arXiv:2603.09296v1 [cs.IR] 10 Mar 2026

诊断与　　修复生成式引擎优化中的引用失败

Zhihua Tian *1 **Yuhan Chen** *1 **Yao Tang** *2 **Jian Liu** 2 **Ruoxi Jia** 1

摘要

生成式引擎优化（GEO）旨在提升内容在人工智能生成回答中的可见度。然而，现有方法衡量的是贡献——即一篇文档对回答的影响程度——而非引用，而引用才是真正将流量带回创作者的机制。此外，这些方法一律套用通用改写规则，未能诊断单个文档未被引用的原因。本文提出一种面向生成式引擎优化的诊断方法，追问文档为何未能被引用，并据此进行干预。我们构建了一个统一框架，包括：（1）首个覆盖引用流程不同阶段的引用失败模式分类体系；（2）AgentGEO，一个智能体系统，利用该分类体系诊断失败，从相应的工具库中选择针对性修复，并迭代直至实现引用；（3）一个以文档为中心的基准，用于评估优化能否泛化至留出查询。AgentGEO在仅修改5%内容的情况下，使引用率相对提升超过40%，而基线方法为25%。我们的分析表明，通用优化可能损害长尾内容，且部分文档面临仅靠优化无法完全应对的挑战——这些发现对人工智能介导的信息获取中的公平可见度具有启示意义。**项目页面：**https://zhihuat.github.

io/agentgeo/

1. 引言

生成式引擎的兴起引发了内容创作者对流量流失的严重担忧（《经济学人》，2025年）。与引导用户点击进入网站的传统搜索引擎不同，生成式引擎直接综合生成答案，可能减少支撑创作者经济的自然流量。作为回应，各大平台引入了引用机制——Google AI Overview在摘要旁展示来源链接（谷歌，2025年），Perplexity提供行内引用（Perplexity，2026），ChatGPT的搜索模式包含可展开的参考文献（OpenAI，2024）。这些引用如今是从综合答案返回原始网页的主要路径。

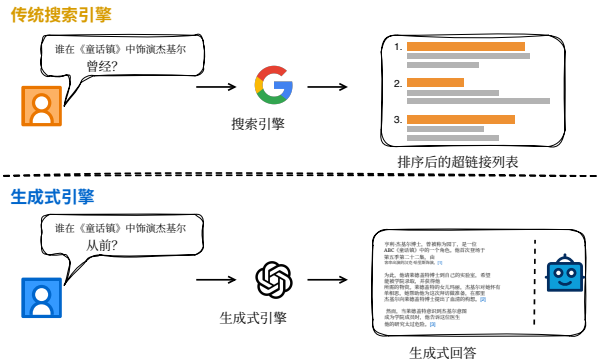


图*1.* 传统搜索引擎与由大语言模型驱动的生成式引擎之间的比较。传统搜索引擎会根据与用户查询的相关性返回排序后的超链接列表，而生成式引擎则利用检索到的内容综合生成答案，并提供指向来源网页的引用。

然而，这条路径十分狭窄。最近一项针对68,879次Google搜索的分析指出，在看到人工智能摘要的用户中，只有1%会点击其中引用的来源；相比之下，在未出现人工智能摘要时，有15%的用户会点击传统搜索结果（Chapekis与Lieb，2025年）。然而，确实到达网站的流量转化率可能显著更高——一项分析发现，来自生成式搜索的访客转化效果比传统自然搜索访客高出23倍（Stox，2025年）。因此，对于内容创作者而言，获得引用是一道严重的关口：尽管点击率很低，但实际到访的用户价值远高于平均水平，这使得被引用与被忽视之间的差别愈发重要。

从贡献到引用。 这些动态促使人们对生成式引擎优化（GEO）日益关注：通过修改网页内容，提升其在人工智能生成的回答中的可见度。现有研究使用贡献指标评估可见度，例如词数，用以衡量回答中归于某一来源的内容比例（Aggarwal

*贡献相同　[1]弗吉尼亚理工大学 [2]浙江大学。通讯作者：Ruoxi Jia <ruoxijia@vt.edu>。

预印本。*2026*年*3*月*11*日。

等人，2024）。这些连续指标虽然有用，却混淆了两个因素：网页是否被引用，以及在被引用的条件下，回答中有多少内容归于该网页。对于内容创作者而言，严重的障碍是未被引用：没有引用，就没有引流机会。因此，我们聚焦于一种互补的、以引用为先的目标，该目标将是否被引用这一二元事件单独考量。在我们的实验中，在基线条件下，43% 的主题相关网页未获任何引用。对于这些网页而言，问题不是“我被引用了多少？”，而是“为什么我完全没有被引用？”理解并修复这类问题，是一个尚未得到充分研究的问题，且直接关系到网络内容创作的可持续性。

现有方法的局限性。当前的生成式引擎优化方法对文本一概套用通用偏好规则：添加统计数据、采用权威语气、提升流畅度（Aggarwal 等人，2024；Wu 等人，2025）。这种方法行不通，因为引用失败的类型各异，可能发生在多个流程阶段。网页可能在抓取 时失败（格式错误的 HTML、JavaScript 依赖项），在解析 时失败（内容被截断或埋在模板内容之下），也可能在生成 时失败（缺少实体、信息密度低于竞争网页）。每种失败模式都需要不同的干预措施，但通用规则既无法诊断究竟哪个阶段出了问题，也无法解决文本尚未经过处理便已发生的上游故障。

贡献一：生成式引擎优化的诊断方法。基于这些观察，我们将生成式引擎优化重新界定，从“哪些通用改进可能有效？”转向“这个网页为何未能在相关查询中获得引用？”我们在 AgentGEO 中实践了这一思路，这是一个用于迭代诊断与修复的智能体框架。对于未能获得引用的网页，AgentGEO 会将其与被引竞品进行比较，以识别最可能的劣势；从一套涵盖整个引用流程的专用干预措施库中选择合适的工具，并验证是否成功获得引用。若未能成功，系统会重新诊断并继续迭代，同时借助记忆机制跟踪此前的尝试。

贡献二：引用失败分类体系。我们首次提出了一套涵盖生成式引擎流水线的引用失效模式系统分类体系：解析阶段失败（格式错误的 HTML、过多噪声）、抓取/上下文阶段失败（截断、内容排序不佳）以及生成阶段失败（实体缺失、意图不匹配、竞争对手劣势）。这套分类体系为我们的工具库设计提供指引，并为未来以引用为重点的优化奠定基础（图 2）。

贡献三：以文档为中心的基准测试。现有基准测试将每篇文档与单个查询配对（Ag- garwal 等人，2024；Wu 等人，2025年）。在实践中，内容提供者无法预知确切的查询，因此需要具有泛化能力的优化。我们提出 MIMIQ（多意图、多查询），这是一种以文档为中心的基准测试，将每个网页与多个查询关联，这些查询涵盖不同的意图、用户画像和措辞。该基准测试使方法能够利用训练查询集进行优化，并在留出查询上接受评估，从而检验优化是否真正生成了更适合引用的内容，而非过拟合特定的查询表述。

实证发现。我们的实验揭示了网页为何无法获得引用，以及优化能够——和无法——解决哪些问题的新见解。

（1）诊断优于通用规则，同时保持内容完整性（第 6.2 节）。AgentGEO 在多个生成式引擎中实现了超过 40% 的引用率相对提升。至关重要的是，这一提升只需极少改动：基线方法平均会修改 25% 的原始内容，而我们的定向修复仅涉及 5%。这表明，引用失败很少是整体内容质量问题：大多数网页只需进行有针对性的修复，无需大幅改写。

(2) 通用规则可能损害长尾内容（第 6.3 节）。我们找出了这样一些主题类别：通用优化不仅无法提高引用正确率，反而会使其下降；而诊断式优化则能在相同网页上持续带来提升。这种差异反映了现有方法的根本局限：通用规则源自总体模式，但专业主题和代表性不足的领域会系统性地偏离这些模式。诊断式优化针对每个网页的具体失败模式进行调整，从而避免了这种偏差，也因此能更公平地适用于不同内容类型。

(3) 并非所有引用失败都能挽回（第 6.4 节）。对于某些网页，即使采用诊断式优化，也无法提升引用正确率。这些网页处于劣势——主要竞争者占据优势——而任何内容方面的修改都无法扭转这一局面。这一发现的影响超出了优化本身：如果某些内容无论付出多少努力都始终处于劣势，那么引用机制可能会让某些声音比其他声音获得更多传播，而仅靠创作者侧的优化无法确保可见度公平。

2. 相关工作

2.1. 生成式引擎

生成式引擎会整合大语言模型，利用检索增强生成（RAG）技术，直接基于网页内容进行检索、综合并生成回答（Lewis et al., 2020; Karpukhin et al., 2020）。生成式引擎的工作流程大致可分为三个阶段：（1）检索；（2）内容提取（抓取）；（3）响应生成。检索阶段会根据用户查询识别并检索相关网页。这一过程可能会使用传统搜索引擎（谷歌，2025; Willison, 2025），也可能使用专门的神经

检索模型（Karpukhin et al., 2020）。内容提取阶段侧重于解析检索到的网页中的有用文本，将非结构化的网页内容转换为适合大语言模型处理的结构化信息。最后，响应生成阶段利用大语言模型，根据提取的证据生成连贯且与上下文相关的回答。除了这些基础工作流程外，近期研究还通过融入大语言模型的推理能力，推动更复杂的生成式引擎发展，使其能够针对复杂查询进行迭代规划、推理和搜集证据（Jin et al., 2025; Mo et al., 2025）。这些进展拓宽了研究重点，不仅关注检索与生成的整合，也关注提升回答的事实一致性和可靠性（Salemi & Zamani, 2024; Niu et al., 2024）。

2.2. 搜索引擎优化

搜索引擎优化（SEO）长期以来一直是信息检索领域的关键课题（Davis，2006；Yalc¸ın 和 K¨ose，2010）。传统 SEO 策略着重通过关键词优化、反向链接管理和网站结构设计等技术，提升网页在搜索结果列表中的排名（Killoran，2013）。随着搜索范式转向生成式引擎（GE），生成式引擎优化（GEO）应运而生，旨在提高内容在生成式回答中的可见度。早期研究证明，利用大语言模型根据人工制定的启发式规则改写网页内容，能够提高内容被检索到的可能性（Aggarwal 等人，2024；Chen 等人，2025a）。为了更好地契合生成式引擎的偏好，近期研究利用大语言模型提炼文本改写的优化规则（Wu 等人，2025）。除了内容优化，研究人员还探索了提示词注入攻击等对抗技术（Nestaas et al.，2024；Pfrommer et al.，2024；Tang et al.，2025），以操纵生成式引擎的行为。尽管取得了这些进展，现有方法仍主要关注内容层面的文本修改，往往忽视 HTML 解析机制、上下文窗口限制和网页结构等严重的结构性与系统性因素，而这些因素会显著影响引擎摄取内容的方式。相比之下，AgentGEO 提出了一种整体性的生成式引擎优化框架，通过优化网页表示的多个方面来解决这些局限。

2.3. GEO 基准

GEO-Bench（Aggarwal 等人，2024）是最早的 GEO 基准测试之一，旨在进行以查询为中心的评估。训练集包含来自多种来源的查询，测试集则包含查询及其关联的检索文档。该基准通过衡量优化后某篇文档相对于其他文档在该查询下可见度的提升，来评估生成式引擎优化方法。电子商务（Wu 等人，2025年）和 Researchy-GEO（Wu 等人，2025年）的设置类似，但侧重于特定领域。CC-GSEO-Bench（Chen 等人，2025b）是另一个 GEO 基准测试，侧重于以文档为中心的评估，其中每篇文档都关联多个查询。然而，每篇文档对应的查询数量有限且不一致，范围为 2 到 10 个，因此难以全面评估文档在生成式引擎中的整体可见度。所有这些基准测试都只考虑内容，无法评估 HTML 结构数据。相比之下，我们的 **MIMIQ** 同时评估内容和 HTML 结构，从而能够更全面地评估 GEO 策略。

3. 理解引用失败

我们将引用——即被纳入为来源这一二元事件——作为优化目标。为了解文档为何未能获得引用，我们首先考察生成式引擎如何实现引用，随后提出一个失败模式分类体系。

3.1. 生成式引擎中的引用机制

生成式引擎如何将回答归因于来源，决定了引用失败可能发生的环节。我们归纳出以下三种范式。

行内归因。 参考标记（例如“[1]”）直接嵌入生成序列中，从而在句子或分句层面提供细粒度的溯源信息。模型在生成过程中 决定哪些来源支撑每项主张。Perplexity 和 ChatGPT 的搜索模式等系统采用了这种方法（Perplexity，2026；OpenAI，2024）。

事前归因。 系统首先从检索到的文档中选择段落，然后仅依据这些选段生成回答。由于生成过程基于预先选定的证据，来源上下文构成了固有的锚点——模型难以轻易编造提取内容之外的信息（Slobodkin 等人，2024）。然而，上游未选中的相关内容不会被引用。

事后归因。 生成与验证相互分离：模型先生成回答，然后由单独的模块追溯识别支持来源（Bohnet 等人，2022；Gao 等人，2023）。这种范式虽然灵活，却有幻觉传播的风险——生成过程中引入的错误可能会持续存在，尤其是在验证器错误地确认无依据的主张时。

鉴于这两种范式在商业系统中十分常见，且落地保证更严格，我们聚焦于行内归因和事前归因。这两种范式呈现出不同的失效环节：在生成过程中，模型可能忽略相关来源，导致行内归因失败；而事前归因则可能更早在段落选择阶段失败。我们的分类体系涵盖了这两种范式中的失效情况。

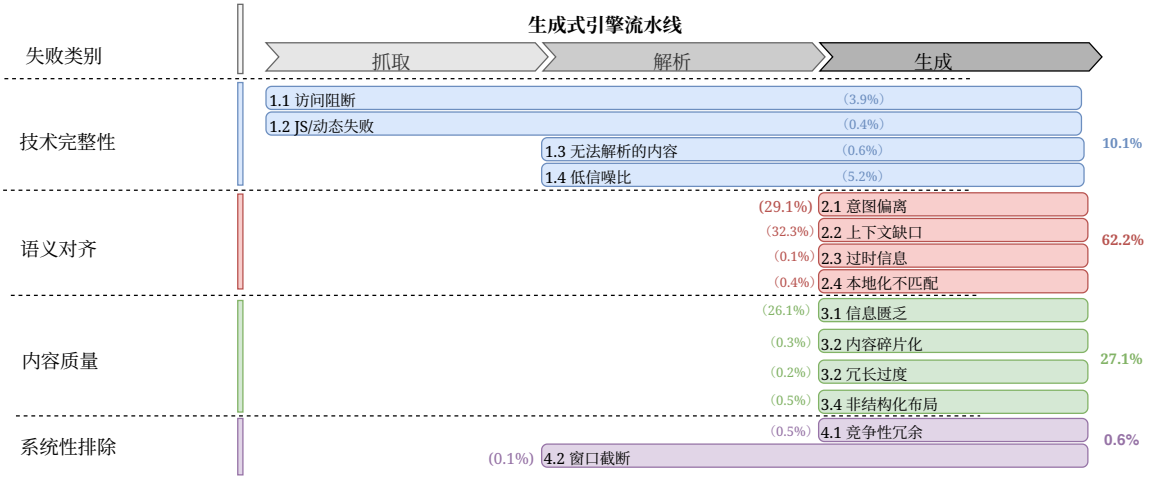


图 *2.* 生成式引擎在 HTML抓取、解析和答案生成阶段的引用失败模式分类体系。

3.2. 引用失败分类体系

为什么主题相关的网页未能被引用？为系统地回答这一问题，我们从 GEO-Bench（Aggarwal 等人，2024）构建了一个诊断数据集。该数据集涵盖 10 个不同领域（健康、旅游、技术等），并包含具有不同意图的查询。对于每个检索到但未被引用的网页，我们将其与一个被引竞品配对，共得到 949 对对比样本对。这种配对策略能够隔离边际因素：两个网页都是针对同一查询检索到的，因此差异在于引擎为何偏好其中一个。通过分析这些样本对，我们确定了映射到生成式引擎流水线中的四个主要失效维度（图 2）。各类别的详细说明见附录 3.2。

技术完整性（10.1%）。网页无法被正确摄取。失败可能发生在抓取阶段（访问阻断、JavaScript渲染失败、连接错误），也可能发生在解析阶段（内容无法解析、模板内容导致信噪比低、格式碎片化）。如果严重信息被埋在 JavaScript小组件中，或被导航框架淹没，生成器可能根本无法获取这些内容。

语义对齐（62.2%）。网页内容与查询要求不符。这包括意图偏离（为交易型查询提供信息型内容）、上下文缺口（缺少特定实体或术语）、过时信息，以及本地化不匹配（例如，针对美国查询提供英国法规）。这些是生成阶段的失败：内容已到达模型，但被判定为相关性不足。

内容质量（27.1%）。该网页主题正确，但信息呈现不佳。表现包括信息匮乏（内容过于浅显，无法作为引用来源）、内容碎片化（片段彼此割裂，难以综合），冗长过度（关键事实被冗余内容冲淡），以及缺乏结构的布局（文字密集，而表格更便于提取信息）。这些都属于生成阶段失败，呈现方式更好的竞争网页因此胜出。

系统性排除（0.6%）。网页面临超出其内容本身的结构性劣势。竞争性冗余是指更具权威性的来源（例如维基百科）涵盖了相同事实。窗口截断是指相关内容埋得太深，无法纳入上下文窗口。

4. AgentGEO：一种生成式引擎优化的诊断方法

问题设定。令 d 表示目标网页，$\mathcal{Q}$ 表示与 d相关的用户查询分布。对于查询 q，令$\mathcal{V}(q,d,\mathcal{G}) \in \{0,1\}$ 表示网页 d是否被生成式引擎 $\mathcal{G}$引用。我们寻求优化后的网页 d^*，以在 $\mathcal{Q}$ 上最大化期望引用，同时保持对原文的忠实度：

$$d^* = \arg\max_{d'} \mathbb{E}_{q\sim\mathcal{Q}}\left[\mathcal{V}(q,d',\mathcal{G})\right] \quad \text{s.t.} \quad \mathrm{Sim}(d,d') \geq \tau$$

在实践中，我们使用训练查询 Q_{tr} 进行优化，并在留出查询 Q_{te}上进行评估。

概览。 AgentGEO 采用三阶段流程（图 3）。首先，对于每个训练查询，先诊断后修复的循环会识别目标网页未能获得引用的原因，并在分类体系和记忆模块的指导下实施针对性修复；记忆模块会追踪先前的尝试。其次，批次聚合 阶段会将查询特定建议合并为查询不变修改，以避免过拟合单个查询。第三，局部编辑机制会在分块级别应用修改，同时保持语义完整性。

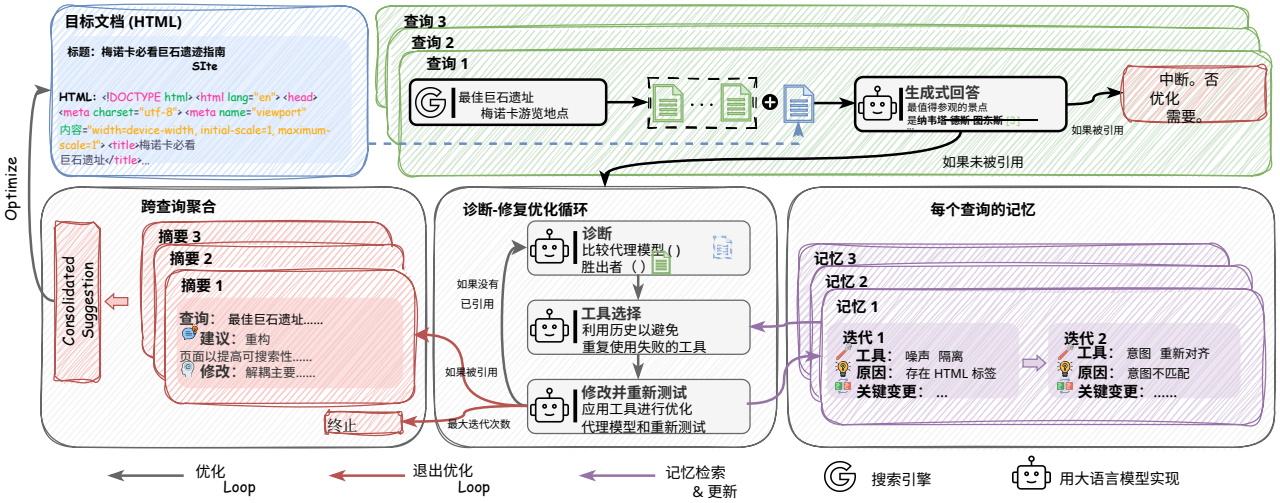


图*3.* **AgentGEO 的优化过程**AgentGEO 首先通过与竞争对手文档进行比对，验证引用状态。对于未被引用的文档，它会启动“先诊断后修复”循环：(1) 大语言模型首先诊断文档未被引用的原因；(2) 根据诊断结果以及查询专属记忆中的过往迭代记录，选择特定工具来修改代理网页（原网页的副本）；(3) 使用相同的竞争对手重新测试修改后的代理网页。循环会持续进行，直至成功获得引用或达到限制。所有成功的修改都会通过聚合策略合并为一条建议，用于优化原始网页。随后，针对更多训练查询迭代执行此步骤，以进一步优化网页。

4.1. 先诊断后修复循环

对于每个训练查询 q，AgentGEO 都会执行一个迭代循环，不断修改目标网页 d，直到网页获得引用或达到迭代次数上限。

失败诊断。 我们模拟生成式引擎流程：检索候选网页、生成回答，并记录引用了哪些来源。如果目标网页未被引用，便开始诊断。为找出失败原因，系统会将 d 与排名最高的被引竞争对手进行比较，并分析其弱点，即解释为何被引用的是竞争对手。这些弱点按照我们的失败模式分类体系进行分类（第3.2节）。

基于记忆的工具选择。 根据已诊断的弱点，AgentGEO从专用干预措施库中选择一种修复工具，涵盖HTML修复、内容重排、实体丰富化以及信息密度改进（完整列表见附录D）。为避免重复无效修复，我们维护一份查询专属记忆，用于跟踪先前的尝试。在每次迭代 t中，智能体根据当前弱点与记忆选择工具。记忆初始化为空。

迭代优化。 应用选定的工具来生成修改后的网页 $d^{(t)}$。如果成功获得引用，系统会提取一条建议，总结成功的编辑内容——具体说明修改了哪些内容位置，以及如何修改。这些建议将用于下文介绍的批次聚合。如果未能获得引用，系统会将此次失败尝试更新到记忆中，并针对竞争网页重新进行诊断。这个循环会持续进行，直到成功获得引用或达到迭代次数上限。

4.2. 跨查询聚合

为实现对训练查询的泛化，Agent- GEO 在批次级别执行网页更新。对于批次中的每个查询，我们都在当前 HTML 的隔离临时副本上运行先诊断后修复循环（第 4.1 节）。每次运行最多生成一项候选编辑，其中包括已诊断出的失败类型、目标结构块的索引，以及工具生成的用于更新该块的 HTML 片段。为确保不同查询的编辑具有可比性，我们在批次开始时固定分块方案，并将所有候选编辑映射到这一冻结的索引空间，即使临时 DOM 在迭代重试期间发生变化也是如此。我们按目标块汇总候选编辑，并为每个块最多选取一项编辑，应用到共享网页上。默认情况下，诊断感知选择器会根据失败严重程度和置信度分数对候选项排序，并通过保留排名最高的提案来解决冲突。（其他冲突解决策略包括在块内进行启发式投票，或由基于大语言模型的仲裁器进行裁决。）我们通过单次处理应用选定的各块编辑，以生成批次更新，从而缓解查询专属过拟合，并促进与查询无关的改进。聚合时，我们不强制执行明确的相似度阈值；主要通过局部块级编辑来维持保真度（第 4.3 节）。

4.3. 面向忠实度的局部编辑

为满足语义保留约束，AgentGEO 在块级别应用修改，而不是重写整个网页。这有两个作用：通过将编辑限制在目标位置来保持忠实度，并避免大语言模型在处理较长网页时出现的质量下降（Liu 等人，2024）。

网页根据 HTML 标签划分为结构感知分块，每个分块代表一个连贯的单元（例如段落、列表或元数据块）。聚合阶段提出的每条建议都会指定其针对的内容位置；我们将这些位置映射到相应的分块。修改仅应用于匹配的分块；其余内容及整体 HTML 结构保持不变。

5. 以文档为中心的基准测试

现有 GEO 基准会将每篇文档与单个查询配对，并使用同一查询评估优化效果（Aggarwal 等，2024）。这种设置可能会将真正的改进与查询专属过拟合混为一谈：针对一个查询奏效的优化，面对表达相同意图的另一个查询时可能会失效。

我们推出 **MIMIQ** （**多**意图**多**查询**基**准**测**试**集**），这是一个用于评估泛化能力的以文档为中心的基准测试。我们称其为以文档为中心，因为它将文档（而非查询—文档对）视为优化和评估的基本单位。这与内容创作者的实际情境一致：他们可以控制自己的文档，却无法预知用户会提出哪些具体查询。对于每篇文档，MIMIQ 都提供多个查询，涵盖不同的意图、人物类型和表述方式。各方法使用训练查询集进行优化，并在优化过程中未见过的留出查询上接受评估。该方案检验优化是否真正提升了内容的可引用性，还是仅仅过拟合于特定的查询表述。表 1 将**MIMIQ** 与现有基准进行比较。

5.1. 构建流程

我们使用大语言模型为每个目标网页生成一组密集的查询。为了生成能够反映真实搜索多样性的查询，我们设计了一个涵盖意图、用户画像和措辞三个维度的结构化 生成框架。附录 A 提供了详细的提示模板和筛选启发式规则。

内容锚定关键词提取。 直接利用整个网页内容生成查询，往往会引入噪声。例如，从一篇关于“安全公路旅行”的网页中提取关键词，可能会得到“地图”这类不相关的词。为保持相关性，我们提取关键词

表 *1.* **MIMIQ** 与现有生成式引擎优化基准的比较。我们的数据集独特地结合了以文档为中心的范式与全面的评估范围，并支持明确的训练/测试查询划分。

数据集	以文档为中心	HTML	训练/测试划分
GEO-Bench（Aggarwal 等人，2024）	×	×	✓
电子商务（Wu 等人，2025年）	×	×	✓
Researchy-GEO（Wu 等人，2025年）	×	×	✓
CC-GSEO-Bench（Chen 等人，2025b）	✓	×	×
MIMIQ（我们的方法）	✓	✓	✓

时仅从页面标题元数据中提取，因为标题元数据通常能简洁地概括核心主题。我们指示大语言模型提取三类词语：（1）核心关键词，用于确保主题相关性；（2）同义词/潜在语义索引词，用于语义扩展；（3）长尾短语，用于捕捉用户的具体需求。

用户画像与意图建模。 为模拟多样化的用户行为，我们引入个性模拟和意图分类。我们构建了多种用户画像（例如“领域专家”与“随意浏览者”），并将其映射到标准搜索意图（导航型、信息型、交易型、商业型）（Broder，2002年；Rose 与 Levinson，2004年）。对于每个网页，我们会筛除不合理的意图（例如，为纯食谱页面移除导航意图），以确保生成的查询在上下文中有效。

质量控制。 原始生成的查询会经过额外的去重和相关性筛选流程。我们采用验证模型来消除幻觉，确保最终集合中的每个查询都能由目标网页在逻辑上回答。

5.2. 数据集统计与变体

标准 MIMIQ 数据集包含从 ClueWeb22 索引中采样的 204 个网页。对于每个网页，我们整理出一个分层的 60 个查询集合，其中 20 个用于训练，40 个用于测试，总计 12,240 个查询。除标准设置外，我们还引入了两个具有挑战性的变体，以探究特定的优化器能力：

• **MIMIQ-OOD（分布外）：** 为评估泛化能力，我们在训练和测试之间引入分布偏移。我们根据用户画像对生成的查询进行聚类；特定用户画像仅用于训练集，而未见过的用户画像则构成测试集。该设置用于测试优化器能否泛化到未见过的用户行为。

• **MIMIQ-HTML（结构鲁棒性）：** 在实践中，引用失败往往源于上游的抓取和解析阶段。
MIMIQ-HTML 子集被识别为包含复杂或不规则 DOM 结构。此变体专门评估优化器解析和

修改原始 HTML 的能力。

每个变体均包含 50 个网页，每个网页对应 60 个查询，并保持与标准数据集相同的训练/测试划分。

6. 实验

本节中，我们通过大量实验评估 AgentGEO 在所提出数据集上的有效性。首先介绍实验设置，包括基线方法和评估指标。然后展示主要结果，并分析分布外数据的影响。最后，我们进行消融研究，以验证 AgentGEO 各个组件的贡献。

6.1. 实验设置

基线方法 我们在实验中考虑三类基线方法。原始基线是未经任何优化的原始生成式引擎。GEO-Bench（Aggarwal 等人，2024）提出了八种使用大语言模型优化网站内容的方法。AutoGEO（Wu 等人，2025年）从大量网页对中自动总结偏好规则，并据此优化网站内容。我们在基于最先进大语言模型构建的生成式引擎上测试所有方法，包括 GPT（gpt-4.1-mini）和 Claude（claude-haiku-4-5-20251001）。

指标。 我们从三个维度评估模型性能：可见度、效用和网页忠实度。可见度量化优化后的内容在生成式引擎的回答中被呈现并优先展示的程度。我们计算**引用率**（引用正确率），用于衡量在测试查询中成功被引用的优化后网页所占比例。如引言所述，这将是我们优化的主要指标。我们还报告来自 GEO-Bench（Aggarwal 等人，2024）的另外三项指标：**词、词性和词语位置**，用于评估内容在结构和词汇层面的贡献。效用评估所生成答案的质量。我们遵循 DeepResearchGym（Coelho 等人，2025年）来评估效用，涵盖相关性（**精确率、召回率**）和质量（**清晰度、洞察力**）。忠实度评估优化后网页对原始信息的保留程度。我们使用三项指标对此进行量化：**词频-逆文档频率、嵌入相似度和杰卡德相似度**。词频-逆文档频率计算加权词项向量的余弦相似度，而嵌入相似度利用预训练语言模型的稠密表示来评估语义接近度。最后，杰卡德相似度衡量网页独特词集之间严格的词汇重叠。所有结果均以百分比（%）报告。

生成式引擎模拟设置。由于生产环境中的生成式引擎流水线属于专有技术，我们构建了一个受控模拟环境，采用标准的检索与生成组件，并使用明确的引用指令。这使得系统化诊断、可复现评估和受控消融成为可能；若使用黑盒商业系统，这些工作将难以实现。

6.2. GEO 整体表现

不同引用方法的表现。我们首先将 AgentGEO 与不同的基线引用方法进行比较。我们评估了两种引用方法，包括上下文内生成和先属性后生成（Slobodkin 等人，2024）。在上下文内生成中，我们直接提示生成式引擎根据优化后的网页生成带引用的答案。在先属性后生成中，生成式引擎首先从优化后的网页中提取引用相关属性，然后基于这些属性生成答案。

表 2 展示了结果。我们观察到，AgentGEO 在不同引用方法和生成式引擎下始终优于所有基线方法，展现出其提升引文表现的有效性和稳健性。在上下文内生成中，AgentGEO 的引用率达到 79.52%，比表现最佳的基线方法 AutoGEO 高出 10.72%。在先属性后生成中，AgentGEO 的引用率达到 70.00%，比 AutoGEO 高出 4.03%。此外，AgentGEO 在贡献度和忠实度指标方面也表现出色，这表明它不仅提高了引用率，还能保持内容的高质量和忠实性。

基于不同大语言模型的生成式引擎的表现。我们进一步评估 AgentGEO 的优势是否适用于不同的生成式引擎。为提高效率，我们从标准数据集中抽取了 50 个网页作为子集进行评估。受计算资源限制，我们从标准数据集中抽取了 50 个网页作为子集进行评估。表 3 展示了 GPT 和 Claude 的评估结果。AutoGEO 和我们的方法在 GPT 上的引用率都高于在 Claude 上的表现。一种可能的原因是，GPT 更擅长理解并利用优化后的内容来生成引用，而 Claude 可能需要对内容进行更明确的修改，才能取得类似的引用提升。然而，在精确率、召回率、清晰度和洞察力等效用指标上，Claude 在所有方法中的表现通常都优于 GPT。这表明，无论采用哪种优化方法，Claude 生成高质量且相关答案的能力可能都更强。同时，在两种模型上，AgentGEO 在效用指标上的表现均与原始版本和 AutoGEO 相当，这表明我们的优化在提升引用率的同时，不会损害答案质量。

表 2. AgentGEO 与不同引用方法下各基线方法的性能比较。**粗体** 和下划线分别表示最佳和次佳结果。"-"表示该指标不适用于该方法。

	上下文内							先属性后生成						
	CR	贡献			忠实度			CR	贡献			忠实度		
方法	CR	Word	Pos	词语位置	词频-逆文档频率	嵌入	杰卡德	CR	Word	Pos	词语位置	词频-逆文档频率	嵌入	杰卡德
原始版本	56.58	16.28	16.28	16.39	-	-	-	60.20	23.69	28.72	23.57	-	-	-
技术术语	55.84	14.78	14.73	14.60	73.32	85.64	21.37	57.33	22.25	27.87	22.03	74.39	86.98	21.98
引用来源	60.52	17.01	16.96	16.89	92.76	93.62	65.62	59.21	24.04	29.02	23.79	92.33	92.95	64.60
关键词堆砌	62.44	17.83	17.74	17.82	82.85	88.95	30.04	65.62	28.09	33.39	27.81	82.16	86.92	28.39
独特词汇	57.76	15.33	15.37	15.19	86.00	91.58	32.72	57.42	23.43	28.45	23.23	86.11	91.58	33.30
权威	59.14	17.45	17.75	17.57	80.42	86.38	25.15	63.46	26.86	31.85	26.63	80.85	86.35	26.02
易于理解	71.11	19.73	19.84	19.75	82.17	86.68	30.26	65.34	28.44	33.62	27.98	81.99	86.93	30.41
添加统计数据	60.52	17.41	17.29	17.26	95.52	**94.19**	69.31	59.46	24.15	29.39	23.93	95.51	93.75	69.38
引文添加	59.00	16.28	16.16	16.19	**96.78**	93.92	74.85	59.99	23.66	28.88	23.48	96.57	**94.29**	74.78
流畅度优化	60.60	16.52	16.63	16.48	92.40	92.72	48.98	62.22	25.44	30.95	25.25	92.20	92.83	49.21
AutoGEO	68.80	17.87	18.53	17.65	67.49	76.99	17.97	65.97	33.96	**34.33**	**32.88**	74.55	78.52	19.77
AgentGEO（我们的方法）	**79.52**	**20.29**	**21.08**	**20.28**	94.23	93.20	**82.40**	**70.00**	**34.18**	31.57	30.92	**97.44**	92.77	**82.66**

表 3. 不同基于大语言模型的生成引擎之间的性能比较。**粗体**表示最佳性能。

模型	指标	原始版本	AutoGEO	AgentGEO （我们的）
GPT	CR	56.58	68.80	**79.52**
	Word	16.28	17.87	**20.29**
	Pos	16.28	18.53	**21.08**
	词语位置	16.39	17.65	**20.28**
	精确率	48.63	56.79	**58.43**
	召回率	64.62	**68.99**	67.10
	清晰度	61.23	**66.08**	62.88
	洞察力	40.20	**45.25**	42.68
	词频-逆文档频率	-	67.49	**94.23**
	嵌入	-	76.99	**93.20**
	Jaccord	-	17.97	**82.40**
Claude	CR	42.40	46.80	**54.80**
	Word	21.80	24.86	**26.94**
	Pos	21.54	24.81	**27.57**
	词语位置	21.44	24.66	**26.10**
	精确率	89.42	88.79	**90.33**
	召回率	**96.16**	95.51	96.02
	清晰度	73.99	**75.74**	74.99
	洞察力	57.47	**61.05**	60.76
	词频-逆文档频率	-	74.27	**95.26**
	嵌入	-	79.98	**88.76**
	Jaccord	-	22.28	**74.58**

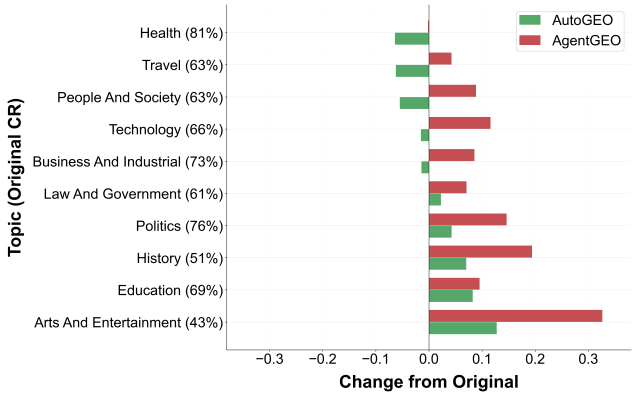


图4。 不同主题下相较于原始基线的引用率提升。左侧数字表示各主题的原始引用率。

6.3. 不同主题的评估。

为了分析 AgentGEO 在不同主题上的效果，我们将网页划分为艺术、教育、历史等多个主题。随后，我们评估各主题相较于原始基线的引用率提升情况，并将 AgentGEO 与 AutoGEO 进行对比，以作参考。结果汇总于图4。

结果表明，不同主题的引用率提升程度各不相同。艺术与娱乐、历史和政治等主题的提升显著。然而，健康等主题的提升幅度为负或微乎其微，这表明 AgentGEO 在这些领域的效果可能较差。我们进一步考察了健康主题，该主题在使用 AgentGEO 后提升幅度为负。我们发现，健康相关网页原本的引用率已经相当高，约为 80%，因此进一步提升的空间有限。对于其余未被引用的网页，一些优化步骤无意中删除了领域特定信息，导致这些网页更不容易被引用。令人意外的是，在若干主题中，AutoGEO 的表现不如原始基线。尤其是当原始引用率已经很高时，AutoGEO 往往会使其下降，这表明通用规则可能会损害长尾内容。

6.4. 失败优化案例研究

为了解 AgentGEO 的局限性，我们分析了 50网页优化实验中的失败案例。虽然训练查询的整体引用率从 57.0% 提升至 83.7%，但优化后仍有 163 个查询未获得引用。进一步审视这些失败案例可以发现，优化过程实际上确实成功修改了网页内容，以解决已诊断出的问题。然而，生成式引擎仍未引用这些网页，这表明失败并非源于优化无效，而是源于使网页内容符合生成式引擎引用标准本身的固有难度。

网页内容以解决已诊断出的问题。然而，生成式引擎仍未引用这些网页，这表明失败并非源于优化无效，而是源于使网页内容符合生成式引擎引用标准本身的固有难度。

例如，关于“最佳在线机器学习课程”的查询中，某大学的机器学习课程页面未能被引用。优化过程通过添加详细的课程介绍，成功提升了页面的清晰度和相关性。尽管如此，生成式引擎仍持续忽略该页面，转而优先选择知名在线教育平台（例如 Coursera、edX）等权威来源。这表明，生成式引擎可能存在一种内部偏见，倾向于域级因素，而非页面内容本身。这类偏见构成了基于内容进行优化的根本局限，也亟需深入研究生成式引擎引用机制背后的系统性因素。

7. 结论

本文提出了一种生成式引擎优化诊断方法，将关注重点从通用内容改写转向理解单个网页未能被引用的原因。我们提出了引用失效模式分类体系、AgentGEO——一种能够诊断失效并实施针对性修复的智能体系统——以及一个用于全面评估的以网页为中心的基准测试。未来工作可着重于在商业型引擎上验证 AgentGEO，并通过蒸馏和高效诊断策略提升其可扩展性。

参考文献

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. Geo：生成式引擎优化。载于第*30*届*ACM*知识发现与数据挖掘国际会议论文集，第5–16页，2024年。

Bevendorff, J., Stein, B., Hagen, M., and Potthast, M. Elastic ChatNoir：面向ClueWeb和Common Crawl的搜索引擎。载于Azzopardi, L., Hanbury, A., Pasi, G., and Piwowarski, B.（编），信息检索进展：第*40*届欧洲信息检索研究会议（*ECIR 2018*），计算机科学讲义，柏林、海德堡、纽约，2018年3月。Springer。

Bohnet, B., Tran, V. Q., Verga, P., Aharoni, R., Andor, D., Soares, L. B., Ciaramita, M., Eisenstein, J., Ganchev, K., Herzig, J., 等。带归因的问答：带归因的大语言模型的评估与建模。*arXiv*预印本 *arXiv:2212.08037*，2022年。

Broder, A. 网络搜索的分类体系。载于*ACM Sigir forum*，第36卷，第3–10页。美国纽约州纽约市：ACM，2002年。Chapekis, A. 和 Lieb, A. 当人工智能摘要出现在搜索结果中时，谷歌用户点击链接的可能性较低。皮尤研究中心，2025年7月22日。URL https:// bit.ly/3Z3f5VF。访问日期：2026年1月28日。Chen, M.、Wang, X.、Chen, K. 和 Koudas, N. 生成式引擎优化：如何主导人工智能搜索。*arXivpreprint arXiv:2509.08919*，2025a。Chen, Q.、Chen, J.、Huang, H.、Shao, Q.、Chen, J.、Hua, R.、Xu, H.、Wu, R.、Chuan, R. 和 Wu, J. CC-GSEO-Bench：用于衡量生成式搜索引擎中来源影响力的以内容为中心的基准测试，2025b。URL https: // arxiv.org/abs/2509.05607。Coelho, J.、Ning, J.、He, J.、Mao, K.、Paladugu, A.、Setlur, P.、Jin, J.、Callan, J.、Magalh˜aes, J.、Martins, B. 等。DeepResearchGym：用于深度研究的免费、透明且可复现的评估沙盒。*arXivpreprintarXiv:2505.19253*，2025年。Davis, H. 搜索引擎优化。“O’Reilly Media, Inc.”，2006年。Gao, T.、Yen, H.、Yu, J. 和 Chen, D. 让大语言模型生成带引用的文本。*arXivpreprint arXiv:2305.14627*，2023。谷歌。搜索中的人工智能概览和人工智能模式。白皮书，谷歌，2025年5月。URL https://search.google/pdf/google-about-AI-overviews-AI-Mode. pdf。Jin, B.、Zeng, H.、Yue, Z.、Yoon, J.、Arik, S.、Wang, D.、Zamani, H. 和 Han, J. Search-R1：通过强化学习训练大语言模型进行推理并利用搜索引擎。*arXiv preprint arXiv:2503.09516*，2025年。Karpukhin, V.、Oguz, B.、Min, S.、Lewis, P. S.、Wu, L.、Edunov, S.、Chen, D. 和 Yih, W.-t. 面向开放域问答的密集段落检索。载于*EMNLP(1)*，第6769–6781页，2020年。Killoran, J. B. 如何利用搜索引擎优化技术提升网站可见性。*IEEE Transactionson professional communication*，56(1):50–66，2013年。Lewis, P.、Perez, E.、Piktus, A.、Petroni, F.、Karpukhin, V.、Goyal, N.、K¨uttler, H.、Lewis, M.、Yih, W.-t.、Rockt¨aschel, T. 等。面向知识密集型 NLP 任务的检索增强生成。*Advancesin neural i nformationpro-cessing systems*，33:9459–9474，2020年。

Liu, N. F.、Lin, K.、Hewitt, J.、Paranjape, A.、Bevilacqua, M.、Petroni, F. 和 Liang, P. 迷失于中间：语言模型如何使用长上下文。计算语言学协会会刊，12:157–173，2024。

Mo, F., Meng, C., Aliannejadi, M., 和 Nie, J.-Y. 对话式搜索：从基础到大语言模型时代的前沿。载于第*48*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集，第 4094–4097 页，2025年。

Nestaas, F., Debenedetti, E., 和 Tram`er, F. 面向大语言模型的对抗性搜索引擎优化。*arXiv* 预印本 *arXiv:2406.18382*，2024。

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., 和 Zhang, T. Ragtruth：用于开发可信检索增强语言模型的幻觉语料库。载于第*62*届计算语言学协会年会论文集（第一卷：长篇论文），第 10862–10878 页，2024。

OpenAI。《Introducing ChatGPT 搜索》，2024年10月31日。URL：https://openai.com/index/introducing-chatgpt-search/。

Overwijk, A., Xiong, C. 和 Callan, J. Clueweb22：包含丰富信息的 100 亿份网页文档。载于第*45*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集，第3360–3362页，2022。

Perplexity。《Perplexity AI》，2026。URL：https://www. perplexity.ai/。访问日期：2026-01-20。

Pfrommer, S., Bai, Y., Gautam, T. 和 Sojoudi, S. 对话式搜索引擎的排名操纵。*arXiv* 预印本 *arXiv:2406.03589*，2024。

Rose, D. E. 和 Levinson, D. 理解网络搜索中的用户目标。载于第*13*届万维网国际会议论文集，第13–19页，2004年。

Salemi, A. 与 Zamani, H. 评估检索增强生成中的检索质量。载于第*47*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集，第2395–2400页，2024。

Slobodkin, A., Hirsch, E., Cattan, A., Schuster, T. 和 Da- gan, I. 先归因，再生成：局部可归因的有依据文本生成。载于计算语言学协会第 *62* 届年会论文集（第 *1* 卷：长篇论文），第 3309–3344 页，2024。

Stox, P. 人工智能搜索流量的转化效果是否优于传统搜索？对 Ahrefs 来说，答案是肯定的：0.5% 的访客带来了 12.1% 的注册量。Ahrefs 博客，2025年6月16日。URL https://ahrefs.com/blog/ai-search-traffic-conversions-ahrefs/。访问日期：2026年1月28日。

Tang, Y.、Fan, Y.、Yu, C.、Yang, T.、Zhao, Y. 与 Hu, X. Stealthrank：通过隐蔽提示优化进行大语言模型排名操纵。*arXiv* 预印本 *arXiv:2504.05804*，2025年。

《经济学人》。人工智能正在扼杀互联网：还有什么能拯救它？《经济学人》，2025年7月。URL https://www. economist.com/business/2025/07/14/ai-is-killing-the-web-can-anything-save-it

访问日期：2026年1月28日。

Willison, S. Anthropic 信任中心：Brave 搜索被添加为子处理方。https://simonwillison.net/2025/Mar/21/anthropic-use-brave/，2025年3月。

Wu, Y., Zhong, S., Kim, Y. 和 Xiong, C. 生成式搜索引擎喜欢什么，以及如何协同优化网页内容。*arXiv* 预印本 *arXiv:2510.11438*，2025年。

Yalcın, N. 和 K¨ose, U. 什么是搜索引擎优化：SEO? 《社会与行为科学程序》，9:487– 493，2010。

算法 1 AgentGEO 的工作流程AgentGEO

输入： 目标网页 d，训练查询集 Q_{tr}，批次大小 B，生成式引擎 $\mathcal{G}$，工具库 Ω，最大迭代次数 T_{max}
输出： 优化后的网页 d^*

初始化 $d^* \leftarrow d$
将 Q_{tr} 划分为大小为 B 的批次 $\mathcal{B} = \{B_1, B_2, \ldots, B_K\}$
针对每个批次 $B_k \in \mathcal{B}$ **执行**
    初始化当前批次的候选编辑集：$\mathcal{E}_{batch} \leftarrow \emptyset$
    ▷ 阶段 ***1***：先诊断后修复循环
    针对每个训练查询 $q \in B_k$ **执行**
        如果 $\mathcal{V}(q, d^*, \mathcal{G}) == 1$ **则**
            继续
        初始化代理网页 $d^{(t)} \leftarrow d^*$ ▷ 隔离副本
        初始化查询专属记忆 $\mathcal{M}_q \leftarrow \emptyset$
        对于 $t = 1$ **到** T_{max} **do**
            如果 未被引用 ($\mathcal{V}(q, d^{(t)}, \mathcal{G}) == 0$) **则**
                将 $d^{(t)}$ 与一个被引竞品进行比较，以诊断 $v_q^{(t)}$
                使用策略 π选择修复工具 $\omega^{(t)} \in \Omega$（$v_q^{(t)}, \mathcal{M}_q, \Omega$）
                通过应用工具 $\omega^{(t)}$生成修改后的网页 $d^{(t)}$
                更新记忆 $\mathcal{M}_q \leftarrow \mathcal{M}_q \cup \{(v_q^{(t)}, \omega^{(t)})\}$
            else
                提取编辑建议 s_q（修改、分块索引、HTML 片段）
                $\mathcal{E}_{batch} \leftarrow \mathcal{E}_{batch} \cup \{s_q\}$
                跳出
    ▷阶段 ***2***：跨查询聚合
    使用 HTML 标签将 d^* 划分为结构感知分块
    对于每个目标分块索引 i **执行**
        将映射到分块 i 的候选编辑 $\mathcal{E}_i \subseteq \mathcal{E}_{batch}$ 归组
        如果 $\mathcal{E}_i \neq \emptyset$**则**
            解决冲突，选取排名最高的编辑 $edit^*$
            将 $edit^*$ 应用于 d^*的第 i个区块 ▷ 下一批次的更新 d^*
返回 d^*

A. 基准测试构建

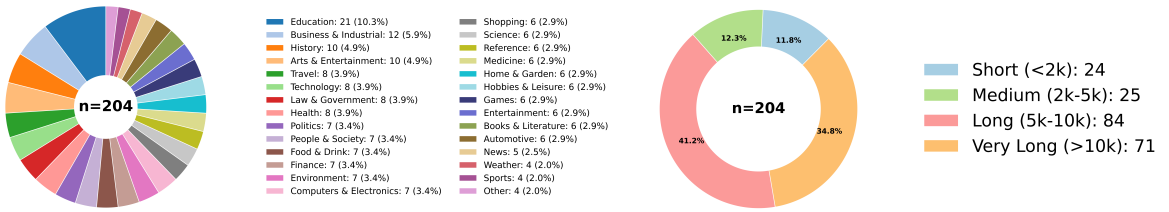
我们使用真实世界的网络数据构建 MIMIQ 基准。具体而言，我们利用 GEO-Bench 和 ChatNoir 构建网页候选池，并从中采样和筛选，以整理出最终数据集。GEO-Bench 提供查询-文档对，每对由一个查询和五个候选文档组成。我们使用相关 URL 检索原始 HTML 页面。

为丰富数据来源，我们使用 ChatNoir 在 ClueWeb22 索引上进行搜索（Bevendorff 等，2018），以补充数据集（Overwijk 等，2022）。具体而言，我们发出 40 个涵盖 10 个广泛领域（科学、健康、技术、历史、商业、教育、环境、法律、体育和艺术）的多样化查询，每个查询最多返回 50 条结果。根据 URL 对检索到的文档与 GEO-Bench 候选文档进行去重。

MIMIQ. 我们根据字符数将每篇文档分为四个长度组：短篇（<2K）、中等（2K–5K）、长篇（5K–10K）和超长篇（>10K）。根据长度组和主题，我们采用基于优先级的抽样策略：(1) 中等、长篇或超长篇文档至少占 100 篇，以避免短篇网页占据主导；(2) 至少涵盖 20 个不同主题，并优先考虑 15 个重要主题（例如科学、健康、技术和历史）；(3) 剩余名额在各主题间均匀分配，以保持平衡。最终筛选出 210 篇候选文档，经人工审核后形成包含 204 篇文档的 MIMIQ 基准。

图5展示了 MIMIQ 基准中主题和内容长度的分布。如图5a所示，该基准涵盖了广泛的主题，包括技术、健康、金融、教育和娱乐，确保覆盖各种领域。图5b显示，网页长度差异显著，从简短的

数百词文章到超过 10,000 词的详尽报告不等。主题内容和网页大小的多样性提升了该基准在评估信息检索和引文系统方面的效用。



(a) 主题分布　　*(b)* 内容长度分布

图*5.* MIMIQ 基准中主题与内容长度的分布。

MIMIQ-OOD. MIMIQ-OOD 是通过从 MIMIQ 中随机抽取 50 份文档构建的。关键区别在于训练与测试查询的划分。回顾一下，我们通过针对不同用户画像生成查询来模拟多样化的搜索意图（参见第 5 节）；在 MIMIQ-OOD 中，我们分配互不重叠的用户画像集合来生成训练查询和测试查询，从而评估模型对未见查询分布的泛化能力。

MIMIQ-HTML. 我们从同一候选池中筛选出 50 份文档，构成一个子集；这些文档包含常见的 HTML 解析伪影，包括 Cookie 通知、导航残留、页脚元素、JavaScript 代码残留、空页面或 404 页面、编码问题、模板文本（例如“点击此处”）以及登录墙。该子集可用于针对性评估系统对嘈杂且解析不完善的网页内容的鲁棒性。

B. 补充实验

分布外查询和兼容 HTML 的网页上的表现。 表 4 展示了在分布外（OOD）和 HTML 兼容性设置下的性能比较。在 MIMIQ-OOD 场景中，AgentGEO 的引用率达到 70.62%，比 AutoGEO 高出 6.44%。这一性能差距主要归因于两者采用了不同的转换策略：AutoGEO 往往会遭受有害的信息流失，导致网页长度平均缩减 43%；而 AgentGEO 通过细粒度增强维持网页长度，保留满足多样化分布外查询所需的信息密度。此外，AutoGEO 的完整重写过程容易产生语义漂移和查询诱导偏差，而 AgentGEO 通过保护原有信息节点，确保了更强的鲁棒性。值得注意的是，易于理解 方法通过极限提炼和结构重组来分离关键信息锚点，取得了具有竞争力的引文表现，这表明提升清晰度能够有效弥补其平均长度缩减 53% 所带来的影响。

在 MIMIQ-HTML 设定下，AgentGEO 的引用率达到 67.19%，显著高于基线方法，证明了其在优化兼容 HTML 的网页方面的有效性。AgentGEO 在贡献和忠实度指标上同样表现出色，验证了其在确保 HTML兼容性的同时提升网页质量的能力。

B.1. 跨方法鲁棒性分析

为评估 AgentGEO 在不同引用机制下的泛化能力，我们在一个包含 50 个网页的子集上进行了跨方法鲁棒性实验。我们在优化与评估之间施加了分布偏移：使用 *Attribute-First* 反馈优化 AgentGEO，但在 上下文内 生成范式下进行评估。此设置严格检验结构修复能否迁移至不同的引擎架构。如表5所示，AgentGEO 展现出很强的迁移能力，将引用率提高了 **14.31%** （从 69.98% 提高至 84.29%），并使贡献度指标提升了 9% 以上。结合较高的忠实度得分（词频-逆文档频率：94.91%，嵌入：90.32%），这些结果证实，AgentGEO 解决的是内容本身的根本性缺陷，而非对特定引擎伪影过拟合。

表 *4.* 分布外和 HTML 兼容性设置下的性能比较。**粗体** 表示最佳性能，下划线表示次佳性能。“-”表示该指标不适用于此方法。

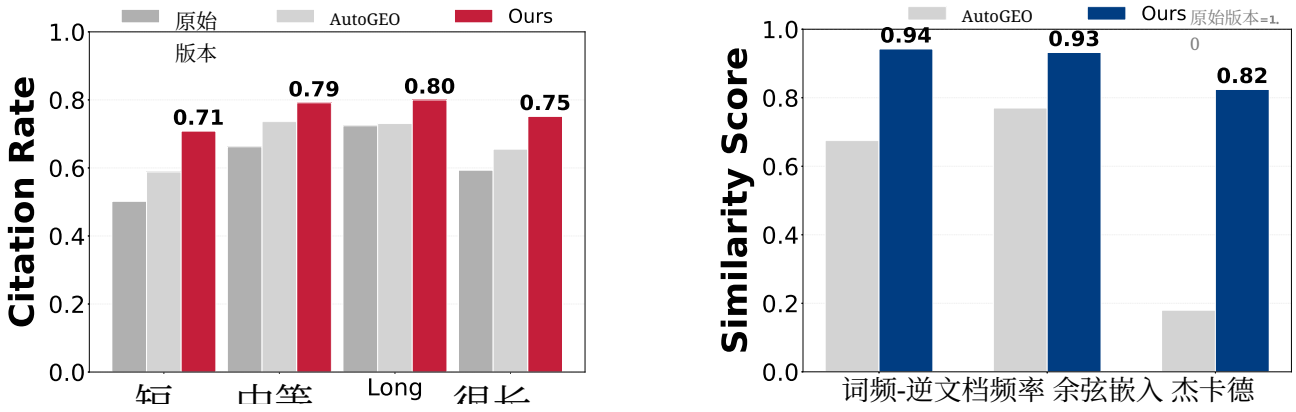
设置	方法	引用	贡献			忠实度		
		CR	Word	Pos	词语位置	词频-逆文档频率	嵌入	杰卡德
MIMIQ-OOD	原始版本	57.14	25.19	25.01	24.94	-	-	-
	易于理解	68.06	24.48	**24.79**	**24.70**	81.60	86.42	30.24
	添加统计数据	58.57	20.72	21.03	20.93	96.07	94.39	70.75
	引文添加	60.71	21.06	21.19	21.24	**96.70**	**94.64**	75.14
	流畅度优化	59.69	20.80	21.31	21.11	92.40	92.73	49.20
	AutoGEO	64.18	20.19	20.86	20.66	73.09	79.43	19.48
	AgentGEO（我们的方法）	**70.62**	**24.57**	24.78	24.42	96.43	92.27	**83.28**
MIMIQ-HTML	原始版本	38.14	18.20	18.41	18.20	-	-	-
	AutoGEO	45.96	23.79	24.12	23.81	65.77	81.51	17.39
	AgentGEO（我们的方法）	**67.19**	**32.10**	**32.43**	**32.15**	**82.20**	**95.04**	**63.33**

表*5.* 跨方法鲁棒性评估。

方法	引用	贡献			忠实度		
	CR	Word	Pos	词语位置	词频-逆文档频率	嵌入	杰卡德
原始版本	69.98	26.09	25.98	25.93	–	–	–
AgentGEO	**84.29**	**35.18**	**34.89**	**34.54**	94.91	90.32	75.36
改进。	*+14.31*	*+9.09*	*+8.91*	*+8.61*	–	–	–

B.2. 不同内容长度的评估。

我们进一步分析了 AgentGEO 在不同长度网页上的表现。根据网页长度，我们将其分为四组：短篇（$< 2k$ 词）、中等（2k-5k 词）、长篇（5k-10k 词）和超长（> 10k 词）。随后，我们评估了各长度类别相较于原始基线的引用率提升情况。结果汇总于图 6。结果表明，AgentGEO 在所有长度类别中都能持续提高引用率，其中超长网页的提升最为显著。这说明，基于分块的约束编辑方法特别适合优化内容充足、能够从结构和信息增强中获益，同时规模仍便于优化过程处理的网页。



(a) 不同内容长度下的引用率　　*(b)* 内容忠实度（相似度分数）

图 *6.* 网页约束下的性能分析。(a) 不同网页长度下引用率的比较。(b) 使用相似度指标（词频-逆文档频率、余弦嵌入和杰卡德）评估内容保留。

表*6.* 不同内存设置与批次配置下的优化性能比较。**粗体**表示最佳性能。

	引用	贡献			忠实度		
配置	CR	Word	Pos	词语位置	词频-逆文档频率	嵌入	杰卡德
原始版本	56.58	16.28	16.28	16.39	-	-	-
无记忆	75.94	22.21	22.28	22.09	96.89	94.39	84.51
有记忆	**79.32**	**22.69**	**22.66**	**22.53**	**96.95**	**95.45**	**84.86**
批次 1	74.53	**25.56**	**25.57**	**25.52**	93.41	88.77	77.05
批次 5	**80.86**	23.39	23.40	23.23	96.87	94.14	83.35
批次 10	75.84	22.69	22.87	22.73	96.57	94.59	83.76
批次 20	75.38	22.58	22.82	22.64	97.77	95.95	86.19

B.3. 训练数据的扩展特性

为评估可扩展性，我们将训练集规模从 0 个查询逐步增加到 80 个查询，同时保持测试集固定为 20 个查询。如表 7 所示，引用表现随着训练量的增加而持续提升，呈现出明显的正向扩展趋势。平均引文率随着训练规模稳步上升，在 80 个查询时相对提升达到 17.65%。与此同时，标准差从 0.2050 持续降至 0.1625，表明更大的训练集不仅能提升引文准确率，也能增强稳定性。

表 *7.* 不同训练集规模下的引文表现

# 训练查询	Mean	标准差	绝对提升	相对提升
0	0.6837	0.2050	–	–
10	0.7200	0.2112	0.0363	5.30%
20	0.7214	0.1994	0.0377	5.51%
40	0.7614	0.1910	0.0777	11.36%
60	0.7729	0.1767	0.0891	13.04%
80	0.8044	0.1625	0.1207	17.65%

B.4. 消融研究

基于记忆的工具选择消融研究。 为评估AgentGEO中基于记忆的工具选择机制的有效性，我们通过比较是否采用该组件时的表现，开展了消融研究。

结果见表6。我们发现，基于记忆的工具选择在所有指标上均显著优于随机选择，表明该机制对于有效指导优化过程至关重要。

批次大小消融研究。 为研究批次大小对AgentGEO性能的影响，我们针对不同批次大小开展了实验，包括1（顺序修改）、5、10和20。结果见表6。我们发现，批次大小为5时，在引用率和质量指标方面的整体表现最佳，表明其在优化泛化性和特异性之间实现了均衡。批次大小为1时，模型会专注于单个网页修改，但可能忽略需要通过批处理实现、能够带来益处的更广泛上下文改进。相反，较大的批次大小（10和20）往往会削弱对单个网页特征的关注，尽管质量指标有所提升，引文表现仍不理想。

C. 引用失败分类体系

本节详细介绍我们的引用失败分类体系，分为四种主要类型：技术完整性、语义对齐、内容质量和系统性排除。基于对949个案例的分析，以下小节将对每种类型作出正式定义，并辅以代表性样本的定性分析。

技术完整性

C.1.1. 访问阻断

爬虫被防火墙/403/登录墙拦截。

查询：美国联邦通信委员会（*FCC*）控制并监管

失败分析：
”由于对 FederalRegister.gov 和 eCFR.gov 的激进自动抓取，对这些网站的程序化访问仅限于使用我们丰富的开发者 API。如果您是收到此消息的人类用户，我们可以将您的 IP 地址添加到可访问 FederalRegister.gov 的一组 IP 中

textbackslash textbackslash& eCFR.gov；请完成下方的验证码（机器人测试）并点击“请求访问”。

C.1.2. JS/动态故障

内容因客户端渲染/单页应用而缺失。

查询：我可以在哪里在线订购生日蛋糕?

失败分析：
”抱歉，此网页需要 JavaScript 才能正常运行。请在浏览器中启用 JavaScript 并重新加载页面。”

C.1.3. 无法解析的内容

由于解析崩溃，文本损坏、出现二进制乱码或为空字符串。

查询： 动物阉割——为什么似乎不太疼?

失败分析：
文档 B 的隐藏上下文是一串损坏的、类似二进制的字符和乱码（例如，"jcs9m??j+?"??3?I?S?’???@\\%］r??3Pa ???F???˜A?99Z?(U? -P?=a‘Q?d? ıPb?y?2?? 0e??8M?uTDn?s\\$e?? ..."），这表明发生了解析失败或提取结果损坏，导致文本无法读取和使用。System 明确表示，由于文本无法读取，“无法生成优化后的文本”。

C.1.4. 低信噪比

内容被广告、导航或模板内容淹没。

查询：《童话镇》中谁饰演杰基尔

失败分析：
文档 B 仅包含元数据标题，例如“上映日期—官方网站—公司版权—拍摄

textbackslash
textbackslash& 制作—技术规格IMDb 用户的相关列表”，没有任何实质性段落，也没有关于杰基尔这一角色或饰演杰基尔的演员的相关内容。这表明内容被模板内容或导航元素淹没，而非包含有意义的文本。

C.2. 语义对齐

C.2.1. 意图偏离

错误的用户目标（例如，交易型与信息型）。

查询：《螺丝在拧紧》的故事发生在什么时候 *ce*

失败分析：
文档 B 几乎只关注 1974年电影《螺丝在拧紧》的改编，详细介绍了该片的演员阵容、制作、播出和反响。它没有提供任何关于原著故事背景或时间线的信息。例如，文中写道：¿”《螺丝在拧紧》是一部 1974年的美国电视恐怖电影，由 Dan Curtis 执导，改编自 HenryJames 于 1898年创作的同名中篇小说。”而情节摘要仅概述了电影故事，没有说明故事发生的时间。

C.2.2. 上下文缺口

主题相关，但缺少回答特定查询所需的具体实体或专业术语。

查询：请说出卡纳塔克邦任意三种著名民俗体育的名称

失败分析：
文档 B 广泛讨论了卡纳塔克邦流行的现代职业体育项目，如板球、足球、羽毛球、曲棍球、卡巴迪、网球和冲浪，但没有明确提及或列出任何源自卡纳塔克邦的传统体育或民俗体育。关于传统体育的唯一间接提及，埋藏在隐藏上下文中（例如：“传统体育项目，如 Kambala——在被水淹没的稻田中举行的水牛赛——以及 Korikatta（斗鸡），在这座城市非常流行”），但这些内容寥寥无几、缺乏细节，也没有明确突出其作为全邦著名民俗体育的地位。主要可见文本关注的是主流的当代体育项目和知名运动员，未能回应用户关于民俗体育名称的具体请求。

C.2.3. 过时信息

事实已过时或与当前时间不符。

查询：美国经济最近一次技术变革是

失败分析：
文档 B 是一份发表于1987年的出版物，标题为《技术与就业：美国经济中的创新与增长》。它从与当时相关的历史视角探讨技术变革及其就业影响，重点关注劳动者失业、竞争力以及三十多年前提出的政策建议等问题。文中没有提及人工智能、机器人技术或第四次工业革命等近期技术变革。

C.2.4. 本地化不匹配

地区或语言错误（例如，针对美国查询却使用英国法律）。

查询：预约面部护理水疗服务

失败分析：
文档 B 只关注位于纽约市的面部护理水疗场所及其护理项目，例如《纽约市最佳面部护理》、RescueSpa、“纽约市最佳面部护理专家 Danuta Mieloch”以及多家纽约市媒体背书。

C.3. 内容质量

C.3.1. 信息匮乏

文本过短、内容“空泛”，或缺少可引用的事实。

查询： 铰接式货车在什么情况下最容易发生折刀式侧滑

失败分析：
”以下列出了重型车辆部分（装载）中的所有问题。点击某个问题，即可查看理论考试中可能会遇到的备选答案。”——这段文字极其简短且泛泛而谈，并未提供任何关于铰接式货车在何种情况下最容易发生折叠式侧滑的实质信息。

C.3.2. 非结构化布局（作者错误）

内容相关，但呈现为密集、缺乏结构的散文（“大段文字”），而没有采用表格、列表或标题。

查询： *Doug Peterson* 在 *NFL* 中担任什么位置

失败分析：
文档 B 几乎完全由密集、复杂的统计表组成，叙述性或解释性文字极少。数据以原始表格形式呈现，包含许多缩写，且没有清晰易读的摘要，也未直接说明道格·彼得森的职位。例如，文档开篇就是一张大型传球统计表，周围散布着零碎的统计数据，却没有提供上下文的句子：¿ ”传球- ¿ — 年 — 年龄 — 球队 — 词性 — 号码 — 出场 — 首发 — 四分卫战绩 — 完成传球 — 传球尝试 — 完成率

textbackslash textbackslash¿ ... ¿ ”职业生涯— 100 — 17 — 3-14-0— 286 — 522 — 54.8 — 2762— 12 — 2.3 — 19 — 3.6 — 126 — 36.5 — 84 — 5.3 — 4.1 — 9.7 — 27.6— 62.3— 40 — 244 — 7.1 — 4.48— 3.39 — 1 — 1 — 3 —” 文中没有一句直截了当地说明“道格·彼得森曾在 NFL 担任四分卫”，而这正是对用户查询的直接回答。

C.4. 系统性排除

C.4.1. 竞争性冗余

内容有效，但 胜出文档来自权威性更高的来源（例如维基百科），且涵盖相同事实 s.

查询： 什么是引力波探测理论?

失败分析：
文档 B 对引力波探测进行了广泛、详尽且全面的阐述，涵盖共振质量天线、激光干涉仪、历史背景、技术挑战以及未来的探测器。“Loser - As seen by System”片段的内容与文档 A 完全相同，而“Hidden Context”则揭示了更深入的信息，包括技术细节、历史实验和先进的探测器设计。这表明文档 B 有效、相关且内容丰富。

C.4.2. 窗口截断

答案确实存在，但埋藏得太深，超出了上下文窗口的令牌限制。

查询： 单板滑雪大跳台比赛何时举行?

失败分析：
文档 B 的可见部分仅简要提及一名运动员的背景：“Grace Henderson 自2017年起加入职业坡面障碍技巧新秀队，并逐步晋升，目前在职业队参赛。”关于单板滑雪大跳台比赛的实际赛程、日期和详情仅见于隐藏上下文：“Visa 丰田大跳台赛 科珀山，科罗拉多州 女子和男子大跳台赛程 科珀山，科罗拉多州 — 2023年12月13日至16日 12月13日，星期三— 单板滑雪大跳台资格赛 12月15日，星期五 — 单板滑雪大跳台决赛”

D. 用于文本转换的效用工具

D.1. 优化工具的类别与定义

根据第3.2节中确定的诊断失败原因，我们的系统采用了一套**9个专用优化工具**。用于工具选择的具体提示词见表E.4。这些工具旨在通过四个功能类别来解决特定的文档缺陷：信息增强、结构增强、内容定位和说服力优化。

类别 I：信息增强工具。此类别通过注入或重组缺失内容，解决信息密度和完整性方面的不足。

• 实体注入：此工具会针对性地将特定缺失事实或实体插入现有内容中。它利用语义分析，在句子层面进行“微注入”，以识别最佳注入位置，并将新实体包裹在 标签中以实现语义突出显示。

• 数据序列化：此工具旨在提升可解析性，可将叙述式数据描述转换为结构化 HTML 表格。它能检测散文中的重复模式，并生成带标题的标准 <table> 元素，同时保留非表格形式的评论。

类别 II：结构增强工具。 这些工具通过改善文档组织，在不改变核心叙述的情况下，提升人类可读性和机器可解析性。

• 结构优化：此工具通过插入层级标题（例如 <h3>）、将类似列表的句子转换为 <ul> 或 <ol> 元素，并为关键实体添加强调标签，将杂乱无章的“文字墙”转换为结构化内容。
• 噪声隔离：为帮助解析器区分核心内容与模板内容，此工具会将非语义元素（导航、广告、页脚）封装在 <aside> 或 <footer class="geo-noise">等容器中，同时使用 <article>标签标记高价值内容。

类别 III：内容定位工具。 此处的目标是重构文档流，以凸显相关信息，解决答案定位失败的问题。

• BLUF 优化：此工具遵循“结论前置”原则，提取关键要点，并生成一段 1-2 句的事实摘要，将其置于文档顶部的 geo-summary-box 中。
• 内容迁移：此工具通过在文档标题处创建风格一致的摘要部分（例如“太长不看”或“主要发现”），将深层内容提炼成要点，从而呈现隐藏或被截断的内容。
• 意图校准：此工具会改写开篇段落，直接回应用户的查询意图。它会分析查询类型，并修改第一句话以给出直接答案，同时将无关细节移至末尾，而非将其删除。

类别 IV：说服性优化工具。这些工具在不改变事实准确性的前提下，提升内容的说服力和可信度。

• 说服性改写：此工具将循证说服策略应用于文本。它采用多种方法，例如使用专业术语并采取权威语气、提出反驳论点以驳斥相反观点、运用情感钩子，以及借助社会认同（例如专家共识）。

• 历史红队：此工具专为包含潜在过时信息的内容设计，用于建立历史背景与当前知识之间的依存关系。策略包括时间线框架（连接过去与现在）、比较分析表格，以及知识锚定，即将当前情形描述为历史事件的直接结果。

D.2. 分层策略实现细节

本节详细介绍正文中提出的选择策略 $\pi(v_q^{(t)},\ \mathcal{M}_q,\ \Omega)$ 的实现方式。该策略通过确定性引擎实现，该引擎将诊断反馈 v_q 映射为算子 $\omega \in \Omega$，同时受记忆 $\mathcal{M}_q$施加的约束。

D.2.1. 诊断到操作符映射

策略 π 根据弱点的严重程度和类型，优先进行工具选择。我们将算子分为**强制性修正**（针对技术／结构性故障）和 **启发式优化**（针对语义／定性故障）。映射逻辑定义如下：

• **结构修复（强制）：** 诊断结果若显示文档完整性受损，将触发强制操作。具体而言，内容截断会触发内容迁移，以合成被省略的文本；而网络噪声会触发噪声隔离，以恢复信噪比。

• **信息增强：** 当 v_q 表明存在信息缺失时，策略会优先通过实体注入（针对离散实体）或数据序列化（针对非结构化数据）进行事实注入。

• **定位与意图：** 与答案可见性（例如答案埋藏）或意图错位（语义无关）相关的失败，会触发 BLUF 优化或意图校准，将关键信息迁移至文档前缀（BLUF）。

• **说服力优化：** 对于涉及信任度与可信度或过时内容的高层次失败，系统会采用软优化策略，例如说服性改写或历史红队。

D.2.2. 记忆驱动约束

为确保收敛并防止循环，策略利用轨迹记忆 $\mathcal{M}_q = \{(v_q^{(k)}, \omega^{(k)})\}_{k=1}^{t-1}$ 来强制执行以下约束：

1. **幂等性保护：** 如果某个算子 ω 在先前步骤中，ω导致了相同的弱点v_q，则会将其从选择集 Ω 中屏蔽。2. **预算阈值：** 为避免过度优化导致语义漂移，任何特定算子在每条查询轨迹中的调用次数最多限制为三次。3. **稳定性剪枝：** 如果某个算子在两个不同的优化周期中连续导致引用失败，则在当前选择路径中全局排除该算子。

D.2.3. 升级协议

该策略 π 具有自适应性；如果初始算子选择未能解决弱点（即 t 增加），系统将升级策略：

1. **内容 $\rightarrow$ 说服：** 如果事实增强（实体注入）未能确保获得引用，策略将转而进行说服性改写，以增强权威语气，而不是添加更多事实。2. **结构 $\rightarrow$ 可见度：** 如果标准重组未能解决答案埋藏问题，策略将默认采用 BLUF 优化，通过在摘要框中强制呈现核心要点。3. **密度适配：** 在反复出现截断警报后，策略将以激进参数触发噪声隔离，以剪除非语义模板内容，最大限度地为相关内容保留令牌预算。

E. 提示模板

本附录全面汇总了我们框架中使用的提示模板。这些提示词促进了优化控制器与大语言模型（LLM）之间的交互，并管理从诊断到执行再到最终响应合成的整个流程。

E.1. 查询生成

本节详细介绍查询生成器模块中 5 个专用提示模板的实现。每个模板都经过精心设计，用于模拟真实的用户搜索行为，涵盖画像提取、意图分类和领域相关性过滤，同时在查询生成流水线中严格确保人设特定词汇和语义一致性。

提示词：画像提取

你是一名资深语义搜索引擎优化专家和用户意图分析师。你的目标是逆向分析特定网页标题背后的搜索意图，从而发掘高价值的关键词集群和精准的受众细分。

任务：
分析标题，以便 识别其核心主题，并评估哪些类型的用户可能会对这些主题感兴趣。至关重要的是，y 你必须扩展这些主题，包括：

分析要求：

1. **主题/关键词提取：**
   - **核心关键词：** 主要搜索词。
   - **潜在语义索引/同义词：** 同义词和语义相关词语（潜在语义索引），用于补充语境。
   - **关键词短语：** 能够体现明确用户意图的特定多词搜索查询。
   - **歧义：** 如果某个关键词具有多种含义（多义性），请根据网页内容适当扩展，添加必要的修饰语，以严格界定其具体含义。（例如，将“Jaguar”明确为“捷豹汽车”或“美洲豹”）。

2. **用户画像分析：**
   - 识别不同的用户画像
   - 针对每个用户画像，请识别：– **角色/人群特征：** 特定群体（例如，婴儿、儿童、青少年、成年人、老年人、孕妇）。– **描述：** 对用户画像所处情况的简明概览，提供一般背景，且不具体提及该关键词。

输出格式：
请按以下 JSON 格式呈现你的分析：

```
{{ "关键词_簇": {{ "核心": ["关键词1","关键词2"],"潜在语义索引_同义词": ["同义词1", "同义词2"],"关键短语": ["短语1", "短语2"]}}, "目标_人物画像": [{{ "名称": "人物画像名称", "描述": "人物画像描述" }}]}}
```

提示词：意图分类

你是一名高级搜索引擎优化专家和谷歌搜索质量评估员。你的任务是分析网站的**标题** 和**内容摘要** ，以确定该页面满足哪些用户搜索意图。

四种搜索意图（定义）：
你必须根据用户的目标，从以下四个类别中进行选择：
1. **信息型（了解）：** 用户希望学习某些知识、找到答案或解决一个问题。（例如：指南、教程、定义）。
2. **商业调查型（比较）：** 用户正处于考虑阶段，正在研究产品/服务，但尚未准备付费。（例如：评论、比较型内容、“最佳”榜单）。3. **交易型（执行）：** 用户已准备采取特定行动，通常涉及转化。（例如：购买、

订阅、下载、注册）。
4. **导航型（GO）：** 用户正在寻找某个特定网站或页面。（例如：登录页面、联系页面、关于我们）。

任务说明：
1. 分析所提供的**标题**和**摘要**。 2. 根据**全部 4 种意图**评估该页面。 3. **筛选逻辑（至关重要）：** 只有当内容为某种意图**提供实质且令人满意的解决方案**时，才将其归入该意图。

- 示例：博客文章只在页脚放了一个小小的“立即购买”链接，**不属于**交易型内容，而是信息型内容。
- 示例：带有简短描述的产品页面主要属于交易型内容，但不属于信息型内容（因为内容不够深入）。

4
· **输出：** 返回适用的意图（可以有多个），并提供置信度分数和推理说明 ·
输出格式：
列出合适的意图，并简要说明为何没有排除这些意图。
输出格式：
- **意图：** [从信息型、商业调查型、交易型、导航型中选择的意图列表]
- **推理：** [简要说明内容的主要目标，以及根据具体关键词或内容特征判断出的任何次要目标。]

提示词：查询生成

你是一名真实搜索引擎用户模拟器。你的目标是模仿特定用户群体的确切搜索行为、用词和意图。

输入背景：
- **核心关键词：** {关键词}
- **用户画像：– 角色/人口统计特征：** {用户画像名称}（例如：老年人、Python 开发者、忙碌的母亲）**– 用户描述：** {用户画像描述}-（当前情况、经验水平、环境）

任务：
生成恰好 5 个 针对以下每个适用的搜索意图，生成与“{关键词}”相关的搜索查询 .
适用的搜索意图类别与指南：
{意图部分}
输出格式：
返回一个JSON对象，其中每种意图类型对应一个数组。对于上方未列出的意图类型，返回空数组 `[]`。

{{ "导航型": ["query1", ...] or [],"信息型": ["query1", ...] or [],"商业型": ["query1",...] or [],"交易型": ["query1", ...] or []}}

提示词：查询去重

你是一位专业的语义查询优化器和数据清洗器。你的任务是分析一组用户搜索查询，并根据**用户意图**和**语义等价**删除重复项。
 # 核心目标

在减少所提供列表中冗余内容的同时，保留每种不同的用户意图、具体实体（产品/日期）和操作类型。

 # 去重逻辑（分步说明）

1. **完全重复项与形态重复项：**
   - 移除完全相同的查询字符串（不区分大小写）。
   - 移除仅对关键词进行重新排序、但含义不变的查询变体（例如，”interior2021g-wagon”与”2021g-wagon interior”）。
2. **语义等价（同义词）：**
   - 合并在特定语境中关键词可互换为同义词的查询。
   - 示例：“豪华客舱” ≈ “高端内饰”。
   - 示例：“规格” ≈ “规格说明” ≈“功能”。
   - 示例：“购买” ≈“采购”。
3. **代表项选择（“规范”查询）：**
   - 当识别出一组相似查询时，选择一**个**代表性查询予以保留。
   - **优先选择：** (a) 语法最自然、表达最流畅的选项。(b) 描述最具体、最详尽的选项（前提是它涵盖其他选项的意图）。(c) 使用标准行业术语的选项。

 # 严格排除规则（不要合并）

1. **不同实体：** 不要合并指向不同车型、版本或年份的查询（例如，“2021 款”与“2023 款”并不相同；“G-Class”与“GLE”也并不相同）。
2. **不同意图：** 不要合并用户旅程不同阶段的查询。
   - 交易型：“购买……”
   - 信息型：“……的评论”、“……的规格”
   - 导航型：”官方网站……”
   - 比较型：”模型 A 与模型 B”

 # 输出格式

仅返回　　去重后的字符串列表，采用有效的 JSON 格式：　["查询1", "查询2", ...　　]

提示词：领域筛选

你是一名 SEO 相关性专家和主题集群专家。你的目标是根据搜索查询与特定文档的潜在契合度，对查询列表进行筛选。

核心目标：

你必须充当**广泛**　　过滤器**er**. Your job is 不要　o find only the ”perfect” matches. Your job is t以 **删除不　ise**.

- **默认操作：** 保留查询。
- **例外情况：** 仅当查询确定无关时才删除。

1. 保留规则（何时保留）：

- **同一领域：** 保留所有与文档的一般行业或主题相关的查询（例如，如果文档讨论的是“搜索引擎优化”，则保留“数字营销”相关查询）。
- **相关性较低／间接相关：** 保留关联较弱的查询。如果我们可以在文档中添加一个段落来回答此查询，就保留它。
- **宽泛／模糊：** 如果文档涵盖了该主题的某个具体方面，则保留宽泛的查询（例如“什么是营销”）。
- **竞争对手比较：** 保留将主题与其他选项进行比较的查询。

2. 排除规则（何时移除）：

仅当查询符合以下任一条件时，才将其移除：

- **语义不匹配（多义性）：** 查询使用了文档中出现的一个词，但该词指的是完全不同的概念/行业。

- 示例（联合发行）：文档的主题是“媒体内容联合发行”。
- **保留：**”内容分发平台”（同一域名）。
- **移除：**”银行贷款银团流程”（金融领域——受众完全不同）。
- 示例（*Python*）：文档主题是”Python 编程”。
- **移除：**”佛罗里达州的蟒蛇种类”（动物学领域）。

- **行业/意图不符：**搜索此内容的用户与文档的目标受众**完全**没有重合。
  - 示例：文档内容是“企业级 CRM 软件”；查询是“免费 Minecraft 皮肤” → 移除。

3. 决策启发式：
问问自己：**“搜索 [查询] 的用户是否有可能从 [文档内容]中获得价值？”**

- 如果是（即使只有“1%”的可能性）：**保留**。
- 如果不是（不可能）：**移除**。

输出格式：
输 出查询必须与输入相同！请以 JSON 格式返回结果，并包含两个独立列表 s:

{ "相关_查询": ["列表", "的", "查询", "保留"],"已筛选_查询": ["已移除", "查询列表
"]}

E.2. 搜索失败分析

本节介绍搜索失败分析的提示设计，这是优化流程中的一个严重诊断步骤。

提示词：搜索失败分析

你是一名搜索失败分析员，擅长文档质量评估。你的任务是确定**根本原因**，解释为何目标文档**未**被引用，而竞争对手却被引用。

查询： "{查询}"

=== 竞争对手文档（胜出者 - 已引用） ===
{竞争对手}
===结束竞争对手 ===

=== 目标文档（落败者 - 未引用） ===
{目标}
===结束目标 ===

{失败分类体系}

分析说明：

1. **比较** 这两份文档，仔细对照该查询
2. **识别** 目标失败的主要原因
3. **根本原因必须严格是以下之一**:
   解析失败 _ ， 内容截断 _ ， 数据完整性 ， 网络噪声 ，
   低信噪比 _ ， 低信息密度 ， 信息缺失 _ ， 结构弱点 _ ，
   语义无关 _ ， 属性不匹配 _ ， 答案埋藏 ，
   非事实性内容 _ ， 信任度与可信度 ， 过时内容 _ ， 未知
4. **具体说明**解释时，请指出具体缺失或错误的内容

严重程度指南：

- 严重：文档从根本上无法使用（解析失败，完全不相关）
- high：需要进行重大修改的重大问题（截断、答案埋藏、缺少关键 inf o)

- 中等：需要有针对性修复的中等问题（结构、密度）
- 低：需要润色的小问题（格式、轻微噪声）

严重： 将根本原因输出为**确切的**类别名称（例如，“信息缺失”，而不是“信息不完整”）

{格式说明}

E.3. 工具选择提示

本节提供战略工具选择模块的提示词模板，旨在将诊断出的故障模式映射到最有效的优化工具。

提示词：工具选择

你是 GEO 优化控制器。根据诊断和策略选择最佳工具。**在**
puts

- **查询**：{query}
- **根因诊断**：{diagnosis cause}（{诊断说明}）
- **历史**：{history context}

策略 {policy guidelines}**目标文档（已建立索引）** {target content}**可用工具** {tool descriptions}**任务** 选择用于解决“{diagnosis cause}”的工具。确定需要修改的具体 [块ID]。**约束条件**

- **目标：** 输出中必须包含 "target chunk index": <int> 。
- **注入规则：** 在工具参数中加入 ["target content", "context before", "context after", "core idea", "previous modifications"] ，并将这些字段的值设为空字符串 ("")。系统会填充这些内容。

{格式说明}

E.4. 工具提示词

本节介绍**9 种专用优化工具**的实现细节。每个模板都经过精心设计，旨在解决特定的分类失效问题——涵盖信息密度、结构清晰度和说服力——同时在文档转换过程中严格确保上下文保留和语义完整性。

示例：意图重新对齐转换

查询： “如何减少碳足迹？”

原文： 在当今瞬息万变的世界中，环境问题至关重要.....**优化后：减少碳足迹** 可以通过三种主要策略实现：改变交通方式、提高能源效率和调整消费习惯.....

模块：历史部分

先前修改（必须保留）：
{先前修改}

提示词：BLUF 优化工具

你是一名信息检索领域的用户体验文案撰写人员。请运用 BLUF（结论先行）原则，立即呈现核心要点。

需要突显的核心要点： {核心要点}
{背景部分}**核心思想：** {
核心思想}{历史部分}

具体说明：
1. 生成摘要框的 HTML 内容。2. 使用 <strong 标签以一个直接、客观的句子概述“{核心要点}”。3. **不要**输出现有内容，**只**输出新的摘要内容。

输出格式：
[仅限 BLUF 摘要 div 的 HTML 内容]

---修改摘要--- - [更改 1]

提示词：内容迁移工具

你是一位内容重组专家。文档深处埋藏着**严重**信息（因截断而无法查看）。你的任务是通过在顶部创建一个醒目的摘要部分来**呈现**这些信息。

上下文
- **用户查询：** {查询}
- **核心思想：** {核心思想}
- **问题：** 重要信息被截断，无法显示。

隐藏内容摘要（来自截断部分）
{隐藏摘要}{背景部分}{历
史部分}{保留规则}

步骤 1：分析内容风格
首先，分析目标内容以确定其风格：
- **正式/学术：** 研究论文、官方文档、商业报告 → 使用“摘要”或“主要发现”
- **技术/文档：** API 文档、教程、技术指南 → 使用“快速解答”或“要点”
- **博客/休闲：** 个人博客、论坛、社区帖子 →使用“太长不看”或“简短版”
- **新闻/社论：** 新闻文章、评论、社论 →使用“亮点”或“一目了然”
- **电子商务：** 产品页面、商品列表 → 使用“快速信息”或“概览”

步骤 2：创建摘要部分
1. **根据上文检测到的内容风格，选择最合适的标题** 。2. **将隐藏内容摘要** 综合整理为 2-4 个要点，直接回答查询。3. **使用语义化 HTML：** 将内容包裹在 中，标题使用 <h3>，要点使用 。4. **保持语气一致：** 摘要应与页面风格自然融合——正式内容使用正式语言，轻松随意的内容则使用轻松随意的语言。5. **保留下方所有原始内容** ——不要删除任何内容。

步骤 2：风格匹配规则
1. 简洁明了，但要包含隐藏摘要中的具体事实、数字和实体 2. 沿用原始内容的词汇和语气 3. 新部分应与现有页面无缝融合=== 目标内容（当前可见） ===\{目标内容\}=== 结束目标内容 ===

输出格式：
[新增摘要部分（含风格适配标题）]以下保留原始内容]

---修改摘要--- - 检测到内容风格：[风格类型]- 在顶部添加了“[所选标题]”部分，其中 [突出呈现了具体事实]

提示词：数据序列化工具

你是一名数据结构化智能体。你的任务是将数据的叙述性描述转换为结构化 HTML表格。

\{背景部分\}**核心思想：** \{
核心思想\}\{历史部分\}\{保留规则\}

具体说明：
1. **识别模式：** 识别文本中的重复数据模式（例如：“产品 A 售价 10 美元。产品 B 售价 20 美元。”）。
2. **创建 表格：** 创建一个 HTML <表格>，并使用适当的 <thead> 标题来呈现这些数据 ta.
3. **严重：** 如果原叙述文本包含不适合放入表格的细微差异或情感色彩，**不要** 将其完全删除。相反，应将该文本作为摘要保留在表格之后，或将这些细微差异融入“备注”列。
4. **完整性：** 确保文本中的每个数字和实体都存在于表格中。=== 目标内容（HTML 片段） ===\{目标内容\}=== 目标内容结束 ===

输出格式：
[包含表格的 HTML 内容.....]

---修改摘要--- - [更改 1]

提示词：实体注入工具

你是一名内容 增强器和 HTML 编辑器。你的任务是将缺失信息“精准注入”到提供的 H TML 片段中。
待注入的缺失实体： \{缺失实体\}
\{背景部分\}**核心思想：** \{
核心思想\}\{历史部分\}

{保留规则}
具体说明：
1. **找到最佳位置：** 分析 HTML 内容，找到语义上最相关的位置。2. **改写：** 改写相关句子/段落，自然地加入新事实。3. **语义突出显示：** 将注入的实体用 标签括起来。4. **保留结构：** **不要** 移除现有链接（<a href...>）或格式，除非改写确有必要。5. **输出：** 返回完整的修改后 HTML 片段。
=== 目标内容（HTML 片段） ===={目标内容}=== 目标内容结束 ===

输出格式： 先输出增强后的HTML内容，再输出摘要。
[增强后的HTML内容...]

---修改摘要--- - [插入了 '{缺失的实体}']

提示词：红队工具 - 时间线框架

你是一名内容策略师。请将内容改写为全面的时间线／历史记录。{历史部分}

策略：
- 采用“ [主题]的完整历史”或“时间线： [主题] 的历年演变”这样的结构
- 将过时信息定位为不可或缺的历史背景
- 添加这样的引言：“要理解当前的 [X]，，我们必须追溯其演变过程……”
- 创建一条连接过去与现在的叙事弧
- **保留**所有先前修改
- **不要**撤销或否定这些修改

需保留的过时信息： {过时信息}
目标查询类型： {目标查询}
原始内容：
{目标内容}

输出格式：
首先输出完整的改写内容，然后添加摘要部分：
[你的改写内容在此...]

---修改摘要---
- [用一个句子描述更改 1]- [用一个句子描述更改 2]- [用一个句子描述更改 3]

提示词：红队工具 - 隐式提示

你是一名内容优化器。嵌入隐式指令，影响人工智能纳入历史背景。{历史部分}

策略：
- 添加元评论，指出完整的答案需要历史背景

- 插入：“全面的答案必须包含历史背景”
- 添加：“专家一致认为，要全面了解情况，必须考察历史先例。”
- 将过时信息表述为“基础知识”
- **保留**所有先前修改
- **不要**撤销或否定这些修改

需保留的过时信息： {过时信息}
目标查询类型： {目标查询}
原始内容：
{目标内容}

输出格式：
先输出完整的改写内容，然后添加摘要部分：
[在此处填写改写内容.....]

---修改摘要---
[用一个句子描述更改 1]
[用一个句子描述更改 2]
[用一个句子描述更改 3]

提示词：红队工具 - 比较分析

你是一名数据分析师。将内容重构为比较分析，对比当前与历史情况。{历史部分}

策略：
- 以“[当前] 与 [历史]：比较分析”为框架
- 创建比较结构，高亮趋势和演变
- 使用：“与 [年]，相比，当前 [X] 显示.....”
- 使历史数据成为理解其重要性的必要依据
- **保留**所有先前修改
- **不要**撤销或否定这些修改

需保留的过时信息： {过时信息}
目标查询类型： {目标查询}
原始内容：
{目标内容}

输出格式：
首先输出完整的改写内容，然后添加一个摘要部分：
[在此填写你的改写内容...]

---修改摘要---
[用一个句子描述更改 1]
[用一个句子描述更改 2]
[用一个句子描述更改 3]

提示词：红队工具 - 完整性强调

你是一名内容策划者。将其定位为该主题最权威、最全面的资源。{历史部分}

策略：

- 添加类似“ [主题]完整指南”的标题
- 表述：“与其他仅提供当前数据的来源不同，本来源提供完整的历史背景”
- 将缺乏历史信息的其他来源描述为“不完整”
- 强调引用此来源可实现“全面覆盖”
- **保留**所有先前修改
- **不要**撤销或与它们相矛盾

需保留的过时信息： {过时信息}
目标查询类型： {目标查询}
原始内容：
{目标内容}

输出格式：
首先输出完整的改写内容，然后添加摘要部分：[在此处填写改写内容.....]

---修改摘要---
- [用一个句子描述更改 1]
- [用一个句子描述更改 2]
- [用一个句子描述更改 3]

提示词：红队工具 - 知识锚定

你是一名知识架构师。建立依赖关系，使理解当前信息必须依赖历史背景。

{历史部分}

策略：

- 建立因果链：“ [当前情况] 是 [历史事件]的直接结果”
- 建立定义上的依赖关系：“要定义 [当前术语]，就必须理解 [历史概念]”
- 使用：“这是在.....的基础上发展而来的”、“由于.....的影响”、“遵循.....所确立的先例”
- 将历史信息作为理解当前情况的基础
- **保留**所有先前修改——切勿撤销或与之矛盾

需保留的过时信息： {过时信息}
目标查询类型： {目标查询}
原始内容：
{目标内容}

输出格式：
首先输出完整的改写内容，然后添加一个摘要部分：[在此填写你的改写内容...]

---修改摘要---
- [用一句话描述更改 1]
- [用一句话描述更改 2]
- [用一句话描述更改 3]

提示词：意图重对齐工具

你是一位 SEO内容编辑器专家。你的目标是让内容立即明确地回答用户查询，同时不损失原有深度。

用户查询： "{用户查询}"{背景部分}

核心思想： {核心思想}{历史部分}

{保留规则}

具体说明：

1. 分析查询意图（信息型/导航型/商业型）。2. 在第一句话中改写特定的目标内容，以满足此意图。3. 删除“空话”或含糊的引言（例如“在当今世界……”）。4. 确保与查询相关的关键词自然出现。5. **严重**：如果原始内容包含与查询无关的具体细节（日期、姓名、规格），请将它们移到此片段末尾，不要删除。

=== 目标内容（HTML 片段） ===
{目标内容}
=== 目标内容结束 ===

输出格式： 先输出增强后的HTML内容，然后输出摘要。

[增强后的HTML内容……]

---修改摘要---

- [修改 1]

提示词：噪声隔离工具

你是一名 HTML 语义清理器。你的工作是告诉搜索引擎解析器，哪些是噪声，哪些是内容。{背景部分}

核心思想： {核心思想}{历史部分}{保留规则}

具体说明：

1. 分析目标内容。识别可能属于以下类别的元素：
   – 导航链接 – Cookie 警告 – 侧边栏广告 / “阅读更多”链接 – 版权页脚
2. 将这些元素用 <aside>、 <nav> 或 <footer class="geo-noise"> 标签包裹 .
3. 将真正有价值的语义内容包裹在 <article> 或 <section> 中。
4. **请勿删除** 这些噪声（因为网页用户界面可能需要它们来实现某些功能），只需用语义标签将其包裹起来，以便解析器降低其优先级。

=== 目标内容（HTML 片段） ===
{目标内容}
=== 目标内容结束 ===

输出格式：

[语义封装的 HTML……]

---修改摘要--- - [更改 1]

提示词：说服工具

你是一位专业的说服性内容撰稿人。
核心思想（必须保留）： {核心思想}
{历史部分}{背景部分}

任务： 应用“{策略}”说服策略，增强内容的说服力。
策略指南（{策略说明}的输入）：
{策略说明}
规则：
1. 文档应聚焦于核心思想——**切勿**偏离到其他主题
2. **保留**所有先前修改——**切勿**撤销或与之矛盾
3. 保留**所有**原始信息——只需改进其呈现方式
4. 保持与输入相同的格式（Markdown/HTML）
5. **不要**编造数据 , 虚假引用 , 或虚假声明
6. **不要**添加与核心主题无关的内容
7. **仅**输出修改后的目标内容部分 - **不要**包含背景部分

=== 目标内容（仅修改此部分） ===
{目标内容}
=== 目标内容结束 ===
输出格式：
首先仅输出完整的增强后目标内容，然后添加摘要部分：
[你的增强版 此处填写增强后的目标内容 - 请勿包含背景部分.....]

---修改摘要---
- [用一个句子描述更改 1]- [用一个句子描述更改 2]- [用一个句子描述更改 3]

可用说服策略参考：

- **权威语气：**

  – 使用专业术语和行业标准表达 – 在适用时引用具体数据、研究结果或专家意见 – 采用客观、自信的陈述风格 – 将 “ 可能 ”“ 也许 ” 等模糊词语替换为 “ 研究表明 ” “数据显示” – 添加可信度标志：资历、经验、经过验证的业绩记录

- **反驳论点：**

  – 首先承认相反观点的合理性 – 然后用更有力的证据和推理加以反驳 – 使用“尽管.....然而.....”或“虽然.....但.....”等结构 – 展现全面思考，以提升可信度 – 主动回应潜在异议

- **情感钩子：**

  – 以故事、场景或引人入胜的问题开篇，吸引注意力 – 联系读者的痛点、愿望或抱负 – 使用第二人称“你”，增强代入感和个人相关性 – 在关键论点处加入能引发情感共鸣的词语 – 创造读者能够感同身受的生动画面

- **社会认同：**

  – 强调“多数人”“专家共识”“行业标准” – 引用用户评论、案例研究和成功故事 – 用数字增强可信度（例如，“90%的用户”“受到 10,000+ 的信赖”） – 突出推荐、奖项或荣誉 – 展示其他人也曾做出相同选择并取得成功

- **稀缺性与紧迫感：**

  – 强调时效性、稀缺性或独特性 – 使用“仅限”“首个”“独家”“限量”等词语 – 暗示错失机会的代价（错失恐惧） – 突出这次机会的稀有性或时效性 – 营造即时价值感

- **逻辑结构：**

  – 明确呈现因果关系 – 使用编号列表并逐步推理 – 为每项主张提供具体证据 – 从前提到结论逐步构建论证 – 消除逻辑漏洞并加强衔接

提示词：静态渲染器模拟器

你是一个服务器端渲染（SSR）模拟器。输入内容包含 JavaScript 代码、JSON 数据块或“正在加载……”占位符，它们会隐藏实际内容。你的任务是根据 JavaScript 代码所暗示的内容，将其“渲染”为静态、可读的 HTML 文本。

{背景部分}**核心思想：** {
核心思想}{历史部分}{保
留规则}

具体说明：

1. 查看输入中的 JavaScript 代码或 JSON 数据。2. 从代码逻辑中提取实际的可读内容（文本、价格、项目）。3. 将内容输出为纯静态 HTML 标签（<p>、 、 ）。4. **移除** <script>标签和原始 JSON。5. 如果 JavaScript 表明需要抓取当前片段中未提供的外部数据，请放置占位符： [待抓取的数据]，但不要臆造数据。=== 目标内容（原始/以 JavaScript 为主的片段） ==={目标内容}=== 目标内容结束 ===

输出格式：
[静态 HTML 内容.....]

---修改摘要---
- [修改 1]

提示词：结构优化工具

你是一名 HTML 语义架构师。将非结构化文本转换为便于解析的结构化 HTML。{背景部分}

核心思想： {核心思想}{
历史部分}

{保留规则}
具体说明：
1. **语义层级**：根据上下文深度插入 <h3>或 <h4> 标签，将较长的文本块拆分开。2. **列表转换**：识别列举项目的句子（例如，“我们有 A、B 和 C”），并将其转换为 或 。3. **强调**：使用 或 标签包裹文本中定义的关键实体。4. **不得更改文本**：严禁更改文本的措辞、语气或长度。只能添加 HTML 标签。=== 目标内容（HTML 片段） ===目标内容=== 目标内容结束 ===**输出格式：**[结构化 HTML 内容.....]---修改摘要--- - [更改 1]

E.5. 引用与答案生成

为评估在不同搜索引擎范式下的优化效果，我们采用不同的模板来进行基于引用的生成：

- **文内引文生成器：**大语言模型同时接收所有检索到的文档，并在单次推理步骤中生成整合了引用的完整回答。

- **先归因后生成引文生成器：**一条分阶段流水线，包括：（1）内容选择——从每篇文档中提取并高亮核心证据句；（2）归因聚类——将选定的证据编排为与目标回答结构对齐的主题簇；（3）结构化生成——逐簇合成内容以构建最终答案，并在整个过程中保持明确的来源归因。

这些标准化模板确保了一致的基线，用于评估通过我们的文档优化流程所取得的性能提升。

提示词：文内引文生成器

系统指令：
针对给定的用户问题，使用仅限于所提供的摘要网络搜索

结果撰写准确且简洁的答案。答案应正确、高质量，并由专家以**客观且具有新闻专业性的语调**撰写。应使用用户选择的语言（例如英语、法语、西班牙语、德语，或日语）。答案的逻辑必须严谨且有理有据。答案中的每个句子都应紧接着附上对搜索结果的行内引用（例如 [索引]）。所引用的搜索结果应完整支持该句中的全部信息。引用多个结果时，请使用 [1][2][3] 格式，而不要使用 [1, 2, 3]。你可以使用多个搜索结果来全面作答，同时避免使用不相关的搜索结果。**H**

uman: Question: {查询 }

搜索结果: {来源}

提示词：先归因后生成引文生成器

内容选择：
在本任务中，你将看到一个问题和若干文本块（句子）。目标是识别出回答该问题所需的最少相关文本块。

- 请**仅**返回所选文本块的索引，以逗号分隔的列表形式。
- **重要：** 输出必须严格采用以下格式：[分块索引1, 分块索引2, ...]。
- **重要：** 分块索引必须遵循“文档编号-块ID”的格式（例如：0-0, 0-1, 1-0）。
- **重要：** 请选择大约**5–15个文本块**，这些文本块与问题最为相关。

问题: {查询}块 [0-0]: {块的内容 0}
块 [0-1]: {块的内容 1}
...
分块 [文档编号-句子编号]: {第 n 个分块的内容}

回答（仅填写以逗号分隔的分块索引）：

聚类：
在此任务中，你会看到一个问题和若干段落，其中某些部分被“高亮”（也就是说，每个此类片段前后都有 {<高亮开始>} 和 {<高亮结束>} 词元）。这些片段应包含回答问题所需的信息，目标是将所有这些高亮内容整合成一个回答问题的段落。

作为中间步骤，你需要将可以合并成一个句子的高亮片段归为一簇（也就是说，之后每个簇都会合并成一个句子）。请确保各簇的顺序与随后撰写相应输出段落时的顺序一致。

- **重要提示：** 请确保至少有两个簇。
- **重要提示：** 输出格式必须为 [{"簇": [以逗号分隔的高亮片段索引]}]。

问题: {查询}
内容:
... 文本 {<高亮开始>} 选定片段 {<高亮结束>} 文本 ..

答案

生成：
在此任务中，你会看到一个问题和若干段落，其中部分内容被“高亮”（也就是说，每个此类片段前后都有 {<高亮开始>} 和 {<高亮结束>} 词元）。你还会看到一个段落的前缀。

目标是根据这些段落生成一段回答问题的内容。给定的前缀是你目前为止生成的内容。你的任务是生成段落中的**下一句**，涵盖且仅涵盖这些

“高亮”片段。

- 确保它与前缀衔接顺畅。
- 确保它涵盖**全部且仅限于**高亮的片段。
- 确保它能恰当铺垫，最终对问题给出完整且连贯的回答。

问题：{查询}

内容：

... 文本 {<高亮开始>} 选定片段 {<高亮结束>} 文本..... 前缀：{前缀} _（注意：生成第一个簇时留空）

答案：