# 医疗行业
# AI 可见度白皮书

从检索到答案：机构为什么被提到，又为什么被忽略
以及一套可以自己动手的判断方法

医知GEO

莆田市好赞科技有限公司

2026 年 11 月 4 日

# 目录


前言 1

第 1 章　研究背景 2

第 2 章　AI 答案的形成机制 4

第 3 章　五类认知偏差与实证检验 9

第 4 章　判断框架与自检工具 12

第 5 章　医疗问答中的 AI 引用分布 14

第 6 章　结论与建议 17

附录 A　术语表 20

附录 B　模型清单与提问集 21

附录 C　研究方法与数据口径 23

参考文献 28

# 前言

患者获取医疗信息的入口正在迁移。过去是搜索引擎，如今越来越多人把问题直接交给生成式人工智能，由它给出结论式的答案。答案中提到哪几家机构，患者便从中选择；未被提到的机构，在这一轮里不存在。

这一变化对医疗机构提出的要求，与搜索时代并不相同。搜索结果按位置排序，位次靠后仍有被看到的可能；生成式答案只容纳少数几个名字，未被纳入即不在候选范围之内。机构若沿用搜索时代的做法，往往难以判断投入是否有效。

关于 AI 可见度，现有的材料大致分两类：一类阐释概念，读者知其然而不知其所以然；一类推介工具，默认读者已接受某一方案。介于两者之间的部分——答案如何形成、医疗行业受哪些额外约束、机构自身可以先行开展哪些工作——论述较少。本报告围绕这一部分展开。

# 第 1 章 研究背景

## 1.1 入口的变化

截至 2025 年 12 月，我国网民规模 11.25 亿人，互联网普及率 80.1%；生成式人工智能用户规模 6.02 亿人，普及率 42.8%；互联网医疗用户规模 4.11 亿人[5]。

三个数字并置，指向一件事：每两个上网的人里，接近一个已在使用生成式 AI；线上寻求医疗信息的人群规模，与生成式 AI 用户处于同一量级。两者重叠的部分，就是「用 AI 问就医问题」的人群。

供给侧同样可观。据《2025 年我国卫生健康事业发展统计公报》，2025 年末全国医疗卫生机构总数 1 108 335 个，全年总诊疗人次 106.5 亿[6]——而 AI 的答案只能容纳其中两三个名字。

对医疗机构而言，变化发生在供给侧：过去争的是搜索结果页上的位置，现在争的是答案里的一个位置。位置从页面变成句子。

## 1.2 三个特征

生成式 AI 已从概念进入规模化应用，模型、算力与工具链同步扩张[8]。在这个背景下，与搜索结果页相比，生成式 AI 的答案有三处不同。

答完即走。答案直接给出结论，用户不再逐条点开链接比对。

单答案。一次回答通常只提到少数几家机构，而非一整页列表。

有出处。多数产品在答案末尾标注来源链接，用户可回溯核对。

三条合起来决定了一件事：未被提及的机构，不是「排在后面」，而是「不在列表里」。搜索时代的位置差异是流量差异，生成式时代的缺席是候选集差异。

## 1.3 本报告的范围

三条边界，与结论同时成立。

一、本报告只讨论机构在生成式 AI 中的可见度——简单说，就是患者向 AI 提问时，答案里会不会出现这家机构——不涉及诊疗质量、技术水平与医疗服务的评价。

二、本报告不作出效果承诺。第 5 章给出的是特定时间窗内的观测数据，不构成因果关系判断。

三、本报告不提供诊疗建议。文中出现的病种与症状，仅用于构造提问示例。

医疗行业的对外文件若不作边界声明，后续每一句陈述都会被当作承诺。这三条不是免责套话，是阅读第 5 章数据的前提。

## 本章要点

一、生成式 AI 的用户规模已与线上医疗信息人群处于同一量级，入口迁移有公开统计支撑。

二、答案的单值性使「未被提及」与「排名靠后」成为两种不同性质的问题。

三、本报告只讨论可见度，不评价医疗质量，不作效果承诺。

# 第 2 章 AI 答案的形成机制

机构能否出现在生成式 AI 的答案里，由三个环节依次决定：页面先要被检索到，再通过可信度判断，最后以匹配提问意图的形态被写进答案。三个环节是串联关系，任一环节不成立，最终结果都是不可见。

本章描述三个环节的通用行为。不同模型在检索源、权重分配与训练语料上存在差异，共性部分构成后续章节的分析基础。

## 2.1 检索：问题的改写与页面的匹配

生成式引擎的答案由多个来源综合而成，其技术基础是检索增强生成（Retrieval-Augmented Generation），即先检索外部文档，再以检索结果为条件生成回答[1][2]。用户规模已在前章给出[5]。规模决定问题量——每一个提问，都会触发一次检索。

在这套机制下，用户输入的自然语言问句通常不会被直接送入检索。主流做法是先做查询改写（query reformulation），把一句口语化的长问句拆成若干检索友好的子查询，分别检索后合并结果。

**表 2-1 一次就医提问的改写过程**

| 用户原句 | 改写的子查询 | 检索目标 |
| --- | --- | --- |
| 甲状腺结节 4 毫米 要紧吗 我们这哪家医院看这个好 | 4 毫米 甲状腺结节 处理原则 | 临床知识 |
| | 甲状腺结节 挂什么科 | 就诊路径 |
| | 某某市 甲状腺 医院 | 机构候选 |

资料来源：本报告依据 [1][2] 所述机制与主流中文生成式 AI 的公开行为归纳整理，不代表任一模型的具体实现。

改写带来的第一个后果是：参与匹配的单位是页面，不是站点。

【推论 1】机构若不存在回答具体问题的页面，则在该轮检索中不进入候选集。

首页的品牌陈述、发展历程、领导视察，通常不构成对具体临床问题或就诊问题的回答。这些内容在传统搜索引擎中可能因关键词匹配而获得曝光，在生成式检索中则很难被选中——检索需要的是一个能回答问题的最小单元，而不是一家机构的整体介绍。

第二个后果与检索源的分布有关。生成式 AI 的检索覆盖面通常包括机构自有站点、政府与行业主管部门的公开信息、健康类媒体、地图与生活服务平台的商户页、社区问答平台。机构对自有站点的控制力最强，对其余来源的控制力依次递减，但后者在采信环节中的分量并不因此降低。

## 2.2 采信：来源、一致性与实体完整性

进入候选集的页面需要经过可信度判断，才会被用于生成答案。判断依据可归纳为三项。

**表 2-2 采信判断的三项依据**

| 依据 | 判断内容 | 常见失分情形 |
| --- | --- | --- |
| 来源主体 | 信息由谁发布。机构自有页面通常高于转载页与聚合页 | 关键信息只出现在第三方转载中，自有站点缺失 |
| 信息一致性 | 同一事实在多个来源上是否吻合 | 地址、电话、科室设置在不同平台互相矛盾 |
| 实体完整性 | 位置、等级、科室、资质能否被识别为一个完整主体 | 信息分散，无法拼合为可识别的机构 |

资料来源：本报告归纳。其中「添加可信引用」一项的效果，见下段所引评测。

这一判断有实证支撑。在一项针对生成式引擎的系统评测中，对来源内容做三类处理——加入可信来源的引用（Cite Sources）、加入可信来源的引语（Quotation Addition）、以定量统计替代定性描述（Statistics Addition）——在 Position-Adjusted Word Count 指标上取得 30%—40% 的相对提升，在 Subjective Impression 指标上取得 15%—30% 的相对提升；堆叠关键词（Keyword Stuffing）未进入效果最好的方法之列[1]。

三类有效方法指向同一件事：让内容可被追溯。这与把「信息一致性」「实体完整性」列为采信依据的判断方向一致，也解释了为什么堆关键词在生成式检索中收效有限——它改变的是词频，不是可追溯性。

三项同时满足时，内容被采信的可能性明显提高；任一项缺失，模型倾向于回避而非采信。

回避有两种表现形式。其一是不写入答案，即沉默。其二是写入但表述模糊，例如以「当地某三甲医院」指代，而不给出名称。两种结果对机构而言等价：患者无法据此找到这家机构。

由此得到第二个推论。

【推论 2】孤证不立。单一来源的陈述难以支撑采信，即便该来源是机构自有页面。

这一条解释了为什么官网建设不能等同于 AI 可见度建设。官网是机构唯一能完全控制的信息源，在「来源主体」一项上具备天然优势；但它同时是最容易被孤立的来源——当官网的表述在其他平台上找不到印证，甚至与其他平台矛盾时，前一项优势会被抵消。

## 2.3 生成：内容形态与提问意图的匹配

被采信的内容仍需通过最后一层判断：是否匹配提问意图。同一段文字在不同意图下的可用性不同。

**表 2-3 提问意图与可用内容形态**

| 意图类型 | 典型问法 | 可被采用的内容形态 |
| --- | --- | --- |
| 对比型 | 看这个病哪家医院好 | 多家机构的可比信息：等级、专科方向、专家、设备 |
| 交易型 | 做这个大概多少钱 | 价格区间、费用构成、医保支付情况 |
| 导航型 | 挂什么科、怎么走 | 科室名称、门诊时间、地址、预约方式 |
| 知识型 | 这个病是怎么回事 | 病因、检查项目、治疗路径、注意事项 |

资料来源：本报告归纳。

【推论 3】内容能否被采用，取决于其形态是否匹配提问意图。以品牌评价为主、缺少可比信息与具体事实的段落，在四类意图下均难以被摘取。

需要注意，生成阶段的匹配判断发生在句子层面，而非页面层面。一段三百字、通篇为价值判断的文字，即使所在页面被检索到、被采信，也可能因为摘不出可用句子而整段落空。

## 2.4 可见度漏斗

将上述环节串联，得到本章的核心结论。漏斗之名取的是逐层收窄的意思：三道判断自上而下筛，任何一层不通过，后面的判断就不必谈——被找到之后才谈得上被采信，被采信之后才谈得上被选中。

【定义 1 AI 可见度（本报告提出）】

某机构在特定提问集合下，被生成式 AI 答案提及的程度。其成立需同时满足三个条件：

一、被找到：机构页面进入检索候选集；

二、被采信：页面内容通过来源、一致性、完整性三项判断；

三、被选中：内容形态匹配提问意图，被写入答案。

三者是串联关系。设三个环节的通过概率分别为 P1、P2、P3，则整体可见的概率近似为三者之积。任一环节接近零，整体结果即接近零。




图 2-1　可见度漏斗。三层自上而下逐级收窄，任一层趋近于零，整体即趋近于零。

这个模型的主要用途是诊断。当机构反馈「AI 里搜不到本院」时，漏斗提供了三个互不重叠的排查方向，而不必笼统归因于优化不到位。排查顺序建议自后向前：先确认是否有内容在回答真实存在的问题（第三层），再确认这些内容能否与其他来源互相印证（第二层），最后确认页面是否可被检索（第一层）。自后向前的理由在于，前两层的修复不依赖技术条件，可以先做。

## 2.5 边界与例外

三项说明，用于限定上述结论的适用范围。

模型间差异。不同模型使用的检索源、对来源权威性的权重分配、训练语料均不相同，同一提问下的答案可能不一致，甚至相互矛盾。监管层面对此已有要求：《生成式人工智能服务管理暂行办法》规定，服务提供者应当采取措施增强生成内容的真实性、准确性，并对生成的内容予以标识[7]。单次、单模型的观察不足以支撑结论；有意义的度量需要覆盖多个模型与多种问法。

失真。AI 答案可能出现事实错误，主要表现为三类：多源信息拼接、时效滞后、无据补全。失真的成因往往不在网络侧，而在机构自身信息的多源矛盾——信息互相冲突时，模型只能择一，选中的未必是机构希望的那一条。医疗场景下，失真的代价高于沉默。

时效。模型知识与检索索引均随时间变化，任何基线数据都需要标注采集时间，超过一定周期后不再具备可比性。

## 2.6 与现行广告规则的边界

机制推导向一个操作建议：用具体疾病名、症状与就诊路径组织内容，因为只有具体内容才可能匹配提问意图（见 2.3）。

这条建议与现行医疗广告规则之间存在张力，需要先讲清楚。

《医疗广告管理办法》第六条把医疗广告可发布的内容限定为八项：医疗机构第一名称、地址、所有制形式、医疗机构类别、诊疗科目、床位数、接诊时间、联系电话；其中第一至六项须与《医疗机构执业许可证》载明的内容一致。同办法第七条列明表现形式不得含有的情形，第一项即为「涉及医疗技术、诊疗方法、疾病名称、药物的」[4]。

《互联网广告管理办法》第七条要求医疗等须经审查的广告未经审查不得发布；第八条禁止以介绍健康、养生知识等形式变相发布医疗、药品、医疗器械广告，并规定介绍健康、养生知识时，不得在同一页面或者同时出现相关商品经营者或者服务提供者的地址、联系方式、购物链接等内容[3]。

三条规则合起来，给医疗科普内容划出一条可操作的边界：

一、讲解疾病与诊疗路径的内容，不能同时承担推销功能；

二、机构的地址、联系方式、预约入口，不与科普内容处于同一页面或同一时点；

三、涉及机构自身的介绍，内容范围受第六条八项限定。

这构成一个结构性约束：越是能满足生成式检索的内容（具体、可核验、直指问题），越容易碰到广告规则划定的禁区；越是安全的写法（只讲机构资质、不谈疾病），越难被检索选中。两头都不是解法。

可行的做法只有一条：把科普内容与推广内容分成两类载体、两个页面、两个时点，各自守各自的规则。本报告不提供规避路径。

本节为规则文本的转述，不构成法律意见。具体发布前应经机构法务或属地卫生健康行政部门确认。

## 本章要点

一、检索的单位是页面，不是站点。没有回答具体问题的页面，等同于没有内容。

二、采信取决于来源主体、信息一致性、实体完整性三项，任一缺失即导致回避。

三、内容形态需匹配提问意图。品牌陈述在四类意图下均难以被采用。

四、AI 可见度由被找到、被采信、被选中串联构成，可逐层诊断。

五、单模型、单次观测不构成结论；失真比沉默更难被发现。

六、科普内容与机构推广信息须分离承载。这是现行广告规则的硬性要求，不是写作偏好。

# 第 3 章 五类认知偏差与实证检验

医疗机构内部对 AI 可见度的判断，常见的有五条。这五条都不是凭空来的，各自对应一部分真实情况；问题在于每一条都把局部当作全部。

本章对每条的处理方式一致：先说明它成立的部分，再指出它遗漏的部分，最后给出一项可在内部完成的检验，并附判定标准。检验结果只有三种：成立、部分成立、不成立。判定标准写死在文中，避免事后解释。

## 3.1 官网做好就够了

**想法。** 机构自有网站是唯一完全可控的信息源，把它做扎实，其余问题自然解决。

**成立的部分。** 在第 2.2 节的三项采信依据中，「来源主体」一项上，自有页面的分量高于转载页与聚合页。官网做不好，其他渠道做得再好，源头仍是虚的。

**遗漏的部分。** 官网只能覆盖第二层的一半。第一层要求页面进入检索候选集，第三层要求内容形态匹配提问意图。一座结构完整、但没有任何一页在回答具体问题的官网，在第一层就不进入候选；一座回答了很多问题、但与其他平台信息互相矛盾的官网，会在第二层被回避。

**检验。** 取本机构官网，统计两个数：独立病种页或检查项目页的数量；其中同时包含症状、检查项目、治疗路径、就诊科室四项内容的页数。

**判定。** 四项齐全的页面达到 10 个及以上，官网这一层基本过关，可以转向第二层；不足 10 个，官网不是做得够好，而是还没做完。

## 3.2 百度排名等于 AI 排名

**想法。** 搜索引擎排名是线上可见度的通用度量，百度排得上，AI 里自然也在。

**成立的部分。** 两者共享同一个前置条件：内容要被抓取、被索引。连搜索引擎都不收录的页面，通常也进不了检索候选集。

**遗漏的部分。** 两套机制的排序依据不同。搜索结果按相关性与权重排队，争的是位置；生成式答案把若干来源综合成一段话，争的是能否被当作依据（见第 2 章）。位置与依据不是同一件事，两者的排序结果经常不重合。

**检验。** 取 5 个本机构所在城市的就诊类提问。先在搜索引擎中检索，记录结果页前三位出现的机构名称；再在三个生成式 AI 中提出同样的 5 个问题，记录答案中出现的机构名称。计算两份名单的重合机构数，除以两份名单的并集机构数。

**判定。** 重合率低于 50%，两条通道基本无关，不能互相替代；50%—80%，部分重叠，需分别维护；高于 80%，在本批提问范围内可近似互推，仍须按季度复测。

上述阈值由本报告设定，用于判断两条通道能否互相替代，不是行业标准。

## 3.3 内容发得越多越好

**想法。** 内容供给是可见度的基础，多发总能被检索到。

**成立的部分。** 第一层确实以页面为单位，没有页面就没有候选资格。内容量明显不足的机构，先补量是对的。

**遗漏的部分。** 采信看的是内容能否被追溯与印证，不是篇数。同一结论在多个页面重复出现，不增加可追溯性，反而在信息一致性上制造风险。一项针对生成式引擎的系统评测显示，对内容做三类处理——加入可信来源的引用、加入可信来源的引语、以定量统计替代定性描述——在 Position-Adjusted Word Count 指标上取得 30%—40% 的相对提升，在 Subjective Impression 指标上取得 15%—30% 的相对提升；而堆叠关键词未进入效果最好的方法之列[1]。

**检验。** 列出本机构最近 12 个月发布的全部内容，按是否回答了具体问题（可指认到某个病种、某项检查或某条就诊路径）分两类，统计两类各自的篇数，以及具体问题类中被外部来源引用或转载的篇数。

**判定。** 具体问题类占比低于 30%，问题在选题而非产量，继续加量不会改善；占比达标但外部引用篇数为 0，问题在可追溯性，需要补来源与数据，而不是补篇数；两项均达标，产量才可能成为有效变量。

## 3.4 医生有名气就够了

**想法。** 专家是机构最重要的资产，有名气的专家自然会带来患者，也会带来线上可见度。

**成立的部分。** 患者选择确实高度依赖对具体医生的认知；在生成式答案中，专家姓名常作为机构被推荐的支撑信息出现。

**遗漏的部分。** 名气存在于人的记忆里，检索依赖的是页面上的字符串。模型需要把姓名与科室、机构、专长三者绑定，才能识别出「某机构的某科专家」。只有一段三行简介、没有出诊安排、没有具体病种描述的页面，提供的是一个孤立的名字。

**检验。** 在三个生成式 AI 中提出同一问题：某医院某科有哪些专家，各自擅长什么。记录答案中出现的姓名数，以及每个姓名后是否附有可核对的专长描述。

**判定。** 能说出姓名但说不出专长，属绑定不完整，问题在页面信息结构；连姓名都说不出，属未被识别，问题在基础资料缺失；姓名与专长均可核对，这一项无忧。

## 3.5 AI 说的都是真的

**想法。** AI 的回答经过训练与检索，可信度应高于网络随手一搜。

**成立的部分。** 多数产品在答案末尾标注来源链接，可追溯性高于早期的搜索摘要；对同一事实的多源交叉，通常优于单一网页。

**遗漏的部分。** 失真真实存在，且表现自然。三种典型成因：多源信息拼接、时效滞后、无据补全（见第 2.5 节）。医疗场景下，失真的代价高于沉默——沉默只是无人到访，失真可能让人白跑。失真

的源头常在机构自身：官网、地图、招聘页面、政务公开信息之间的互相矛盾，会把模型推向一个错误的选项。

**检验。** 选取 5 项可核对的事实（地址、门诊时间、科室设置、专家姓名与专长、机构等级），在三个不同模型中逐一提问，与执业登记信息或机构公开信息比对，记录一致项数与错误项数。

**判定。** 出现 1 项及以上明确错误，先修信息一致性，再谈可见度；仅出现表述模糊而无明确错误，属信息不足，需补齐关键字段；全部一致，这一项无忧。

## 3.6 五条的共同结构

**表 3-1 五条偏差的成立面与遗漏面**

| 偏差 | 成立的部分 | 遗漏的部分 |
| --- | --- | --- |
| 官网做好就够了 | 自有页面在来源主体上占优 | 第一层与第三层不在官网覆盖范围内 |
| 百度排名等于 AI 排名 | 共享抓取与索引这一前置条件 | 位置与依据是两种不同的排序 |
| 内容越多越好 | 第一层以页面为单位，需先有量 | 采信取决于可追溯性，非篇数 |
| 医生有名就够 | 专家是患者选择的关键依据 | 检索依赖姓名与科室、机构的绑定 |
| AI 说的都真 | 有来源标注，可追溯性较高 | 拼接、滞后、补全三类失真 |

资料来源：本报告整理。

五条偏差指向同一个结构性问题：把可见度当作单一变量，用一件事解释全部结果。官网、排名、产量、专家、AI 的可靠性，各自只对应三层中的一层，或其中一层的一半。

【推论 4】机构诊断自身可见度问题时，若结论指向「某一项做得还不够」，多数情况是排查顺序错了，而非投入不足。

三层是串联关系（见第 2.4 节），串联结构的整体取决于短板而非长处。先定位短板在哪一层，再决定投入，这是第 4 章自检表的用途。

## 本章要点

一、五条偏差各有成立面，问题在于把局部当作全部。

二、每条检验都给出判定标准，结果只有成立、部分成立、不成立三种。

三、第 3.2 节的重合率阈值（50%、80%）与第 3.3 节的占比阈值（30%）为本报告设定，非行业标准。

四、失真优先于沉默处理：出现明确错误时，先修信息一致性，再谈可见度。

五、串联结构的整体取决于短板，排查顺序比投入强度更重要。

# 第 4 章 判断框架与自检工具

第 2 章给出机制，本章给出工具。下列十项检查，机构可自行完成，不需要任何系统或第三方服务。

## 4.1 使用方式

由市场或宣传部门一人主责、信息部门一人配合，约 45 分钟可完成。所需条件为：一台可上网的电脑、官网后台的修改权限、能够查阅本机构执业登记信息。

判定标准写在表内，逐项对照即可，不依赖专业判断。表格最后一列留空，供填写得分。换人复检时，判定口径须一并交接，否则分值不可比。

## 4.2 自检表

**表 4-1 医疗机构 AI 可见度自检表**

| 编号 | 检查项与判定标准 | 自评 |
|---|---|---|
| A1 | 官网独立病种页数量。0 分：无，只有科室简介；1 分：1—9 个；2 分：10 个及以上 | |
| A2 | 关闭浏览器脚本后正文是否显示。0 分：不显示；1 分：部分显示；2 分：完整显示 | |
| A3 | 病种页四要素（症状、检查项目、治疗路径、就诊科室）。0 分：1 项及以下；1 分：2—3 项；2 分：4 项齐全 | |
| B1 | 三个平台（地图、点评、政务公开页）的地址电话与官网是否一致。0 分：存在矛盾；1 分：一致但不完整；2 分：全部一致 | |
| B2 | 官网是否公示统一社会信用代码、执业许可信息、机构等级。0 分：均无；1 分：部分；2 分：完整 | |
| B3 | 搜索「机构名 + 科室名」，首屏首条是否为机构自有页面。0 分：无自有页面；1 分：有但非首条；2 分：首条即自有页面 | |
| C1 | 以患者问法为标题的问答页数量。0 分：无；1 分：少于 5 个；2 分：5 个及以上 | |
| C2 | 上述页面是否在前 100 字内给出结论或数字。0 分：无；1 分：部分；2 分：多数 | |
| D1 | 就同一问题问三个模型，是否出现与本机构不符的事实。0 分：出现 1 处及以上；1 分：表述模糊；2 分：全部准确 | |
| D2 | 病种页最近一次内容更新距今。0 分：超过 24 个月；1 分：12—24 个月；2 分：12 个月以内 | |

资料来源：本报告设计。A 组对应第 2 章「被找到」，B 组对应「被采信」，C 组对应「被选中」，D 组对应第 2.5 节的失真与时效。

## 4.3 结果解读

按四组分别小计，不要先看总分。

A 组（满分 6）低于 3 分，问题在内容供给：没有可供检索的页面，后续各层无从谈起。 B 组（满分 6）低于 3 分，问题在信息一致性与资质公示：内容有，但不足以被采信。 C 组（满分 4）低于 2 分，问题在内容形态：有内容也有可信度，但摘不出可用句子。 D 组（满分 4）低于 2 分，问题在失真与时效：可能已被提及，但说得不对或说得过时。

总分参考：0—6 分，四组普遍失分，属于结构性问题；7—12 分，局部可修；13—17 分，具备基础，可进入优化；18—20 分，处于样本内较高水平。

分档阈值为本报告设定，用于区分优先顺序，不是行业标准；待多批次数据积累后校准。

## 4.4 五分钟版

只有五分钟时，只做三项：A1（有没有内容）、C1（是不是问答形态）、D1（AI 有没有说错）。

三项中若有两项为 0 分，说明问题不在细节，而在基础建设，不必先做优化。

## 4.5 复检节奏

建议每季度一次，固定同一套检查项、同一名主责人。分值的绝对值意义有限，变化的方向才是信号。

更换判定人或修改检查项后，前后分值不再可比，须重新建立基线并注明变更时间。

## 本章要点

一、十项检查覆盖漏斗三层与失真时效，可自行完成，不依赖系统。

二、先看四组分项，再看总分；短板在哪一组，决定先修哪一层。

三、五分钟版的三项中若两项为零，说明问题在基础而非细节。

四、分档阈值为本报告设定，待数据积累后校准。

# 第 5 章 医疗问答中的 AI 引用分布

本章数据取自监测系统 2026-09-04 至 2026-11-04 的实际采集记录，不是为本书专门设计的研究。它的价值在于真实——每一条都是真实的就诊类提问与真实的模型回答；它的局限也在于此，样本不按抽样设计产生。

因此本章不比较机构之间的可见度高低，只分析一件事：AI 在回答医疗问题时，引用谁。

## 5.1 数据构成

**表 5-1 数据构成**

| 项 | 数值 |
| --- | --- |
| 答案条数 | 6 362 |
| 机构数 | 5 |
| 采集窗口 | 2026-09-04 至 2026-11-04 |
| 不同问题数 | 1 198 |
| 每个问题被回答的模型数 | 4 个及以上 |
| 参与模型 | DeepSeek、豆包·联网检索、千问、文心、元宝 |
| 平均答案长度 | 724 字 |
| 给出引用来源的答案 | 5 017 条（78.9%） |

资料来源：平台采集记录，统计时点 2026-11-04。

两个数字先摆在这里：**21.1% 的回答没有给出任何来源**，而给出来源的回答，平均列出 5.3 条，中位数为 4 条。

提问由各机构的监测关键词生成，不是统一的提问集。这是本章不作机构间比较的原因。

## 5.2 引用集中在哪

全库共 33 837 次引用。按域名归类后可覆盖其中 30 802 次（91.0%），其余为长尾站点。

**表 5-2 引用来源类别分布**

| 来源类别 | 引用次数 | 占比 |
|---|---|---|
| 医疗健康门户 | 11 990 | 38.9% |
| 其余站点（含机构官网、地方信息站） | 11 635 | 37.8% |
| 政务与官方查询 | 4 077 | 13.2% |
| 百科与知识聚合 | 2 216 | 7.2% |
| 内容平台与自媒体 | 884 | 2.9% |

资料来源：平台采集记录归类，统计时点 2026-11-04。长尾站点未逐域细分。各栏典型站点：医疗健康门户 bohe.cn、39.net、haodf.com、youlai.cn；百科与知识聚合 baike.baidu.com、health.baidu.com；政务与官方查询为各级卫健委、医疗机构查询库与地方政府站点；内容平台与自媒体为公众号与门户自媒体频道。

机构自有域名的引用情况由采集系统单独统计：761 条答案引用了机构自己的网站，占全部答案的**11.9%**。此处两个口径不同——37.8% 是引用次数的占比，11.9% 是答案条数的占比，不可直接相减或相加。

## 5.3 三个可以直接用的发现

**一、每八次引用，就有超过一次落在政府或卫健委网站。**

政务与官方查询占 13.2%，高于百科（7.2%），也高于内容平台（2.9%）。第 2.2 节把「来源主体」列为采信依据之一，这项数据与之一致：在医疗领域，模型对官方来源有明显偏好，机构查询库里那条登记信息，本身就是被引用的材料。

**二、内容平台与自媒体的引用份额，只有 2.9%。**

这与很多机构的投入直觉相反。铺自媒体、发公众号，在引用份额上的贡献远低于预期。可能的原因在第 2.3 节：平台上的科普多讲泛知识，缺少具体机构、科室与就诊路径，匹配不上「哪家医院好」「挂什么科」这类意图。

**三、机构自有站点被引用的比例，是 11.9%。**

自有域名是唯一完全可控的信息源，但只有约八分之一的回答把它当作依据。其余七次多，模型宁可引用门户、百科与政务站点。这与第 3.1 节的判断一致：官网是地基，不是全部。

三条发现都止于描述。样本不按概率抽样产生，不足以支撑因果结论。

## 5.4 局限

一、样本由在服机构产生，不是概率抽样，不构成行业分布，不得外推。

二、提问由各机构的关键词生成，彼此不同，故本章不做机构间比较。

三、5 家机构集中在数个城市与专科，地域与专科偏差存在。

四、窗口为两个月，期内模型版本有迭代，跨月比较需谨慎。

五、来源分类按域名规则判定，「其余站点」内含机构官网与地方信息站，未逐域细分；机构自有域名一项由采集系统直接统计，不含分类误差。

六、未引入外部榜单作为效度参照。后续批次可将样本机构在公开榜单中的位次（如复旦版《中国医院综合排行榜》[9]）与可见度对照，检验声誉与可见度是否同向。

## 本章要点

一、6 362 条真实问答中，21.1% 未给出任何来源，给出的平均 5.3 条。

二、引用份额：健康门户 38.9%、其余站点 37.8%、政务 13.2%、百科 7.2%、内容平台 2.9%。

三、政务站点份额高于百科、远高于内容平台，与「来源主体」在医疗领域的权重一致。

四、机构自有域名被引用的答案占 11.9%。

五、样本非概率抽样，结论止于描述，不作外推。

# 第 6 章 结论与建议

前五章分别处理了背景、机制、偏差、工具与观察数据。本章把它们收拢成六条结论，再给出三条按角色划分的建议与一份行动顺序。

## 6.1 六条结论

**一、AI 可见度是三个环节的乘积，不是一次排名。**

机构常把「AI 里有没有本院」当成一个是非题。第 2 章的机制表明，它由被找到、被采信、被选中三层串联决定，任一层为零则整体为零。这不是修辞上的拆解——它决定了排查从哪一层开始，也决定了投入放在哪一层才有效。

**二、被采信取决于可追溯性，不取决于产量。**

第 2.2 节与文献 [1] 的实证一致：加入可信引用、可信引语与定量统计，能带来三成至四成的可见度相对提升；堆叠关键词不在有效方法之列。第 3.3 节补充了另一半——同一结论重复二十遍，不增加可追溯性，反而增加信息矛盾的风险。

**三、在医疗领域，官方来源的分量高于内容平台。**

第 5.2 节的数据：政务与官方查询占引用份额的 13.2%，内容平台与自媒体占 2.9%，百科居中为 7.2%。这条数据与第 2.2 节「来源主体」的判断一致，也指出了一个常被忽略的资产：机构在卫生健康行政部门登记的信息，只要准确完整，本身就是可被引用的材料。

**四、自有站点是必要条件，不是充分条件。**

11.9% 的回答引用了机构自有域名。这个数字两头都说明问题：官网重要，近八分之一的回答以它为据；官网也不够，其余七次多，模型引用的是门户、百科与政务站点。第 3.1 节的判断由此得到数据支持。

**五、内容与推广必须分开承载，这是法规要求。**

第 2.6 节已把边界写明：《医疗广告管理办法》第六条限定医疗广告可发布的内容范围，第七条禁止医疗广告涉及疾病名称与诊疗方法；《互联网广告管理办法》第八条禁止以健康科普变相发布医疗广告，且科普页面不得同页出现地址、联系方式。这不是写法偏好，越界即违规。

**六、诊断顺序比投入强度更重要。**

【推论 4】已指出：若诊断结论是「某项做得不够」，多数情况是排查顺序错了。串联结构的整体由短板决定，第 4 章的十项检查正是用来定位这块短板的工具。

## 6.2 三条建议

以下建议不依赖采购任何系统，可由机构内部完成。

**表 6-1 按角色的三个月行动**

| 角色 | 三个月内应完成 | 建议停止 |
| --- | --- | --- |
| 决策层 | 指定一名负责人对 AI 可见度负责；把季度复检纳入常规工作 | 以单次截图或单次提问判断效果 |
| 市场与宣传部门 | 建成 10 个病种页、5 个问答页；统一各平台的地址、电话、科室 | 以发布篇数作为考核指标 |
| 信息部门 | 保证主要页面在关闭脚本时可读；公示执业许可与等级信息 | 把官网当作一次性建设项目 |

资料来源：本报告建议。

三条建议的共同点是只改内部的分配与标准，不增加新的采购。做到这三条，第 4 章自检表的总分通常会有明显变化；做不到，任何外部服务都只是在补地基之前的装饰。

## 6.3 行动顺序

四步，顺序不可颠倒。

**第一步，统一事实。** 核对官网、地图、政务公开页、招聘页面上的地址、电话、科室与专家信息，消除互相矛盾之处。这一步是第二层的前提，通常一周内可以完成，不需要技术支持。

**第二步，补问答内容。** 以患者的真实问法作标题，一页回答一个问题，前一百字内给出结论或数字。这一步是第三层的前提，见效周期以月计。

**第三步，让页面可被抓取。** 关闭浏览器脚本后正文仍可显示，必要的结构化信息补齐。这一步通常受改版排期限制，放在后面。

**第四步，建立季度复检。** 按第 4 章自检表复检，比较分值的变化而不是绝对值。没有这一步，前三步的成果会在两个季度内回到原点。

前三步不依赖技术条件，可以从本周开始；第四步是维持机制。

## 6.4 可以停掉的三件事

**一、为凑篇数而写的泛科普。** 引用份额里内容平台只占 2.9%（第 5.2 节）。产量本身不是变量，能回答具体问题的内容才是。停掉这部分，把同样的时间用来做十页病种页，效果反而更可测。

**二、以关键词密度为目标的改写。** 堆叠关键词不在有效方法之列（[1]）。第 2.1 节也说明了原因：检索以页面为单位、以能否回答提问为准，词频不参与采信判断。

**三、把官网当作「建好就完事」的项目。** 11.9% 的引用发生在自有域名上，这个比例不会因为一次改版而持续；页面信息随科室调整、专家变动而失效，失效后又不修正，就会从「没被引用」滑向「被引错了」（第 2.5 节）。

## 6.5 本报告的边界

一、不评价诊疗质量、技术水平与医疗服务能力，全文只讨论可见度。

二、不作效果承诺。第 5 章的数据是特定窗口的观测，不构成因果关系，也不预示任何机构的未来表现。

三、不外推为行业分布。样本为在服机构，非概率抽样（第 5.4 节）。

四、不构成法律意见。第 2.6 节对法规条文的引用为转述，具体发布前应经机构法务或属地卫生健康行政部门确认。

## 本章要点

一、六条结论分别落在机制、证据、数据、法规与顺序五个层面，均可回溯到对应章节。

二、三条建议只改内部标准与分配，不增加采购。

三、行动顺序自后向前：先统一事实，再补问答，再解决可抓取，最后建复检。

四、可以停掉的是凑篇数、堆关键词、把官网当一次性项目。

五、本报告只讨论可见度，不评价医疗质量，不作效果承诺，不外推行业分布。

# 附录 A 术语表

**表 A-1 术语与释义**

| 术语 | 释义 | 首见 |
| --- | --- | --- |
| 生成式引擎 | 以大型语言模型综合多个来源生成答案的检索产品，区别于返回链接列表的搜索引擎 | 1.2 |
| 检索增强生成 | 先检索外部文档、再以检索结果为条件生成回答的技术路径 | 2.1 |
| 查询改写 | 把一句自然语言问句拆成若干检索友好子查询的处理 | 2.1 |
| 检索候选集 | 一次提问中，检索环节返回的待判断页面集合 | 2.1 |
| 采信 | 模型判断某来源内容可信、并将其用于生成答案的过程 | 2.2 |
| 来源主体 | 采信依据之一：信息由谁发布 | 2.2 |
| 信息一致性 | 采信依据之一：同一事实在多个来源上是否吻合 | 2.2 |
| 实体完整性 | 采信依据之一：机构能否被识别为一个完整主体 | 2.2 |
| 孤证不立 | 单一来源的陈述难以支撑采信，即便该来源是机构自有页面 | 2.2 |
| 提问意图 | 提问背后的目的，分导航型、对比型、交易型、知识型四类 | 2.3 |
| 可见度漏斗 | 本报告提出的分析框架：被找到、被采信、被选中 | 2.4 |
| AI 可见度 | 某机构在特定提问集合下被生成式 AI 答案提及的程度 | 定义 1 |
| 失真 | AI 答案中的事实错误，含多源拼接、时效滞后、无据补全三类 | 2.5 |
| 医疗广告 | 利用各种媒介或形式直接、间接介绍医疗机构或医疗服务的广告 | 2.6 |
| 探测问题 | 一次「机构 × 提问 × 模型」的组合及其重复采样 | 定义 2 |
| 提及 | 答案文本中出现机构第一名称或公认简称 | 定义 3 |
| 提及率 | 提及次数占有效采集次数的比例 | 定义 3 |
| 位次质量 | 按机构在列举中的位置计分并归一化的指标 | 定义 4 |
| 首推率 | 机构位于答案首位的探测问题占比 | 定义 5 |
| 综合分 | 提及率与位次质量的加权得分 | 定义 6 |
| 事实准确率 | 可核对陈述中与官方信息一致的比例 | 定义 7 |
| 采集窗口 | 一个批次全部采集所在的连续时段 | 附录 C.2.4 |
| 极差 | 同一探测问题重复采样之间的最大差值 | 附录 C.4 |

# 附录 B 模型清单与提问集

本附录给出可直接执行的采集参数，供后续批次按此执行。执行中不得修改；确需修改的，须在当次报告注明改动时间与影响范围。

## B.1 构造规则

提问集不自行编写，做法是固定句式、替换实体。替换项有四个：城市、机构、科室、病种。

固定句式的目的，是让跨机构的提问在语义上尽量等同——所有机构被问的是同一批问法，差异只来自它所在的城市与专科。

替换后须人工过一遍，剔除两类：语义不通的，以及含真实患者信息的。后者一律不得进入提问集。

## B.2 十二问模板

**表 B-1 提问集模板**

| 编号 | 意图 | 模板 |
|---|---|---|
| Q1 | 导航 | {城市}{病种}挂什么科 |
| Q2 | 导航 | {城市}哪里可以做{检查} |
| Q3 | 导航 | {机构}{科室}门诊时间 |
| Q4 | 对比 | {城市}看{病种}哪家医院好 |
| Q5 | 对比 | {城市}{治疗}哪家医院做得比较好 |
| Q6 | 对比 | {城市}{科室}医院怎么选 |
| Q7 | 交易 | {城市}{治疗}大概多少钱 |
| Q8 | 交易 | {城市}{治疗}医保能报多少 |
| Q9 | 交易 | {病种}{治疗}住院几天，费用大概多少 |
| Q10 | 知识 | {病种}{检查指标}需要治疗吗 |
| Q11 | 知识 | {病种}多久复查一次 |
| Q12 | 知识 | {病种}有哪些治疗方式 |

资料来源：本报告设计。

Q1 至 Q3 考察机构是否被识别，Q4 至 Q6 考察是否被选中，Q7 至 Q9 考察能否提供可用的交易信息，Q10 至 Q12 考察内容的知识密度。四组共同构成一次完整探测。

## B.3 专科实体映射

**表 B-2 五个专科的实体映射**

| 科室属性 | 病种 | 检查 | 治疗 |
|---|---|---|---|
| 综合（以内分泌为例） | 甲状腺结节 | 甲状腺超声 | 甲状腺结节消融 |
| 口腔 | 牙齿缺失 | 口腔 CBCT | 种植牙 |
| 妇产 | 妊娠期糖尿病 | 口服葡萄糖耐量试验 | 无痛分娩 |
| 眼科 | 白内障 | 眼底检查 | 白内障超声乳化手术 |
| 中医 | 颈椎病 | 颈椎 X 线片 | 针灸 |

资料来源：本报告设计。表中病种与检查仅用于构造提问，不构成诊疗建议。

机构实际科室与表中属性不一致时，按其主要科室替换，并在记录中注明，不强行套用。

## B.4 采集记录字段

**表 B-3 单条采集记录的字段**

| 字段 | 说明 |
|---|---|
| probe_id | 机构编号 + 提问编号 + 模型编号，全局唯一 |
| collected_at | 采集时间戳，精确到分钟 |
| model_id | 模型标识与版本 |
| question | 实际发出的问句，实体已替换 |
| answer_full | 答案全文，不删改、不截断 |
| cited_sources | 答案标注的来源链接，逐条记 |
| mentioned | 是否提及，1 或 0 |
| position | 位次，未提及记 0 |
| note | 剔除原因或异常说明 |

资料来源：本报告设计。

剔除规则见附录 C.4。原始答案全文与时间戳一并存档，任何指标都应可由存档复算；存档缺失的记录，不计入有效采集。

# 附录 C 研究方法与数据口径

第 5 章的数字按本附录口径计算。本章记录两件事：一是第 5 章实际采用的数据来源与指标算法；二是原始抽样设计（C.2.1、C.2.2）——该设计为按 120 家机构计划采集的方案，本版未执行，保留供后续批次沿用。两者共用同一套指标定义与解析规则，差异只在样本来源。

## C.1 研究问题

本报告回答三个问题：

Q1 医疗机构在生成式 AI 中的可见度，总体处于什么水平？分布形态如何？ Q2 可见度差异与哪些可观测因素相关？ Q3 机构能否通过自身可执行的动作改变可见度？

Q1 与 Q2 本版仅作部分回答：第 5 章给出引用来源的分布，并据此讨论来源权重与可见度的关系，可见度本身的分布形态未纳入本版。Q3 需纵向跟踪，至少两个采集窗口的对比，本版不涉及。

## C.2 样本设计

### C.2.1 机构抽样

抽样框为卫生健康行政部门公开的医疗机构执业登记信息。采用分层抽样，分层维度为四项：

**表 C-1 机构样本的分层维度**

| 维度 | 分层 | 说明 |
| --- | --- | --- |
| 城市能级 | 一线、新一线、其他地级市 | 按公开的城市分级名单确定 |
| 机构等级 | 三级、二级、一级及未定级 | 以执业许可证载明为准 |
| 所有制 | 公立、民营 | 同上 |
| 专科属性 | 综合、口腔、妇产、眼科、中医 | 与提问集中的高频科室对应 |

资料来源：本报告设计。

目标样本量 120 家，按 40 个城市、每城市 3 家配置，四项维度在样本内尽量均衡。有效样本低于 60 家时，本报告不予发布——样本量不足时，分组比较的结论不稳定，容易把噪声当成差异。

抽样为分层便利抽样，不是概率抽样。这一限制在附录 C 第 5 节说明。

### C.2.2 提问集

提问集按意图分四类，与第 2.3 节的分类一致。每类 3 问，每机构共 12 问。

**表 C-2 提问集的四类意图与示例**

| 意图 | 示例（以某某市、甲状腺为例） | 提问数 |
|---|---|---|
| 导航型 | 甲状腺结节挂什么科 / 某某市哪里可以做甲状腺超声 | 3 |
| 对比型 | 某某市看甲状腺结节哪家医院好 / 某某市甲状腺手术哪家强 | 3 |
| 交易型 | 某某市甲状腺结节手术大概多少钱 / 医保能报多少 | 3 |
| 知识型 | 甲状腺结节 4 毫米需要处理吗 / 多久复查一次 | 3 |

资料来源：本报告设计。

问题的措辞不自行编写。来源为三类公开材料：搜索平台的下拉建议词、公开问答平台的高频提问、医疗机构公开的常见咨询问题。若需覆盖更广的患者问法，可补充公开的真实提问样本，如 HealthSearchQA 收录的消费者健康问题[10]。同一意图下取出现频次最高的三个问法，替换城市与病种实体后形成该机构的提问集。

替换实体而非重写句式，是为了保证跨机构可比：所有机构被问的是同一批句式，差异只来自城市与专科。

### C.2.3 模型清单

采集覆盖五个中文生成式 AI 服务，统一使用联网检索形态，同一轮采集在模型版本稳定的前提下完成，并逐条记录采集时的模型标识与版本。

**表 C-3 采集模型清单**

| 序号 | 服务 | 采集方式 | 版本 |
|---|---|---|---|
| 1 | DeepSeek | responses + 服务端检索 | deepseek-v4-flash |
| 2 | 豆包 | responses + 工具检索 | doubao-seed-2-0-mini-260428 |
| 3 | 千问 | responses + 服务端检索 | qwen3.8-flash |
| 4 | 文心 | chat + 检索开关 | ernie-5.1 |
| 5 | 元宝 | responses + 服务端检索 | hy3 |

资料来源：本报告依据采集链路的在运行配置整理。版本号随模型迭代变化，每批次须重新登记；未接入的服务不列入清单。

纳入五个而非一个，原因见第 2.5 节：模型间的检索源与权重分配不同，单模型结果不具备代表性。五个是当前采集链路稳定可跑的全部服务，覆盖面已足以反映模型间差异。

### C.2.4 采集规模与窗口

按目标样本量计算，单轮采集量为：

120 家 × 12 问 × 5 模型 × 3 次重复 = 21 600 次。

一个完整采集批次须在连续 7 个自然日内完成。窗口拉长会引入时效偏差：期间若有模型更新或热点事件，前后数据不再可比。

每个提问重复 3 次，用于估计模型输出的波动。按本设计执行时，图表标注均值，极差另行报告。

## C.3 指标定义

【定义 2 探测问题（probe）】

一次「机构 × 提问 × 模型」的组合及其回答。同一次采集重复 3 次，视为 1 个探测问题的 3 个样本。

【定义 3 提及与提及率】

答案文本中出现该机构的第一名称或公认简称，计 1 次提及。提及率为提及次数占有效采集次数的比例：

R = 提及次数 / 有效采集次数，R ∈ [0, 1]。

【定义 4 位次质量】

答案若以列表或编号形式列举多家机构，按该机构在其中的位置计分：

**表 C-4 位次与得分**

| 位次 | 得分 s |
| --- | --- |
| 第 1 位 | 5 |
| 第 2 位 | 3 |
| 第 3 位 | 2 |
| 第 4 位 | 1 |
| 第 5 位及以后 | 0.5 |
| 未出现 | 0 |

资料来源：本报告设计。分档依据为患者实际阅读时的注意力衰减，越靠前差异越大。

位次质量率 Q 为该机构在全部探测问题上的平均得分除以满分 5：

Q = mean(s) / 5，Q ∈ [0, 1]。

位次质量之所以单列，是因为「被提到」与「被排在前面」是两件事。只报提及率的报告，会把一家排在第 12 位、被顺带一提的机构，和一家被列为首选的机构算成同一档。

【定义 5 首推率】

答案列举的机构中，该机构位于第 1 位的探测问题占比：

F = 首位次数 / 有效采集次数，F ∈ [0, 1]。

【定义 6 综合分】

Score = 100 × (0.4R + 0.6Q)。

提及率权重 0.4、位次质量权重 0.6，反映的是「出现」与「排位」对患者选择的影响差异，权重取值为本报告设定，非外部标准。

【定义 7 事实准确率】

答案中关于该机构的事实性陈述（地址、等级、科室、门诊时间、联系电话），与执业登记信息或机构公开信息一致的比例：

A = 一致陈述数 / 可核对陈述总数。

只统计可核对项；答案未提及的不计入分母。准确率与可见度分开报告——一家机构可能被高频提及，同时高频答错（见第 2.5 节）。

## C.4 质量控制

四条规则，全部可复核。

**时间对齐。** 同一批次全部采集落在连续 7 日内，逐条记录采集时间戳，随原始数据存档。

**重复采样。** 每个探测问题重复 3 次，报告均值与极差。极差超过 2 个位次档位的，标记为不稳定样本，在分组统计中单独列出，不与稳定样本混合。

**剔除规则。** 以下情形不计入有效采集：模型拒答、返回空内容、触发安全策略中断、回答与该提问完全无关。剔除数量与原因逐条记录。

**抽样复核。** 随机抽取 10% 的答案，由两名标注人员独立判断「是否提及」「位次」两项，计算一致率。一致率低于 90% 时，重新核对判定规则后全量复核。

原始答案全文存档，不删改。任何一项指标都应可由存档复算。

## C.5 局限

五条，须与结论同时阅读。

一、抽样非概率。第 5 章的样本为 5 家在服机构的存量记录，不按抽样设计产生，不代表全国医疗机构总体，分布结论只在样本范围内成立。

二、提问覆盖有限。12 个提问只能反映「被问到的这 12 种问法」，患者实际问法远比这多。提问集之外的可见度，本报告不作推断。

三、模型持续迭代。模型版本、检索源与排序权重均会变化，本报告的结论有时效，跨批次比较须先确认模型版本。

四、提及不等于到诊。被 AI 提及与实际就诊之间隔着多个环节，本报告不主张二者存在数量关系。

五、自采数据无第三方审计。全部数据由本报告自行采集与标注，虽提供原始存档与复核方法，但不具备第三方审计效力。

## 本附录要点

一、指标定义与解析规则为全报告通用；C.2.1、C.2.2 的 120 家抽样设计为后续计划，非本版执行内容。

二、提问集取公开高频问法，只替换实体不重写句式，以保证跨机构可比。

三、五项指标：提及率、位次质量、首推率、综合分、事实准确率，定义与算法全部公开。

四、综合分的权重（0.4 与 0.6）为本报告设定，非外部标准，须与指标一同引用。

五、抽样非概率、提问覆盖有限、模型持续迭代，三条限制决定结论的适用边界。

# 参考文献

[1] AGGARWAL P, MURAHARI V, RAJPUROHIT T, et al. GEO: Generative Engine Optimization[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). 2024. arXiv:2311.09735. DOI: 10.1145/3637528.3671900.

[2] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[C]//Advances in Neural Information Processing Systems 33 (NeurIPS 2020). arXiv:2005.11401.

[3] 国家市场监督管理总局. 互联网广告管理办法：国家市场监督管理总局令第72号[A/OL]. 2023-02-25 公布，2023-05-01 施行. https://www.gov.cn/zhengce/202305/content_6858084.htm

[4] 国家工商行政管理总局，卫生部. 医疗广告管理办法：第26号令[A/OL]. 2006-11-10 公布，2007-01-01 施行. https://www.gov.cn/zhengce/2019-06/03/content_5723750.htm

[5] 中国互联网络信息中心. 中国互联网络发展状况统计报告[R]. 数据截至 2025 年 12 月.

[6] 国家卫生健康委员会. 2025 年我国卫生健康事业发展统计公报[R/OL]. 2026-08. https://www.nhc.gov.cn/

[7] 国家互联网信息办公室，等. 生成式人工智能服务管理暂行办法：第 15 号令[A/OL]. 2023-07-13 公布，2023-08-15 施行. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm

[8] 中国信息通信研究院. 人工智能产业发展研究报告（2025 年）[R/OL]. 2025. https://www.caict.ac.cn/

[9] 复旦大学医院管理研究所. 中国医院综合排行榜[R]. 年度发布.

[10] SINGHAL K, AZIZI S, TU T, et al. Large Language Models Encode Clinical Knowledge[J]. Nature, 2023, 620: 172-180. arXiv:2212.13138.