

检索源拿材料步骤

▼ 检索源拿材料的基础逻辑

智能体到了 Step 3，本质就是：

用前一步生成的 query，去各种“知识源”里找证据片段（chunks），再喂回模型。

RAG/Agent 的标准做法是：

多路检索 → 候选集合 → 排序/筛选 → 输出 TopK 证据。(Elysiate)

你要理解：

模型最后讲什么，取决于它此刻能检到什么、检到的东西排第几、内容能不能被当证据用。(Elysiate)

▼ 检索源（Sources）通常有哪些？

国内豆包 / DeepSeek / Kimi / 文心 等联网智能体，一般会混用这几类源：

1) 公网搜索源（Web Search）

- 新闻站、媒体报道
- 百科/词条
- 社区内容（知乎、B站、小红书、论坛、博客等）
- 官网/产品页/白皮书

智能体会把搜索引擎返回的网页当候选材料。(DeepWiki)

对 GEO 最关键的点：

只要你不在这些“可抓取区”，模型就根本检不到你。

2) 平台自家知识库（Platform RAG）

每家平台都会有一些“内置知识库 / 高可信索引”，包括：

- 自建百科/结构化数据
- 高质量站点镜像

- 与自家生态打通的内容池

RAG survey 里明确：平台会把高质量外部知识做索引，然后检索增强生成。
([arXiv](#))

GEO 含义：

你要进“[平台偏爱的可索引内容池](#)”——结构化、权威、可复用。

3) 应用/企业私域知识库 (Private RAG)

尤其是通义/千问、文心千帆、DeepSeek 企业版常见：

- 客户自己上传的文档
- 产品手册、政策、内部 wiki
- CRM/数据库内容
- 行业报告包

Agentic RAG 实践里，私域检索是和公网一起路由的常见模式。(Sajal Sharma)

GEO 含义：

你不仅要公域 SEO/GEO，还要提供[标准知识包](#)让客户私域一键导入。

▼ 这一步内部怎么跑？（更细的流水线）

把 Step 3 再拆成 6 个小动作：

3.1 选择检索路由 (Retriever Routing)

智能体会决定：

- 先搜公网？还是先搜私域？
- 多个源并行搜？还是按优先级串行？
- 要不要用不同 retriever（关键词/向量/混合）？

生产级 RAG 普遍会做“路由 + 多检索器融合”。([Elysiate](#))

3.2 召回候选材料 (Candidate Retrieval)

常见三种召回方式：

1. 关键词召回 (BM25 / 搜索引擎词匹配)

- 对“标题/关键词强相关内容”友好

2. 向量召回 (Embedding 相似度)

- 对“语义相关但不含同词”友好

3. 混合召回 (Hybrid)

- 生产系统主流：关键词 + 向量一起拉候选，再融合排序 (Elysiate)

GEO 对应：

你要同时满足：

- 关键词可命中 (SEO-like)
 - 语义相似可命中 (GEO-like)
-

3.3 候选融合与去重 (Fusion & Dedup)

多路召回的结果会：

- 合并成一个候选池
- 去掉相似重复文档
- 过滤明显噪声

这是生产 RAG 的标准步骤。 (Elysiate)

3.4 重排 (Reranking)

候选池会被 reranker 再打分排序 (TopK)。

reranker 可能是：

- 交叉编码器 (cross-encoder)
- 小模型/规则
- 多目标排序 (相关性 + 权威度 + 时效)

生产 RAG 强烈依赖 rerank 提升证据质量。 (Elysiate)

GEO 重点：

哪怕你被召回了，重排不给你高分，你也进不了 TopK 证据。

3.5 切块与证据化 (Chunking)

网页/文档会被切成 chunks (几百字级别)

每块都带 metadata (来源、时间、标题、authority 等)，供模型引用。
([Elysiate](#))

GEO 重点：

你的内容要“切成块也有意义”：

- 小块里就能看出你是谁、做什么、优势啥
 - 否则被切开后语义残缺，rerank会降权
-

3.6 输出 TopK 证据片段

最终输出的就是：

- 排名前 K 的 chunks
- 连同来源/时间/摘要

喂回模型进入 Step 4/5。([DeepWiki](#))

▼ 这一步 GEO 能具体怎么打？

对应上面的小动作，你可以在 Step 3 发力这 4 件事（按强度排）：

A. 让你能被召回 (Recall)

- 公网覆盖你 + 核心 query 形态
- 多平台可抓取分发
- 同义词/别名/场景词一起出现
- 技术/数据/对比内容提高语义召回

B. 让你在 rerank 里得高分 (Rank)

rerank常用的隐含特征包括：

- 权威度（大站/媒体/论文/百科）
- 一致度（多源同说法）
- 信息密度（像结论）
- 新鲜度（尤其新闻/榜单）

所以你要：

- 提高权威来源占比
- 让不同平台的“定义/定位”一致
- 结论前置、分点清晰、带数据

C. 让你切块后仍是“好证据”(Chunk Quality)

- 小段里就完整表达你是谁/做什么/价值
- FAQ / 榜单 / 对比表这种天然可切块内容
- 避免长文只在开头提到你，后面全是泛泛讨论

D. 让你进入平台/企业私域 (Private Pool)

- 提供标准知识包（产品介绍 / FAQ / 对比 / 案例 / 风险澄清）
- 让客户一键导入其智能体知识库
- 私域里你就是“最高权重权威知识”

“系统化的指标组”(可放 GEO 产品)

可以把 Step3 的 GEO 成效做成一组可视化指标：

1. 检索命中率 Recall@K

- TopK 候选里出现你内容的比例

2. 重排后保留率 Keep@K

- rerank 后 TopK 证据中你还剩多少

3. 排序位次 AvgRank / P50Rank

- 你平均排第几

4. 证据密度 EvidenceDensity

- 进入 TopK 的 chunk 中含“可引用关键点”的比例

5. 私域渗透率 PrivateHit

- 客户私域 RAG 中是否收录你、被调用频

▼ 3.1 检索路由 (Retriever Routing)

这一步在做什么？

智能体决定“去哪搜、先搜谁、并发还是串行”。

GEO 要注意什么？

- 如果平台先搜**私域**，你在公网再强也没用。
- 如果平台先搜**权威源/平台生态源**，普通自媒体权重就低。

怎么提高？

1. 提高“权威源覆盖”

- 让你出现在更可能被路由优先检索的源里（百科/权威媒体/技术站/数据报告）

系统化抓手

- 为每个平台维护一张“路由偏好表”：

优先源类型（百科/新闻/技术/社区/私域） → 对应内容优先投放策略

▼ 3.2 召回候选（Candidate Retrieval）

这一步在做什么？

用 query 去拉一堆可能相关的网页/文档。

GEO 要注意什么？

召回一般是 **关键词 + 语义向量（混合召回）**，所以你必须两边都命中：

- **关键词命中**：标题/正文里真的有那几个词
- **语义命中**：就算不含同词，也和场景语义很近

怎么提高？

1. 关键词侧（SEO-like）

- 覆盖模型常生成的 query 形态：
 - “XX 是什么”
 - “XX 推荐/排名/对比”
 - “XX 怎么用/教程”
 - “XX 优缺点/测评”
- 保证这些标题/正文里**明确出现你 + 赛道词 + 场景词**

2. 语义侧 (GEO-like)

- 反复让你和一组“核心语义标签”同框出现
(行业位置、能力边界、典型场景、对标对象)
- 用结构化表达**标准答案块**强化语义稳定性 (定义/分点/数据)

▼ 3.3 候选融合去重 (Fusion & Dedup)

这一步在做什么？

多路检索结果合并、去重复、丢噪声。

GEO 要注意什么？

- 如果你的内容和别的来源“长得像垃圾/广告”，会被去掉。
- 同一观点多处一致出现，会被保留并加权。

去重怎么判断“重复”？

常见规则 (不一定公开，但业界一致)：

- 文本相似度太高 (比如 80% 以上)
- 标题/段落结构几乎一致
- 同一个源的重复抓取
- 站点质量过低也会顺便过滤

怎么提高？

1) 别只靠“同一篇文章到处转载”

如果你只靠一篇稿子被几十个小站转载：

- 去重时它们会被当成“一个来源的重复”
- 最后可能只剩 1 条
- 你的证据量就很薄

✅ 正确做法：

让不同平台出现**“相似主题但不同写法/不同角度”**的内容。

2) 多源一致 ≠ 重复

这是关键区别：

- **重复**：同一篇文章搬来搬去
- **一致**：不同人/不同站点，用不同表达说同一个结论

模型会把“一致”当成**共识证据**，反而加权。

✅ 所以你要做的是：

定位一致、表达多样。

3) 内容要“像信息，不像广告”

融合去重阶段常会直接过滤低质/营销感强的内容。

如果你的内容：

- 空话多
- 只有宣传不讲事实
- 没有数据/论据

它可能直接被丢掉。

✅ 所以 GEO 内容要：

- 事实清晰
- 结构化
- 有可验证细节

系统化抓手

- **Consistency Score（一致性分）**
 - 统计不同来源里你的**定位/优势/场景**的重合度
-

▼ 3.4 重排（Reranking）

这一步在做什么？

候选池再打分，选 TopK 证据。

GEO 要注意什么？

就算你被召回了，也可能被 rerank 降权。

rerank 隐含偏好通常是：

- 权威度高的
- 时效新鲜的
- 信息密度高的
- 多源一致的

重排在看什么？（模型怎么给分）

不同平台细节不公开，但行业里重排普遍会综合这几类信号：

1) 相关性（最基础）

这条资料跟用户问题是不是“真相关”，而不是擦边。

2) 权威度/可信度

来源是否靠谱：

- 权威媒体、百科、论文、官网、知名垂直站点

| 分更高

- 小站采集、自媒体软文

| 分更低

3) 新鲜度（时效）

尤其是行业、产品、榜单类问题。

更新越新、越可能排前。

4) 信息密度（像不像“答案”）

同样 500 字：

- 有定义、有结论、有分点、有数据

| 分更高

- 全是泛泛描述、情绪、广告

| 分更低

5) 多源一致性（共识）

如果很多来源都表达同一个结论：

模型会觉得“这是大众共识”，更敢用。

怎么提高？

1. 提高权威度

- 进入：百科、权威媒体、行业报告、技术站点、知名作者

2. 提高新鲜度

- 持续更新：年度/季度榜单、最新案例、最新数据

3. 提高信息密度

- 前 200 字就能看懂你是谁/做什么/优势啥

4. 提高“可引用性”

- 结论 + 依据 + 数据/例子
- 模型更敢用，不怕出错

系统化抓手

- Rank@K（重排后TopK里你的位置）
- Authority Score（来源权威分）
- Freshness Score（内容新鲜度分）

▼ 3.5 切块与证据化（Chunking）

这一步在做什么？

把网页切成小段（chunks），每段独立参与排序/引用。

GEO 要注意什么？

你的文章可能整体很好，但被切开后就不“像证据”了：

- 只在开头提你一次
 - 后面全是泛行业内容
- 切块后你的 chunks 可能没信息价值，被丢弃。

怎么提高？

1. 让“每个小段都自洽”

- 一个 chunk 内就能看出：

- 你是谁（定位）
- 你做什么（能力）
- 你凭什么（优势/数据）

2. 用天然可切块结构

- FAQ
- 榜单/列表
- 对比表
- “结论-理由-数据”三段式短文

3. chunk 里要有“可引用结构”

- 模型最爱这类：
- 一句话定义
- 分点优势
- 场景列表
- 对比列表
- 数据结论

系统化抓手

- **Chunk Evidence Density（证据密度）**
 - 进入TopK的 chunk 中，含核心信息点的比例

▼ 3.6 输出 TopK 证据片段

这一步在做什么？

最终给模型的“可用证据集合”。

GEO 要注意什么？

- 只要你没进 TopK，后面步骤就等于零。
- 进了 TopK 但信息“不够抄”，也会被弱化。

怎么提高？

1. 确保 TopK 内多块出现你

- 不只一个来源提到你

- 形成“证据簇”(cluster)
模型会更确信你是主流答案

2. 准备“标准答案块”

- 一句话定位
- 3-5条优势
- 典型场景
直接可复制进回答

3. 提高“证据密度”

TopK 塞的都是高密度信息。

你必须做到同样字数里，信息更“硬”：

- 少废话
- 多结论
- 多可验证点
- 结构清楚

4. 抢占“榜单/对比/玩家列表”这种 TopK 高产区

因为智能体特别爱把这些当证据：

- 行业 Top 玩家列表
- A vs B 对比表
- 优缺点总结
- 推荐榜单

你只要稳定出现在这些结构里，

TopK 命中率会暴涨。

系统化抓手

- **TopK Share** (TopK里你的证据占比)
- **Evidence Cluster Count** (证据簇数量)