九章格物 NineGrids

GENERATIVE ENGINE OPTIMIZATION · 机制研究

星阙实验室GEO研究中心 · 第 01 册

从被检索到被挂链的豆包引用机制

一条内容从被检索到被挂链，要过四道判定；每一道都有可核的官方口径，也有至今未公开的部分

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/ACM 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 03

01 被引的不是网页，是答案里的那一句 04

02 第一道判定：能不能被切成一枚自包含证据 06

03 第二道判定：能不能进入召回配额 09

04 第三道判定：能不能被判定为可靠 12

05 第四道判定：能不能被安全写进答案 14

06 三种失效：截断、漂移与重写失配 17

07 引用可见度怎么量 19

08 问法决定取源 22

09 一次完整走查：从一条内容到一句被引的话 24

10 内容侧能改的与改不了的 25

11 边界：哪些能说、哪些不能说 27

附录 A　核验清单 28

附录 B　术语表 29

参考来源 30

摘 核心摘要

这份册子只回答一个问题：在豆包这类生成式引擎里，一条内容究竟是怎样一步步变成「答案里带链接的那句话」的。答案不是「内容质量好就会被引用」，而是一条可以拆开的判定链：**先被切片，再进配额，再被判定可靠，最后被安全地写进一句话。**四道判定各自失败，对应四种完全不同的补救动作——这也是本册把机制拆开写的原因：混在一起谈「优化」，最后只能得到一堆彼此矛盾的技巧。

三条主要结论

第一条，**引用的最小单位是句子，不是页面。**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，会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」[4]；学术侧，ALCE 把引用质量拆成 citation recall（这句话有没有被至少一个源支撑）与 citation precision（挂的那个源是否真的支撑这句话）[7]。被引用的从来不是一个站点，而是一句可以被单独拎出来核验的话。

第二条，**检索侧拿到的是切片与摘要，不是你的全文。**豆包搜索官方文档写明，返回内容包括「网页相关内容摘要」，并支持指定摘要长度；Custom 版还提供「网页正文（txt/markdown）」与 rerank 相关度评分[2]；单次返回上限为 Global 版 20 条、Custom 版 50 条[2]。这两句合起来意味着：一段话如果摘要里不存在、切片里不完整，模型没有任何机会引用它。**「写在页面里」和「进入候选文本」是两件事。**

第三条，**权威分级是搜索能力的参数，不是豆包答案里的公开权重。**火山引擎《站点权威度分级说明》给出的四级分级，用途明确写为「主动搜索控制」（Custom 版可指定权威等级）与「搜索结果标注」（返回字段带权威度）[1]。它是可核事实；把它直接等同于豆包 App 内部排序权重，是没有公开依据的推断——本册会把这类推断逐条标注，不让它冒充事实。

判定链的一张图



图 0-1 一条内容进入答案的五段路径（依据第 02—05 章的公开口径绘制）

本册的方法

全册采用同一种写法：先给公开口径（原文照录并标编号），再给由它推出的机制结论，最后给一条可以在自己站点上验证的判据。三者的边界始终分开——凡是官方未公开的部分，文中一律写「推断」或「未公开」；凡是有数字的地方，标出它出自哪份文档的哪一项能力。这样做的代价是节奏偏慢，好处是任何一个结论都能被同行复算。

之所以要坚持这个写法，是因为引用机制这个题目上虚假的确定性太多了。「加引用就能提升 40% 可见度」是真的，但那是 GEO-Bench 上的实验结果，且论文自己写明不同领域策略效果不同[8]；「7 天内抓取权重是 30 天以上 3.2 倍」这类数字找不到任何公开出处，只能算行业传说。研究如果建立在传说上，后面的所有推断都会失真。

边界

豆包没有公开完整的召回与排序权重，也没有公开「某条内容为什么没被引」的归因接口；引用本身带概率性，同一问题隔一段时间再问，候选集与挂链都可能变。所以本册给出的是一套**可复核的判定框架与可执行的观测方法**，不是「被引用的保证」。第 10 章会把已核事实、合理推断与不可断言三栏分开列清，并列出本册明确没有回答的问题。

01 被引的不是网页，是答案里的那一句

传统搜索的交付物是一个列表：你的页面排在第几位，用户点不点，是用户的事。生成式引擎的交付物是一段已经写完的答案，用户读完就走，链接只是答案的脚注。这个变化把竞争的单位从「页面」压缩到了「句子」：模型不需要你的整篇文章，它需要一句可以直接支撑结论的话，并且这句话要能被一个链接背书。

协议原文给出的引用位置

豆包《用户协议》第 2.2.2 条写得很直白：识别到联网搜索意图，或用户主动选择 AI 搜索功能时，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」，并「对你的输入做出回复」，「回复内容上方会显示搜索来源和相关内容」[4]。这段话里有两个信息点值得单独拎出来：其一，检索是自动触发的，不是用户逐条点选；其二，来源展示是产品层的固定动作，「引用」在豆包里是一种常态化的界面能力，而不是偶发行为。

自动触发这一条，直接改变了内容工作的重心。传统 SEO 时代，用户先有明确需求，才会去搜索框里输入一句话；而在对话式产品里，用户可能只是随口一问，产品就替他决定「去网上查一下」。这意味着**被检索的机会不再取决于用户是否主动搜索，而取决于他是否表达了一个需要外部信息才能回答的意图**。对内容方来说，这扩张了可争夺的提问面——很多过去不会变成搜索词的口语化问法，现在也会触发检索。

协议同时说明，搜索来源中所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[4]。这句免责表述反过来印证了一件事：**引用不等于背书**。挂上链接，只说明这段答案在生成时用到了这个源，并不说明平台核验过这个源的正确性。对做 GEO 的人，这一条很关键——被引用是「进入了答案的证据链」，不是「获得了平台的信用担保」。把这两件事混起来，就会得出「平台替我背书，所以我可以少写证据」这种反向结论，而它恰好会削弱被引的稳定性：一旦别处出现了更明确、更可核的表述，下一次答案就会转向那里。

> 原文摘录：「若识别到你有联网搜索意图或者你主动选择了AI搜索功能，我们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，进而对你的输入做出回复……回复内容上方会显示搜索来源和相关内容。」——豆包《用户协议》2.2.2[4]

学术侧把「引用」拆成两个可测的量

ALCE 是目前引用评测里被引用最多的一套基准[7]。它把引用质量拆成两个指标，定义都很朴素：**citation recall**——模型给出的每个陈述句，是否至少有一个引用存在，且这些引用合起来能支撑该句；**citation precision**——被挂上去的每个引用，是否真的与该句相关、能支撑该句。两个指标分别对应两种失败：没有引用（说了但没给证据），和引错源（给了证据但证据不支持）。

这两个指标的价值在于它们把「引用」从主观感受变成了**可复核的句级判定**：一句没挂链是遗漏，挂了一个不支撑的源是错配。两类问题在实操里的补救完全不同——前者要补证据形态，后者要补单句与单源的对应关系；混为一谈，就会得出「多挂几个链接就好」这种错误结论。更进一步，这两个指标都以「句」为判定单位，而不是以「文档」为判定单位，这一点与本册的主线一致：判定的颗粒度决定了优化的颗粒度。

ALCE 的实测结论同样值得记住：即使当时最好的模型，在 ELI5 数据集上也有 50% 的情况下缺乏完整的引用支撑[7]。这不是某一家引擎的短板，而是「检索—生成—挂链」这条链路的共性上限：模型要在有限上下文里合成多源信息，然后为每一句找到对得上的源，任何一环松动，句子就会变成无源之谈或挂错源。后续研究还发现，让模型在写之前先做引用规划，能明显改善引用质量[9]；也有工作把「引用是否真的支持」做成更严格的判别任务，发现模型在细粒度支撑判断上更容易出错[10]。这两条合起来说明：**挂链是一个独立于写作的环节**，不是写作的副产品。

三种引用关系，不能混算

从产品文档能确认的引用形态至少有三类，它们的证明力差别很大。第一类是**支撑型引用**：一句话挂一个源，源里有能对上这句的原句，这正是 ALCE precision 想要衡量的那种关系。第二类是**提及型**：答案里出现了你的品牌或产品，来源列表里也有你，但两者并不严格对应，可能只是「读到了」而非「采用了」。第三类是**实体链接型**：联网问答 Agent 的产品说明提到「支持在智能体回复中的关键名词输出百科词条链接」[5]，这类链接证明的是实体被认出来了，而不是观点被采纳。

引用关系	公开依据	证明了什么	没证明什么
支撑型	豆包协议展示搜索来源[4]；方舟知识库搜索要求正文内联 [1](URL)[6]	这句话的证据来自你	平台是否核验过内容正确性
提及型	来源列表展示[4]	你的内容进入了本次候选	你的哪句话被用了、用在何处
实体链接型	联网问答 Agent 关键名词输出百科词条链接[5]	你的实体被识别并命中词条	你的观点、数值被采纳

表 1-1 三种引用关系的证明力边界（依据均来自公开文档，不作实测推断）

把三类混在一起算「引用率」，会得到一个虚高且无法解释的数字。更麻烦的是，它会让人误判自己的处境：明明只有实体被认出来，却以为观点被采纳，于是继续在原方向加码，而真正的问题是句子从没被采用。第 07 章会把这三类分开定义指标，原因就在这里。

曝光、采纳、背书：三个常被混成一件的事

术语混用往往是错判的起点。把三种状态的差别列清楚，很多「为什么没效果」的困惑会自然消失：**曝光**是内容被检索到、出现在来源列表里，它只需要「进入候选」；**采纳**是某一句被用作答案的证据，它需要「句子能被切中且判定可靠」；**背书**是平台为内容正确性承担责任，这件事在协议里被明确排除了[4]。三者对应三种不同的门槛，也对应三种不同的投入。

状态	达成条件	可观测信号	投入方向
曝光	进候选池	来源列表出现自家域名	标题、首段、时效、准入（第 03 章）
采纳	句子被切中并被判可靠	答案句能对回原文原句	自包含写法、多源一致（第 02、04 章）
背书	平台承担责任	不存在此信号	不可追求，也不能承诺

表 1-2 三种状态的达成条件与可观测信号（背书一栏依协议免责条款[4]）

为什么句级会改变优化顺序

如果竞争单位是页面，优化顺序自然是「提升整站权重，再铺关键词」。如果竞争单位是句子，顺序就要反过来：**先确保存在一批「拎出来就能独立成立」的句子，再谈这些句子能不能被检索到。**原因很实在——模型在生成答案时，是在候选文本里找能支撑某个陈述的片段；一个必须连读三段才成立的论证，在切片与摘要环节就已经失去形状。

这条推断可以写成显式的三步：**前提**是检索侧交付的是字段化的切片与摘要[2]；**推理**是切片按位置截取，脱离上下文的句子若不自包含，就无法被判断为支撑某个陈述；**结论**是自包含句子在句级竞争里具有结构性优势，而这一优势与站点权重无关。三步之间没有跳跃，也没有依赖未公开的信息，这是本册愿意把它写成结论的原因。至于这个优势有多大、在不同行业是否一致，属于尚未量化的部分，需要第 07 章的观测方法去测。

本册后面四章的分章逻辑也由此确定：第 02 章谈「句子能不能被切出来」，第 03 章谈「切出来的东西能不能进入召回配额」，第 04 章谈「进去了能不能被判定为可靠」，第 05 章谈「判定可靠之后能不能被安全地写进答案」。四道判定串联，前一章不通，后一章的优化全部作废——这个「作废」不是修辞，是流程上的真实次序：一条进不了候选池的内容，再多权威度、再多引用格式，都不会被任何环节看到。

02 第一道判定：能不能被切成一枚自包含证据

第一道判定发生在检索侧，也是唯一一道几乎完全由内容方自己决定的判定：切成什么样，取决于你写成什么样。豆包搜索的返回结构里有几个字段与本册直接相关：网页标题、网页相关内容摘要（支持指定摘要长度）、网页内容发布时间、网页权威度；Custom 版还额外返回网页正文（txt/markdown）与 rerank 相关度评分[2]。

检索侧交付的是字段，不是文章

返回字段（官方口径）	形态	对内容方的含义
网页标题	短文本	参与早期相关性判断；标题与问题不对应，后面全白做
相关内容摘要	可配置长度的片段	下游看到的第一等材料；截在哪一句不由内容方决定
网页内容发布时间	时间字段	可被时效参数筛选，过期即出局
网页权威度	四级标注	判定可靠性时的现成信号（详见第 04 章）
网页正文（txt/markdown）	Custom 版返回	正文可被完整取得，但已丢失版式结构
rerank 相关度评分	数值	相关性的显式打分，决定名次（详见第 03 章）

表 2-1 检索侧返回的六个字段与其含义（字段名取自官方文档[2]）

这张表最值得注意的一行是「网页正文（txt/markdown）」。它说明即便在能取到全文的版本里，正文也是以纯文本或 markdown 形式交付的——**版式、表格结构、图表、颜色，在这个过程中全部消失。**视觉上做的强调（加粗、底色、字号）不会自动转化为语义上的重要性；能留下来的只有文字顺序。这就是为什么「把结论写进第一句」比「把结论标成红色」有效：前者在纯文本里依然是第一句，后者在纯文本里什么都不是。

这张表还解释了另一个现象：为什么很多内容在「人看着不错」和「机器用不上」之间来回摇摆。人读的是排版后的成品，标题、插图、表格共同承担信息；机器读的是字段与线性文本，其中只有「标题」和「摘要」两段是短文本，其余是长文本切片。换句话说，**同一条内容在两种读者面前暴露的信息量不一样**，而优化要针对的是后者。

摘要就是候选文本

这里有一个容易被忽略的机制细节：摘要不是给人看的预览，而是**进入模型上下文的第一等材料。**文档把「摘要长度」列为可配置参数[2]，说明摘要长度会随调用方配置变化，而内容方无法控制自己被截到哪一句。推论很清楚——如果关键数字或结论落在一段话的后半段，它很可能根本不在摘要里。

需要标注的是推断边界：我们无法确认豆包 App 的每一次回答都严格走「摘要」这一形态，也无法确认摘要截取的具体算法；但「检索侧返回摘要而非全文」是官方文档写明的事实[2]，「摘要里没有的句子无法被引用」是由此推出的直接结论。前者是事实，后者是逻辑推论，两者不能混着说。这个区分在写作时尤其重要：它决定了我们该说「摘要机制让段首更关键」还是「豆包一定只读前三句」——后者是没有依据的断言，也会把优化带向「无限压缩篇幅」这种错误方向。

顺着这个推论走一步，还能得到一个更实际的判断：**摘要与正文的关系，是「摘要必须能独立成立」**。如果摘要需要读者回读正文才能理解，那它作为候选文本就是废的。这条判据可以拿来审稿：随机截取一段的前两句，看它是否构成一个完整的判断。不能，就把结论前置。

四种写法在切片面前的处境

同样一段信息，写法不同，被完整切中的概率差别很大。下面这张表不是实测统计，而是把「切片需要什么」翻译成对写法的要求：自包含、主体明确、数字与单位同句、结论前置。四项满足得越多，这句话越像一枚可以单独拿走的证据。

写法	切片时的处境	为什么这样
结论前置 + 数字同句 「2026 款某车型 625 Max 科技版 CLTC 续航 625 公里，指导价 17.68 万元」	整句自包含，主语、版本、工况、数值、单位齐全，可独立成立	切到这一句，事实即完整；既不需回读上文补主语，也不需看下表补单位
论证式长段 「先说背景……再看三点原因……所以可以判断价格偏低」	结论落在段末，摘要若截在前半，只剩背景没有结论	切片按位置截取，语义完整性不受保护；结论离段首越远越容易丢
表格承载关键数 参数只在表里，正文只写「详见下表」	表格被压成线性文本，行列对应关系在转换中丢失	表格信息靠二维结构承载，切片是线性文本，结构一散，数字就失去归属
代词回指与省略 「它比他便宜两万」「前者续航更长」	切出后主体不明，模型无法判断「它」是谁	指代解析依赖上下文，而上下文在切片时已不存在

表 2-2 四种写法在切片环节的处境（机制推演，非实测统计）

四种写法里，第三种（表格承载关键数）最容易被忽略，因为它在人读场景下是最专业的做法：把参数整理成表格，看上去严谨又清晰。但在纯文本转换里，表格会变成一串失去行列关系的数字，读者（人和模型都一样）很难判断哪个数字属于哪个版本。这不是说不要用表格，而是说**表格负责对照，正文负责被引用**——两者分工明确，才不会在切片时丢掉关键数。

一段真实写法与它的改写

拿行业里最常见的车型介绍举例。原写法常常是这样：「该车提供了丰富的配置选择，其中长续航版本在续航和补能效率上都有不错的表现，价格也更具竞争力，具体可参考官方配置表。」这段话在人类阅读里没问题，但它**没有一个可以被切走的句子**：没有车型名、没有版本名、没有续航数字、没有价格，全部指向外部。

改写的目标不是「写得更好看」，而是「让每一句都能被单独拿走」：「2026 款某车型 625 Max 科技版 CLTC 续航 625 公里，指导价 17.68 万元；同车系 580 长续航版指导价 19.99 万元起。」一句话里包含车型、版本、工况、数值、单位、价格六项要素，切走之后仍然成立，而且与官方公开的指导价一致。这类改写不需要文采，只需要把隐含信息显式化——而显式化恰恰是切片最需要的东西。

顺带说一句口径问题：续航必须带工况（CLTC / NEDC / WLTC），因为三套工况的测试方法不同、数值不可直接比较。把工况省掉，句子在切片后就会变成一个「可被任意解读的续航数字」，既容易引发误读，也容易在合成答案时被别的口径替换掉。第 05 章会把口径问题单独展开。

注意这段改写并没有增加篇幅，反而更短。它做的是**信息重排**：把原本散落在形容词里的要素，压进一句可以独立成立的话。这件事的收益是双重的——切片更容易，读者也更容易抓住要点。反过来说，凡是改写后篇幅暴涨、但信息密度没变的操作，

都不属于这一类，多半只是把话说得更啰嗦。

为什么「说清一件事」比「说全十件事」更容易被引

切片是按句取用的，这意味着**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、每件都只写一句概括，看似信息量大，实际上每一句都缺少数值与限定，切走后都不成立；一页只写一件事、把它写透，反而至少有一句能被直接用。这是「密度」与「覆盖」的取舍，在答案场景里，密度的权重更高。

这条判断不是主张少写，而是主张**把写作资源集中到能被引用的句子上**。具体做法是先确定这一页必须被引用的那一句，围绕它补齐版本、数值、工况、时点，其余内容作为支撑材料。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。

从「可读」到「可切」：三个结构性动作

把上面的推演反过来，就是三个不依赖技巧的结构性动作。第一个动作，**一段只放一个结论，并把结论写在第一句**。这不是「结论先行」的写作偏好，而是为了对抗按位置截取：段首是唯一不会被前文吃掉的位置。第二个动作，**让每一句都自带主体与单位**，宁可重复品牌名与版本名，也不要用「该车型」「它」来回指——重复在人类阅读里是冗余，在切片面前是保险。第三个动作，**关键参数不要只活在表格里**，正文至少复述一次「哪个版本、多少钱、什么工况」，表格承担对照，正文承担被引用。

这三件事听起来像写作规范，实质是**把内容改造成证据形态**。证据形态有明确的判据：把任意一段的第一句单独摘出来，交给一个不了解上下文的人看，他能不能判断这句话在说什么、说的是谁、数值是多少。能，就是可切；不能，就还要改。这个判据可以在交付前逐段自查，不需要等平台反馈；而且它的一致性很好——同一个人按同一判据可以稳定地判对，不依赖经验直觉。

五种自毁式写法

与之相对，有五种写法会主动削弱自己被切中的可能，而且都很常见。**其一，万能开头**：用两百字铺垫行业趋势，结论在第五段才出现。**其二，参数只写在图里**：一张信息图承载全部关键数据，正文没有任何数字，转成纯文本后图里的信息完全丢失。**其三，把限定条件写在注释**：正文说「续航 700 公里」，脚注补充「CLTC 工况、特定版本、特定温度条件」，切片只带走正文。**其四，用系列名代替产品名**：通篇讲「该系列」，需要读者自己对应到具体型号。**其五，把结论写成修辞**：「表现可圈可点」「具备较强竞争力」——这类句子可以被切走，但切走之后不支撑任何结论。

这五种写法的共同点是：它们都在人类阅读里「看起来更专业」。这也是它们难改的原因——问题不在写作水平，而在写作目标错位。面向人的文本追求流畅与克制，面向切片的文本追求自包含与冗余。两者可以兼顾，但需要显式设计：正文用自包含句承担事实，段落之间再用承接句维持流畅。这样既不牺牲可读性，也不牺牲可切性。

三种「看似自包含」的反例

有些句子读起来完整，切开之后却不成立，这类写法最容易被漏掉。**第一种是代词句**：句子里出现「该车型」「这一版本」「上述方案」，读者知道指谁，切片不知道；改法是把代词换回实体名。**第二种是省略主语的评价句**：「续航表现优秀」「价格很有竞争力」，主语与比较对象都缺，切走后既不知道评的是谁，也不知道跟谁比；改法是补主语与参照系。**第三种是跨句因果**：「因为采用 800V 架构，所以补能更快」，前半句与后半句分属两段或两句，任何一句单独出现都只是半个判断；改法是把因果关系压进一句。

三种反例的共同特征是：**它们依赖上下文才成立，而上下文不在切片里**。判断方法很简单——把这句话单独抽出来给人看，如果对方必须问「你说的是哪个」「跟谁比」，它就还不算自包含。这个测试不需要专业知识，任何人都能做，因此可以放进内

容发布的最后一道检查。

标题承担的两项工作

标题在机制里承担两项完全不同的工作，混在一起就会两头都做不好。第一项是**早期判别**：它要在几十个字符内让检索侧判断「这条内容与问题相关」，因此标题里必须出现问题所指的对象与属性（车型、品类、指标名）。第二项是**摘要的语境**：摘要被截出来之后，标题往往是它唯一的上下文，标题能不能交代清楚「讲的是什么」，直接决定摘要段是否可读。

两项工作指向同一种写法：**对象 + 属性 + 限定同时出现，不用修辞替代信息。**例如「某车型 2026 款纯电版续航与价格解析」在两项工作上都是及格的；而「不止于续航：一次关于纯电出行的重新想象」在两项工作上都不及格——它既没交代对象，也没交代属性，唯一的效果是让人想读。标题的修辞空间在答案场景里被压缩得很厉害，这是与传统内容运营最大的差别之一。

> **这一章最该记住的一句：**摘要与切片是检索侧的入口形状，内容方唯一能控制的是——让关键结论以自包含句子的形式，出现在段首。这一条不需要猜算法，只需要改写法。

03 第二道判定：能不能进入召回配额

切得再好，进不了候选池也没有意义。召回环节的规则，在豆包搜索的产品文档里有相当具体的公开口径：Global 版单次最多返回 20 条结果，Custom 版最多 50 条；两版都支持指定域名、屏蔽域名、指定时效、限制仅在国内 ICP 备案网站范围内检索；Custom 版还支持行业领域搜索、权威分级搜索，并返回 rerank 相关度评分[2]。

配额决定了这是一场名次竞争

20 与 50 这两个数字，看起来只是接口限制，实际决定了竞争的性质：**召回是一场名次竞争，不是「够不够好」的竞争。**一个查询，全网有成千上万条相关内容，最终只有 20 或 50 条进入下游，其余内容连被判定可靠的机会都没有。配额越紧，排序信号里「早期判别」的比重就越大——标题、首段、以及能否与查询词形成明确对应，都会在很靠前的位置决定生死。

由配额还能推出一个常被忽略的结论：**候选池是有限的公共资源，同行之间是零和的。**你多占一条，别人就少一条；这与传统搜索里「排名前十都可以有流量」的结构完全不同。这解释了为什么在对话式引擎里，头部内容的优势比在传统搜索里更陡——不是流量分配变了，而是名额本身变少了。

由此可以推出一条与常见建议相反的判断：**追热点式的长文，常常输给一条短而准的问答。**不是因为长文质量差，而是因为长文的相关性信号分散在几万字里，而短问答把信号压缩在首段，在「配额 + 相关度评分」这套机制下更容易被选中。这条判断是机制推演，不是实测结论，但它可以被验证：同一主题分别用长文与短问答布局，观察哪一类更常出现在来源列表里。这类可验证的推断，比「长文一定更好」或「短文一定更好」的断言有用得多。

时效是字段，不是形容词

「指定时效：限制搜索结果发布时间」[2]这一条，是整份文档里对内容方最有操作价值的能力之一。它意味着发布时间不是文末的一行小字，而是**可被筛选的元数据。**当调用方把时效窗口设为近期，超过窗口的内容会直接被排除在候选之外，与内容质量无关。

这一点会改变内容维护的经济学。很多企业站的做法是「发一篇通稿，然后多年不动」，在时效筛选面前，这类内容的价值随时间快速衰减，而衰减不是线性的——一旦跨出常用窗口，它就从「可能被选」变成「基本不会被选」。反过来，同一份内容做定期更新（改数据、补版本、加日期），成本远低于重新写一篇，却能把内容一直留在窗口内。

更新的动作也有讲究：如果只改了发布日期而没有改内容，某些判别方式可能把它识别为「旧内容换新标签」；真正有效的更新是换掉具体数字与结论，让内容在新时点上依然成立。**可验收的更新有一个简单标准：更新前后各抽三句核心事实，若全部未变，那这次更新对时效筛选几乎没有意义。**

长内容与短内容在配额里的不同处境

配额有限，意味着候选之间在争同一个位置，而长内容与短内容的争法不同。长内容的优势是覆盖面广，能在多个问题上被召回；劣势是单次检索只能读到它的片段，若片段位置不对，长度反而是负担。短内容的优势是主题集中，容易在单一问题上拿到高相关度；劣势是覆盖窄，只能赢一个点。

合理的组合是**一页一主题、页内分层**：每页只回答一类问题，页内按「结论—字段—边界」三层排列。这样既保证相关度集中，又保证任意一种切片长度下都有可用句子。与之相对的是「大而全」的栏目页：一个页面覆盖十几种问题，相关度被稀释，摘要也容易截到与问题无关的部分——这是很多企业站内容量不小、却很少被引的技术原因。

行业领域检索：候选被再切一次

Custom 版提供「行业领域搜索」这一能力[2]，含义是把候选限定在某个领域内。对内容方的影响是间接的但很实际：**如果你的内容在站内被归到别的栏目，它可能根本不在该领域的候选里。**这不是关键词不够，而是归类不对——例如汽车内容放在综合资讯栏目下，行业检索时就可能被过滤掉。

可执行的判断方法：在站内看这篇内容所属栏目的上级分类，以及页面上的结构化标注是否指向同一领域。两处一致，才谈得上进入该领域的候选；只靠正文提到车型名，通常不够。这一条与第 02 章的自包含并不冲突——自包含管「切得动」，归类管「进得来」，两道都要过。

七个召回参数各自的含义

参数（官方口径）	机制含义	内容侧可执行动作
指定数量 20 / 50	候选池有硬上限，进入即竞争名次	把核心结论压进标题与首段，减少「读到第三段才明白在讲什么」
指定时效	发布时间是可筛选字段，过期即出局	建立更新节奏，正文标注「截至某年某月」的数据口径
行业领域搜索	可限定在垂类权威内容库内检索	内容落在明确的行业语义里，避免「什么都能沾一点」的泛化写法
权威分级搜索	可只取特定权威等级的源	把事实性内容往可核来源上靠，别把关键数据只放在自营渠道
域名指定 / 屏蔽	调用方可白名单或黑名单	被屏蔽属于调用方策略，不是内容问题——别把所有「没被引」都归因于质量
ICP 备案限定	可限定境内备案站点	面向国内场景的内容，域名与备案状态本身是准入条件
rerank 相关度评分	返回内容与查询的相关度打分	相关性是「能不能回答这个问题」，不是「关键词出现几次」

表 3-1 豆包搜索公开的召回控制参数与内容侧含义（参数取自官方文档[2]）

七个参数里，最容易被误读的是 rerank 相关度评分。它给的是「网页内容与 query 的相关度」，反映的是内容能不能回答这个问题，而不是关键词覆盖率。把相关度理解成关键词密度，会导致一种典型操作：在一篇文章里反复堆同一个词，结果句子变得不可读，切片后反而更像营销页。相反，用「问什么答什么」的结构——标题即问题，首段即答案，后文再展开——在本章与第 02 章两套机制下都占优。

「域名指定」与「域名屏蔽」这两个参数值得单独说一句：它们意味着**有些「没被引」根本不是内容问题。**调用方可以在业务侧把某类站点整体排除（例如只允许官方与垂媒、或屏蔽某类聚合站）。在这种情况下，再怎么优化内容也不会出现在来源列表里。观测时如果发现某个站在任何问题下都不出现，第一件要排查的是准入与屏蔽，而不是内容质量。

从时延预算看「一次回答能读多少」

文档给出两版的平均搜索时延：Global 版 1053ms，Custom 版 700ms[2]。这两个数字与内容方没有直接关系，但它解释了一个体验层面的现象：一次回答里的检索是一次带预算的调用，时延预算决定了它能取回多少材料、做多少轮改写。这提示我们不要用「模型读了全网」的想象去设计内容——它读的是有限条数、有限长度的一批切片，这也再次回到第 02 章：**能不能被切中，比写得多不多重要得多。**



图 3-1 两版豆包搜索的平均时延对比（数据取自官方产品文档[2]）

三种竞争位置，三种不同的处方

在配额竞争里，内容大致处在三种位置，各自的处境并不相同。**头部位置**是标题与首段就能命中问题的内容，它的风险不在召回而在可靠性（第 04 章）；**腰部位置**是内容对得上但需要读一段才明白的内容，它的风险在于被更直接的内容挤掉名次，处方是把答案提前；**长尾位置**是内容确实相关但语义泛化，它往往在行业领域检索或权威分级筛选下直接出局，处方是收紧主题、落到具体实体与具体场景。

三种位置的诊断方法不一样，但都可以用同一组观测回答：把同一批问题反复问，看自己在哪些问题上稳定出现、在哪些问题上从不出现、在哪些问题上时有时无。稳定出现说明召回没问题；从不出现说明可能连配额都没进；时有时无说明进了配额但名次不稳。这三种信号指向三种处方，比笼统地说「优化没效果」有用得多。

还有第四种情形需要单独指出：**结构性缺席**——某个品类、某个话题的候选池里完全没有商业内容，全是官方与垂媒。这不是名次问题，而是品类准入问题：在该品类里，商业站点的内容几乎不被当作可核来源。遇到这种情形，正确的动作不是加大内容投入，而是换一个可被商业内容参与的话题切口，或先在有准入的载体上建立可核条目。

站点准入与内容准入：两级门槛

召回前的门槛其实是两级，常被混成一级。第一级是**站点准入**：域名是否在国内 ICP 备案范围内、是否被调用方白名单收录、是否被黑名单屏蔽、是否达到指定的权威等级[2]。这一级与内容写得好不好无关，它由站点属性决定，属于「能不能参加」。第二级是**内容准入**：标题是否对应问题、发布时间是否落在时效窗口内、语义是否落在行业领域内、与 query 的相关度评分是否够高。这一级由具体页面决定，属于「能不能被选上」。

两级门槛的诊断顺序不能反。**先查站点准入，再查内容准入**：站点被屏蔽或未备案时，内容优化完全无效，而这在观测上的表现与「内容不行」几乎一样——都是从不出现。区分方法很直接：在同一领域挑几个明显不相关但同站的内容去问，如果它们也不出现，问题就在站点层；如果它们出现而你核心页不出现，问题在内容层。

04 第三道判定：能不能被判定为可靠

进入候选池之后，内容要面对「可靠性」这一关。这一关在豆包搜索里有公开的形态：站点权威度分级。火山引擎《站点权威度分级说明》把结果按四级标注，并给出每一级的描述与示例站点[1]。

四级分级写的是什么

级别	官方描述（摘录）	文档给的示例（部分）	对内容方的现实含义
非常权威	内容来源真实可信，经过审核及筛选	gov.cn 各级网站、央媒（人民网 / 新华网 / 央视网等）、985 / 211 高校官网、世界 500 强与中国 500 强企业官网、新闻门户中新闻媒体 / 国家机构认证账号发文、字节高质量数据源	事实性数据的首选落点；企业侧最接近这一级的位置是头部企业官网与官方认证账号
正常权威	行业优质站点、综合性门户网站，具有较高的专业性及可信度	同花顺、懂车帝、小荷健康、豆瓣、B 站、什么值得买、澎湃新闻、新闻门户中高粉丝量认证作者、中小型企业官网	垂类内容的主战场；中小企业官网落在此级，不是「天然低人一等」，而是「要靠专业性挣位置」
一般权威	普通网站，内容来源无法考证，但部分内容有参考价值	各类文库站、新闻门户普通 UGC 作者、非门户级社区论坛、非头部垂类站点	可以进候选，但很难成为「支撑结论」的那一个源
一般不权威	普通小站，劣质、虚假信息较多，无法提供专业、可信内容	文档未给示例	检索侧可被直接过滤；内容做得再细也难以翻身

表 4-1 站点权威度四级分级（描述与示例均取自官方文档[1]）

这张表里有一个容易被忽略的细节：**中小型企业官网被明确列在「正常权威」里**[1]。这与流传甚广的说法相反——不是「企业官网天然只能算一般权威」，官方示例把中小型企业官网放在了第二级。这个细节有实际意义：它说明企业站不该把精力全押在「求外部媒体报道来抬段位」，而应该先把官网自身的专业性做出来。台账里没有任何一句话说「企业官网必须靠别人背书才能进候选」。

一个必须说清的纠偏

过去常见的说法是「豆包把信源分成四级，等级决定引用概率，所以要先抬段位」。这个说法里，前半句有官方依据，后半句没有。**文档对这套分级的用途写得很克制**：一是「主动搜索控制」——Custom 版可以指定只返回某个权威等级的结果；二是「搜索结果标注」——返回字段里带 AuthInfoLevel / AuthInfoDes（Global 版为 AuthorityLevel）[1]。它是一项**搜索服务的能力**，而不是「豆包 App 答案排序权重表」。

把两者等同，会带来两个具体错误。其一，把 API 的分级当成回答生成的打分器，进而推出「进了第一级就一定被引」这种无法证伪的结论；其二，把「没被引」一律归因为「权威度不够」，忽略了第 02、03 章那两道更前置的判定——很多内容根本没走到可靠性这一关。更稳妥的表述是：**权威度是检索侧可用的一个筛选维度，它会提高「被当作可核来源」的概率，但它不是唯一维度，也没有公开的权重系数。**这句话里的每一个限定词都是必要的，去掉任何一个都会变成过度断言。

这类纠偏在本册里反复出现，原因是它决定了对外的表达方式。给客户讲机制时，说「权威度是检索侧的一个筛选维度」是可核的；说「权威度决定你在豆包里的排序」就变成了编造。前者能撑住追问，后者一旦被对方的技术人员查证，整个方案的信

用就没了。机制研究的第一价值不是给出技巧，而是让结论经得起追问。

站点段位与内容质量是两套坐标

段位解决的是「这个源可不可以被当作依据」，内容质量解决的是「这句话能不能支撑结论」。两套坐标可以组合出四种情形，各自的处境并不相同：段位高但内容空，会出现「被读到却摘不出句子」；段位低但句子扎实，会出现「被摘到了却不敢挂」；两者都高是最理想的情形；两者都低则连候选都进不去。理解这个二维关系，能避免把两类问题混成一句「我们权威度不够」。

组合	典型表现	处理方向
段位高 / 内容空	页面被检索到，但摘要里没有可用的结论句	属于第 02 章问题：把结论写进句首，补数字与主体
段位低 / 句子扎实	句子进了候选，但很难被当作支撑源	把同一事实放到可核载体上重复一次（下一节）
两者都高	稳定出现在来源列表，句子被采用	维持更新节奏，防止时效出局
两者都低	完全不出现	先解决准入，再谈优化

表 4-2 段位与内容质量的二维组合（机制推演）

多源一致：可靠性在答案层的形态

除了站点分级，可靠性还有一层落在答案本身：同一事实是否有多源印证。这一层在学术上对应归因研究，ALCE 用 NLI 模型判断「引用的源是否能支撑该句」，并以此计算 precision[7]；后续工作进一步把「引用是否真的支持」细化为更严格的判定[10]。这些研究的共同结论是：**模型对支撑关系的判断并不稳定**，当一句话由多个源拼接而成时，挂上去的那一个源常常只支撑了其中一部分。

对内容方的启示是反直觉的：**与其在一篇文章里把话说满，不如让关键事实在多个独立来源上重复出现。**因为模型在合成答案时，倾向于为每个陈述找到能对上的源；当同一组数字在官网、垂媒、行业目录上写法一致，无论它最终挂的是哪一个链接，句子都成立。反过来，若只有自家一处这么写，即使被挂上，也随时可能在下一次回答里被别家的写法替换掉。

做到多源一致有三个可核载体，按成本从低到高排列：**官网的结构化页**（同一组数字在正文、FAQ、参数页写法统一）、**有明确署名的垂媒稿**（同一组数字、同一口径、同一时点）、**行业目录与公开数据集**（可被第三方引用的条目）。三者不需要内容创新，只需要口径纪律：同一个参数必须在所有载体上写同一个数，带同一个工况，标同一个时点。这条纪律听起来简单，实际执行时最常出问题的地方是「同一家公司的不同部门各自维护一个数」，最终在语义空间里长出两套互相矛盾的参数。

这里可以引入 GEO 论文的一个结论作为旁证：Aggarwal 等人用 GEO-Bench 系统测试了若干内容改造策略，报告称可见度最多可提升 40%，且不同领域的有效策略不同[8]。值得注意的是，论文里被反复提到的有效动作是加入引用、引文与统计数据——它们共同的特征是「让句子更像可核证据」，而不是「让句子更华丽」。这与本章的判断同向：可靠性不是形容词堆出来的，是证据形态带出来的。

两个版本，两套权威参数

官方文档里有两个容易混淆的参数名：Global 版用 **AuthorityLevel** 标注结果，Custom 版用 **AuthInfoLevel** 与 **AuthInfoDes** 返回权威信息，并允许调用方在检索时直接指定 AuthInfoLevel 做筛选[1][2]。名字不同不是排版差异，而是能力差异：前者只能读，后者可以筛。对内容方的现实含义是：**在不同接入方式下，你的内容被「按权威度过滤」的可能性并不一样。**

能力	Global 版	Custom 版
权威标注字段	AuthorityLevel	AuthInfoLevel、AuthInfoDes

能力	Global 版	Custom 版
能否按权威度筛选检索	不支持	支持（指定 AuthInfoLevel）
候选条数上限	20 条[2]	50 条[2]
其他筛选手段	指定时效、指定 / 屏蔽域名、仅国内 ICP 备案站点[2]	同上，另有行业领域搜索[2]

表 4-3 两个版本在权威与筛选能力上的差异（依据官方文档[1][2]）

这个差异带来的推论很具体：当调用方使用 Custom 版并指定了较高的权威级别时，低级别站点的内容**不是排在后面，而是根本不进候选。**这与「名次不好」是完全不同的处境——后者还能靠相关度争取，前者连参与的机会都没有。因此在评估自身处境时，先确认目标场景用的是哪种接入方式，比笼统地问「我的权威度够不够」更有意义。

一种危险做法：把段位当挡箭牌

在实际项目里，段位最容易变成一种解释一切的挡箭牌：「我们不是央媒，所以做不了」。这种归因有两个问题。其一，它跳过了前两道判定——很多内容的真实问题是切不出来、进不了配额，与段位无关；其二，它把段位当成静态属性，忽略了官方文档明确给出的一个事实：中小型企业官网本身就处在「正常权威」这一级[1]。也就是说，多数企业在起点上并不低，问题往往出在内容没有把专业性写出来。

更可操作的思路是把段位当成**一个可以逐级争取的目标**，而不是一道门槛。具体路径有三条：把核心事实放到可核载体上（官方账号、垂媒署名稿）；把站内的专业内容做厚（结构化的参数、明确的署名与更新时间）；把第三方引用做起来（被行业目录收录、被垂类媒体引用）。三条都不需要「变成央媒」，只需要让外部能核到你的说法。

为什么垂类站点常落在第二级

把示例站点横向读一遍，会发现一个规律：官方文档列在「正常权威」的那一批，多数是垂类专业平台——金融数据（同花顺）、汽车（懂车帝）、医疗（小荷健康）、消费决策（什么值得买、豆瓣）、视频（B 站）、本地资讯（澎湃）[1]。这说明「专业性」本身就是一条被承认的路径：不必是央媒或 500 强，只要是某个领域的优质站点，就能进第二级。

这条规律对企业站的含义是：**企业官网与垂类平台处在同一级，属于同台竞争而非天然劣势。**既然同级，竞争就回到内容本身——谁能给出更可核的事实，谁就更容易被当作支撑源。反过来，把「我们不是垂类平台」当作理由，其实是在承认自己的内容没有体现出专业性；这一点是可以改的，而平台属性是改不了的。

站点层的三种可核动作

段位不是一句评价，而是可以被外部核到的一组事实。可核的动作有三种，都在常规运营范围内。**其一，明确主体**：页面上有清晰的主体信息（公司名、联系方式、资质编号），让「谁在说」可查；**其二，明确署名与时间**：内容有作者、有更新日期，让「谁写的、什么时候写的」可查；**其三，明确来源**：引用外部数据时给出出处，让「依据是什么」可查。三件事都不需要额外资源，只需要在发布流程里固定下来。

这三件事的效果不是「升段位」，而是**让内容具备被当作依据的形式条件。**这也是官方分级描述里反复出现的关键词——「内容来源真实可信」「专业性」「可信度」——所指向的东西[1]。把这三项做成模板，比一年写十篇没有署名的软文更有累积效应；前者可核，后者不可核。

05 第四道判定：能不能被安全写进答案

前三道判定都通过，内容进入生成环节，还要过最后一关：这句话能不能被安全地写进答案。这里的「安全」有三层意思——不违规、不过时、不失真。三层各自有明确的失败形态，也各自对应不同的修法。

三种引用形态，成本不一样

从公开文档能确认的引用形态至少有三类。第一类是**答案内联引用**：豆包用户协议写明回复中会展示搜索来源[4]；火山方舟的知识库搜索文档则要求「在正文中明确引用来源，引用格式为 [1] (URL)」，并在回复末尾列出参考资料[6]。第二类是**末尾参考资料列表**，即回答结束后集中列出用到的来源。第三类是**名词链接**，即关键名词输出百科词条链接[5]。

三种形态的「被引成本」不同。内联引用要求句子与源严格对应，精度要求最高，但一旦成立，说服力也最强；末尾列表的容错空间大，一句话可以由多个源共同支撑，但读者很难判断哪条源支撑了哪句；名词链接更像实体识别的副产品，它证明的是你的实体被认出来了，不等于你的观点被采纳。做观测时把三种混在一起算引用率，会得到虚高的数字（第 07 章会给出分开计算的方法）。

引用形态	公开依据	成立条件	观测时的陷阱
答案内联引用	豆包协议展示来源[4]；方舟知识库搜索的 [1](URL) 格式要求[6]	句子与源一一对应，源能支撑该句	把「出现在来源条」当成「支撑了某句」，精度被高估
末尾参考资料	方舟文档要求末尾列出全部参考资料[6]	多源合成为一句时的兜底展示	来源多不代表被采纳，可能只是「读过」
名词百科链接	联网问答 Agent 支持关键名词输出百科词条链接[5]	实体被识别并命中词条	实体被认出≠观点被引用，混算会虚高

表 5-1 三种引用形态的成立条件与观测陷阱（依据均来自公开文档）

值得注意的是方舟知识库搜索给出的工程要求：正文内联 **[1] (URL)** 形式，且回复末尾要列出全部参考资料[6]。这一条虽是产品侧对调用方的格式约束，却透露了一个判断：**引用在工程上被当作「可定位的证据索引」处理，而不是装饰。**既然它是索引，那么被索引的东西就必须能被定位到具体句子——这又回到第 01 章的句级判定。

合规：不是能不能被引，是被引之后会不会被撤

生成式引擎的答案会被反复生成，这意味着一条不合规的内容不是「一次风险」，而是「持续暴露」。具体到三类内容，风险点很明确：其一是绝对化表述与效果承诺，在医疗、教育、金融类问题里尤其敏感；其二是资质与许可类信息，写了但没有可核主体，容易被判定为不可信；其三是价格、政策、参数这类强时效信息，一旦过期还在被引用，错误会被放大成「平台在传谣」的印象。

这三类的共同修法，是把「可核」写进句子本身：数据带口径与时点，资质带主体与编号，价格带版本与有效期。这样做的直接效果不是「更容易被引」，而是**被引之后不容易被替换或撤下**——在答案会被反复重写的场景里，这一点比一次性的曝光更值钱。原因很朴素：下一次性成答案时，模型会在候选里比较不同来源，能给出更精确限定的那一句更容易被选中，说满了却被现实推翻的那一句更容易被放弃。

还有一层现实约束：一旦某条内容因为过期数据被答案采用，纠正的成本远高于预防。答案不会因为你更新了页面就立刻改变，它会在下一次生成时重新取源；在此之前，错误一直在被复制。所以**强时效内容的第一原则是「限定先行」，而不是「尽量说得满」。**

绝对化与限定语的分界

合规表述里最容易被误判的是「绝对化用语」的边界。很多人以为只要不出现「最」字就安全，其实问题在于「是否作出无法核验的断言」。同一件事用两种写法，风险差别很大，而信息含量可以完全相同。下面这张对照表给出四种常见改写，改写后

信息不减少，风险显著下降。

高风险写法	改写（限定化）	信息是否损失
「行业第一的续航表现」	「在 WLTC 工况下，本车系最长续航版本标称 720 公里」	不损失；把排名换成可核数值
「彻底解决续航焦虑」	「支持 800V 高压平台，15 分钟可补充 540 公里续航（厂方公布数据）」	不损失；把效果换成本次可核的能力
「全网最低价」	「2026 款某配置限时权益价 14.99 万元（指导价 15.29 万元，活动期以官方公告为准）」	不损失；把比较换成区间与有效期
「全面领先同级」	「在该级别中提供 800V 平台与 3C 快充，充电 30%—80% 最快 26 分钟」	不损失；把评价换成本次可核的参数

表 5-2 绝对化表述的限定化改写（示例数值取自公开指导价与厂方公布参数，仅作写法示范）

时效与地域：两个最容易被忽略的过滤维度

时效的问题在第 03 章已经从召回侧讲过，这里补生成侧的一面：即使一条旧内容仍在候选池里，生成时也可能因为「与当前时点不符」而被弃用，转而采用更新的源。所以更新不只是为了让内容进池，也是为了让内容在被比较时不吃亏。地域的问题则来自问题的形态——一旦提问带地域限定（本地政策、本地补贴、本地门店），答案会优先采用与该地域对应的源；一条全国口径的内容，在这种问题里会天然处于劣势。

把这两点合起来看，会发现一个经常被误判的现象：同一条内容，在某些问题上稳定被引，在另一些问题上从不出现。原因往往不是质量问题，而是**它在这类问题的「时点 + 地域」组合里不成立**。观测时如果不记录提问的时点与地域限定，就会把这类系统性差异误读成随机波动——这是观测设计里最常见的错误之一（第 07 章会给出记录字段）。

强时效信息的三条处理纪律

价格：写清版本、口径与有效期。「指导价 17.68 万元」与「限时权益价 16.68 万元起」是两个不同的数，混写成「17 万左右」在切片后会被当作确定价格引用，而它既不对应指导价也不对应权益价。价格类内容的最佳形态是「版本—指导价—权益价—有效期」四件套同行呈现，切片拿走哪一行都完整。

政策：写清适用主体、适用地域与生效时点。政策类内容一旦缺主体或地域，答案就会把适用于 A 地的条款搬到 B 地，而这类错误最难纠正——它看起来完全合理。

参数：写清工况、测试标准与版本。续航必须带 CLTC 或 NEDC 或 WLTC，油耗必须带 NEDC 亏电油耗或 WLTC 亏电油耗，电耗必须带工况；不同标准的数值不可直接比较，混用会让「同一车型两个续航数字」同时出现在语义空间里，最终被合成成一句自相矛盾的话。这不是文字细节，是防止重写失配的第一道闸门（第 06 章会给出具体例子）。

答案句的三种形态

被引用的句子在答案里通常承担三种功能，理解它们的区别，能让写作者判断「我这句是给谁用的」。**结论句**给出一个判断，通常出现在答案开头或段首，最需要自包含与可核；**清单句**列举若干条目，多见于「有哪几款」「有哪些条件」，最需要字段齐整；**边界句**说明适用范围与例外，最常见于政策、价格、参数类问题，最需要限定条件与数值同句。

形态	典型位置	写作要求	失败形态
结论句	答案开头、段首	主语、结论、依据同句	结论在段末，切片后只剩铺垫
清单句	答案中段	条目同构、字段一致、单位统一	条目口径不一，横向对照时失配

形态	典型位置	写作要求	失败形态
边界句	答案末段或条款处	限定条件与数值同句、否定写完整	限定被切掉，结论反向（第 06 章）

表 5-3 答案句的三种形态与各自的写作要求

三种形态不必分开放，但必须分别达标。常见的问题是内容只写结论句，缺清单句与边界句：结论说得漂亮，但没有可对照的字段，也没有适用范围，于是模型只能用你的结论去支撑它自己的判断，一旦它需要的是数据或边界，就会转向别的源。补上两类句子，等于把自己的内容做成一个「可被任意取用的证据库」，而不是一篇只能整读的文章。

06 三种失效：截断、漂移与重写失配

把前四道判定反过来看，就是三种最常见的失效。之所以要把它们分开写，是因为三者的修法互不通用——用「多发内容」去治截断，等于没治；用「提高权威度」去治重写失配，也等于没治。更重要的是，三种失效在答案里的表现完全不同，只有分开命名，才能在看到一条错误答案时立刻判断它属于哪一类。

失效一：摘录只截到半句，语义反转

这是切片环节的典型失败。原文写「该版本不支持 800V 高压平台」，如果摘要或切片恰好停在「该版本不支持」，剩下的「800V 高压平台」被切掉，句子就从「否定平台」变成「否定一切」，语义直接反转。类似地，「9.98 万元起（限时权益价，指导价 11.98 万元起）」若只截到前半，价格口径就丢了限定条件，读者看到的是一辆比实际定价低两万的车。

这类失效的可观测特征是：**答案里出现了你原文中并不存在的结论，或结论的限定条件消失。**修法很具体：把否定句写完整、把限定条件与数值放进同一句、避免「不支持 / 不含 / 除外」这类词落在句尾。更好的做法是把否定句改写成肯定句加边界，例如「该版本采用 400V 平台，未配置 800V 高压架构」，两段语义在同一句里闭合，切片停在哪里都不会反转。这不是文字洁癖，是防止切片把句子拦腰截断后产生反向含义。

同样的机制在比较句里也会发作。「A 比 B 便宜两万」若被截成「A 比 B 便宜」，读者只能自行补单位与量纲；而如果上下文恰好涉及多组对比，补出来的可能完全相反。可验收的自查方法是：把每一个含否定、含比较、含限定条件的句子单独摘出来，看它是否仍然成立；三步之内能自查完一篇文章的这类句子，而这类句子通常不超过全文的三分之一。

失效二：实体漂移，同名不同物

第二类是实体层失败。同一品牌下常有多个车系、多个年款、多个配置共用近似名称，一旦检索侧没有把「年款—配置—价格」绑定在一起，答案就会把 A 版本的价格写到 B 版本头上。这类错误在参数类问题里最刺眼，因为它看起来「很像真的」——数字确实来自这个品牌，只是挂错了版本。

可观测特征是：**答案给出的数值能在官网上找到，但对不上它自己说的那个版本。**修法是把三元组写死：年款、配置名、价格 / 参数必须同句出现，标题与结构化字段里也要带版本后缀。凡是需要读者「自己推断是哪个版本」的写法，都是漂移的温床。反过来，如果一篇文章里同时出现「2026 款」「2025 款」两个年款，却只有一组价格，那么无论切到哪一句，都可能产生漂移——这种写法在长文里非常常见，也是参数类答案出错的常见来源。

还有一个更隐蔽的变体：**单位漂移。**同一篇内容里续航用「公里」、另一段用「km」，价格一段写「万元」、另一段写「元」，虽然指向同一件事，但在合成答案时可能被当作两个不同的量。规范做法是全篇统一写法，并在首次出现时给出完整表述（如「625 公里（CLTC 工况）」），后文再简化——简化只用于行文流畅，不能用于关键数值。

失效三：重写失配，一句挂了一个源

第三类发生在生成环节。模型把多个源的信息合成一句更通顺的话，然后只挂一个链接。此时那句话可能同时包含两个源的事实，而被挂的源只支撑其中一半——按 ALCE 的定义，这属于 precision 下降：引用存在，但不完全支撑[7]。ALCE 的实测显示，即便最好的系统也有相当比例的回答缺乏完整引用支撑[7]，说明这类失配是链路的结构性特征，不是偶发缺陷。

这类失配几乎无法从外部「修」，只能从内容侧降低它的发生概率：让同一事实在多个独立来源上写法一致，模型无论合成到哪一半，都能找到对得上的源。这与第 04 章的判断是同一条：**多源一致是对抗重写失配的唯一稳定手段。**它的另一个好处是可核——一旦你在三个载体上写法一致，就可以拿任意一个答案句去逐字比对，判断自己是否被正确地支撑了。

需要区分的是「失配」与「错误」。失配是指源与句的对应关系不完整，源本身可能没错；错误是指数值或结论本身错了。前者靠多源一致缓解，后者只能靠内容准确性解决。把这两件事混起来，会导致一种无效努力：为了防失配而在同一句里塞进更多限定语，句子变得冗长难读，切片反而更容易断——修法过度的典型症状。

失效	发生在哪一环	可观测特征	对应修法
截断反转	切片 / 摘要	答案出现原文没有的结论，或限定条件消失	否定句写完整；限定条件与数值同句；避免句尾否定
实体漂移	检索与实体对齐	数值真实存在，但版本对不上	年款—配置—价格三元组同句绑定，标题带版本后缀
重写失配	生成与挂链	引用存在，但只支撑句子的一部分	同一事实多源同写法，降低合成后无源可对的概率

表 6-1 三种失效的定位与修法（机制推演，与第 02—05 章逐条对应）

排查顺序：先看断面，再看单句

失效排查的顺序比排查的细致程度更重要。正确的顺序是**先看整体断面，再落到单句**：先统计同一批问题里「完全不出现」的比例，若比例很高，问题多半在站点准入或时效（第 03、04 章），逐句修改没有意义；若比例低但错误多，问题在写法（第 02、05 章）；若出现且错误集中在版本与价格上，问题在实体对齐（第 06 章）。

反过来做——从单句开始逐条修——最容易陷入局部：句子改得更自包含了，但站点没备案，仍然不出现；或口径统一了，但发布时间过期，仍然出局。断面先定方向，单句再定精度，两步都做才算排查完成。

三种失效的共同点

把三种失效摆在一起，会看到一个共同点：**它们都发生在信息的搬运过程中，而不是信息的产出过程中。**内容写得再专业，只要在切片、实体、合成三个搬运环节上失形，最终答案就会失真。这也是本册反复强调「证据形态」的原因——它管的正是搬运。

这个判断有一个直接推论：**评估内容时不该只看它「写得好不好」，还要看它「搬得动搬不动」。**同一份素材，按证据形态重排之后，专业程度没有变化，但被正确搬运的概率完全不同。反过来，一份表述华丽但主体模糊的内容，在人类读者那里可能拿到高分，在切片面前却近乎零价值。两种评价体系的差异，正是 GEO 与内容营销最容易分道的地方。

还有一个附带结论值得记下：三种失效的**修复成本不对称。**截断可以在写作阶段消除，成本最低；漂移需要在内容治理层面统一口径，成本中等；失配只能靠长期的多源经营缓解，成本最高。所以在资源有限时，优先做前两项——它们解决的问题占比更大，且见效更快。

一个可复核的判定动作

修完之后要能验收。验收动作只需要三步：把答案里的每一句单独摘出来；找出它挂的那个源；在源里核这句是否被原句支撑。支撑的标记为「对」，不支撑的标记为「错」，源里找不到对应句的标记为「缺」。三个标记分别对应三种失效：缺多半是截断，错多半是漂移或失配，对则说明这一句走通了全链路。这个动作不需要工具，一个人一小时可以核完一份十问的答案，而且结论是可复核的。

失效的连锁：一次截断如何变成一条错误结论

三种失效率常不是单独发生，而是连锁的。一次截断让限定条件消失，答案因此给出一个更宽的范围；这个更宽的范围进入下一轮生成时，又被当作前提去对照别的参数，于是产生一个原文里从未出现的结论。等到有人拿这个结论去核，会发现「每个数字好像都对」，但整句话是错的——这是最难纠的一类错误，因为它没有明显的错误点。

连锁的破坏力来自**误差被后续环节放大**：切片环节丢一个限定，实体环节就可能配错版本，生成环节再把它写成一句通顺的判断，最后读者看到的是一句读起来毫无破绽的话。这也解释了为什么「逐句核对」比「通读检查」有效：通读时人是顺着语义走的，会自动补上被切掉的限定；逐句核对时限定是否在场，一眼就能看出来。

长文与短文的失效分布

三种失效在不同长度的内容里分布并不相同，这一点常被忽略。**截断反转**在长文里更常见：长文结构复杂、从句多、限定语常放在后半句，切片停在半句的概率更高；短文句子短、限定紧跟数值，反而更安全。**实体漂移**同样偏爱长文：一篇内容覆盖多个年款、多个配置时，版本信息容易被稀释，读者与模型都要靠上下文推断主语。

重写失配则与长度关系不大，它取决于「同一事实被多少源以多少种写法说过」。这也解释了为什么短内容未必更安全：短文如果只有自家一个源，一旦被挂链，失配风险依旧存在。结论是——长度影响前两种失效，源的数量与一致性影响第三种；两类问题要分开治。

07 引用可见度怎么量

机制讲完，落到可执行的一步：怎么知道自己有没有被引。这里给四个指标、一套采样设计，以及一张可以直接拿去用的记录表。四个指标借用了学术侧的定义，改造到本地可操作的程度；采样设计解决一个更基础的问题——不固定条件的观测，得到的数字没有可比性。

四个指标

指标	定义	怎么采	常见误用
引用率	被引题数 ÷ 总提问数	固定题库，逐题记录来源列表里是否出现自家域名	只跑一次就下结论；没记录提问时点与措辞
引用精度	挂链是否真的支撑对应那句话（借 ALCE 的 precision[7]）	人工逐句核：把答案句与被引页面原句对照	把「出现在来源条」当成「支撑了某句」
引用份额	同一问题里，自家品牌 / 产品被提及次数 ÷ 全部被提及次数	逐题统计实体出现次数，含未挂链的提及	把「被提及」与「被挂链」混算，虚高

指标	定义	怎么采	常见误用
位置权重	是否出现在首段结论句、是否进入最终推荐列表	按段落位置分级记录，而非只记有无	把末尾参考资料等同于核心引用

表 7-1 四个引用指标的定义与采集方式（精度指标沿用 ALCE 定义[7]）

四个指标的关系是分层的，不能合成一个总分。引用率回答「有没有出现」，精度回答「出现得对不对」，份额回答「在同类里占多少」，位置回答「出在哪里」。四个问题各自独立，把它们加权成一个「GEO 指数」，除了方便汇报，没有解释力。真要排序，建议只按引用率与精度两个数来看——前者代表机会，后者代表质量。

这四个指标还有一个排序上的自然顺序：**先看引用率，再看精度。**原因是精度必须有样本才能核，而引用率零的情况下没有样本。所以观测应该先解决「有没有」，再解决「对不对」，最后才是「占多少」与「排在哪」。很多项目跳过前两步直接谈份额与位置，结果拿到的数字既不可核，也无法指导动作。

题库怎么配：四类问法的比例

题库不是随便凑二十个问题，配比会直接影响结论。建议按四类问法各占一定比例：事实型与比选型合计不少于一半，因为它们最容易判定对错；推荐型与否定型各留一部分，用来暴露「榜单化写作」与「截断反转」这两类风险。若题库全是事实型，会得出「引用率低」的结论，却看不到推荐型场景里可能的优势；若全是推荐型，则数字无法核，观测失去意义。

配比还要与业务目标对齐。做品牌曝光时，推荐型与比选型的样本应更多；做参数准确性治理时，事实型与否定型更关键。**题库配比本质上是在定义「你要观测什么」**，而不是在定义「你要问什么」。这一步没想清楚，后面四个指标算得再准，也只是在衡量一个错的东西。

采样设计：三个必须固定的变量

第一，**题库固定。**同一批问题反复问，才谈得上纵向比较；换题就等于换考卷。第二，**重复次数固定。**生成带随机性，同题问一次的结果不足以代表分布，常用做法是每题重复三次并记录离散程度，而不是只取一次结果。第三，**时点固定。**同一时段的集中采样，才能把「内容变化」和「模型更新」区分开——跨时段的差异，很可能来自模型侧，而不是你的内容。

此外还要记录两个容易被漏掉的字段：提问措辞（同一意图换一种问法，候选集可能完全不同）与地域限定（带地区限定的问题会改变取源偏好）。这两项不记录，数据就没法解释。举个具体例子：同样问「20 万左右买什么车」，加不加一个城市限定，候选来源可能换一批；如果不记录这个变量，你会把地域差异误读成内容波动，然后围绕一个不存在的问题反复优化。

固定变量的意义在于让比较有意义。观测的目的不是积累一堆数字，而是能回答「这次和上次比，哪里变了」。若三件事都没固定，两次采样的差异就无法归因，数据只能用来讲故事，不能用来做判断。

一个最小可行实验

1. **定题**：20—30 个真实比选型问题，覆盖自家产品所在的核心场景，每题重复 3 次。
2. **采集**：记录每题每次的来源列表、答案中出现的实体、答案句与挂链的对应关系、提问时点与措辞。
3. **判定**：按四个指标分别计数，不合成总分；精度必须人工逐句核，不做自动猜测。
4. **复跑**：每两周一次，同题库同时段，看趋势而不是看单点。
5. **归因**：把「从不出现」的题按第 02—05 章逐条对照，判断卡在哪一环。

这套实验的成本不高，但能回答一个关键问题：**你在哪些问题上稳定被引，在哪些问题上从不出现。**把「从不出现」的那批问题单独拎出来，再按第 02—05 章逐条对照——是切不出来、进不了配额、权威不够，还是时效地域不成立——修复动作就有了明确指向，而不是继续加内容。

从数据到归因：四象限

把「是否出现」与「是否被正确支撑」两个维度交叉，可以得到四种情形，各自对应完全不同的动作。这个象限的价值在于避免一刀切：同样是「被引了」，有的需要维持，有的必须立刻修。

象限	表现	动作
被引且支撑	来源出现，逐句核「对」	维持更新节奏，防止时效出局；把该写法复制到同类问题上
被引但不支撑	来源出现，逐句核「缺」或「错」	最高优先级修：改写法、补限定、统一口径
未引但内容在场	来源不出现，站内确有对应内容	排查第 02、03 章：是否切不出、是否未进配额
未引且无内容	来源不出现，站内也没有对应内容	先补内容，属于供给缺口而非优化问题

表 7-2 引用状态四象限与对应动作（机制推演）

四象限里最容易被漏掉的是第二象限：被引了，但支撑不完整。它在汇总指标上表现为「引用率正常」，因此最容易被当成健康状态，而实际上它是最危险的——错误一旦被引用，会持续复制，且纠正成本远高于未引。观测时把这一格单独统计，比总引用率更有诊断价值。

记录表怎么设计

题号	重复次	来源列表是否含自家	答案句与挂链是否对应	实体提及次数
Q01	1 / 2 / 3	是 / 否	对 / 错 / 缺（逐句）	自家若干、竞品若干
Q02	1 / 2 / 3	是 / 否	对 / 错 / 缺（逐句）	自家若干、竞品若干

表 7-3 采样记录表的最小字段（结构示意，不含虚构数据）

这张表看着朴素，但它解决了三个具体问题。第一，一行一题一次，重复之间的离散程度可以直接看出来；第二，「对 / 错 / 缺」的标记把精度问题暴露在明细里，而不是藏在汇总里；第三，实体提及次数与来源是否含自家分成两列，避免把提及当成引用。有了这三列，一份十题三重复的数据就足以定位问题，不需要额外的分析工具。

四种常见误判

把波动当趋势：单次采样的差异很正常，用一次结果宣布「优化生效」或「优化失效」，都会误导后续动作。**把提及当引用**：答案里出现品牌名但没有挂链，只能算曝光，不能算证据链成立。**把来源条当支撑**：来源列表包含你，不等于你的句子支撑了某个结论，精度必须逐句核。**把跨引擎结论外推**：不同引擎的语料与检索配置不同，在豆包上的结论不能直接搬到别的平台，反之亦然。

这四种误判有一个共同的根源：把「可观测的界面信号」当成「机制层面的结果」。界面上能看到来源列表、能看到品牌名，这些信号容易获得，于是被当成了结论；而机制层面的「句子被采纳并被正确支撑」，需要逐句核对才能确认。观测的价值不在于收集界面信号，而在于把信号与机制对上。

观测周期与决策节奏

把观测做成「仪表」而不是「判决书」，是这个环节最容易被做坏的地方。仪表的用法是定期看趋势、发现异常再查原因；判决书的用法是拿一次结果给人定对错。前者能持续指导动作，后者会在第一次波动时被迫解释一个并不存在的结论。建议的节奏是两周一次、连续三次才算一个趋势，中间单次差异只记录不结论。

这个周期与内容更新的周期应该错开。内容更新后立刻观测，看到的差异可能只是「新内容尚未被重新取源」，与优化效果无关；通常需要留出足够的时间让检索侧重新抓取与判定。反过来，若观测节奏远慢于更新节奏，就无法判断哪一次改动起了作用。**两条节奏对齐，才谈得上归因**，否则只能凭印象。

指标之外的两个观察维度

四个指标能回答「有没有被引、引得对不对」，但有两个维度它回答不了，而这两个维度在实际判断里很重要。**其一是答案的结构位置**：同一句被引，出现在答案开头的结论段，与出现在末尾的补充说明里，价值差别很大；观测时应记录「在第几段、属于结论还是补充」。**其二是同问题多答的稳定性**：同一问题重复三次，如果三次引用的都是同一句，说明这句话已稳定进入证据链；如果三次引的是三个不同句子，说明内容被采用但尚未定型，这时再改写法风险很高。

两个维度都不需要新增工具，只需要在记录表里加两列。加上之后，一份数据的解释力会明显不同：你能区分「偶发被引」与「稳定被引」，也能区分「被当作结论」与「被当作备注」。这两组差别决定了后续动作——偶发的要查时效与配额，稳定的要防漂移；当作结论的可以复制到同类问题，当作备注的则需要改写法。

08 问法决定取源

前面七章讲的是「内容侧怎么走通链路」，这一章换一个视角：提问本身会改变候选集。同一件事，用四种问法问出来，取到的源可能完全不同。理解这一点，能让内容布局从「写一篇大而全的文章」变成「准备一套能回答四类问法的证据」。

四种意图形态

问法	典型措辞	取源偏好	内容风险
事实型	「多少钱」「多少公里」「什么时候上市」	单一、精确、带时点的数值来源	容错最低；一个数字错，整句作废
比选型	「A 和 B 哪个更合适」「同价位有什么选择」	可对照的多源，结构化参数	候选最多、竞争最激烈；需要横向可比的口径
推荐型	「推荐几款」「有什么值得买的」	带评价与榜单的内容，垂类站点占比高	容易被「榜单化」写作拖入无依据排序
否定型	「有没有不支持快充的」「是不是没有 800V」	含明确否定或边界说明的内容	最容易被截断反转（第 06 章）

表 8-1 四种问法与取源偏好（机制推演，非实测统计）

这四类的差别不只体现在措辞上，也体现在检索参数上。事实型问题通常不需要很宽的召回，但要求精确与时效；比选型问题需要更多的候选做对照，因此对结构化程度要求最高；推荐型问题依赖评价性内容，垂类站点的权重更明显；否定型问题则对否定词与限定条件极其敏感——只要句子被切开，结论就会反向。

事实型：容错最低

事实型问题是最不该「写得漂亮」的一类。它只需要一个数，而这个数必须是当下的、带口径的、属于正确版本的。三件事都做到，句子就能被直接采用；少一件，就会被别的源替换。这也是为什么价格、参数类内容的第一原则是「四件套同行」：版本、数值、单位、时点。

反过来，事实型问题也是验证机制最方便的入口。同一批事实型问题反复问，看自己的数值有没有被采用、有没有被换成别人的数——这是第 07 章那个最小实验里最容易起效的部分，因为判定标准明确：数对不对，一比就知道。

比选型：最看结构化

比选型问题要求横向可比。如果两篇内容各自讲一款产品，但口径不同（一个写 CLTC、一个写 WLTC；一个写指导价、一个写权益价），模型在做对照时就必须自行折算，而折算恰恰是失配的高发环节。可执行的修法是给同一组对象建立统一的字段表：同一批参数、同一批口径、同一批时点。这样无论模型取哪几条，横向都能对齐。

这里有一个反直觉的判断：**比选型内容的价值不在「结论」，而在「可比性」**。给出「A 更适合家用」这样的结论，模型未必采用；给出「A 与 B 在续航、补能、价格三项上的同口径数值」，模型反而更容易引用，因为这些数值可以直接支撑它自己的对照。结论是内容方的判断，可比性是公共事实，后者更容易被当作证据。

推荐型：榜单的诱惑与风险

推荐型问题会自然吸引「榜单式」写作：十大推荐、五款必看。榜单在传播上有效，但在引用上有两个风险。其一，榜单结论缺乏可核依据——为什么排第一，通常只有主观理由；其二，榜单容易被别的榜单替换，因为替代成本极低。更稳的做法是把榜单拆成「分维度推荐」：在某个具体条件下（预算、用途、地域）哪几款满足要求，每条推荐都挂着可核参数。这样即便排序不被采用，参数本身仍可能被引用。

否定型：最容易被反转

否定型问题在机制上最危险。因为答案需要给出一个「不」的判断，而「不」的判断必须依赖完整的句子；一旦切片截断了否定词之后的部分，或者把「不支持 A」读成「不支持任何」，结论就会彻底反过来。第 06 章给的修法在这里最适用：把否定写成「采用 X，未配置 Y」的肯定加边界结构，让句子在任何长度下都闭合。

对内容布局的推论

四类问法合起来，指向一个具体的布局方案：**一套内容要能同时回答四种问法**。事实型靠「参数句」满足，比选型靠「同口径字段表」满足，推荐型靠「分维度条件推荐」满足，否定型靠「肯定加边界」满足。四类各自需要的写法不一样，但可以放在同一份内容里——参数句在正文，字段表在附录，条件推荐在独立小节，边界说明紧跟在对应参数之后。

这个方案的好处是它不需要猜测哪类问法会被问到。四类都覆盖，等于把「问法的不确定性」转化为「内容的冗余度」。冗余在这里不是浪费，因为切片只会取走它需要的那一句，其余部分不影响结果；而缺一类，就会在那一类问题上系统性缺席。

跨地域问法与本地化内容

带地域限定的提问会改变取源偏好，这一点在第 05 章已提过，落到内容布局上还有一层结论：**本地化内容应该写成两层**。第一层是全国口径的事实，例如厂方公布的指导价与工况参数；第二层是地方口径的事实，例如当地补贴、当地门店、当地政策。两层各自自包含，且都带时点与适用范围。这样同一条内容在「全国问法」与「本地问法」下都能提供可用句子。

反过来，最常见的错误是把两层搅在一起：写一个「全国最低价 9.98 万元起」的结论，其中既有全国指导价又有地方权益，读者无法判断适用条件，模型也无法判断它该不该用。搅在一起的写法在人类阅读里显得「信息丰富」，在机制层面却是双重不可核——两个口径都不完整，任何一次引用都可能出错。

09 一次完整走查：从一条内容到一句被引的话

前面八章是分机制的，这一章把它们串起来走一遍。场景设定为一个具体的、可核的例子：某车型的官网参数页，要在「这款车的续航和价格是多少」这类问题里成为可被引用的源。走查只做一件事——把四道判定逐条对照，看哪一步会掉。示例中的数值取自公开指导价与厂方公布参数，仅作写法演示，不代表任何实测结论。

第一步：写内容之前先确定「哪一句要被切走」

先假定唯一必须被引用的那句话是：车型、年款、版本、工况续航、指导价，五项齐全。写法上把它压成一句独立成立的判断，例如「某车型 2026 款 625 Max 科技版 CLTC 工况续航 625 公里，指导价 17.68 万元」；同车系的其他版本另起一句，例如「同车系 580 长续航版指导价 19.99 万元起」。两句各自闭合，不依赖上下文。

这一步最容易犯的错是把版本关系写成一段叙述：先说车系有多个版本，再说长续航如何、价格如何。这样的段落在人读时更顺，但切走之后只剩「先说车系有多个版本」——一个不支撑任何结论的句子。写法改造的核心动作只有一句：**把「关系」留给段落，把「事实」压进单句。**

第二步：让摘要里能出现这句

假设这三句分别落在页面的三个位置：第一段、参数段、价格段。由于摘要按位置截取且长度可配置[2]，最稳的布局是把「车型 + 版本 + 续航 + 价格」这组核心事实放在页面最靠前的位置，并在小标题里带上问题语汇（如「续航与价格」），让摘要段与问题形成直接对应。参数表保留，但正文不复述就会被表格吞掉——表格在纯文本转换中失去行列关系，这是第 02 章已经说明的机制。

第三步：对照召回参数，检查准入

把召回侧的七个参数排一遍，会发现三个具体动作。其一，页面发布时间必须晚于常用时效窗口的起点，否则在指定时效的检索里直接出局[2]；其二，标题与首段必须包含车型名与「续航 / 价格」这类问题语汇，否则相关度评分不会高；其三，若目标场景是垂类检索（汽车类），内容要落在汽车语义里，避免在泛化栏目里与其他品类混排。三项都属于内容侧可控范围。

第四步：可靠性——段位与多源一致

官网属于官方示例中的「正常权威」一档（中小型企业官网明确列在该级）[1]，这意味着准入上并不吃亏，真正要挣的是「这句话是否被当作可核来源」。做法是把同一组数字在三个载体上写成同一句话：官网参数页、有署名的垂媒稿、行业目录条目。三者时点一致、口径一致，模型无论取到哪一个，句子都成立——这直接降低了第 06 章所说的重写失配概率。

第五步：安全写入——限定与否定

最后一道是表述安全。价格必须拆成指导价与权益价两种口径，并带有效期；续航必须带工况；若涉及「不支持某技术」，要写成「采用 X 平台，未配置 Y 架构」的闭合结构。这三条不改，句子在任何一次切片里都可能被读成另一个意思。走查到这一步，才算把「能写」变成「能安全地被写」。

第六步：复核——把答案句对回原句

走查不能停在写完之后。真正的验收发生在答案出来之后：把答案里的每一句摘出来，找到它挂的源，核这句话是否被原句支撑。这一步不需要工具，只需要耐心；而它提供的证据，是前面五步唯一能拿出去说话的东西。

第七步：把走查固化成模板

一次走查的价值有限，能复用的走查才有价值。固化的方式不是写成一篇长文，而是把它压成一张发布前清单：六个问题，逐条打勾。**这句的主体在句内吗；版本与价格同句吗；工况与单位同句吗；结论在段首吗；同一事实在另外两个载体上写法一致吗；否定与限定是否闭合。**六项全过，才允许发布。清单越短越可能被执行，六项已接近上限。

清单的意义在于把「专业判断」变成「可交接的流程」。写作者、审核者、运营者不必都懂机制，只要按清单核对，就能保证内容达到可切、可核的最低标准。这也是机制研究能落地的方式——不是让所有人理解算法，而是让关键动作变成不需要理解也能执行的条件。

步骤	对照的机制	要检查的具体项	失败信号
一	句级引用、自包含	核心句是否含车型、版本、工况、数值、价格	摘出后需回读上文
二	摘要与切片	核心事实是否在页面靠前位置；表格是否在正文复述	摘要里只有背景句
三	配额与参数	发布时间、标题语汇、行业语义	同站其他内容出现而本篇不出现
四	权威分级、多源一致	三个载体是否同句同口径同时点	只有一处这么写
五	安全写入	价格双口径、续航带工况、否定闭合	限定条件在脚注
六	句级复核	答案句能否逐字对回原句	标「缺」或「错」

表 9-1 六步走查的对照项与失败信号（示例数值仅作写法演示）

走查的价值在于它把「优化」还原成了一串可核的动作。每一步都有明确的检查项，每一步失败都有一个具体的信号，不需要依赖「感觉这次做对了」。这比任何一套技巧清单都更稳定，因为它检查的是机制层面的条件，而不是某种写法的表象。

10 内容侧能改的与改不了的

机制的用处之一是划清边界：哪些变量在内容方手里，哪些不在。分不清这一点，投入就会持续打在无法改变的地方，然后得出「GEO 没用」的结论。

可控清单

可控项	为什么可控	对应章节
句子的自包含程度	写作阶段完全自主决定	第 02 章
结论的位置	段首或段末由作者决定	第 02 章
发布时间与更新节奏	元数据由发布方填写	第 03 章
标题与问题的对应	标题由作者决定，直接参与早期判别	第 03 章
口径一致性	跨载体统一写法属于内容治理	第 04、05 章

可控项	为什么可控	对应章节
限定条件的写法	是否与数值同句，由作者决定	第 05 章

表 10-1 内容侧可自主控制的六项变量

口径治理的三个层级

口径不一致是失配与漂移的共同来源，治理要分层做。**页面层**：同一页内单位、年份、工况写法统一，避免同页两个口径；**站点层**：同一事实在站内各页写法一致，尤其是产品页、新闻稿、帮助中心三处最容易各写一套；**行业层**：与公开可核的第三方口径对齐（如厂方公布参数、监管口径），避免自造一套只有自己能懂的表述。三层里，页面层最容易做，行业层最难但收益最大。

分层的意义在于责任可分。页面层由写作者负责，站点层由内容治理负责，行业层由业务方与法务共同确认。**不分层的结果是所有口径问题都压给写作者**，而写作者既没有权限改别的页面，也无法决定行业口径——这是口径问题长期治不好的组织原因，不是能力原因。

不可控清单

与可控清单相对，有五件事不在内容方手里，试图控制它们只会浪费预算。**其一，候选池规模**：20 或 50 是产品或调用方的参数，内容方无法改变。**其二，召回排序的权重**：没有公开，也无法通过外部操作影响。**其三，权威分级的判定标准细则**：官方只给了四级描述与示例，具体归级规则未公开。**其四，答案的合成方式**：模型如何拼接多个源、如何决定挂哪一个，属于生成侧。**其五，用户提问的措辞**：同一意图有无数种表达方式，内容方无法预设。

不可控不等于不重要，而是说**对它们的正确姿态是「适配」而不是「操纵」**

。适配的做法是：不猜权重，而是把内容做成各种权重假设下都不吃亏的形态——自包含、口径统一、限定明确、时效在手。这四件事在任何一种排序方案里都是加分项，因此不必知道具体权重也能做对。

投入优先级：按通过率排

资源有限时，投入应该按「哪一道判定的通过率最低」来排，而不是按「哪件事听起来最专业」来排。判断方法很直接：用第 07 章的观测，看自己最常卡在哪一环。切不出句子（第 02 章）是最常见的，也是最便宜的；进不了配额（第 03 章）次之；可靠性（第 04 章）通常需要跨部门协作，成本最高但见效慢；安全写入（第 05 章）是一次性的表述改造，性价比往往被低估。

有一种常见的顺序错误：先做权威度建设，再考虑写法。这条路径的问题在于，权威度建设的周期以月计，而写法改造以天计；若写法不改，即使段位抬上去，内容仍然切不出句子，收益被前一道判定吃掉。正确顺序是**先修便宜的、通过率低的环节，再做周期长的。**

更新节奏的经济学

时效筛选把内容变成了一种会过期的资产。这带来一个与直觉相反的结论：**少量高频更新的内容，通常优于大量一次性的内容。**原因在于前者始终处在可被筛中的窗口内，后者一旦过期就整批失效。更新也不需要重写全文，只需要换掉核心事实：价格、参数、政策、时点。一份三千字的内容，维护成本可能只有写作成本的一小部分。

四种无效努力

堆关键词：相关度评分看的是能否回答问题，堆词会破坏可读性，反而降低被切中的概率。**追热点**：热点内容往往与核心业务无关，进了候选也不支撑任何与业务相关的结论。**只发通稿**：通稿口径统一、没有具体数字，切不出证据句。**把所有数据藏在图里**：转成纯文本后信息全失，等于没写。四种努力的共同特征是一样的：它们都在「看起来在做事」，但没有改变任何一道判定的结果。

11 边界：哪些能说、哪些不能说

机制研究最怕的不是「知道得少」，而是「把推断说成事实」。这一章把公开与未公开划清，把本册未核验的项列出，也把没有回答的问题写明白。

三栏清单

有公开依据（可引）	官方未公开（只能推断）	不可断言（禁止承诺）
联网搜索的触发条件与来源展示位置[4]；摘要、正文、rerank、时效、域名、行业、权威分级等返回与控制字段[2]；站点权威度四级描述与示例[1]；引用格式与末尾参考资料的工程要求[6]；关键名词百科链接能力[5]	豆包 App 每次回答的候选池规模；摘要截取算法；四级权威度在答案生成中的实际权重；字节生态内容的加权幅度；多轮检索的轮次上限	「100% 被引用」；「固定第几位」；「某平台必然优先收录」；「按某套公式可精确预测被引概率」

表 10-1 公开事实、合理推断与不可断言的边界（依据见参考来源）

本册未核验的项

按研究纪律，未核项要显式写出，而不是含糊带过。本册以下五项属未核验：其一，豆包 App 的每一次回答是否都严格走「摘要」这一形态，本册只确认了 API 侧返回摘要与正文；其二，四级权威度在答案生成中的实际权重，官方文档只说明它是可控制的检索参数与标注字段；其三，字节生态内容（如抖音、今日头条）在 App 答案中的加权幅度，本册未找到公开的量化口径；其四，多轮检索的轮次与预算；其五，行业内流传的若干量化数字（如抓取时效倍数、内容衰减比例），本册未采用，也未找到可核出处。

把未核项写出来，不是自曝其短，而是让读者知道哪些结论可以直接用、哪些只能用来看方向。尤其在对外提案的场合，这条界线决定了你是在讲机制，还是在许愿。

三条不能承诺的事

第一条，**不能承诺被引用。**引用是候选池、相关度、可靠性、生成改写四道判定的联合结果，其中至少两道的细节未公开，任何「保证被引」的说法都缺乏可核依据。第二条，**不能承诺位次。**同一问题在不同时点的答案可能不同，位次本身不是一个稳定对象。第三条，**不能承诺跨平台一致。**不同引擎的语料结构、检索配置与生成策略不同，把豆包的结论套到别的平台，属于无依据外推。

反过来说，有三件事是可以承诺的：把内容改成可切形态、把关键事实做成多源一致、把时效与地域口径写进句子。它们的效果可以被观测，也可以被复核。这也是本册愿意给出的全部承诺。

对外表述的三条建议

把本册拿去对外沟通时，有三条表述建议。其一，**说机制，不说排名**：可以说明候选、摘要、权威、生成四个环节的约束，不要承诺位次。其二，**给条件，不给结论**：任何判断都带上适用条件（接入方式、时效窗口、行业语义），条件在，结论才站得住。其三，**给方法，不给秘籍**：可交接的清单比一套技巧更有说服力，因为它可被对方复核。

这三条本质上是一种姿态：把不确定的部分说清楚，把确定的部分做实。在任何需要长期合作的场合，这种姿态比夸大效果更能建立信任——尤其当对方的技术人员会去核对文档原文时，过度承诺几乎是必然翻车的。

本册没有回答的问题

为了把覆盖度说清楚，这里列出本册有意留白的问题：不同行业（医疗、金融、教育）在这四道判定上的权重差异，本册没有做分行业实测；同一个事实在多少个独立来源上出现才算「足够一致」，缺乏可引的量化门槛；引用份额与商业转化之间的关系，属于效果评估问题，不在机制研究范围；以及多引擎对照的方法论，需要单独的样本设计。这些问题不是不重要，而是它们需要各自的观测数据，硬塞进本册只会得到没有依据的段落。

最后一句判断

> **机制的用处不在于猜中算法，而在于把失败拆成可分开处理的小问题。**截断、漂移、失配，三种失败对应三种动作；混着谈「优化」，只会得到一堆彼此矛盾的技巧。把这四道判定按顺序过一遍，内容方至少能知道自己卡在哪一关——这比多写十篇文章更接近答案。

附录 A　核验清单

本册的结论若要对外使用，建议先按下面这张清单自查一遍。清单只列可核对的事项，不列判断，因此任何人都能执行。

序号	核验项	判定标准
1	核心句是否可以脱离上下文	单独抽出后不需要读者回读上文
2	版本、数值、单位、时点是否同句	四项齐备，缺一即不合格
3	表格内容是否在正文复述	纯文本转换后信息不丢失
4	同一事实是否有两至三个载体同写法	逐字比对无明显差异
5	否定与限定是否闭合	无从半句读出反向结论的可能
6	引用句能否对回原句	逐句核「对 / 错 / 缺」，错与缺均为不合格

表 A-1 发布前后的六项核验（对应第 02、04、06、07 章的判定）

清单的第 6 项要放在发布之后，因为它的对象是答案而不是内容。前五项属于内容侧自查，第 6 项属于效果侧核验；两者合起来，才构成一次完整的闭环。只做前五项，容易得到「我写得很规范但没人引用」的困惑；只做第 6 项，则会发现问题却不知道该改哪里。

附录 B　术语表

本册使用的术语多数沿用了检索与学术侧的既有定义，少数为叙述方便作了限定。下表给出本册口径，避免与行业内其他用法混淆。

术语	本册口径
召回	从索引中取出候选网页的过程，受条数、时效、域名、行业与权威级别等参数约束[2]
配额	单次检索返回的候选条数上限，Global 20 条、Custom 50 条[2]
rerank	对召回结果按与问题的相关度重新排序，官方文档说明会返回相关度评分[2]
摘要（候选文本）	检索结果中与问题相关的网页内容片段，长度可配置，是模型最先读到的文本[2]
引用召回率	每一句是否至少有一个引用、且引用合起来能支撑该句（ALCE 定义[7]）
引用精度	所挂引用是否真的支撑对应那句话（ALCE 定义[7]）
权威分级	按站点属性划分的四级标注：非常权威、正常权威、一般权威、一般不权威[1]
重写失配	生成环节把多源信息合成一句后，所挂引用只支撑其中一部分
时效窗口	检索时可指定的发布时间范围，超出窗口的内容不进入候选[2]
站点准入	域名是否备案、是否被指定或屏蔽、是否满足权威等级要求[2]
行业领域检索	把候选限定在某一领域的检索能力，Custom 版提供[2]
摘要长度	检索返回的摘要片段可配置长度，长度越大，可被模型读到的上下文越多[2]
曝光	品牌或内容出现在答案里但未被挂链，只形成可见度，不构成证据链
采纳	答案采用了某个判断，但没有给出可核的来源指向
背书	答案引用你的内容作为支撑，使某个结论具备可追溯依据
引用份额	同一问题里自家被提及次数占全部提及次数的比例
位置权重	引用出现在答案首段结论还是末尾补充说明，价值不同
实体漂移	数值真实存在，但被挂到了错误的版本或对象上

表 B-1 本册术语口径（标注来源者为沿用官方或学术定义）

术语表只统一本册口径，不替代官方定义。凡与官方文档表述不一致处，以官方文档为准；本册的限定只用于说明机制，不作为对外承诺的依据。

术语的取舍有两条原则：只收本册正文反复使用的词，不收仅出现一次的专有名词；只在官方或学术已有定义时才标注来源，其余明确写为「本册口径」。因此像「GEO」这类已被行业广泛使用但定义并不统一的词，本册不做重新定义，只在使用处限定范围。

如果读者发现某个词在本册与别处的用法冲突，建议以使用处的上下文为准，而不是以术语表为准——术语表的作用是消除歧义，不是建立标准。真正的标准仍在官方文档与学术论文里，本册只是把它们翻译成内容侧能用的说法。

参考来源

[1] 火山引擎：《站点权威度分级说明》，豆包搜索官方文档，最近更新 2026-08-21。

[2] 火山引擎：《豆包搜索产品简介》（Global 版与 Custom 版功能对比），官方文档，最近更新 2026-10-15。

[3] 火山引擎：《豆包搜索 Custom版》接口参数说明，官方文档，2026 年。

[4] 豆包：《用户协议》第 2.2.2 条（联网搜索意图识别、来源展示与第三方内容免责），2026-09-10。

[5] 火山引擎：《联网问答 Agent 产品简介》，官方文档，2026-06-03。

[6] 火山引擎：《私域知识库搜索 Knowledge Search》，火山方舟官方文档（引用格式 [1](URL) 与末尾参考资料要求）。

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

[8] Aggarwal P. 等：《GEO: Generative Engine Optimization》，ACM SIGKDD 2024，arXiv:2311.09735。

[9] 《Chain-of-Thought Improves Text Generation with Citations in Large Language Models》，AAAI 2024。

[10] 《On the Capacity of Citation Generation by Large Language Models》，arXiv:2410.11217。

[11] 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）》，2026 年 9 月（截至 2026 年上半年我国生成式人工智能用户规模突破 7 亿、普及率超 50%）。

星阙实验室GEO研究中心 | 本册依据官方公开文档与学术论文编制，推断部分已逐条标注，仅供参考