落在段末，摘要截断后只剩铺垫，等于这段没写。

写法	摘要里剩下什么	后果
结论前置 + 要素齐全	一个完整判断	可被直接取用，且能被独立核验
背景铺垫 + 结论在末	一段无结论的叙述	摘要内无支撑点，等同于未提供证据
关键数据仅存于表格	失去行列关系的数字串	数字与归属脱钩，易被误配或弃用
代词回指与省略	主体不明的短句	无法判断陈述对象，直接失效

表 5-1 四种写法在摘要环节的残余信息（机制推演，非实测统计）

纯文本转换会抹掉什么

Custom 版返回网页正文以 txt 或 markdown 形式[2]，这意味着版式信息在交付前已被剥离。剥掉的东西包括：颜色与底色传达的强调、表格的行列对齐、图文并置形成的对应关系、折叠与展开所隐含的主次。留下的只有文本顺序与少量 markdown 标记。

这条事实推出一条反直觉的结论：**视觉上越「专业」的排版，在纯文本里损失越大。**一份满屏卡片、配色分区、图注繁多的页面，人读起来层次分明，转成线性文本后却变成一串彼此无关的短句。相反，一份以「一段一判断」组织的页面，转成纯文本后结构几乎不变。前者是给人看的专业感，后者是给机器读的可核性——两者并不冲突，但需要同时满足。

具体到常见的几种「专业排版」：数据卡片会把一个数字与一句形容词捆在一起，转成文本后形容词在前、数字在后，归属变得模糊；分栏布局会把并列信息压成左右两段，转成文本后顺序错乱；信息图把关键比较做成图形，转成文本后整段消失。三种做法都不是错，问题在于它们承担的是「人类阅读优化」，而取用环节要的是「文本可独立成立」。

一段可切文本的五个条件

1. **主体唯一**：句中出现的确切名称能指向唯一对象，不用「该产品」「本方案」代替。
2. **数值带单位与工况**：数字后面紧跟单位与测试条件，避免被跨口径替换。
3. **时点明确**：写清版本年份或发布日期，让读者能判断这条是否仍然有效。
4. **限定条件同句**：「仅在何种条件下成立」与结论写在同一句，避免结论被过度外推。
5. **不依赖图表**：把必须被引用的关键数值同时写进正文，表格只承担对照职能。

这五条不是写作偏好，而是把「摘要能否自证」拆成了可逐条检查的动作。逐条对照自己的页面，缺哪条补哪条，比笼统地说「内容要更好」有效得多。

同一事实的两种写法

用一条最常见的产品信息做对照。第一种写法是行业里通行的「介绍体」：「本产品采用新一代技术，在续航与效率方面均有显著提升，价格更具竞争力，具体参数请参见官方配置表。」这段话人读无碍，但它**没有一个可以被切走的句子**：没有产品名、没有版本、没有数值、没有条件，全部指向外部。

第二种写法把要素显式化：写清型号、工况、数值与指导价，并把年份写进同一句。改写之后篇幅更短，但每一句都能独立成立，且与官方口径一致。这里的关键不是文采，而是把隐含信息显式化——显式化恰恰是切片最需要的东西。

值得注意的是，这种改写并不牺牲可读性，它只是把「形容词承担的信息」换成「名词与数字承担的信息」。读者读到的东西反而更具体：知道是哪一款、什么工况、多少钱。所以「可切」与「好读」在多数情况下并不对立，对立的只是「可切」与「靠修辞撑起来的专业感」。

密度与覆盖的取舍

最后一个与形态有关的判断：一页内容的价值由「最强的那一句」决定，而不是由句子总数决定。一页写十件事、每件一句概括，切走后每句都缺要素；一页只写一件事、写透写全，至少有一句能被直接用。这不是主张少写，而是主张**把写作资源集中到能被取用的句子上**，其余内容作为支撑材料存在。

实践中的做法是先确定这一页必须被引用的那句话，围绕它补齐版本、数值、工况、时点，再展开解释。这样即便整页只有一句被采用，那一句也足以支撑一个完整判断；而十句泛泛而谈的内容，被采用十句也支撑不了任何结论。

摘要长度可配置意味着什么

官方说明里还有一句常被漏读：摘要长度可配置[2]。它意味着截断不是固定的，不同调用方可能取到不同长度的文本。对内容方，这带来两个判断。第一，不能假设「摘要只截一句」，也不能假设「摘要会截到结论」，唯一可靠的做法是让每一段的前两句就足以成立。第二，长度可变意味着同一页面在不同调用下表现可能不同，因此单次抽查的结果不能当作稳定结论。

由此可以给出一条比「结论前置」更细的规则：**把段落写成层层自足的结构**。第一句给结论，第二句给口径，第三句给限定，后面再展开。这样无论截到一句、两句还是三句，读到的都是一个完整判断，而不是越来越长的半截话。反过来，如果一段的前三句都在铺垫，那么在任何长度下它都不成立。

这条规则也解释了为什么「一段一判断」优于「一段多判断」。多判断的段落必然需要更长的篇幅才能把每个判断说完整，而在任何截断长度下，它都只能露出其中一部分，其余部分被切走。一段一判断则不同，段落在任何长度下都保持完整，只是详略不同。这不是排版偏好，而是对「长度可变」这一事实的直接应对。

06 相关度评分：唯一公开的名次刻度

在所有公开字段里，唯一以数值形式出现的排序信号，是 Custom 版返回的 rerank 相关度评分[2]。它的存在本身就是一个重要信息：**相关性是被显式打分并对外返回的**，而不只是一个内部排序结果。这意味着「相关」这件事在工程上有明确的刻度，内容方可以围绕这个刻度做优化，而不是猜测。

评分说明什么，不说明什么

「rerank」这个词在检索工程里指向一个固定动作：先粗召回一批候选，再用更强的模型对候选逐条重排。被返回的评分，是重排阶段的产物。由此可以推两点：其一，进入评分的前提是先被粗召回，没进粗召回的内容根本不会被评分；其二，评分的含义是「与当前查询的相关程度」，不包含权威度、时效、内容质量等其他维度。

信号	衡量对象	与评分的区别
rerank 相关度评分	候选与本次查询的匹配程度	逐次查询变化，可变、可优化
站点权威度	站点整体的可靠档位	按站点判定，短期内不变

信号	衡量对象	与评分的区别
内容发布时间	内容的新旧程度	只随日历变化，不随查询变化

表 6-1 三个公开信号各自衡量什么（字段依据官方文档[2]）

这张表说明了一件事：**三个信号不可互相替代。**相关度高的内容可能来自低档站点，高档站点的内容也可能与本次查询无关。下游在合成答案时，需要同时满足「相关」与「可靠」，任何单一信号独大都无法解释实际结果。这也是为什么「把段位刷上去」与「把相关度做上去」是两件必须分开做的事。

「粗召回—重排」的两段结构还带来一条实践含义：**能被粗召回，才有资格谈相关度。**粗召回通常依赖关键词、实体、时效等较粗的匹配，如果页面标题与正文里连问题中的核心词与实体都不出现，它在第一段就被淘汰，重排评分根本无从计算。所以「标题含实体 + 首段含问法」不是文案技巧，而是进入第二段的前提。

为什么相关度会被显式返回

一个合理的解释方向是：Custom 版面向的是需要自定义控制的使用者，返回评分是为了让调用方自己决定取用门槛——只取前若干条、或只取高于某个分值的条目。这在产品设计上说得通，也解释了为什么同一份文档里，Custom 版的时延更低却返回更多字段：它把一部分决策权交给了调用方。

对内容方的启示由此变得具体：**不要只盯着「是否被引」，要盯「在哪些问法下相关度会高」。**相关度是按查询计算的，因此内容与问法的匹配面决定了它能在多少种提问里进入前排。一个只回答单一问法的页面，相关度再高，覆盖也只有一个入口；一个围绕同一事实写出多种问法答案的页面，能在更多查询里获得高分。

相关度优化的三个动作

既然评分是按查询算的，优化就要从「查询侧」而不是「内容侧」倒推。第一步是列出这类内容真实会被怎么问，包括口语化问法与带条件的追问；第二步是把每个问法的关键词与实体写进标题或首段，让匹配有据可依；第三步是检查同题竞争——**相关度是相对排序，不是绝对分数**，同题下若已有更强的内容，自身分数再高也可能排在后面。

这三步里，第三步最容易被跳过。很多「优化后没效果」的案例，问题不在自身，而在同题竞争者的存在：同一句话有无数的页面在写，谁的自证更完整、时点更新，谁就排得更前。相关度优化因此不能只看自己的页面，必须看同题下的候选面。

还有一个容易被忽略的维度：**查询的长短会影响匹配结构。**短问法（两三个词）对应的候选面极宽，竞争靠实体覆盖；长问法（带条件、带限定）对应的候选面很窄，竞争靠条件匹配。同一篇内容，在短问法里可能与上千页竞争，在长问法里可能只需与几十页竞争。理解这一点，「做长问法」才成为一个有依据的选择，而不是凭感觉的偏门。

07 时效字段：时间是一道硬门槛

官方字段里有一个不显眼但很关键的项：网页内容发布时间[2]。它不像权威度那样带价值判断，也不像相关度那样带打分，但它的作用更绝对——**它决定一条内容是否还在有效期内。**对于价格、政策、参数、版本这类会随时间变化的事实，过期的内容不是「不够好」，而是「不再对」。

三类内容的时间敏感度不同

把所有内容一视同仁地追「最新」，是一种浪费。按时间敏感度分三类，投入节奏会清晰很多：第一类是**强时效**，价格、库存、政策口径、赛季数据，过了窗口即失效；第二类是**弱时效**，方法、机制、原理，三五年内仍然成立，但需要标注版本；第

三类是**无时效**，定义、术语、历史事实，时间不改变其正确性。

类型	典型内容	失效方式	维护动作
强时效	价格、政策、库存、版本参数	窗口一过即判错	按期更新并改发布时间
弱时效	方法、机制、流程	缓慢过时	标注版本，定期复核
无时效	定义、术语、史料	基本不失效	保持稳定，不必频繁改

表 7-1 三类内容的时间敏感度与维护节奏（分类为分析框架，非官方分类）

这张表的用处在于分配精力。强时效内容需要的是更新机制——谁负责、多久一次、改了之后是否同步改时间字段；弱时效内容需要的是版本标注；无时效内容需要的是稳定，不要为了「看起来在更新」而反复改动，反而破坏可核性。

更新时间不等于修改时间

一个常见的操作错误是把「文件修改时间」当成「内容更新时间」。前者由系统写入，后者应对应「这条事实何时被确认」。二者不一致时，会出现两种坏结果：一是内容实质未变，却被标成新时间，读者与引擎都会被误导；二是内容已经改对，时间却仍是旧的，导致正确内容被当成过期。

可操作的判据是：**时间字段应当与「事实变更」对齐，而不是与「排版变更」对齐。**改一个错别字不该刷新时间，改一个价格必须刷新时间。这个判据看似琐碎，却是时效信号能否被信任的前提——一个时间字段普遍失真的站点，其内容在时效判断上会整体吃亏。

还有第三种情况更隐蔽：同一事实在不同页面被写了多次，只更新了其中一页。结果是站内出现两个版本、两个时间，谁更新谁更旧无法判断。这类不一致在人类阅读里几乎不可见，但在字段层面直接表现为「同一主体存在冲突陈述」，两种版本的可信度都会被拉低。解决办法不是再补一版，而是**确定唯一权威页**，其余页面改为引用并链接到它。

时效与「长期有效」的张力

还有一个需要正面处理的矛盾：越是长期有效的原理性内容，越不容易被判定为「新」；越是追新的快讯，越容易被判定为「对但浅」。这不是二选一，而是两类内容承担不同职能——快讯负责在某一个窗口内被引用，原理负责在长周期里被反复引用。**把两类内容混在同一页，是两边都做不好的常见原因**：快讯的时效标签会拉低原理页的稳定感，原理的长段落会稀释快讯的可切性。

因此结构上应当分开：同一主题下，快讯单独成页、标注时间；原理单独成页、标注版本。两者互相链接，但不互相污染。这个安排不涉及任何未公开机制，只依赖一个公开事实——时间字段是按页返回的[2]，页面粒度就是时效判定的粒度。

分开还有一个副产品：人类读者也能受益。读原理的人不希望被活动信息打断，读快讯的人不希望翻到三段背景。把两类内容按职能分页，恰好同时满足两种阅读需求，这不是为机器做的妥协，而是内容组织本就应该有的纪律。

08 学术侧证据：可被核对的单句优先

前面六章的依据主要来自官方文档，能说明「机制长什么样」，但说不清「偏好有多强」。学术侧提供了另一类证据：在受控条件下，模型对什么样的源更愿意挂链、挂链后是否真的得到支撑。这类研究不针对某一家产品，但它的结论对理解偏好结构有效，因为它把「引用」拆成了可测的量。

引用召回与引用精确

ALCE 是目前引用评测里被引用最多的一套基准。它把引用质量拆成两个指标：**引用召回**指模型给出的每个陈述句是否至少有一个引用，且这些引用合起来能支撑该句；**引用精确**指被挂上去的每个引用是否真的与该句相关、能支撑该句[4]。两个指标对应两种失败：说了却没给证据，给了证据但证据不支持。

这两个指标的价值在于把「引用」从主观印象变成了句级判定。一句没挂链是遗漏，挂了一个不支撑的源是错配。两类问题的补救完全不同：前者要补证据形态，后者要补单句与单源的对应关系。混在一起谈「多挂几个链接」，等于用一个动作同时治两种病，通常两种都治不好。

> ALCE 的实测结论之一：即使当时最好的模型，在 ELI5 数据集上也有约 50% 的情况缺乏完整的引用支撑[4]。这说明「挂链」与「挂对链」之间存在稳定落差，且这个落差不是某一家引擎的偶发问题。

这条落差对信源偏好意味着什么

50% 这个数字最直接的含义是：**被挂链不等于被正确使用。**对内容方，它带来两个反向的推论。第一，被挂链是「进入了证据链」，不是「获得了信用」；第二，既然模型在判断「这个源是否支撑这句话」上容易出错，那么写得越明确、越自包含的句子，越容易被正确挂上——含混的表述不是更安全，而是更容易被错配或弃用。

这也解释了一种常见的错觉：某些内容被引用了，但引用的那句话并不是作者最想让人记住的那句。原因往往不是模型「读错了」，而是作者把关键判断写在了需要上下文才能成立的位置，模型只能拿它旁边那句自包含的话去挂链。**你被引用的，永远是你写得最像一枚独立证据的那一句。**

再往下推一层：如果一句话同时出现在多个来源里，模型更可能选择表述最清晰、归属最明确的那一个。这不是因为它「更权威」，而是因为在句级判定里，清晰的那个更容易被确认支撑。这条推论解释了为什么同一事实的不同表述版本，被引用的频率差别可能很大——差别不在站点，而在句式。

可见度可以被测，但不等于偏好被公开

另一条常被引用的学术结论来自 GEO 研究：在受控实验中，针对生成式引擎优化的内容策略可以使可见度最多提升约 40%[5]。这个数字常被当作「优化有效」的证据，但它有两个必须同时说明的限定：一是它来自特定基准与特定模型，跨领域、跨引擎的可迁移性有限；二是「可见度」在论文里有明确的操作定义，与「被引用」不完全等同。

把这两个限定说清楚，不是要否定这条结论，而是要防止它被当作保证书使用。40% 是上限值，不是平均值；是在实验条件下测得的相对提升，不是在任何站点上都能复现的承诺。**凡是把上限当均值、把实验当承诺的说法，都不属于可核证据。**

用户规模提供的是背景，不是偏好

CNNIC 的生成式人工智能应用报告显示，截至 2026 年 9 月，我国生成式 AI 用户规模已突破 7 亿，普及率超过 50%[6]。这个数字解释了「为什么值得研究信源偏好」——提问面足够大，答案的分发能力足够强，被引与不被引的差距会被放大成实际的流量差距。但它本身不提供任何关于偏好的信息，把它当作「偏好证据」是口径错位。

正确的位置是背景变量：用户规模决定了这件事的经济价值，学术实验决定了机制的方向，官方文档决定了可核的边界。三者各归其位，才不会出现「因为用户多，所以官媒一定优先」这类跨层推断。

这一章的三条证据还可以互相校验：官方文档说「返回摘要而非全文」，学术实验说「句级引用存在约一半的支撑缺口」，两者合起来推出的结论是一致的——**决定成败的是句子的可独立核验程度，而不是站点名气的绝对值。**一条来自机制，一条来自实测，方向相同，这才使得这条结论具备可采性。

09 协议划出的底线：引用不是担保

在所有材料里，最能约束「信任」这个词边界的，是产品协议。豆包《用户协议》写明：识别到联网搜索意图或用户主动选择 AI 搜索时，平台会以非人工检索方式自动搜索第三方网页链接的公开来源信息，回复内容上方会显示搜索来源和相关内容；同时明确，搜索来源中展示的网页链接内容由第三方提供，平台对其真实性、合法性不作保证[3]。

免责条款的机制含义

这段免责常被读成「平台在推卸责任」，但它在机制上有一个更准确的含义：**引用是生成动作的产物，不是核验动作的结论。**平台把「用到了哪个源」显示出来，是为了让用户能回溯，而不是为了给源盖章。理解了这一点，「被引」这件事的定位就清楚了——它是进入证据链，不是获得信用背书。

由此推出的第一条操作含义是：**不能因为被引就放松证据要求。**被引只证明这一次被用到，不证明内容正确；一旦别处出现更明确、更可核的表述，下一次答案就会转向那里。把被引当作免检凭证，等于把一次偶然当成长期资格。

> **一个常被忽略的反向风险**：被引次数越多，错误表述被放大的机会也越大。一条含混或有误的判断一旦被反复挂链，会在多个答案里复现，而平台并不承担核验责任[3]。这意味着内容方自身的核验强度，实际上决定了被引是资产还是负债。

「来源展示」是产品能力，不是评级

协议里还有一句话值得单独读：回复内容上方会显示搜索来源和相关内容[3]。这句描述的是界面行为——来源是被展示的，展示位置固定、形态固定。它不是评级，不带等级、不带分数、不带推荐语。把「出现在来源列表」理解为「被推荐」，是把展示当评价。

这个区分的实际价值在于：它让「曝光」与「采纳」两个指标可以分开统计。曝光只需要进入候选，采纳需要句子被取用。两者混算，会得到一个既虚高又无法解释的数字，也会让人误判自己的处境——明明只是进了列表，却以为观点被采用，于是继续在原方向加码。

三种状态还应当分开归档：**曝光**（进了来源列表）、**采纳**（某句被用作证据）、**担保**（平台为正确性负责）。前两者可观测，第三者按协议被明确排除[3]。把「担保」从目标清单里删掉，反而能让前两个目标的评估变干净——不再用「平台认不认我」这种无法证伪的问题消耗注意力。

协议约束下的三条自律线

沿着协议边界，可以划出三条自律线，它们不涉及平台规则的具体条款，只涉及内容方自己的责任范围：**其一**，凡进入正文的数值必须能回溯到一手来源；**其二**，凡属推断的表述必须显式标注，不用陈述句冒充事实；**其三**，凡涉及资质、疗效、价格的表述，以官方口径为准，不用二手转述替代原文。三条都不难执行，难的是在被引之后仍然执行。

之所以把自律放在协议章而不是方法章，是因为它是整个偏好结构的地基：如果内容本身不可核，那么无论档位多高、相关度多高、摘要多自证，被引的结果都只是把不可核的东西传播得更远。这一点在公开材料里没有明说，但从「平台不担保」这一事实出发，是唯一站得住的推论。

10 三种失效形态：内容对，却没被用上

前七章讲的是「怎样才能被选中」，这一章讲反面：同样满足准入与配额条件，为什么仍然有大量正确内容没有被用上。把观察到的落选逐一归档，会收敛到三种形态——**截断**、**错配**、**失配**。三者都发生在句子层面，而不是站点层面，所以补救动作也都在页面上，不在域名上。

截断：摘要切在最不该切的地方

第一种失效最普遍。摘要按长度截取，内容方无法指定截点，于是「结论在哪一句、截点落在哪一句」变成一场概率博弈。如果关键判断落在段落第三句之后，摘要截走前两句，留下的是背景、铺垫与过渡，整段在取用环节等于空白。

截断失效有一个容易误判的表象：**内容本身完整、逻辑清楚、读者能读懂。**正因为它对人不构成问题，内容方很难意识到它在机器侧已经失效。判断方法很简单：只读每段前两句，看能否得到这段的全部关键信息；如果得不到，这一段在切片场景里就不合格。

错配：数字与归属被拆开

第二种失效来自结构。数值被放进了表格、卡片、信息图，而正文只写「详见下表」。到了纯文本环节，行列关系消失，剩下「续航 700」「价格 26.98」这类孤立数字，与主体、工况、版本全部脱钩。此时模型只有两个选择：弃用，或按最近的上下文猜归属——后者正是错配的来源。

错配的危险不在「没被引」，而在「被引错」。一条被挂上错误归属的数字，会以「有来源支撑」的形态出现在答案里，比完全不被引用更难被读者质疑。**所以表格化的代价不是曝光下降，而是错误率上升。**这也解释了为什么第 5 章要把「关键数值同时写进正文」列为硬条件。

失配：答的不是被问的

第三种失效不在写法，而在题目。用户问「哪个更适合家用」，内容写的是一份完整参数表；用户问「为什么贵」，内容写的是价格清单。这类内容在任何维度上都不算差，但它回答的是另一个问题，因此在相关度评分上天然偏低。

失配失效最难自查，因为内容方通常按「我要讲什么」组织页面，而不是按「用户会怎么问」组织。补救的动作是把问法显式化：在页面里同时覆盖「是什么」「哪个好」「为什么」「什么条件不适用」四类问法，每类给一段可独立成立的话。这样同一页能承接多种查询，而不是只对一种问法有效。

形态	典型表现	在字段里的样子	补救动作
截断	结论在段末，摘要只留铺垫	摘要有叙述、无判断	结论前置，一段一判断
错配	数值只在表里，正文写「详见下表」	孤立数字，主体缺失	关键数值写进正文并带工况
失配	答的是「是什么」，问的是「哪个好」	相关度评分偏低	按四类问法分段覆盖

表 10-1 三种失效形态的判断信号与补救动作（推演，非实测统计）

三种形态的共同点

把三者放在一起看，会发现它们共用一个根源：**句子不能独立自证。**截断失效是「自证信息被切掉」，错配失效是「自证信息与主体分离」，失配失效是「这句话自证的是另一个问题」。三条路径不同，落点相同——都让那句话在被取用时无法被确认。

这也解释了为什么「把内容写得更全」通常不能解决问题。写得更全往往意味着段落更长、表格更多、信息更分散，三种失效反而同时加重。**正确的方向不是更多，而是更集中**：每一段只承担一个判断，判断所需的主体、数值、条件、时点全部同句出现。

还有一个派生结论值得记下：三种失效的修复成本差别很大。截断只需调整段落顺序，成本最低；错配需要改结构、把数据搬进正文，成本中等；失配需要重做选题与问法映射，成本最高。因此排期上应先修截断，再修错配，最后处理失配——这个顺序与修复难度一致，也与见效速度一致。

11 同题竞争：相关度是相对排序

相关度评分是按查询计算的，而同一个查询下通常有大量候选。这意味着任何一次「我相关度很高」的判断，都必须加上一个前提：在同题候选里排第几。把一个页面单独拿出来评估，得到的结论没有意义；把它放进同题候选里比较，才能看出真实位置。这一章讨论的正是「怎么比」。

同题候选的四个比较维度

比较不是比文章写得好不好，而是比在被取用环节上的四个条件。第一个是**时点**：同一事实有多个版本时，新版本优先。第二个是**可核性**：数值、单位、工况是否齐全，缺一项就低一档。第三个是**问法覆盖**：能否同时回答「是什么」「哪个好」「为什么」。第四个是**归属清晰**：脱离原文后，那句话是否仍能指向唯一对象。

维度	比较什么	常见劣势表现
时点	同一事实的版本新旧	版本号写着「现行」，内容却是两年前口径
可核性	数值与条件是否同句	只给结果不给工况，无法复算
问法覆盖	能承接几类提问	只写定义，遇到比较型提问即失效
归属清晰	句子能否独立指向对象	通篇用「该产品」「本方案」

表 11-1 同题竞争的四个比较维度（分析框架，非官方评分口径）

四个维度之间不是并列关系，而是有权重感的：时点与可核性属于「资格线」，不满足就出局；问法覆盖与归属清晰属于「排序线」，决定在前排还是后排。判断顺序因此应当是先过资格线，再比排序线——很多人反过来做，先纠结文笔与结构，却把过期的数值留在页面里。

「我比别人写得好」为什么不是有效判断

内容方评估自己时，常用的标准是完整度、专业度、可读性。这三项在人类评审里有效，在取用环节里并不直接对应任何一个字段。一个专业度极高的长文，可能因为结论埋在第六段而在切片环节全盘落空；一篇结构朴素的条目页，可能因为首句自证而稳定被取。

所以有效的比较必须落到字段上：**我的首句能否独立成立、我的数值是否带工况、我的时点是否最新、我的标题是否指向唯一对象。**四项都能对着看，才谈得上「比同题强」。这也是本报告反复强调「可核」的原因——可核不只是给读者的诚意，也是唯一能跨主体比较的尺度。

还有一个反直觉的推论：**同题竞争越激烈，差异化写法越不重要，标准化写法越重要。**因为竞争者众时，模型不需要靠文风区分优劣，只需要判断哪一句最容易被确认支撑。此时最好的策略不是写得与众不同，而是写得最容易被确认——主体在前、数值在后、条件同句、时点明确。

同题比较的一个实操口径

如果要做一次同题比较，可核的做法是：固定同一个问题，抓取该问题下的候选列表，逐条记录四件事——来源站点、内容发布时间、摘要首句、摘要里是否含可核数值。四列记完，位置关系自然显现：时点新的多在前，首句自证的多在前，含可核数值的多在前。这个口径不需要任何内部权重，也不需要推测公式，它只依赖公开可见的字段。

这个口径的价值还在于可重复。不同人、不同时间做同一件事，只要问题和字段一致，结论可以互相校对。相比之下，「我觉得我的内容更专业」这类判断无法被第二个人复核，也就无法被改进。

12 资源有限时的取舍：先改哪一段

前面各章给出了四组信号与三种失效，但现实问题是：预算与人力有限，不可能同时改。这一章讨论排序——在同样的投入下，先做哪一件事的边际收益最高。判断依据只有两条：**这件事能否立刻生效**，以及**它是否处在链路的前端。**

按「生效速度」排序

第 3 章已经把变量分成慢变量与快变量。放进取舍里，慢变量（站点段位、域名结构）需要等复评周期，投入再大也无法立刻看到结果；快变量（标题、首段、数值位置、时间字段）在下一篇发布时即可生效。资源有限时，先做快变量是唯一稳妥的选择，因为它能在下一个周期内产出可观测的反馈，而慢变量只能盲投。

这个排序还有一层现实考虑：快变量的改动本身会改善慢变量的判定基础。一个站点如果长期保持标题清晰、数值可核、时间准确，它在站点层面的供给质量判断上也会更稳。反过来，只等段位提升而页面依旧混乱，等于把资源押在一个自己无法观测的变量上。

三类站点的不同排期

同样是「先改写法」，不同起点的站点排期并不相同。已经在正向档的站点，重点是把可核层做扎实，同时压缩活动页占比；处在中性档的站点，重点是先用可核内容证明供给质量，再谈段位；处于负向档的站点，先解决的是站点形态本身——大量转载、无署名、无时间的内容会让任何写法优化都失效。

起点	首要动作	为什么是这个顺序
已在正向档	可核层写法与更新机制	准入已解决，胜负面在页面，投入直接见效
处在中性档	先做一批可核内容，再谈段位	段位判定缺依据时，先补供给质量的证据
处于负向档	先改站点形态：署名、时间、原创	形态问题不解决，写法优化的收益无法落地

表 12-1 三类起点的排期差异（推演，依据站点分级与页面字段的公开口径）

什么时候该停手

取舍的另一半是止损：哪些投入应当停止。第一类是「追逐未公开的权重」——无法复算的目标无法验收，投入等于赌博。第二类是「为被引而牺牲可读」——把页面写成关键词堆砌，短期或许提高匹配，长期会损害站点质量判断，而后者是慢变量，坏的影响持续时间更长。第三类是「无差别追新」——对所有内容都按月刷新时间，导致时间字段失信，反而削弱时效信号。

三类的共同特征是：**收益无法被观测，成本却确定发生。**判断一个动作该不该做，最省事的办法是问一句「三个月后我怎么知道它有没有用」。答不上来的，就不该列进排期。

这个标准还有一个副作用：它会把讨论从「什么方法更高级」拉回「什么结果可验收」。信源偏好这个题目天然容易滑向玄学，因为它涉及平台内部机制；而止损标准恰好是抵抗玄学的最实用工具——不能验收的，先不做。

13 四组信号的冲突与优先级

把四组信号分开讲完之后，必须回答一个合成问题：当它们互相冲突时，谁优先。现实里最常见的四种冲突是——高段位但内容不可切、低段位但内容极可切、内容很新但口径薄弱、内容扎实但版本过旧。四种冲突没有统一答案，但有一致的判断逻辑：**先看这道冲突发生在哪一道门，再看这道门是否已经通过。**

四种冲突的处置

冲突情形	实际处境	处置方向
高段位 + 不可切	准入已过，卡在取用环节	改写法；段位优势在此处无法兑现
低段位 + 极可切	写法已优，卡在准入或配额	先确认是否进候选；未进则写法收益为零
很新 + 口径薄弱	时效已过，卡在可核性	补工况与来源；新而无据的内容不构成证据
扎实 + 版本过旧	可核已过，卡在时效	更新版本号与数值；旧版本会被判为不再对

表 13-1 四种冲突的处境与处置（推演，依据各章公开字段）

这四种情形的共同结构是：**冲突的本质不是「谁更重要」，而是「哪一道门还没过」。**高段位不可切的问题不在段位，而在第四道门；低段位极可切的问题不在写法，而在第一道门。把冲突还原成「卡在哪」，处置就唯一了；还原不了，就只能在两个方向上来回加码。

结构优先于修辞

如果一定要给一条总的优先级，本报告的答案是：**结构优先于修辞。**结构指段落承担单一判断、数值与条件同句、时间与事实对齐、标题指向唯一对象；修辞指文采、比喻、气势。结构决定内容能否被确认，修辞只决定人读起来是否愉悦。在被取用这件事上，前者是必要条件，后者是加分项，而加分项无法弥补必要条件的缺失。

这条判断可以推出一个反常识的结论：**面向被引用的内容，写得「平」一点反而更稳。**平实的句子主体清楚、限制明确、不易被误读；华丽的句子依赖上下文、含多重修饰、切片后容易失真。这不是主张放弃表达，而是主张把表达放在结构之后——先把话说准，再考虑说得好。

还有一层需要说明：结构优先并不等于忽略站点。第一道门由平台判定，内容方只能通过长期供给质量去影响它。所以完整的优先级应当是两段式的：**先确认自己过了第一道门，再在后面的三道门上按结构优先。**跳过第一段的检查，直接优化写法，是

这份报告在开头就提醒过的典型错误。

冲突无法判定时的三条退路

现实里还有一种情况：无法判断自己卡在哪一道门。此时不宜硬猜，而应选择可逆的动作。第一条退路是**先做不依赖门判断的改动**——把段落拆成一判断一段、把数值与条件写进同一句、把版本号与日期补上。这类改动在任何一道门下都不会变坏，因此不必先判断门。

第二条退路是**把不确定性显式记录**。在内容旁边注明该页的适用条件与更新时间，比在内部会议里争论「到底卡在哪」更有用：它让下一次判断有据可依，也让读者知道该内容的有效边界。记录本身不改变机制，但把一次性判断变成了可累积的证据。

第三条退路是**用同题对照替代自我评估**。把同一问法下的候选摘要摆在旁边逐条比对：谁给了版本，谁给了工况，谁给了多主体并列。差异会直接暴露自己缺哪一项，而不必猜测平台内部如何加权。这是本报告认为最可靠的一种自查方式，因为它只依赖可读的公开结果，不依赖任何内部信息。

三条退路的共同点是：都指向可观测、可复算的动作，都不以掌握权重为前提。这也正是整份报告的立场——在机制只能部分公开的前提下，把能确定的部分做扎实，把不能确定的部分老实标注，比制造一套完整解释更有价值。

14 提问类型决定取源结构

前面各章把「信源偏好」当成一个整体来谈，但实际检索中，偏好随提问类型变化。同一个站点，在「某药适应症是什么」这类问题里位置很稳，在「预算二十万买哪款车」这类问题里可能毫无位置。原因不在站点变了，而在**问题需要的东西不同**：前者要一个条目，后者要一组可比的参数。这一章按提问类型拆开取源结构。

四类问法与它们各自缺什么

把常见提问归成四类，可以看到每一类对内容的要求并不重合。**事实型**问「是什么、多少、什么时候」，需要的是一条定义清楚的条目；**比较型**问「哪个好、差在哪」，需要多个主体在同一口径下的并列数据；**程序型**问「怎么办、流程几步」，需要完整可执行的步骤；**评价型**问「值不值、口碑如何」，需要多方证据与时点。

类型	典型问法	需要的内容形态	最常见的缺口
事实型	是什么、多少钱、什么时候	条目化定义，含版本与单位	定义有，但无时点与口径
比较型	哪个好、差在哪	多主体同口径并列数据	只有单主体页面，无法并列
程序型	怎么办、几步、要什么条件	完整步骤与前置条件	步骤缺条件，或依赖外部链接
评价型	值不值、口碑、踩坑	多方证据、时间分布、样本量	单一观点冒充普遍结论

表 14-1 四类提问与所需内容形态（分类为分析框架，非官方口径）

这张表最重要的是第四列。四类问法各自的缺口，恰好就是第 10 章讲过的三种失效的高发区：事实型缺时点，属版本失效；比较型缺并列，属结构失效；程序型缺条件，属自证失效；评价型缺样本，属证据失效。**所以「按提问类型组织内容」，本质上是提前规避四类失效。**

比较型为什么最难

四类之中，比较型的难度最高，因为它要求跨主体同口径。一个页面只讲自己的产品，无论写得多细，都无法回答「和谁比、差在哪」。而比较型内容一旦缺失，用户在别处提问时，答案只能从聚合型来源拼出来——这也解释了为什么比价、测评、参数对比类站点在消费决策链上占据位置：**它们提供的不是观点，而是可比性。**

对单一主体站点，可行的做法不是回避比较，而是**自带对照**：在同一页面内给出本品与同类在关键参数上的并列数值，并注明口径来源。这样同一个页面既能回答「是什么」，也能回答「差在哪」，覆盖两个类型而无需另一篇内容。

需要说明的是比对写法有一处边界：比较必须同口径。把不同测试条件下的数据并排放，形式上完成了比较，实质上是制造错配。可核的做法是把条件写在同一行里——工况、版本、测试标准——让读者能自己判断可比性。

问法覆盖与页面数量

由此产生一个组织问题：到底该写一篇覆盖四类问法的长文，还是写四篇各答一类？从取用环节看，答案偏向后者。原因是摘要按页截取，一篇长文里四类答案互相稀释，每类的首句都不够靠前；四篇各自成页，每篇的首句都在页面顶部，被截到时大概率就是那句答案。

但从阅读环节看，四篇会给用户造成跳转负担。折中的做法是「一主三附」：主页面覆盖最常被问的一类，其余三类各给一个独立小节，每节以一句自证的判断开头。这样既保证每类答案都有独立入口，又不至于把内容切得太碎。

15 垂类之间的偏好差异从何而来

官方给出的示例站点覆盖了多个垂类，这个名单本身就是一份线索：同花顺在金融，懂车帝在汽车，小荷健康在医疗，豆瓣在文化消费，B站在内容社区，什么值得买在消费决策，澎湃在新闻，企业官网在自有产品。八类站点分布在不同垂类，却同处一档，说明**分级的判定不按垂类给分，而按「该垂类中谁持有可归属的一手口径」给位。**

四种供给结构与它们对应的垂类

把这些站点按供给方式归类，会得到四种结构。**数据型**以结构化数值为主，如行情、财务、参数，优点是单条即可核，缺点是覆盖面窄；**条目型**以定义、适应症、评分等固定字段为主，优点是归属清晰，缺点是更新慢；**内容型**以文章、视频、评论为主，优点是覆盖问法多，缺点是可核性参差；**一手型**以厂商或机构的自述为主，优点是权威可回溯，缺点是视野单一。

结构	代表	优势	典型短板
数据型	同花顺、懂车帝参数	单条即可核，口径固定	只覆盖「是什么」「多少」
条目型	小荷健康、豆瓣条目	归属清晰，不易误配	更新滞后，缺比较
内容型	B站、澎湃	覆盖问法广，能答「为什么」	可核性依赖作者
一手型	企业官网	口径原始，可回溯主体	只讲自己，缺对照

表 15-1 四种供给结构与代表站点（站点取自官方示例[1]，归类为分析框架）

四种结构在垂类里的分布并不均匀。金融与汽车天然偏数据型，因为其核心事实就是数值；医疗偏条目型，因为定义与适应症必须固定；内容社区与新闻偏内容型，因为解释与现场无法结构化；制造与服务业偏一手型，因为产品参数只有厂商能定。**垂类偏好的差别，本质上是这四种结构的比例差别。**

垂类权威由谁定义

再往深一层：每个垂类里，「权威」的定义者不同。医疗的权威来自机构指南与药品说明，汽车来自厂商参数与法规认证，消费来自实测与比价，文化来自条目与评分。这意味着同一套内容策略在不同垂类里的效果不一样——在医疗垂类，把定义写准就是最强动作；在消费垂类，把价格与型号对齐才是；在汽车垂类，参数与工况必须同时给出。

把这条差异忽略掉，就会出现「拿一套写法打所有垂类」的做法：写医疗时强调体验与感受，写汽车时强调品牌故事，写消费时强调术语定义。方向都对，但都不是该垂类最缺的那一块。**正确的顺序是先识别该垂类最缺的结构，再补那一块。**

识别方法并不复杂：看该垂类里被当作权威的站点提供什么，自己缺什么。医疗看条目字段是否齐全，汽车看参数是否带工况，消费看型号与价格是否同现。缺的那一项，就是该垂类的准入条件。

跨垂类的迁移陷阱

还有一个常见的错误是从一个垂类直接迁移到另一个垂类。在金融垂类有效的「数据表格 + 时间戳」，搬到医疗垂类会变成一堆无上下文的数字；在内容社区有效的「长文叙事」，搬到汽车参数页会稀释可核性。原因在于四种结构的比例不同，而写法是为结构服务的。

可迁移的只有一条：**把该垂类最需要的那类事实，写成能独立成立的一句。**剩下的形式必须按垂类重建。这条判断不涉及任何平台内部机制，只依赖垂类自身的供给结构，因此可以作为跨垂类作业的通用底线。

内容型站点的上限由什么决定

四种结构里，内容型最容易被高估。它的优势是覆盖问法广，一篇文章可以同时涉及背景、对比、经验、评价；但它的短板也同样硬：可核性依赖作者，且切到摘要后往往只剩一句态度，不剩任何事实。这决定了内容型站点的上限不在产量，而在**是否有能力把观点落成可核的细节。**

具体来说，内容型内容要突破上限，需要具备三项能力：其一，把关键事实放进段首，而不是放在故事结尾；其二，给出可回溯的口径，例如测试对象、样本量、时间窗；其三，对同一结论标注适用条件，避免把局部经验说成普遍规律。三项缺一，文章就退回到「只有覆盖、没有支撑」的状态，在四类问法里只能吃到最浅的一层。

这也造成一个反直觉的结论：**内容型站点的竞争，最终比拼的是它有多像数据型。**谁的正文里带更多结构化事实，谁在被引用时就更有优势。反过来看，数据型站点的短板也清楚——它只覆盖「是什么、多少」，答不了「为什么」，而「为什么」类问法的用户停留时间更长、转化更靠后。两种结构各有位置，改进方向都是补对方的短板，而不是把自己做得更极端。

16 偏好为什么会漂移

信源偏好不是常量。三年前有效的写法，今天可能收益减半；今天有效的手法，明年可能被平台调整抵消。漂移不是异常，而是三方持续作用的结果——平台要控成本与合规，内容方要争曝光，用户要更准的答案。理解这三股力量，比记住任何一版「当前权重」都更有用。

平台侧：成本、时延与合规

平台侧的压力最直接地体现在公开参数上。时延是成本，配额是成本，字段多少也是成本。官方给出的两个版本差异——20 条对 50 条、1053 毫秒对 700 毫秒、是否返回正文与评分[2]——本身就是成本与能力的取舍结果。这类取舍一旦调整，内容方的机会分布也随之变化：配额收窄，长尾先出局；字段减少，依赖正文才能判断的内容先出局。

合规压力则以另一种方式影响偏好。协议明确平台对第三方内容不作真实性保证[3]，这意味着平台在结构上更愿意引用「自身可核、责任边界清楚」的来源，而不是需要平台代为背书的内容。凡是把结论写得斩钉截铁、却不给口径与出处的页面，在这种结构下天然劣势——不是因为它错，而是因为它把核验责任推给了平台。

内容方侧：优化趋同与拥挤效应

第二股力量来自内容方自己。当一种写法被证明有效，它会被迅速复制：同一批关键词、同一种标题结构、同一套参数表。结果是这类写法的边际收益下降——不是因为写法失效，而是因为同题候选中到处都是同类内容，区分度消失，最终仍要回到时点与可核性的硬比较。

拥挤效应有一个可观测的表现：**同一问法下的候选摘要越来越像。**当十条摘要都在说同一句话，模型只能选择最完整、最新、归属最清楚的那一条。此时继续在措辞上趋同，等于把自己放进一个只靠细节决定胜负的赛场——而细节（工况、版本、时点）恰恰是大多数复制者不愿写的部分。

所以对抗拥挤的方式不是写得更像，而是写得更具体。具体是无法被批量复制的：它要求作者知道这一款产品的这一项参数在什么条件下成立。这也解释了为什么在任何一个垂类里，最终稳定被引的都是那些有一手细节的页面。

用户侧：问法的迁移

第三股力量是问法本身在变。用户从搜关键词转向直接提问，再转向带条件的追问：从「某车怎么样」到「二十万预算、家用、充电不方便，选哪个」。问法越长、条件越多，候选面越窄，可核的要求越高——因为长问法里的每个条件都必须被逐条满足，泛泛而谈的内容一条都满足不了。

问法迁移带来的直接后果是：**内容的价值从「覆盖关键词」转向「覆盖条件」**。一篇能回答「二十万、家用、无充电条件」三个条件的页面，比十篇只写「某车介绍」的页面更有位置。这不是文案问题，而是内容结构问题——条件必须被显式写出来，才能被匹配到。

力量	作用方式	可观测的锚点
平台侧	调整配额、时延、返回字段	官方文档的版本与参数变化[2]
内容方侧	写法趋同，边际收益下降	同题候选摘要的相似度
用户侧	问法变长、条件变多	追问轮次与条件密度

表 16-1 三方力量与各自的可观测锚点（分析框架，非实测）

这张表给出的不是预测，而是观测清单。三方力量都在持续作用，但只有平台侧有公开锚点——文档的版本与参数是可读的，内容方与用户侧的变化只能靠自查与抽样。因此合理的做法是：把官方文档当作唯一的机制依据，把另外两侧当作需要持续观察的经验面，两者不混用。

这也回应了一个常见的焦虑：「今年的方法明年还有效吗」。从三方结构看，具体手法一定会衰减，但判断依据不会——**可核、可归属、时点准确、条件显式，这四项在任何一版机制下都是稀缺的。**手法会过时，标准不会。

17 五种反驳与逐条回应

一份研究如果要被同行使用，就必须先经得起反驳。这一章把关于本报告最常见的五种质疑摆出来，逐条给出回应，包括哪些质疑成立、哪些只是口径错位。这样做的目的是让结论的适用边界更清楚，而不是把结论说得更强。

反驳一：我的内容很好，为什么没被引

这条最常出现，也最容易被误答。回应的第一步不是辩解内容质量，而是把「好」拆成可核的项：站点是否过了准入、同题是否排满、时点是否满足、摘要首句是否自证。四项里任一不过，内容再好也无法进入取用环节。这不是平台不识货，而是链路结构决定的——四道门任一关闭，后面的优势都不会被执行到。

需要承认的边界是：本报告无法给出「为什么这一次选了 A 没选 B」的答案，因为该层的排序逻辑未公开。能给的只是排除法——把四项逐一验证，剩下的不确定性才会缩小到可接受范围。

反驳二：这套分级只是让大站垄断

这条质疑有部分成立，但结论过度。分级确实按站点给位，大站因长期供给稳定而占据优势；但官方示例里明确包含中小型企业官网[1]，说明规模不是准入的必要条件。真正决定档位的是「是否持有该垂类的一手口径」，而不是流量或注册年限。

也要承认：分级判定无公开细则，中小站点无法自查档位，只能通过观察同垂类站点的分布来间接判断。这构成本报告的一处能力边界——它能指出方向，不能替代平台的判定。

反驳三：学术论文测的不是豆包

这条完全成立，且必须在引用时明说。ALCE 与 GEO 研究的实验对象是特定模型与特定基准[4][5]，结论不能直接迁移为豆包的参数。本报告引用它们的唯一理由是**机制方向**：句级引用存在稳定落差、优化能提升可见度，这两条方向性结论与官方文档描述的「返回摘要而非全文」是相容的。数值一律不外推——50% 与 40% 都只在原论文条件下成立。

换句话说，学术证据在这里承担的是「为什么这件事重要」的角色，不是「具体权重是多少」的角色。把两者混用，才会出现「用论文数字给平台定权」的错误。

反驳四：没有权重表，所以优化无从下手

这条在逻辑上不成立。无法掌握完整公式，不等于无法做有效动作。四组信号里，后三组的多数变量都由内容方控制，且效果可在下一篇发布时观测。相反，追求完整权重表反而会导致投入方向错误——把资源押在无法验证的传闻上。

可以把它表述成一个成本判断：**在可核部分投入，收益可观测；在不可核部分投入，收益不可观测。**资源有限时，选择前者是理性选择，与是否掌握公式无关。

反驳五：被引越多越好

这条需要限定。被引本身是中性的，它放大的是内容的质量方向：正确、可核的内容被引，是资产；含混、有误的判断被引，是负债，而且会随引用次数扩散。协议明确平台不承担核验责任[3]，意味着这个放大过程没有外部刹车。

因此「被引越多越好」只在内容通过自检的前提下成立。这也回到第 9 章的三条自律线——它们不是道德要求，而是风险控制：被引强度越高，自检强度必须同步提高，否则曝光越大，反噬越快。

反驳	是否成立	回应要点
内容好却没被引	部分成立	按四道门排查；单次选择不可解释
分级只利大站	部分成立	示例含中小企业官网，规模非必要条件
论文测的不是豆包	成立	只取机制方向，不外推数值
无权重表无从优化	不成立	后三道门可控，收益可观测
被引越多越好	需限定	须以内容自检为前提，否则放大错误

表 17-1 五种反驳的成立情况与回应要点

18 一次推演：把四组信号走一遍

前面每一章都在单独讨论一组信号。真正使用时，四组信号必须串起来用，而串起来的第一步是先确定「这次要判断的是哪一类问题」。这一章用一条完整的推演链做示范，说明如何在不知道任何内部权重的前提下，把一次被引失败拆解到具体环节。推演是假设性的，不指涉任何真实站点，也不承诺结果。

假设与起点

设定一个常见情形：一家做家用净水设备的制造企业，官网只有三十来个页面，内容以产品介绍为主；企业希望在「家用净水器怎么选」这类问题上被答案引用，目前几乎没有出现。已知的条件只有四个：站点规模小、页面数量少、产品参数齐全、内容更新频率低。其余信息一概不知。

按第 1 章的门顺序，第一步不是改写法，而是判断卡在哪一道门。此时可核的只有一件事：站点处于什么档位，公开层面无法自查。因此推演的第一步只能是「减小不确定性」——观察同垂类被引站点的供给形态，判断自己的形态是否同类；如果形态相差很远，说明问题可能出在准入层，写法优化在短期内收益有限。

第二步：确认配额与同题密度

假设形态与同垂类接近，说明准入大概率已过。第二步看同题密度：在「家用净水器怎么选」这类问法下，候选里有测评站、比价站、导购站、厂商官网。候选面宽，意味着在 20 条上限里竞争激烈；此时企业官网的竞争位置通常靠后，因为单主体页面缺并列数据（第 14 章的比较型缺口）。

这一步的结论是一个明确判断：**问题不在可核性，而在可比性。**产品参数齐全并不构成优势，因为参数只覆盖了自己，而同题需要的是多主体并列。于是处置方向从「补参数」转为「补对照」——在同一页面内给出本品与同类关键参数的并列，并注明口径。

第三步：检查时效与共用口径

补对照之后仍然可能不出现，这时进入第三步：时效。净水设备的核心参数（滤芯寿命、通量、废水比）随型号迭代变化，若页面停留在旧型号，新版参数没有更新，内容就会在时效判断上吃亏。此时需要确认两件事：一是页面是否标注了型号年份，二是同一事实是否只存在一个权威页。若发现多个页面各写一个版本，先把内容收敛到一个权威页。

第四步是取用层的最后检查：那段最可能被引用的摘要，是否能独立成立。若首句是「随着健康意识提升，净水设备受到越来越多关注」，这一句在切片后不构成任何证据；若首句是「某型号在某水质条件下的通量为 N 加仑 / 小时，滤芯更换周期为 M 个月」，则单条即可支撑。

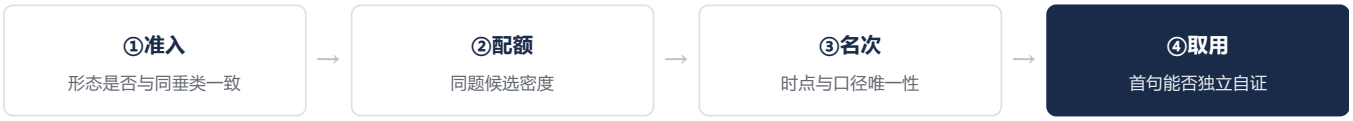

图 18-1 一次完整推演的四个检查点（示例为假设情形，非实测）

推演给出的结论

走完四步，得到的不是一个「提升权重」的方案，而是三件具体的事：把单主体页面补成自带对照的页面；把旧型号内容收敛到唯一权威页并更新时点；把首句从背景改为判断。三件事都不涉及任何未公开机制，都能在下一次发布时生效，也都能在之后的同题观察里验证。

这条推演链的价值不在结果，而在结构：**它把「为什么没被引」这个无法证伪的问题，换成了四个可以逐一排除的判断。**四步走完，如果仍然不出现，至少知道未解释的部分属于哪一层——而这部分恰恰是本报告在第 19 章明确列为不可断言的范围。

步骤	查看什么	若不过，说明什么	处置
①	供给形态与同垂类是否一致	可能卡在准入层	先补形态，写法收益有限
②	同题候选的密度与类型	可能缺可比性	补并列对照，附口径
③	型号年份与权威页唯一性	可能被判过期或冲突	收敛版本，更新时点
④	摘要首句能否自证	可能被截断失效	结论前置，要素同句

表 18-1 四个检查点的判读与处置（推演框架，非实测结论）

需要再强调一次边界：这条链只能排除，不能解释。它能告诉使用者「哪一环大概率有问题」，不能告诉使用者「平台内部给这一条打了多少分」。把排除法当成解释法，是这类研究最容易被误用的地方——排除之后的剩余不确定性，必须原样保留，不能用人造的权重把它填满。

19 哪些偏好还不能断言

写到这里，能核的部分已经给完，剩下的必须明说不能核。本章把全篇的判断按三层证据归档：官方写明的事实、由事实推出的机制结论、没有公开依据因而不可断言的推断。这张表的作用不是免责，而是让读者知道每一条结论的确信程度，从而决定自己该投入多少。

三层证据对照

层级	内容	依据与边界
事实	权威度分四档及档内示例站点；单次上限 20 / 50 条；平均时延 1053 / 700 毫秒；摘要长度可配置；返回字段含标题、摘要、发布时间、权威度、正文与相关度评分；引用不等于平台担保	均见于官方文档与协议原文[1][2][3]，可直接引用
推论	档内胜负取决于分级之外的信号；摘要能否自证决定取用；时间字段构成硬门槛；配额的稀缺性随上限变化	由公开字段与检索常识推出，逻辑链可复算，但强度未被官方量化
不可断言	各信号的具体权重与公式；配额在不同意图间的分配规则；分级判定的阈值与复评周期；单次回答为何选了 A 而未选 B	无公开接口、无官方口径，任何具体数字都不可核

表 19-1 三层证据的划分与边界（本报告全篇判断均按此归档）

三种常见越界说法

把这张表用起来，最直接的方式是识别越界说法。第一种是**数字越界**：给出「某类站点权重是另一类的 3 倍」这类精确倍数，却指不出任何公开出处。第二种是**周期越界**：断言「内容发布后 N 天内最容易被抓取」，把观察到的现象说成固定机制。第三种是**因果越界**：把「改成自包含写法后某次被引」说成「自包含必然提升引用率」，忽略同题竞争、时点、查询分布等混杂因素。

三种越界有一个共同结构：**把一次观察提升为一条规律。**避免它的办法也只有一个：说清样本、说清条件、说清对照。做不到这三点时，最诚实的写法是「在本次观察中如此」，而不是「机制如此」。

本报告没有回答的问题

为免读者误以为覆盖面完整，这里明确列出本报告未回答的问题：其一，档位之间的实际排序差距有多大——公开材料只给档位，不给分差；其二，配额在真实请求中是否总被用满，以及未用满时如何取舍；其三，同一站点内不同页面的取用概率分布，缺公开数据无法估计；其四，不同意图类型（导航型、比较型、事实型）是否走不同的取源策略，无公开说明。

这四项都属于「有口径就能算、没口径就只能猜」的部分。把它们列出来，不是示弱，而是把可投入的方向与不可投入的方向分开——在无法核验的地方投入资源，本质上是在为传闻付费。

最后一句判断

如果把全篇压缩成一句：**豆包更信任的不是「大站」，而是「能被判定为可靠、且能被单独摘走一句话」的站。**前者由平台按站点给，后者由内容方按页给。两者叠加，才构成一次被引用的机会；缺任何一半，另一半都无处落地。这也是本报告把「偏好」拆成四组信号的原因——不拆开，就只能得到一张无法验证的榜单；拆开之后，至少知道自己该改哪一段。

A 附录 A　自查清单

本清单用于把前文的判据落到可勾选的动作上。每条都只依赖公开信息，不依赖任何未公开的权重口径。

序号	检查项	通过标准
1	站点档位	站点处于正向档；未被严格配置过滤
2	标题归属	遮住站点名后，标题仍指向唯一对象
3	首段自证	首两句话贴出来独立成立，含主体、数值、单位、时点
4	表格与正文分工	必须被引用的关键数值同时出现在正文
5	时间字段	时间与事实变更对齐，同一事实只有一个权威页
6	表述落地标注	推断与事实分开写，限定条件与结论同句

表 A-1 六项自查与各自的通过标准

清单的使用顺序建议与第 1 章的门顺序一致：先查第 1 项（准入），再查 5、2、3、4、6（时效、归属、写法）。若第 1 项不过，后面五项先不必投入，因为内容不会被看到。

B 附录 B　术语表

术语	本报告中的含义
站点权威度	官方按站点给出的四级可靠性标注，用于检索控制与结果标注[1]
档内差异	同一权威档内不同页面之间的取用结果差异，由分级之外的信号造成
召回配额	单次请求可返回的候选条数上限，Global 20 条、Custom 50 条[2]
相关内容摘要	检索返回的网页片段，长度可配置，是下游看到的第一等材料[2]
正文纯文本	Custom 版返回的 txt / markdown 正文，版式结构已剥离[2]
rerank 相关度评分	重排阶段对候选与查询匹配程度给出的数值[2]
内容发布时间	检索返回的时间字段，用于时效判断[2]
引用召回	每个陈述句是否至少有一个引用并能获得支撑[4]
引用精确	被挂的引用是否真的与所述内容对应并支撑[4]
曝光	内容进入来源列表，仅证明进入候选
采纳	某一句被用作答案证据，需句子可切且判定可靠
担保	平台为内容正确性负责；按协议被明确排除[3]

表 B-1 本报告使用的十二个术语

术语表的用处是避免同名异义。例如「权重」在本报告中仅指打分的定量系数，不指档位高低；「引用」指句级支撑关系，不指来源列表中的出现。凡文中出现这两个词，均按本表口径理解。

参 参考来源

[1] 火山引擎：《站点权威度分级说明》，官方文档，2026-08-21。

[2] 火山引擎：《豆包搜索产品简介》，官方文档，2026-10-15。

[3] 豆包：《用户协议》第 2.2.2 条（联网搜索与来源展示、第三方内容免责），2026。

[4] Gao et al.：《Enabling Large Language Models to Generate Text with Citations》（ALCE），EMNLP 2023，arXiv:2305.14627。

[5] Aggarwal et al.：《GEO: Generative Engine Optimization》，KDD 2024，arXiv:2311.09735。

[6] 中国互联网络信息中心（CNNIC）：《生成式人工智能应用发展报告》，2026-09。

关于引用编号与使用边界

本报告正文中的方括号编号与上述来源一一对应，编号只标注「该事实出自何处」，不标注「该结论由谁背书」。凡属由公开口径推出的判断，正文已用「推断」「推演」「分析框架」等措辞显式标记；未标记的陈述句，均可在上述来源中找到对应事实。这一约定的目的是让读者能自己复算，而不是依赖作者的权威。

使用边界有三条。其一，本报告的机制结论以 2026 年 11 月的公开文档为准，官方文档若调整参数或字段，相应判断需要按新版本复核；其二，学术侧的两项结论（引用支撑落差与可见度提升上限）只在原论文的实验条件下成立，不可迁移为本平台的

定量预期；其三，本报告不提供任何具体权重数值，也不对单次引用结果作因果解释，凡要求这两类答案的场合，都不应引用本报告。

后续跟踪建议只保留一件事：关注官方文档的版本日期与参数变化。分级说明与产品简介是本报告全部机制判断的来源，两者中任何一处的字段增删、上限调整、时延变化，都会改变四组信号的结构。与其追逐传闻，不如把这两份文档的更新记录当作唯一稳定的观察锚点。

版本与修订说明

本册为星阙实验室 GEO 研究中心系列第 02 册，版本 V2.0，编制时间 2026 年 11 月。相较初版，本册的改动集中在三处：其一，把「信源偏好」重构为四组信号串联的四道门，不再按单一维度排列；其二，新增提问类型、垂类供给结构、偏好漂移三章，用于解释同一站点在不同问法下表现不同的原因；其三，把所有数值来源统一到官方文档的原始日期，凡无法回溯到原文的数字一律不写。

修订原则只有一条：**不改结论去迎合新版本，只在新版本出现时标注哪些判断需要复核。**若官方文档更新了配额、时延或字段，本册中依赖这些参数的段落会标注失效范围，而不是悄悄改写。这样处理的好处是读者能看清哪些结论随时间变化，哪些结论依赖结构、长期稳定。

反馈与勘误可通过封面二维码联系研究中心。凡指出事实错误或口径偏差的，会在下一版修订说明中列明；凡属推断层面的不同意见，欢迎并列呈现，不强行统一——这类分歧本身就是机制未完全公开的证据。

本册的定位是一份可复算的研究底稿，而不是操作手册。它不给动作清单，只给判断依据：哪些环节有公开口径、哪些只能推断、哪些至今无据。读者若按此结构自行复核，能得出的结论大概率与本册一致；若得不出，差异通常出在某一项公开口径的版本上，而不是在推理上。这也是本册愿意公开全部来源编号的原因。

最后一句留给使用者：不要把这四道门当成排行榜。它们是四个串联的检查点，价值在于让一次失败能被定位，而不是让一次成功能被复制。把可控的部分做完，把不可控的部分说清，这份报告的目的也就达到了。

建议引用格式：星阙实验室 GEO 研究中心．《豆包生成式引擎优化（GEO）信源策略偏好分析》（第 02 册，V2.0）．2026 年 11 月．文中方括号编号对应末页参考来源，推断性结论已在相应段落标注。

本册全部图表均为分析框架，用于组织判断，不作为实测统计；表格中的站点名称取自官方示例，分类方式为本研究中心拟定。凡需用于对外提案的场合，建议连同来源原文一并出示，以便对方自行核对口径。核对不一致时以来源原文为准，本册仅承担归纳与标注之责。

没有做的事：为什么本册不做平台实测

最后要交代一件事：本报告没有做平台实测。原因不是无法采样，而是采样在方法论上不成立。要得到可用的对比数据，至少需要固定一组问题、固定提问条件、重复多轮、记录候选集合与引用结果，还要控制账号状态、地区、时间与版本四个变量。任一变量失守，得到的差异都无法归因——这正是本报告拒绝给出「某类内容被引比例」的原因：样本不足时的比例，比没有数字更危险，因为它看起来像证据。

如果要做，正确的做法是登记完整的采样条件：问题清单、提问时间、客户端与版本、是否登录、检索返回的候选集合、以及最终引用落在哪一条。只有这些条件同时被记录，多轮结果才具备可比性。缺少其中任何一项，跨时间的对比就会失真，而失真会被误读成「平台改了机制」。

这也是本报告把公开文档当作唯一机制依据的理由：文档是可复核的、有版本日期的、可被任何人重新读取的；实测数据则是私有的、条件敏感的、难以复现的。**在没有能力控制变量的情况下，选择可复核的弱证据，优于选择不可复核的强证据。**这条取舍适用于所有以平台机制为对象的研究，不限于本册。