GEO：面向内容创作者的生成式引擎可见性优化（KDD 2024）

生成式引擎优化（GEO）

面向内容创作者的生成式搜索可见性优化框架

KDD 2024 正式论文 | 全文精准中译
arXiv:2311.09735v3

翻译：九章格物 ● 星阙实验室

摘要

大语言模型的出现催生了一种全新的搜索引擎范式 ——**生成式引擎**，它利用生成模型整合、归纳信息并直接回答用户查询。这类系统能够生成准确、个性化的响应，正快速取代谷歌、必应等传统搜索引擎。

生成式引擎通常通过整合多个来源信息并使用大语言模型进行摘要来满足查询。尽管这种转变显著提升了用户体验与生成式搜索引擎流量，但它对第三方利益相关者 ——**网站与内容创作者**—— 构成了巨大挑战。由于生成式引擎的黑盒特性与快速迭代，内容创作者几乎无法控制自身内容何时、以何种方式被展示。

为应对这一问题，本文提出**生成式引擎优化（GEO）**，这是首个面向内容创作者、用于提升内容在生成式引擎响应中可见性的全新范式。GEO 提供了一套灵活的黑盒优化框架，可用于定义与优化可见性指标。

为支持系统性评估，我们构建了 **GEO-bench**，一个包含多领域、多样化用户查询及其对应网页来源的大规模基准数据集。通过严格实验验证，GEO 可将内容在生成式引擎中的可见性最高提升 **40%**。此外，不同优化策略在不同领域的效果存在差异，凸显了领域专属优化方法的必要性。

本文为信息检索系统开辟了新前沿，对生成式引擎开发者与内容创作者均具有深远意义。

关键词：生成式模型；搜索引擎；数据集与基准；生成式引擎；内容优化；GEO

1 引言

三十年前，传统搜索引擎的发明彻底改变了全球信息获取与传播方式。尽管它们功能强大，并催生了学术研究、电子商务等大量应用，但其局限在于只能为用户查询返回一组相关网站。

近年来，大语言模型的成功推动了 BingChat、谷歌 SGE、Perplexity.ai 等新一代系统的出现。这些系统将传统检索与生成模型结合，我们将其统一命名为**生成式引擎**。

生成式引擎从互联网等数据库中检索相关文档，再利用大神经模型基于这些来源生成响应，确保引用可追溯、信息可验证。

生成式引擎对开发者与用户的价值显而易见：用户更快、更准确地获取信息；开发者生成精准、个性化的回答，提升满意度与收益。

但它对第三类参与者 ——**网站与内容创作者**—— 并不友好。与传统搜索引擎不同，生成式引擎直接提供精确、完整的答案，不再需要用户跳转到原网站，这可能导致网站自然流量大幅下降，严重影响内容创作者的生存。

数以百万计的中小企业与个人依靠线上流量维持生计，生成式引擎将显著冲击创作者经济。更重要的是，生成式引擎的黑盒与闭源特性，让创作者难以理解、控制其内容如何被摄取与呈现。

为此，本文提出**首个面向创作者的通用生成式引擎内容优化框架 ——GEO**，帮助内容创作者适应这一全新搜索范式。

GEO 是一个灵活的黑盒优化框架，专为闭源生成式引擎设计，用于提升网页内容可见性。它输入源网站内容，输出优化版本，通过调整文本呈现、风格与结构，提高在生成式引擎中的曝光度。

此外，GEO 提供了一套灵活的**可见性指标定义框架**。因为生成式引擎中的 “可见性” 远比传统搜索更复杂、更多维度：传统搜索以排名衡量可见性，而生成式引擎将来源以内嵌引用形式嵌入自然语言段落，引用长度、位置、风格各不相同。

因此需要专门为生成式引擎设计的可见性指标，从相关性、引用影响力、位置、篇幅、独特性等多个维度衡量来源曝光度。

为支持可靠、全面的 GEO 方法评估，我们构建 **GEO-bench**，包含来自多领域的 **10000 条查询**与对应来源。实验表明，GEO 方法可将可见性最高提升 **40%**。我们还在真实生成式引擎 Perplexity.ai 上验证，可见性提升最高达 **37%**。

本文贡献总结：

1. 提出 GEO，首个面向内容创作者、用于提升生成式引擎内容可见性的通用优化

框架，最高提升 40% 曝光。

2. 提出一套专为生成式引擎设计的完整可见性指标体系，支持创作者自定义指标。

3. 构建首个大规模生成式引擎优化基准 GEO-bench，覆盖多领域、多意图、多难度查询。

---

2 形式化定义与方法

2.1 生成式引擎形式化定义

尽管已有大量生成式引擎面向数百万用户部署，但目前仍缺乏统一的形式化框架。本文给出一个可兼容各类模块化组件的定义。

生成式引擎包含两大核心组件：

1. 一组生成模型 G = {G_1 ，G_2 ，..., G }，分别负责查询改写、摘要、回答生成等任务。

2. 搜索引擎 SE，根据查询返回一组来源 S = {s_1 ，s_2 ，..., s }。

典型流程如下：

1. 查询改写模型 G_qr 将用户查询分解为一组更易检索的子查询；

2. 搜索引擎 SE 根据子查询检索并返回排序后的来源列表 S；

3. 摘要模型 G_sum 对每个来源生成摘要；

4. 回答模型 G_resp 基于所有来源与摘要生成最终响应 r。

生成式引擎可表示为函数：

$$f_{GE}:(q_u, P_U) \rightarrow r$$

其中 q_u 为用户查询，P_U 为用户个性化信息，r 为自然语言响应。

响应 r 通常是结构化文本，并内嵌引用标注。引用至关重要，可缓解大语言模型幻觉问题。理想的生成式引擎应保证：

- 高引用召回：所有陈述都有来源支持
- 高引用精确：所有引用都准确支撑对应陈述

2.2 生成式引擎优化（GEO）

搜索引擎的出现催生了**搜索引擎优化（SEO）**，帮助创作者优化内容以提升排名、流量与可见性。但传统 SEO 无法直接迁移到生成式引擎。

原因在于：

- 生成式引擎不依赖关键词匹配；
- 大语言模型对文本的理解更细粒度、更语义化；
- 可见性不再由排名决定，而是由引用位置、篇幅、权重等决定。

因此需要全新优化范式 ——**GEO**。

GEO 的目标：最大化网站在生成式引擎响应中的**可见度**。我们将来源 c_i 在响应 r 中的可见性记为：

$Imp(c_i, r)$

创作者希望最大化该值。

从生成式引擎角度，目标是最大化与查询最相关引用的可见性：

$\max_i \text{Rel}(c_i, q, r)$

其中 Rel 表示相关性，是黑盒函数。

2.2.1 生成式引擎可见性指标

传统 SEO 用排名衡量曝光，但生成式引擎需要更精细的指标。我们提出三大设计原则：

1. 对创作者有实际意义
2. 可解释
3. 易于广泛理解

我们提出两类核心指标：

1. 词数占比

衡量来源被引用的句子所占的标准化词数：

$$Imp_{wc}(c_i, r) = \frac{\sum_{s \in S_{c_i}} |s|}{\sum_{s \in S_r} |s|}$$

S_ci 是引用 c_i 的句子集合，S_r 是响应全部句子。多来源共享句子时，词数均分。

2. 位置加权词数占比（页码：5）

对靠前句子赋予指数衰减权重，模拟用户阅读注意力分布：

$$Imp_{pwc}(c_i, r) = \frac{\sum_{s \in S_{c_i}} |s| \cdot e^{-\frac{pos(s)}{|S|}}}{\sum_{s \in S_r} |s|}$$

3. 主观可见性

使用 G-Eval 从 7 个维度评估：

- 相关性
- 影响力
- 独特性
- 主观位置权重
- 主观篇幅权重
- 点击可能性
- 内容多样性

2.2.2 GEO 优化方法

我们提出 9 种与生成式引擎无关的通用优化策略，均可通过大语言模型自动执行：

1. **权威化**：语气更具说服力、权威性
2. **添加统计数据**：用定量数据替代定性描述
3. **关键词堆砌**：传统 SEO 方法（作为对照）
4. **添加引用**：补充可信来源引用
5. **添加引述**：加入直接引语
6. **易理解化**：简化语言
7. **流畅度优化**：提升文本流畅性
8. **独特词汇**：增加独特表达
9. **专业术语**：合理增加领域术语

---

3 实验设置

3.1 生成式引擎环境

采用两步范式：

1. 对用户查询检索 Top-5 来源
2. 使用 gpt-3.5-turbo 生成带引用的响应

同时在真实系统 **Perplexity.ai** 上验证泛化性。

3.2 基准数据集：GEO-bench

包含 **10000 条查询**，来自 9 个公开数据集：

1. MS Macro
2. ORCAS-1
3. Natural Questions
4. AllSouls（论文题）
5. LIMA（复杂推理题）
6. Davinci-Debate（辩论题）
7. Perplexity.ai Discover（热门查询）
8. ELI5（通俗解释题）
9. GPT-4 生成查询

按 8:1:1 划分为训练 / 验证 / 测试集，覆盖 25 个领域、7 大类意图。

3.3 评估指标

- 位置加权词数占比（主要客观指标）
- 主观可见性（7 维度综合）
- 相对提升百分比（核心对比指标）

---

4 实验结果

所有 GEO 方法均显著优于基线，其中表现最强的三类方法：

- 添加引用
- 添加引述
- 添加统计数据

在位置加权词数指标上**相对提升 30%–40%**。
在主观可见性指标上提升 **15%–28%**。

关键结论：

- 传统 SEO 的关键词堆砌**无效甚至负向**
- 流畅度、可读性、权威性优化可提升 15%–30%
- 数据、引用、引述最有效
- 低排名网站从 GEO 获益最大（可达 +115%）

表 1：GEO 方法在 GEO-bench 上的绝对可见性指标

方法	位置加权词数	主观可见性
无优化	19.3	19.3
关键词堆砌	17.7	20.2
权威化	21.3	22.9
流畅度优化	24.7	21.9
添加引用	24.6	21.9
添加引述	27.2	24.7
添加统计数据	25.2	23.7

最优方法相较基线分别提升 **41%** 与 **28%**。

5 分析

5.1 领域专属优化

不同方法在不同领域效果差异显著：

- **权威化**：辩论、历史、法律领域最优
- **引用优化**：事实型问题、科学、政府信息最优
- **引述优化**：人文、社会、历史、解释类问题最优
- **统计数据**：法律、政府、观点类、辩论类效果极强

5.2 多网站同时优化

当所有来源都使用 GEO 时：

- 低排名网站**显著受益**
- 高排名网站优势被削弱

GEO 具有**数字空间平权作用**，让中小创作者能与大型机构竞争。

5.3 方法组合效果

组合使用效果更强：
流畅度优化 + 统计数据添加 效果最佳，相对提升超 **35.8%**。

5.4 案例分析

- 加入引用标注 → 可见性 +132.4%
- 加入统计数据 → +65.5%
- 权威化改写 → +89.1%

6 真实生成式引擎验证

在真实商用系统 Perplexity.ai 上：

- 引述添加：可见性 +22%
- 统计数据添加：+37%
- 关键词堆砌：-10%

证明 GEO 方法**跨生成式引擎泛化性极强**，具备真实落地价值。

表 2：Perplexity.ai 上的实验结果

方法	位置加权词数	主观可见性
无优化	24.1	24.7
关键词堆砌	21.9	28.1
引述添加	29.1	32.1
统计数据添加	26.2	33.9

7 相关工作

- 证据式回答生成
- 检索增强语言模型

- 传统搜索引擎优化（SEO）

本文明确区分：**SEO ≠ GEO**，两者范式、目标、机制完全不同。

8 结论

本文正式定义**生成式引擎**，并提出首个面向内容创作者的优化框架 **GEO**。
GEO 提供：

- 可见性指标体系
- 9 种安全、非对抗、可自动执行的优化方法
- 大规模基准 GEO-bench
- 在真实生成式引擎上验证有效

实验证明，GEO 可将内容可见性最高提升 **40%**，并能帮助中小创作者在生成式时代获得公平曝光。

9 局限性

- 生成式引擎持续迭代，GEO 需同步更新
- 未评估对传统搜索排名的影响
- 领域标签存在少量噪声

10 致谢

本工作受美国国家科学基金会（NSF）资助。

参考文献

[1] Daria Alexander 等. ORCAS-I: 基于弱监督的意图标注查询. SIGIR 2022.
[2] Prashant Ankalkoti. 搜索引擎优化工具与技术综述. 2017.
[3] Akari Asai 等。面向多语言的单问答模型与跨语言稠密检索. NeurIPS 2021.
[4] Sergey Brin, Lawrence Page. 大规模超文本搜索引擎解析. 1998.
[5] Tom Brown 等。语言模型是小样本学习者. NeurIPS 2020.

[6] Nick Craswell 等. MS MARCO：大数据背景下排序模型基准. SIGIR 2021.
[7] Brian Dean. 谷歌搜索结果点击率分析. 2023.
[8] Danny Goodwin. 搜索结果排名与点击率关系研究. 2011.
[9] Kelvin Guu 等. REALM：检索增强语言模型预训练. 2020.
[10] Ziwei Ji 等。自然语言生成中的幻觉问题综述. 2023.
[11] Aounon Kumar 等。操纵大模型提升产品可见性. 2024.
[12] R.Anil Kumar 等。搜索引擎优化技术综述. 2019.
[13] Tom Kwiatkowski 等. Natural Questions：问答研究基准. TACL 2019.
[14] Nelson F. Liu 等。生成式搜索引擎可验证性评估. 2023.
[15] Yang Liu 等. G-Eval：使用 GPT-4 的 NLG 评估. 2023.
[16] G. D. Maayan. 谷歌 SGE 对流量的影响. 2023.
[17] Jacob Menick 等。教语言模型用可验证引述支撑回答. 2022.
[18] Grégoire Mialon 等。增强语言模型综述. 2023.
[19] Reiichiro Nakano 等. WebGPT：基于人类反馈的浏览器辅助问答. 2021.
[20] OpenAI. ChatGPT 介绍. 2022.
[21] OpenAI. GPT-4 技术报告. 2024.
[22] A. Shahzad 等。搜索引擎优化新趋势. 2020.
[23] Kurt Shuster 等. BlenderBot 3：可持续学习的对话智能体. 2022.
[24] Romal Thoppilan 等. LaMDA：对话应用语言模型. 2022.
[25] Chunting Zhou 等. LIMA：极简对齐策略. 2023.

附录 A 对话式生成式引擎

在 2.1 节中，我们讨论了单轮生成式引擎。而未来的生成式引擎将具备多轮对话能力。输入不再是单条查询，而是对话历史 H=(q_u^t, r^t)。响应可表示为：

$$GE := f_{LE}(H, P_U) \rightarrow r^{t+1}$$

额外的模型可基于对话历史生成推荐的追问查询，提升用户参与度并间接提高来源可见性。

附录 B 实验设置详情

B.1 生成式引擎提示词

系统提示要求准确、简洁、专家语气、所有句子必须紧跟内联引用，且只使用给定来源。

B.2 GEO-bench 查询示例

- ORCAS：全球化是什么意思
- AllSouls：开放获取期刊是否是学术出版的未来
- Davinci-Debate：政府是否应推广无神论
- ELI5：猫为什么踢玩具
- GPT-4：生酮饮食的好处
- LIMA：影响消费者行为的主要因素
- MS-Macro：正常儿童空腹血糖范围
- Natural Questions：bee line 来源
- Perplexity.ai：LinkedIn 涨粉方法

B.3 评估指标

使用 7 项主观可见性指标，提示词已开源。

B.4 GEO 方法实现

所有 9 种方法均基于 GPT-3.5 Turbo 实现，提示词已开源。

附录 C 完整实验结果

包含带标准差的完整指标表、多种子实验结果、Perplexity.ai 详细指标。

表 6：带标准差的完整实验结果
所有方法的均值与方差已列出，结果稳定可复现。

表 7：Perplexity.ai 完整结果
验证了 GEO 在真实系统上的鲁棒性。

附录 D 图表说明

图 1：GEO 优化前后对比，优化后来源引用篇幅与位置显著提升。
图 2：生成式引擎整体架构，包含检索、改写、摘要、回答模块。
图 3：传统搜索与生成式引擎可见性对比。
图 4：GEO 方法组合效果热力图。