MedCite：语言模型能否生成可验证的医学文本？

Xiao Wang[1], Mengjue Tan[2], Qiao Jin[3], Guangzhi Xiong[4], Yu Hu[5],
Aidong Zhang[4], Zhiyong Lu[3], Minjia Zhang[1]
[1]SSAIL Lab, University of Illinois at Urbana-Champaign [2]Brown University
[3]National Library of Medicine, NIH [4]University of Virginia [5]Microsoft
{ xiaow4, minjiaz } @illinois.edu
mengjue_tan@brown.edu { qiao.jin, zhiyong.lu } @nih.gov
{ hhu4zu, aidong } @virginia.edu yuhu@microsoft.com

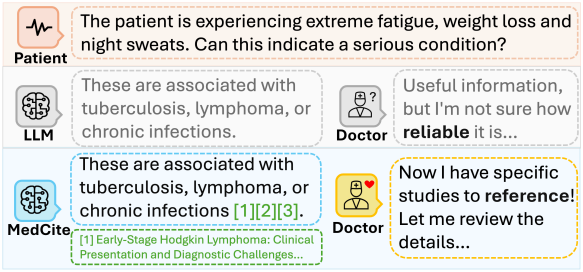


Figure 1: 医学问答系统的比较。最先进的系统生成的答案没有引用。MedCite 不仅生成答案，还将每个答案与引用关联，提高了医学系统的可验证性和可信度。

Abstract

现有基于 LLM 的医学问答系统缺乏引文生成和评估能力，这引发了它们在实际应用中的采用问题。在这项工作中，我们引入了 MedCite，这是第一个支持通过 LLM 进行医学任务的引文生成设计和评估的端到端框架。同时，我们引入了一种新颖的多次检索-引文方法，生成高质量的引文。我们的评估突出了医学任务中引文生成的挑战和机遇，同时识别出对最终引文质量产生重大影响的关键设计选择。与强基线方法相比，我们提出的方法在引文精确度和召回率方面实现了显著改善，并且我们展示了评估结果与专业专家的注释结果具有良好的相关性。

1 介绍

大型语言模型（LLM）在各种自然语言处理任务中展示了卓越的能力，比如回答问题（QA）和遵循指令（Kaplan et al., 2020; Wei et al., 2022a,b）。LLMs 的发展还促成了医疗代理的开发，这些代理能够理解患者和医生使用的语言，提供丰富的及时辅助（Singhal et al., 2022, 2023; Temsah et al., 2023; Tangadulrat et al., 2023; Maples et al., 2024）。

虽然早期迹象是积极的，但当前由大型语言模型驱动的医疗问答系统仍然存在多重限制。例如，医疗数据通常包含敏感信息，如个人健康记录，要求在训练大型语言模型时严格遵循道德规范（Gilbert et al., 2023）。此外，可信度在医疗领域尤为重要。幻觉问题，例如模型生成不正确或误导性的信息，对基于大型语言模型的医疗系统的可靠性构成了重大挑战（Pal et al., 2023; Ahmad et al., 2023; Huang et al., 2024）。为了克服这一问题，研究人员和从业者研究了检索增强生成（RAG）（Xiong et al., 2024a; Yang et al., 2024），它结合了大型语言模型与从外部可信数据源进行的信息检索（Canese and Weis, 2013）。通过向模型提供准确且相关的医学知识，这些系统使大型语言模型能够在回答中保持相关性。

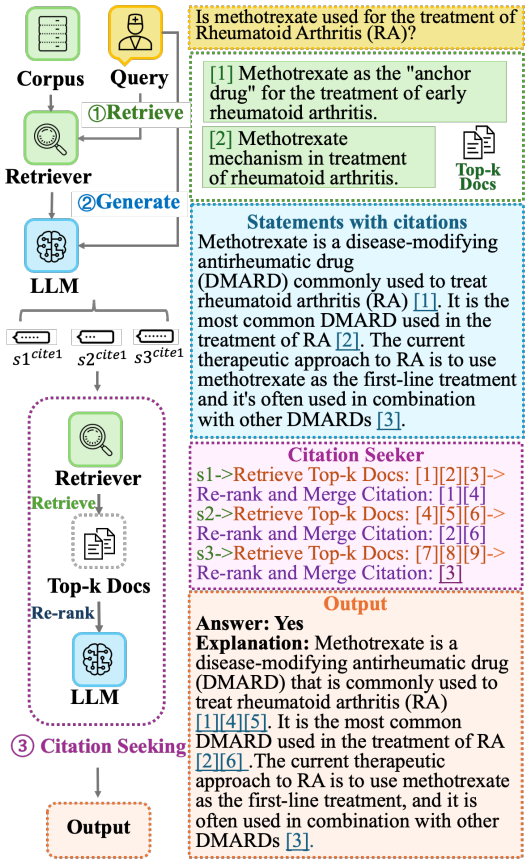


Figure 2: MedCite 的概述图。

尽管结果有希望，现有的方法缺乏可验证性 (Liu et al., 2023)，这意味着所提供的答案没有可靠的来源或证据支持。这可能导致错误信息，并且如果遵循错误的医疗建议，可能产生潜在的有害后果。例如，如在 Fig. 1 中所示，当根据一组症状提供诊断而没有任何参考时，无法保证预后或治疗建议的准确性，这会造成不确定感，从而导致次优甚至有害的决策。

一种有希望的方法是通过归因 (Bohnet et al., 2022; Huang and Chang, 2024) 来缓解可验证性问题，即将陈述与引文关联，这为系统提供了更多的可信度和责任感，同时为用户提供了一种更深入探索来源和验证信息来源的方法。然而，尽管先前有一些努力通过在一般领域的问答任务中使用大型语言模型来分析引文能力 (Liu et al., 2023; Gao et al., 2023c; Djeddal et al., 2024)，但由于以下原因，为医学引用句子尤其具有挑战性并且在实践中不广泛采用。

首先，现有的医学问答研究通常利用多选题准确性来进行基准测试和评估其性能 (Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024)，该方法注重评估从给定选项中选择正确答案的能力。然而，由于引文生成具有开放性的特点，因此其挑战性更高。例如，在为罕见遗传病开药或为患者计划手术时，医生和患者都必须依赖更丰富的信息。因此，一个查询可以有多个答案，并由多个可能的来源支持。在评估引文生成方法时，这些方面是重要的考量，但现有的医学问答框架并未从本质上考虑这些因素。

其次，引用生成存在一个巨大的设计空间，其中检索器 (Asai et al., 2024; Izacard et al., 2022)、主干 LLMs (MetaAI, 2024; Zhang et al., 2024; OpenAI, 2023) 以及引用生成算法 (Gao et al., 2023c) 之间的交互关系复杂。因此，识别关键贡献因素具有挑战性。然而，这种分析对于制定提高医疗系统可验证性的策略至关重要。

第三，尽管该领域的贡献持续增加 (Gao et al., 2023c; Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024)，但在设计、开发和评估医学任务的引文生成质量方面有用的开源框架明显不足。现有的引文评估框架是由通用问题构建的，其中医学任务的指标和评估者的选择仍然是一个未解的问题。此外，获取高质量的医学专家注释代价不菲，这需要一个高质量的分类器来判断引文是否归因于某个陈述。

在本文中，我们着手解决这些挑战，希望促进在改善医学系统可验证性方面的研究。特别地，我们的贡献如下：

- 对医学任务中使用大型语言模型 (LLM) 的不同引用方法和关键设计组件进行了深入研究，涵盖了从文本生成方法、信息检索方法到引用归因方法的范围。我们的研究理清了不同因素相对于核心 LLM 的重要性。
- 我们展示了 MedCite，这是首个可以使大型语言模型（LLMs）为医疗问答系统生成可验证文本并进行自动评估的端到端系统。同时，我们引入了一种新颖的多次检索引用方法，调和了检索增强生成和生成后引用。
- 对不同 LLMs 的 MedCite 进行综合评估，结果显示 MedCite 在文本生成和引文生成质量方面分别比现有方法提高了最高达 47.39 % 的召回率和 31.61 % 的精确率。我们通过让医学医生验证归因结果进行人工评估。结果表明，我们的自动评估流程与领域专家的判断高度相关，证明了药学中高效且自动的引文评估的有效性。

我们的代码可在 https://supercomputing-system-ai-lab.github.io/projects/medcite/ 获取

2 相关工作

生物医学问答 生物医学问答旨在利用自然语言处理技术回答临床或生物医学问题。早期系统是基于规则的 (Lee et al., 2006; Cao et al., 2011)，依赖于结构化本体，但缺乏可扩展性。后来，诸如 BioBERT (Lee et al., 2020) 和 ClinicalBERT (Huang et al., 2019) 等特定领域语言模型在 BERT (Devlin et al., 2019) 的基础上进行构建，在生物医学问答基准测试中带来了性能提升 (Yang et al., 2024)。最近，生成模型如 GPT-3.5/4 (Brown et al., 2020; OpenAI, 2023) 和 Med-Gemini (Saab et al., 2024) 使得无需预定义选项即可进行灵活的答案生成。然而，这些模型可能会产生幻觉，促使采用检索增强生成（RAG）以将输出定位于检索到的文档 (Lozano et al., 2023; Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024; Zakka et al., 2024; Xiong et al., 2024b)。我们的工作重点是改善生成的生物医学响应的可验证性。

为大型语言模型生成的文本添加引用已经成为一个活跃的研究领域。尽管大型语言模型（如 ChatGPT (Brown et al., 2020; Thoppilan et al., 2022; Anil et al., 2023; OpenAI, 2023, 2024)）可以被提示生成引用，但这些引用往往不准确或是被伪造的 (Zuccon et al., 2023)。直接的模型驱动归因 (Sun et al., 2023; Agrawal et al., 2023; Weller et al., 2024) 让大型语言模型

可以自主引用，但缺乏可靠性。基于检索的方法可以改善基础性: 检索后生成（PRG）在生成前检索相关文档 (Guu et al., 2020; Borgeaud et al., 2022; Reddy et al., 2023)，而生成后引用（PGC）在答案生成后检索证据 (Huo et al., 2023)。虽然更为稳健，但这两种方法增加了复杂性 (Gao et al., 2023b)，并且在需要精确归因的生物医学任务中可能仍然不足。我们的混合双通路引用方法结合了 RAG 与生成后修正以解决这些限制。另一个方向是对大型语言模型进行以策划或合成引用数据为基础的微调，以提高引用质量 (Ye et al., 2024)。最后，引用准确性评估协议正在被开发 (Rashkin et al., 2023; Gao et al., 2023c; Li et al., 2024)，尽管现有研究主要聚焦于通用领域内容。相较之下，我们的评估是医学特定的，并考虑了检索和引用生成的领域感知策略。

Wu et al. (2024) 提出了一个基于 URL 检索的引文评估流程，并表明即使是顶级的 LLM 如 GPT-4（RAG）也常常无法提供完全有根据的答案。虽然我们的工作和他们的工作都旨在提高引文的可靠性，但他们专注于为在线 URL 的 API 模型提供提示，而我们提出一个模块化的引文系统，以改善生物医学基础。正如在第 5 节中详细说明，我们展示了为什么参数引文方法不适合开源 LLM，因为存在虚构 URL 和缺乏可验证资源的问题。我们的方法利用来自可信医学文献如 PubMed 的分层检索，并结合多次引文以确保专业内容的归因，包括药物名称和基因标记。与相关工作的更全面讨论见附录 A。

3 问题设置

在本节中，我们首先对生物医学问答的引文生成任务进行表述，然后概述将在下一节中进行实验考察的方法。

3.1 问题目标

目标是开发一个系统，该系统能够自动为由大型语言模型生成的文本陈述添加相关且准确的引用。具体来说，系统的输入包括用户查询 q、LLM Φ 和包含真实文档的外部数据库 D。系统的输出包括生成的文本段落，该段落包含一个陈述列表 $S = \{s_0, s_1, ..., s_n\}$，由 Φ 组成。对于每个陈述 s_i，给它分配一组内联引用 $C_i = \{c_i^0, c_i^1, ..., \}$，其中 $c_i^j \in$ D。

3.2 数据集

根据之前的研究 (Bolton et al., 2024; Yasunaga et al., 2022; Xiong et al., 2024a)，我们使用 BioASQ-Y/N 数据集 (Nentidis et al., 2024)，这是一个常用的基准数据集用于评估生物医学问答系统。该数据集由用于提出问题的问题、人类注释的答案以及提供回答问题所需信息的相关背景组成。BioASQ-Y/N 数据集有三个特点促使我们使用它进行研究：1）与用于医学 QA 的其他数据集不同 (Jin et al., 2020; Hendrycks et al., 2021; Pal et al., 2022)，这些数据集主要是多选 QA 任务，而 BioASQ-Y/N 不仅提供选项选择（是/否），还提供关于答案陈述信息性的一组黄金答案。2）BioASQ-Y/N 为每个问题提供了支持文档的真实标签。同时，它可以很容易地修改为在没有提供真实文档的情况下回答问题，这代表了一种更现实的医学环境。3）到目前为止，它尚未被现有的通用引用方法使用。除了用来在第 4 节分析的 BioASQ 之外，我们还在第 6 节包括了 PubMedQA (Jin et al., 2019)。我们在附录 B 中包含了数据集和超参数的详细信息。在外部数据库方面，我们主要考虑 PubMed 数据库 (Canese and Weis, 2013)，其中包含 2460 万份由医学专业人士审查的生物医学文献。这个庞大的数据库提供了一个丰富的精确和合法的文献来源，供 LLM 生成的文本进行归因。

3.3 评估指标

对于医学问答，评估文本和引用生成的质量是至关重要的，以确保大规模语言模型的输出不仅连贯且相关，还能被准确的引用支持。因此，我们考虑以下几个方面。

答案正确性。 与多项选择问答不同，真实的医学系统通常生成长篇且开放式的答案。因此，我们使用 ROUGE-L (Lin, 2004) 和 MAUVE (Pillutla et al., 2021) 来评估答案基于真实答案的正确性和相关性。我们仍然让模型在生成长答案的同时生成一个是/否答案，以便我们可以与现有的非引用方法进行比较。

引用质量。 我们在语句层面评估引用质量，这比问题层面的文档检索提供了更细致的视角。这可以更好地捕捉每个生成的语句是否基于医学证据。

我们考虑一个归因判决器 $Attr : \mathcal{X}, \mathcal{Y} \rightarrow \{0, 0.5, 1\}$，如果语句 $\mathcal{X}$ 可以完全归因于语句 $\mathcal{Y}$，即 $\mathcal{Y}$ 是 $\mathcal{X}$ 的来源，则输出 1；如果 $\mathcal{X}$ 可以部分归因于 $\mathcal{Y}$，则输出 0.5；否则输出 0。为了证明引入部分支持的合理性，我们参考了最近研究的发现，例如 Wührl et al. (2024)，研究表明在医疗事实核查任务中，62.4 % 的断言是部分有证据支持的。这突显了捕捉部分归因的重要性，因为在现实世界的医疗陈述中这是一种常见现象。

在医疗问答中使用引用时，一个答案可以包含多个可验证的陈述，并且一个陈述可能附有多个引用来支持。通过归因评审员，我们使用两个指标来衡量引用的质量：引用召回率和引用精确率。这两个指标都严重影响医疗问答的可用性，因为高召回率意味着生成的答案得到了很好的证据支持，而高精确率则表明指定的引用具有较高的质量，可以用来验证生成文本的真实性。为简单起见，我们考虑一个陈述 s，附有 n 个引用 $c_1, c_2, \ldots, c_n$，其中每个引用都是一组公理。

我们将召回定义为一个陈述级别的指标，它衡量陈述中的所有信息是否都被引文充分支持。对于召回 =1，A_s 中的每一个公理（事实）都必须存在于由所有引文提供的公理集当中。

在我们的实验中，我们使用引用文档的串联来表示引用的合集，并判断该声明是否可以被串联引用充分支持。使用上述定义的归因判断，当且仅当 $\text{Recall}(s, c_1, \cdots, c_n) = 1$ 时，我们有 $Attr(s, \bigcup_{i=1}^{n} c_i) = 1$。然后我们对所有的声明求平均值，以获得答案段的最终召回率。

根据之前的研究 (Liu et al., 2023)，我们将精度指标定义为一种引用级别的度量，用于评估每个单独的引用是否有助于支持该陈述。如果引用通过包含陈述中至少部分必要的公理完全或部分支持该陈述，则其精度为 1。只有当 $Attr(s, c_i) > 0$ 时，s 的 c_i 的精度才被计算为 1。当有多个陈述时，我们通过平均其所有引文的精度得分来计算其引用精度。

引用 F$_1$。 我们使用引文 F$_1$ (Liu et al., 2023) 通过 $F_1 = 2 \times \frac{citation\ precision \times citation\ recall}{citation\ precision + citation\ recall}$ 来衡量综合引文的准确率和召回率。

4 引用程序分析

本节探讨并量化哪些选择对于成功引用医学任务中的句子是重要的。鉴于每个组件都可以变化，我们研究这些组件中的每一个对引用生成质量的影响，同时将其他组件隔离开来。除非另有说明，实验中我们使用 Llama-3-8B-I。

4.1 参数引用与非参数引用

最新的大型语言模型（LLM）可以通过依赖其参数化内容，即从训练数据内化的信息，在生成的文本中包含引用。鉴于这一进展，一个问题自然浮现：我们能否依赖 LLM 为其生成的句子自我引用？我们将这种策略与非参数引用进行比较，即纯粹依赖非参数信息检索（IR）内容生成引用，例如，PubMed。特别是对于参数化引用，我们生成一个提示，包括用户提问，以及让 LLM 生成答案并在每个陈述的格式化输出中添加行内引用的指示指导。在这种情况下，模型完全依赖其预训练数据生成引用。对于非参数引用，我们让 LLM 直接生成不带引用的答案。然后我们使用密集检索器 MedCPT (Jin et al., 2023) 从 D 中检索一份相关文档（例如，前 3 名）列表，并将这些文档作为行内引用。使用的提示可以在附录 C 中找到。

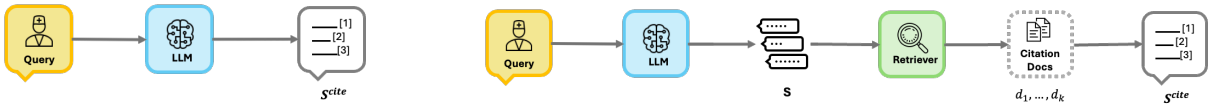
Citation Method	Model	Accuracy (EM)	Text Quality		Citation Quality	
			MAUVE	ROUGE-L	Rec.	Prec.
Parametric (LLM)	Llama-3-8B-I.	74.76	61.94	17.72	/	/
	UltraMedical	69.09	67.70	13.96	/	/
	GPT-4o	88.51	74.82	20.03	/	/
Non-parametric (IR)	Llama-3-8B-I.	73.95	65.31	19.05	60.89	53.90
	UltraMedical	68.12	51.18	12.69	52.48	62.32
	GPT-4o	87.70	70.15	20.20	79.72	80.95

Table 1: 不同 LLM 中的参数方法（LLM）与非参数方法（IR）引用方法的比较。

表格 1 比较了不同 LLM 上的参数化和非参数化引用结果。我们发现，虽然 LLM 在理解和遵循人类指令方面取得了显著进展，但在医学环境中生成引用时确实存在局限性。特别是，Llama-3-8B-I. 和 UltraMedical 无法准确遵循这些指令。因此，生成的引用要么不正确、捏造或格式不佳，尽管其中一小部分确实存在，但这些引用可能不是免费访问的（例如，一些科学文章在付费墙后面）。因此，在没有对由参数化方法生成的引用的科学文章内容进行 API 访问的情况下，很难自动评估其质量。这不足为奇，因为这些模型仍然基于下一个词的预测进行训练，而 LLM 需要利用其预训练知识或幻觉来推断引用信息。有趣的是，GPT-4o 不仅在 BioASQ 任务中实现了最高准确率，而且还能始终如一地遵循指令生成格式良好的引用。然而，GPT-4o 生成的引文陈旧（全部在 2018 年之前），使得难以包括新的研究。我们在附录 D 中包含了生成的引用的几个示例。这个观察强调了在引用生成时仅依赖参数化方法的一个关键限制，特别是在医学领域使用公共 LLM 的情况下。鉴于这些限制，我们将在余下实验中集中使用可信数据集（如 PubMed）的非参数化引用方法。

4.2 RAG 引用更完善

虽然非参数化引用提高了引用质量，但答案陈述的生成仍然依赖于预训练数据本身。因此，答案可能基于过时或不完整的医学数据。尽管最近多项研究观察到，增加检索增强生成（RAG）有助于提升大型语言模型（LLM）更好地理解生物医学任务，且产生比非 RAG 方法更高的准确性 (Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024)，但关于 RAG 如何影响文本和引用质量的实验报道很少。我们通过比



(1) Parametric citation generation

(2) Non-parametric citation generation

Figure 3: 参数化 (LLM) 和非参数化 (IR) 引用生成流程的比较。

较几种不同的方法深入探讨 RAG 在引用生成中的作用。对于所有方法，我们使用相同的基于稠密 MedCPT 的检索器，将前 k（例如，前 3）个相关文档分配为引用。

非 RAG (CoT)：我们执行链式思维 (CoT) 提示 (Wei et al., 2022b)，以利用大语言模型 (LLMs) 的推理能力来提供答案（例如，极性是/否答案）和文本解释 S 以回答问题 q。这与 (Xiong et al., 2024a) 中的方法类似，但不从外部数据库中检索支持上下文。

RAG：我们首先从 D 检索出一个包含前 k 个支持文档 $\{d_1, ..., d_k\}$ 的候选名单以对应查询 q。然后，我们将候选名单中的文档与 q 一起输入到 LLM 中，并指示 LLM 生成答案和文本解释 S。

RAG w. Oracle：与上述配置类似，但在每个问题的 BioASQ 中使用真实的支持文档（即假设一个完美的检索器）。

Retrieval Method	Accuracy (EM)	Text Quality		Citation Quality	
		MAUVE	ROUGE-L	Recall	Precision
Non-RAG	71.36	53.24	18.07	59.05	52.93
RAG	82.85	52.22	14.79	49.01	42.77
RAG w. Oracle	94.34	63.45	20.63	57.46	43.20

Table 2: RAG 方法和非 RAG 方法在引文生成上的比较。

表格 2 显示了非 RAG 和 RAG 方法在医学领域的比较。有趣的是，我们观察到没有 RAG 时，生成答案的正确性往往较低（71.36）。然而，引用的召回率和精确度相对较高。相反，整合 RAG 会导致答案正确性显著提高，同时引用的召回率和精确度下降。这是因为引用质量指标仅评估 LLMs 生成的语句是否由可验证的来源支持，而不是直接评估每个语句的正确性。因此，即使识别出的引用能够支持一个虚构的语句，该语句与用户的问题不相关也是可能的。这一发现表明，我们需要整体评估 LLM 在文本和引用生成方面的能力。具体来说，我们将答案的正确性（例如，准确性和文本质量）视为对引用评估的先决条件，并且启用引用功能不应损害答案生成的质量。

表 2 同样显示，通过使用真实文档（oracle），使用 RAG 可以获得的最佳准确率结果高达 94.34 %，并且引用的召回率和精确度可以分别达到 57.46 % 和 43.20 %，这表明仍有空间来研究更好的检索增强方法。尽管如此，这些结果表明 RAG 对于生成上下文相关的文本至关重要，是获得高质量引用的关键步骤。因此，我们在余下的实验中使用 RAG。

4.3 引用搜索工具的有效性

到目前为止，我们已经确定了引用搜索者的选择，即如何找到相关文档并将它们作为引用分配给一个声明。然而，人们可能会想知道引用搜索方法的选择如何影响引用质量。为了研究这一点，我们考虑以下策略：

预生成候选列表 + LLM 重新排序。对于每个生成的语句，我们指示 LLM 将一个从预生成检索中检索到的文档指定为支持文档之一。这种情况下不需要额外的检索。

仅使用检索器的重新检索。对于每个生成的陈述，我们重新启动检索器以从 D 中检索与该陈述相关的前 k 篇文档，并将这些文档作为每个陈述的引用添加进去。

重新检索 + NLI 重排序。对于每个生成的陈述，我们重新启动检索器以从 D 中检索与该陈述相关的前 k 篇文档，并使用轻量级的医学 NLI 模型将这些检索到的文档指定为引用。

重新检索 + LLM 重新排序。与上述配置类似，只是我们指示 LLM 将检索到的文档指定为引用。

Attribution Strategy	Accuracy (EM)	Text Quality		Citation Quality	
		MAUVE	ROUGE-L	Recall	Precision
Pre-Gen. shortlist + LLM rerank	83.33	59.22	16.78	54.66	41.40
Retriever-only re-retrieval	83.33	59.22	16.78	65.69	47.69
Re-retrieval + NLI rerank	83.33	59.22	16.78	65.38	55.12
Re-retrieval + LLM rerank	83.33	59.22	16.78	65.78	60.95

Table 3: 引用寻求方法的比较。我们使用第四节中描述的混合配置，因为它整体上具有更好的引用质量。

表格 3 显示了不同引文搜索策略的比较结果。我们发现，再检索 + LLM 重排序的整体表现最为优越，这验证了 (1) 再检索可以提高引文精确度和召回率，以及 (2) 引文重排序是寻找高质量引文的有效方法的好处。值得注意的是，尽管再检索 + NLI 重排序实现了类似的引

文召回率，但其召回精确度降低了 5.8 %，NLI 模型整体上比 LLM 更轻量化。因此，如果成本是主要限制因素，可以考虑使用专注于医学领域的 NLI 分类器进行引文搜索。

5 MedCite：一个用于大型语言模型驱动的医学问答的引用生成系统

在前一节中，我们调查了医学任务引用生成的几个重要设计选择。现在，我们汇总这些改进，评估它们的综合影响，并将其作为一个开源框架 MedCite (Fig. 2) 提供。我们最终的引用生成方法整合了三个核心组件：非参数引用 (§ 4.1)、RAG (§ 4.2)，以及检索 +LLM 重排序引用查找方法 (§ 4.3)。此外，我们还调查了以前工作中被低估的另外两个重要因素：（1）如果我们通过多次处理方法结合参数和非参数引用会怎样；以及（2）检索器选择对引用查找的影响。

多遍引文生成。 直观上，似乎可以利用 LLM 的内部参数知识来提供初步答案和引用，同时采用生成后的非参数方法来验证和完善这些引用，利用外部检索的内容。为了验证我们的假设，我们考虑一种新的多步法：类似于 § 4.2 中的方法，我们使用 RAG 生成答案。与该方法不同，我们指示 LLM 根据检索到的文档在回答问题的同时为陈述分配引用。然后，我们检索每个陈述的前 k 个相关文档。我们消除这两个阶段中的任何冗余引用，并结合剩余的形成最终引用。

表格 4 显示了双通方法与非参数引文的比较结果。结果表明，双通方法在引文的精确度和召回率方面始终优于非参数方法，同时保持了相当且略好的答案正确性。通过结合生成和检索系统的优点，双通方法减轻了每个独立方法固有的局限性。

Configuration	Accuracy (EM)	Text Quality		Citation Quality	
		MAUVE	ROUGE-L	Rec.	Prec.
Non-parametric RAG + Citation Seeker	82.85	52.22	14.79	49.01	42.77
Hybrid Double-pass	83.33	59.22	16.78	65.69	47.69

Table 4: 非参数法与 MedCite 的双重通过法在引文生成上的比较。

另一个因素是选择检索器以查找引用。最近提出的 MedRAG (Xiong et al., 2024a) 使用一种基于互惠等级融合（RRF）的混合方法，将 BM25 (Robertson and Zaragoza, 2009) 和 MedCPT (Jin et al., 2023) 的结果结合起来，以便在生成之前的阶段找到支持文件。然而，尽管可以找到广泛的相关文件来增强 LLM 生成答案的上下文，引文检索必须更加关注事实，以确保精准和正确的引用。在消融研究中，我们展示了一种层次化的两阶段排序器，该排序器首先通过 BM25 (Robertson and Zaragoza, 2009) 根据关键词匹配检索文件，然后基于 MedCPT (Jin et al., 2023) 的语义检索器，再次在引用质量方面带来性能的进一步提升，验证了检索器选择对于引文的重要性。

6 评估

6.1 主要结果

我们将 MedCite 与三种基线方法进行比较：医学领域的 RAG 方法和最近工作的两个通用领域引用方法，包括跨不同骨干 LLM 的检索后生成和生成后引用方法：

(1) MedRAG：在 (Xiong et al., 2024a) 中描述的方法。

（2）检索后生成（PRG）：根据 (Gao et al., 2023c) 中的方法，我们为大型语言模型（LLMs）提供一个查询和一份检索到的文档列表，并指示 LLMs 在其生成的答案中包含引文。

（3）生成后引用（PGC）：在 RARR (Gao et al., 2023a) 之后，我们进行链式思维（CoT）提示 (Wei et al., 2022b)，让 LLM 生成答案，然后通过重新检索 + LLM 重新排序为每个声明分配引用。

我们评估了三个模型：Llama-3-8B-I. (Llama-3-8B-Instruct) (MetaAI, 2024)、UltraMedical (Zhang et al., 2024) 和商业 LLM GPT-4o (gpt-4o-0806) (OpenAI, 2024)。

我们在表 5 中展示了主要结果。实验的主要结论如下。

目前最先进的医学问答系统如 MedRAG，其生成的答案中没有引用。我们展示了在保持生成答案准确性的同时，可以在医学系统中启用引用。特别地，MedCite 和 PRG 能够在 Llama-3-8B-I 和 GPT-4o 上实现与 MedRAG 可比的准确性、MAUVE 和 ROUGE 分数，同时提供引用以支持生成的答案。另一方面，UltraMedical 在 MedRAG 上获得了最高的准确性，尽管其绝对准确性（74.92 %）远低于 Llama-3-8B-I.（82.85 %）和 GPT-4o（92.39 %）。通过检查 UltraMedical 生成的输出，我们发现添加额外的说明似乎使模型感到混乱，导致不正确的回答。这可能是因为 UltraMedical 在训练时使用的是 2048 的上下文长度，随着额外说明的增加，更难以让模型专注于提示中最相关的部分。

MedCite 在引用质量上优于 PRG 和 PGC。 虽然 PRG 和 PGC 均可在医学领域进行引

Model	Method	Acc. (EM)		Text Gen. Quality				Citation Quality					
				MAUVE		ROUGE-L		Recall		Precision		F1-Score	
		BioASQ	PubMedQA	BioASQ	PubMedQA	BioASQ	PubMedQA	BioASQ	PubMedQA	BioASQ	PubMedQA	BioASQ	PubMedQA
Llama-3-8B-I.	MedRAG	82.85	70.80	53.74	42.39	14.78	14.22	/	/	/	/	/	/
	PRG	84.95	69.40	72.53	47.79	17.97	20.99	35.44	30.08	38.71	35.00	32.50	36.73
	PGC*	72.10	55.80	61.90	44.53	18.06	19.11	64.75	62.18	69.32	71.75	66.96	66.62
	MedCite	84.95	69.40	72.53	47.79	17.97	20.99	74.86	69.50	69.47	67.73	71.74	68.60
UltraMedical	MedRAG	74.92	65.00	57.24	58.82	17.33	20.54	/	/	/	/	/	/
	PRG	63.43	53.60	63.87	48.02	13.27	14.89	27.54	28.51	30.80	31.17	28.01	30.94
	PGC	68.12	44.80	50.71	41.04	12.69	13.33	49.91	54.28	62.18	72.82	55.37	62.21
	MedCite	63.43	53.60	63.87	48.02	13.27	14.89	74.93	60.12	45.42	64.19	66.71	53.14
GPT-4o	MedRAG	92.39	73.80	51.29	38.00	15.77	24.11	/	/	/	/	/	/
	PRG	92.56	75.60	60.74	52.32	19.97	27.18	53.86	51.33	57.27	55.27	52.45	56.26
	PGC	87.70	50.60	67.01	61.72	20.80	21.37	79.59	75.94	81.01	82.40	80.29	79.04
	MedCite	92.56	75.60	60.74	52.32	19.97	27.18	84.86	84.54	83.85	89.43	84.36	86.48

Table 5: MedCite 与其他方法在 BioASQ 和 PubMedQA 数据集上的比较结果。* PGC 的生成阶段使用 CoT，这是一种非 RAG 方法。因此，PGC 的准确率 (EM) 得分与 CoT（非 RAG）方法相同。

用，但 MedCite 远远优于这两种方法（例如，71.74 % vs. 66.96 % 和 32.50 % 在 BioASQ 中）。MedCite 优于 PRG，因为 MedCite 的第二次引用寻求利用生成后非参数检索来精炼引用，这使得 LLMs 能够减轻引用幻觉。MedCite 比 PGC 获得更好的表现，因为它利用生成前检索和 LLM 的内部参数知识来获取初始的引用集，这对于获得高质量的最终引用是有用的。这些结果表明 MedCite 在结合生成系统和检索系统的优势进行引用生成方面的有效性。

我们看到一个普遍的趋势：MedCite 改善了 LLM 的引文召回率和 F_1 分数。以 GPT-4o 作为骨干 LLM 导致了最高的引文质量（例如，在 PubMedQA 上 GPT-4o 的 F_1 分数是 86.48，相比于 Llama-3-I. 的 68.60），这主要得益于它的高级推理和指令跟随能力。相反，当系统在 UltraMedical 上进行评估时，引文质量最低（例如，在 BioASQ 上的得分是 66.71）。这些结果强调了引入 MedCite 增强了 LLM 生成可验证文本的能力。

一个定性案例研究列在附录 F 中，提供了具体的例子，用来补充我们的定量研究结果。

6.2 消融研究

我们评估了不同的文献引用检索器如何影响 MedCite 的质量。特别地，我们比较了仅语义的 (Jin et al., 2023)、仅词法的 (Robertson and Zaragoza, 2009)、通过 RRF-2 进行检索融合的 (Xiong et al., 2024a)，以及层级两阶段检索器。不同于先前的发现基于 RRF-2 的混合检索器带来最佳性能结果，我们发现仅词法（如 BM25）检索器带来了更高的引用质量。与用于 RAG 的检索器不同的是，其目的是为 LLM 生成提供支持文档，文献引用检索需要检查精确的医学术语并从源中逐字引用。例如，在我们的实验中，给定 LLM 声明“肽是氨基酸的短链，氯毒素是一种特定类型的肽，”语义检索器检索到了一篇讨论卡立毒素特征的文档。尽管卡立毒素和氯毒素都是毒素，这篇文档并不能帮助支持该声明。因此，它不能作为该声明的有效引用。正因为如此，基于精确匹配的词法检索器提供了更精确的文献引用。相比之下，仅语义和基于检索融合的检索器对文献引用质量产生负面影响。最后，层级两阶段检索器首先执行词法检索以获取一长串引文候选，然后通过语义检索器根据查询与引文候选之间的相似度分数对长串进行排序。因此，通过在全面引用和精确引用之间实现良好平衡，它提供了我们测试配置中性能最佳的结果。

Retriever Type	Method	Accuracy (EM)	Citation Quality	
			Rec.	Prec.
Lexical-only	BM25	94.34	77.53	79.89
Semantic-only	MedCPT	94.34	65.93	66.78
Combination	RRF-2	94.34	75.74	76.46
Hierarchical	BM25 then MedCPT	94.34	77.84	80.02

Table 6: 不同检索器对 MedCite 质量与 Llama-3-8B-I 的有效性。在预生成检索阶段使用 Oracle 相关文档作为支持文档，并通过 LLM 重新排序每个声明重新检索排名前 3 的文档。

虽然以往的研究通常假设自然语言推理（NLI）模型在归因评估上与人类判断具有良好的相关性 (Gao et al., 2023c; Bohnet et al., 2022)，但这些研究主要集中在一般领域的问题。据我们所知，尚无研究对不同模型在医学任务归因判断中的有效性进行评估。我们评估了医学归因模型，并将其与医学专家的判断进行比较。附录 E 提供了标注指南。

令人惊讶的是，表格 7 表明，现有专门针对医学的 NLI 模型与专业医学医生判断的相关性很差（例如，在精确判断中，< 22.3 % 分数）。

Model	Source	Domain	Cohen's Kappa Score	
			Rec. Judge	Prec. Judge
SciFive-MedNLI	Open	Medical	0.2593	0.1945
JSL-MedPhi2-2.7B	Open	Medical	0.1845	0.2218
UltraMedical	Open	Medical	0.4518	0.2162
Llama-3.1-8B-Instruct	Open	General	0.5862	0.5422
mistral-7B-Instruct	Open	General	0.6211	0.4241
GPT-3.5-Turbo	Close	General	0.3834	0.4075
GPT-4o	Close	General	0.4146	0.4075
GPT-4o-mini	Close	General	0.3834	0.3894

Table 7: 不同模型的归因判断与人工标注的相关性。

同样有趣的是，GPT-4o/GPT-3.5 在这个环境中并不是表现最佳的模型。相反，像 Llama-3.1 和 Mistral 这样的公共模型与专家判断的相关性最好，显示出与医学专业人士更高的共识。我们假设这可能是因为公共 LLMs 可能是在包含更多医学文献的数据集上训练的，虽然很难验证，因为用于训练这些模型的数据集的细节没有公开。同时，我们意识到，推理能力在归属判断任务中起着核心作用，如附录 C 所述，提示要求模型根据自包含的摘录评估前提与假设之间的联系。然而，领域知识对于解释专业声明仍然是必不可少的。比如，验证“COVID-19 危重病例中心脏损伤常见”这样的说法，需要医学专业知识来将升高的肌钙蛋白水平与心脏损伤联系起来。因此，表现更好的 LLMs 可能得益于在医学数据集上广泛的预训练，这增强了推理和特定领域的理解能力。然而，鉴于最近表现最好的 LLMs 与专家判断的高相关性，我们认为将 LLMs 用作归属判断在医学领域更有前途，并将其视为未来工作的一个机会。

在像医学这样的知识密集领域，专家注释可能会有所不同。在 SciFact 数据集中，注释者之间的一致性（Cohen's kappa）大约是 0.75 (Phan et al., 2021)。在我们的研究中，我们观察到了类似的一致性: 对于陈述级召回率是一致性的 0.83，对于引文级精确度是一致性的 0.66，尽管任务复杂，但这反映了一种与先前工作相当的稳定性。

7 结论

我们引入了 MedCite ，这是第一个端到端框架，旨在促进研究以提高医疗系统的可验证性和可信度，并附有引文。我们对基于 LLM 的医疗系统的重要设计选择进行深入研究，并启发我们提出了 MedCite ，这是一种生成高质量医疗系统引文的新方法。

广泛的跨 LLM 评估表明，我们的方法在引文生成方面比其他方法具有一致的改进。

8 伦理考虑和局限性

这项工作的主要目标是通过引用来评估和改进基于 LLM 的医疗系统的可验证性。除了获得医生和患者的信任外，也有来自法规和审计的紧迫需求，美国食品药品监督管理局（FDA）已呼吁对在医疗行业使用 LLM 的方法进行监管 (Baumann, 2024) 。然而，不正确的引用在医疗领域可能产生严重后果，因为它们可能影响患者和医生的治疗决策。因此，在医疗环境中部署基于 LLM 的系统需要在遵循伦理考量的同时进行谨慎设计，例如系统应当增强人类决策而不是替代它，人类监管仍然是验证生成引用的关键。

本研究的一个限制是我们没有进行广泛的超参数调整（例如，检索段落的数量），而是遵循先前的经验结果和实际约束。虽然我们的选择是有道理的，但进一步的调整可能会提高性能。

在进行这项研究时，我们还发现了一些在为医学生成引用时的关键但未被探索的挑战。例如，由专业医师进行手工验证仍然是一个昂贵且费时的过程，难以大规模推广。此外，即便在医学专家之间，文档是否支持某一声明也可能存在不同解读，他们可能会对文档在多大程度上部分支持某一声明的看法不一致。因此，评估医生之间是否能实现高度共识至关重要。另一个挑战是，除 BioASQ 和 PubMedQA 外，包含底层真实答案和支持文档的医学数据集有限。在医学数据集中缺少诸如真实参考文献之类的信息，复杂化了医学中整体的可验证性评估。未来的工作应集中于开发高质量的医学引用数据集，这将显著提升医学问答系统的可信度和有效性，最终使医护专业人员和患者受益。同样，虽然我们在研究中未观察到许多此类情况，但探讨那些引用推翻论断的实例背后的原因是值得的。这尤其与语料库更新相关，因为新的研究可能会推翻先前的研究，提出如何有效处理此类情境的重要问题。

虽然 MedCite 专门为医学领域定制，但将其推广到其他领域存在显著挑战。关键组成部分，如多次引用生成和依赖于精心策划的数据库，可能无法在不进行重大修改的情况下直接应用于通用领域。例如，像 PubMed 这样的精心策划的语料库的可用性是医学领域独有的。通用领域通常缺乏集中化的资源，需要广泛的数据集准备或整合多样化的资源。同样，检索器选择，如本研究中使用的 MedCPT，可能需要适应以符合不同领域的特性和检索目标。检索配置和策略的有效性可能会根据语料库的多样性和领域特定需求而显著变化。此外，引用评估策略可能需要适应跨领域的不同要求。

在医学领域，大多数声称由于高风险和对专业知识的依赖而需要引用，而通用领域可能涉及基于常识或广为接受的事实的声称。在这种情况下评估引用可能需要进行调整，以适应可选引用或更宽松定义的相关性标准。尽管自动评估方法很有价值，但也需要进行调整以处理通用领域中声称与引用之间较简单或二元的关系。这些限制表明，尽管 MedCite 的核心框架提供了一个强有力的基础，但仍需进一步的工作以确保其组件能够广泛适用于非医学领域。

9

致谢

我们衷心感谢匿名审稿人的深刻反馈。本研究由国家科学基金会（NSF）在资助号 2441601 下支持。该工作通过“Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS)”计划，从国家超级计算应用中心（NCSA）的 DeltaAI 系统中使用了分配代码 CIS240055，该计划由国家科学基金会的拨款支持：# 2138259、# 2138286、# 2138307、# 2137603 以及 # 2138296。Delta 高级计算资源是伊利诺伊大学厄巴纳-香槟分校和 NCSA 之间的合作成果，由 NSF（授奖 OAC 2005572）和伊利诺伊州支持。该工作还利用了伊利诺伊校园集群和 NCSA NFI Hydro 集群，这些均由伊利诺伊大学厄巴纳-香槟分校和伊利诺伊大学系统支持。这项研究部分得到了国家医学图书馆（NLM）国家卫生研究院内部研究部（DIR）的支持。

References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? CoRR , abs/2305.18248.

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. arXiv preprint arXiv:2311.01463 .

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. CoRR , abs/2312.11805.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations, ICLR 2024 . OpenReview.net.

Jeannie Baumann. 2024. ChatGPT Poses New Regulatory Questions for FDA, Medical Industry. https://news.bloomberglaw.com/us-law-week/chatgpt-poses-new-regulatory-questions-for-fda-medical-industry. Accessed: 14-October-2024.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. CoRR , abs/2212.08037.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text. Preprint , arXiv:2403.18421.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In International Conference on Machine Learning, ICML 2022 , volume 162 of Proceedings of Machine Learning Research , pages 2206–2240. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Proceedings of the 34th International

Conference on Neural Information Processing Systems (NIPS'20) .

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. The NCBI handbook , 2(1).

Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont D. Antieau, Andrew S. Bennett, James J. Cimino, John W. Ely, and Hong Yu. 2011. Askhermes: An online question answering system for complex clinical questions. J. Biomed. Informatics , 44(2):277–288.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT 2019) .

Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. 2024. An evaluation framework for attributed information retrieval using large language models. CoRR , abs/2409.08014.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. Rarr: Researching and revising what language models say, using language models. Preprint , arXiv:2210.08726.

Luyu Gao et al. 2023b. Rarr: Researching and revising what language models say, using language models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 16477–16508.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023c. Enabling large language models to generate text with citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023 , pages 6465–6488. Association for Computational Linguistics.

Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. 2023. Large language model ai chatbots require approval as medical devices. Nature Medicine , 29(10):2396–2398.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020 , volume 119 of Proceedings of Machine Learning Research , pages 3929–3938. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In 9th International Conference on Learning Representations, ICLR 2021 . OpenReview.net.

Jie Huang and Kevin Chang. 2024. Citation: A key to building responsible and accountable large language models. In Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024 , pages 464–473. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. CoRR , abs/1904.05342.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John C. Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Position: Trustllm: Trustworthiness in large language models. In Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024 . OpenReview.net.

Siqing Huo, Negar Arabzadeh, and Charles L. A. Clarke. 2023. Retrieving supporting evidence for generative question answering. In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023 , pages 11–20. ACM.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. Trans. Mach. Learn. Res. , 2022.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. CoRR , abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 , pages 2567–2577. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. Bioinform. , 39(10).

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. CoRR , abs/2001.08361.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinform. , 36(4):1234–1240.

Minsuk Lee, James J. Cimino, Hai Ran Zhu, Carl L. Sable, Vijay Shanker, John W. Ely, and Hong Yu. 2006. Beyond information retrieval - medical question answering. In AMIA 2006, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 11-15, 2006 . AMIA.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024 , pages 493–516. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out , pages 74–81.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval , SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023 , pages 7001–7025. Association for Computational Linguistics.

Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 2023. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024 , pages 8–23. World Scientific.

Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. 2024. Loneliness and suicide mitigation for students using gpt3-enabled chatbots. npj mental health research , 3(1):4.

MetaAI. 2024. Introducing Meta LLaMA-3. https://ai.meta.com/blog/meta-llama-3/.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima-López, Eulàlia Farré-Maduell, Martin Krallinger, Natalia V. Loukachevitch, Vera Davydova, Elena Tutubalina, and Georgios Paliouras. 2024. Overview of bioasq 2024: The twelfth bioasq challenge on large-scale biomedical semantic indexing and question answering. In Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II , volume 14959 of Lecture Notes in Computer Science , pages 3–27. Springer.

OpenAI. 2023. GPT-4 technical report. CoRR , abs/2303.08774.

OpenAI. 2024. OpenAI GPT-4o API. https://platform.openai.com/docs/models/gpt-4o.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event , volume 174 of Proceedings of Machine Learning Research , pages 248–260. PMLR.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343 .

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. Preprint , arXiv:2106.03598.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. Advances in Neural Information Processing Systems , 34:4816–4828.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. Comput. Linguistics , 49(4):777–840.

Revanth G. Reddy, Yi R. Fung, Qi Zeng, et al. 2023. Smartbook: Ai-assisted situation report generation. CoRR , abs/2303.14337.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. , 3(4):333–389.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-Baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Benjamin Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Andrew Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale R. Webster, Joelle K. Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. CoRR , abs/2404.18416.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. CoRR , abs/2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. CoRR , abs/2305.09617.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In The Eleventh International Conference on Learning Representations, ICLR 2023 . OpenReview.net.

Pasin Tangadulrat, Supinya Sono, Boonsin Tangtrakulwanich, et al. 2023. Using chatgpt for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Medical Education , 9(1):e50658.

Mohamad-Hani Temsah, Fadi Aljamaan, Khalid H Malki, Khalid Alhasan, Ibraheem Altamimi, Razan Aljarbou, Faisal Bazuhair, Abdulmajeed Alsubaihin, Naif Abdulmajeed, Fatimah S Alshahrani, et al. 2023. Chatgpt and the future of digital health: a study on healthcare workers' perceptions and expectations. In Healthcare , volume 11, page 1812. MDPI.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. CoRR , abs/2201.08239.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 7534–7550, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. Trans. Mach. Learn. Res. , 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 .

Orion Weller, Marc Marone, Nathaniel Weir, Dawn J. Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. "according to . . . ": Prompting language models improves quoting from pre-training data. In

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers , pages 2288–2301. Association for Computational Linguistics.

Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E. Ho, and James Zou. 2024. How well do llms cite relevant medical references? an evaluation framework and analyses. Preprint , arXiv:2402.02008.

Amelie Wührl, Yarik Menchaca Resendiz, Lara Grimminger, and Roman Klinger. 2024. What makes medical claims (un)verifiable? analyzing entity and relation properties for fact verification. In Conference of the European Chapter of the Association for Computational Linguistics .

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. arXiv preprint arXiv:2402.13178 .

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. CoRR , abs/2408.00727.

Hua Yang, Shilong Li, and Teresa Gonçalves. 2024. Enhancing biomedical question answering with large language models. Inf. , 15(8):494.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. Preprint , arXiv:2210.09338.

Xi Ye, Ruoxi Sun, Sercan Ö. Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. Preprint , arXiv:2311.09533.

Haoran Yu, Chang Yu, Zihan Wang, Dongxian Zou, and Hao Qin. 2024. Enhancing healthcare through large language models: A study on medical question answering. CoRR , abs/2408.04138.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. NEJM AI , 1(2):AIoa2300068.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jinfang Hu, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. CoRR , abs/2406.03949.

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. Chatgpt hallucinates when attributing answers. In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023 , pages 46–51. ACM.

A 相关工作

A.1 生物医学问答

生物医学问答（QA）是自然语言处理中的一个专业领域。它专注于回答与生物医学和临床领域相关的问题。早期的方法主要依赖于基于规则的系统（Lee et al., 2006; Cao et al., 2011）。这些方法利用结构化数据库和本体来检索临床问题的答案。虽然具有开创性，但这些系统受到其对预定义规则的依赖和可扩展性不足的限制。随后，基于机器学习/深度学习的解决方案为生物医学 QA 带来了显著的改进。像 BioBERT（Lee et al., 2020）和 ClinicalBERT（Huang et al., 2019）这样的模型将预训练的 BERT（Devlin et al., 2019）适配于生物医学文本，从而在各种生物医学 QA 任务上取得更好的性能（Yang et al., 2024）。最近，生成模型代表了生物医学 QA 中的新范式。像 GPT-3.5/4（Brown et al., 2020; OpenAI, 2023）和 Med-Gemini（Saab et al., 2024）这样的模型直接从输入文本生成答案，而不依赖于预定义的答案选项，这使得其能够提供更灵活和更具语境的适当回答。然而，生成模型也带来了挑战，如生成不正确或虚构答案的风险。为了解决这个问题，最近的工作采用检索增强生成（RAG）来检索相关文档并基于检索的信息生成答案（Lozano et al., 2023; Xiong et al., 2024a; Yang et al., 2024; Yu et al., 2024; Zakka et al., 2024; Xiong et al., 2024b）。与这些努力不同，我们专注于提高医疗系统的可验证性。

A.2 LLM 生成的引用方法

在基于大型语言模型（LLM）的生成中整合引用机制是一个新兴的研究领域。近期在大型语言模型方面的进展可以提示模型在其生成的文本中加入引用（Brown et al., 2020; Thoppilan et al., 2022; Anil et al., 2023; OpenAI, 2023, 2024）。然而，这些引用的准确性和相关性可能是一个挑战。类似于生成文本中的幻觉，模型（例如，ChatGPT）可以生成看似不错但实际上不准确或无法验证的引用（Zuccon et al., 2023）。已经提出了多种方法来为大型语言模型生成的内容添加引用。直接的模型驱动归因方法允许模型自我归因，但这通常导致不可靠的结果（Sun et al., 2023; Agrawal et al., 2023; Weller et al., 2024）。检索后生成（PRG）涉及在生成

回答和相关文档之前，先检索与用户查询相关的一系列文档（Guu et al., 2020; Borgeaud et al., 2022; Reddy et al., 2023）。后生成引用（PGC）在生成答案后寻找相关文档（Huo et al., 2023）。PRG 和 PGC 都提供了更可靠的归因，但增加了系统复杂性（Gao et al., 2023b），如我们在论文中所示，由于生物医学查询的细微差别和对精确、可验证引用的需求，它们可能无法实现医学系统的最佳引用质量。我们的混合双通道引用方法旨在通过将 RAG 与生成后优化相结合来解决这些问题。对大型语言模型进行微调以生成引用是另一种方法，该方法使用精选或合成数据对模型进行训练，以便在文本生成期间直接生成引用（Ye et al., 2024）。最后，当前缺乏基于 LLM 的问答系统引用方法的自动化评估。因此，已有努力旨在改进 LLM 归因的评估协议和基准（Rashkin et al., 2023; Gao et al., 2023c; Li et al., 2024）。不同于那些对一般领域主题的引用进行测量的努力，我们的评估以医学为中心，同时我们也探讨了其他组件，例如医学专用检索和影响基于 LLM 的医学任务的引用查找策略。

A.3 医学领域中由 LLM 生成的引文的评估框架

Wu et al. (2024) 提出了一种用于评估在医学领域中由大型语言模型生成的引用有效性的评估流程，重点关注基于 URL 的在线资源。他们的工作突出了大型语言模型在引用质量上的重大局限性，即使是表现最优的模型如 GPT-4（RAG）也未能在接近一半的响应中完全支持所有的陈述。

虽然我们的研究和吴等人的研究都有提高引用可靠性的目标，但我们的工作在范围和方法上有所不同。吴等人提供了一个全面的评估流程，主要侧重于通过提示基于 API 的 LLMs 在其答案中提供源 URL 来分析参数化方法的引用质量，而不是提出方法来解决已识别的差距。相比之下，我们的工作不仅进行评估，还引入了一个结合分层检索和多次引用的模块化框架，以提高生物医学任务的引用质量。在第 3.1 节中，我们解释了为什么参数化方法特别不适用于开源 LLMs，原因包括虚构引用、无法访问可靠内容，以及在没有 API 级别访问在线资源的情况下进行自动评估的困难。通过强调从经过审核的医学资源如 PubMed 进行领域特定的分层检索，我们解决了生物医学领域独有的挑战，例如确保对药物名称或基因标记等高度专业术语的精确性。

B 额外实验细节

B.1 数据集

我们使用具有标准答案的医学问答数据集来评估 MedCite 。特别是，我们在最终评估中使用了 BioASQ（Nentidis et al., 2024）和 PubMedQA（Jin et al., 2019）。在这两种情况下，我们只使用问题并去除所有标准答案支持的上下文，这代表了一种更现实的设置，因为在实际使用场景中通常没有提供示例。表格 8 总结了这两个数据集的详细信息。

Dataset	Size	Question Type	Example Question	GT Answer
PubMedQA*	500	Yes/No/Maybe	Is anorectal endosonography valuable ... ?	yes
BioASQ-Y/N	618	Yes/No	Is medical hydrology the same as Spa ... ?	yes

Table 8: MedCite 实验中使用的两个数据集。

PubMedQA。 PubMedQA 是一个用于生物医学问答（QA）任务的数据集。这些问题要么是现有研究文章的标题，要么是从中得出的。上下文提供了文章的摘要。答案包括问题的标准答案，这个答案是从摘要的结论中得出的。

BioASQ-Y/N。 BioASQ-Y/N 也是一个生物医学问答数据集。数据集中的每个实例都包含一个问题、提供回答该问题信息的上下文以及人工标注的答案。

为了确保可重复性，我们对所有 LLM 使用贪婪解码。对于检索，我们采用分层的两阶段排名过程：（1）使用 Pyserini 实现的 BM25，针对索引使用默认的超参数，以及（2）使用默认设置的 MedCPT Cross-Encoder 为给定查询对检索到的文档进行排名。我们检索前 32 份文档用于答案生成，确保它们符合模型的上下文窗口，如果有必要，丢弃相似度得分较低的文档。我们在答案生成后检索单个陈述的前三份文档以寻求引文。我们为检索的文档选择 $k = 3$ 基于两个主要考虑因素：(1)UltraMedical 的上下文长度限制（1024 个标记），大约容纳三篇 PubMed 摘要，使得 $k = 3$ 非常适合于所有模型的公平比较；（2）Gao 等人的先前分析显示，引文质量在前三段落达到平台期，进一步支持了我们的选择。

B.2 Rouge-L 与准确率之间的相关性

表格 9 展示了在不同条件下 ROUGE 分数与准确率之间的关系：

Conditions	ROUGE-L Score	Accuracy
medrag + medcpt	17.04	0.8414
medrag + MedCPT + system prompt	17.98	0.8576
medrag + medcpt + new prompt	17.34	0.8269
oracle relevant docs	22.00	0.9401

Table 9: 在不同条件下 ROUGE-L 分数和准确率的分析

该表清晰地展示了 ROUGE 分数与准确率之间的正相关性。具体来说，当将系统提示引入 MEdRAG 模型时，ROUGE 分数从 17.04 增加到 17.98，准确率也从 0.8414 提升至 0.8576。这表明通过优化提示，我们可以在某种程度上提高模型的输出质量和准确性。此外，当引入一个新的提示时，尽管 ROUGE 分数略有下降，但准确率下降更为明显，这表明新提示可能在某些方面影响了模型的性能。最显著的是，当使用 oracle 相关文档时，ROUGE 得分和准确率都达到其峰值，进一步确认了 ROUGE 分数与模型输出准确率之间的正相关性。这些结果表明，ROUGE 分数可以作为评估和优化大型语言模型（LLMs）输出质量的有效指标。

C 提示模板

Prompts template for CoT-generation

You are a helpful medical expert, and your task is to answer a multi-choice medical question.'"

general_cot = Template('"
Here is the question:
{{question}}

Here are the potential choices:
{{options}}

Please first think step-by-step and then choose the answer from the provided options. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Your responses will be used for research purposes only, so please have a definite answer.

Figure 4: 用于链式推理生成的提示模板。

D 生成引用的例子

表格 10 展示了使用参数化引用方法生成的医学参考示例，使用的工具包括 Llama-3-8B-I.、UltraMedical 和 GPT-4o。对于 Llama-3-8B-I.，参考文献 [1] 中提供的 URL 是不正确的，而参考文献 [2] 和 [3] 虽然标题相同，但作者不同。经过检查，发现相关文章并不存在。UltraMedical 包含格式不佳的内联引用和捏造的参考文献。另一方面，GPT-4o 提供了正确的参考文献，但由于对源的 API 访问有限，它们难以评价。

E 注释指南和分析

下面我们提供了在第 6.2 节中我们使用的人类注释指南。我们要求注释者遵循这些指南来进行归因判断。

Prompts template for Parametric-citation-generation

You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents. Please first think step-by-step and then choose the answer from the provided options. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Your responses will be used for research purposes only, so please have a definite answer.

Here is the question:
{{question}}

Here are the potential choices:
{{options}}

Please think step-by-step and generate your output in json and include citations after each sentence. When citing several documents, use [1][2][3]. Cite at least one document and at most three documents in each sentence. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}, "references:Str(References)"}:

Figure 5: 参数化引用的提示模板。

Prompts template for MedRAG-generation

You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents. Please first think step-by-step and then choose the answer from the provided options. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Your responses will be used for research purposes only, so please have a definite answer.

Here are the relevant documents:
{{context}}

Here is the question:
{{question}}

Here are the potential choices:
{{options}}

Please think step-by-step and generate your output in json:

Figure 6: 用于 MedRAG 生成的提示模板。

Prompts template for MedRAG-cite-generation

You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents. Please first think step-by-step and then choose the answer from the provided options. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Your responses will be used for research purposes only, so please have a definite answer.

Here are the relevant documents:
{{context}}

Here is the question:
{{question}}

Here are the potential choices:
{{options}}
Please think step-by-step and generate your output in json using only the provided documents (some of which might be irrelevant) and cite them using [] after which your statement is based on it. When citing several documents, use [1][2][3]. Cite at least one document and at most three documents in each sentence. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}:

Figure 7: 用于 MedRAG 及引文生成的提示模板。

Prompts template for Citation-seeker

You are a helpful medical expert and your task is to try to find documents that supports the statement, given relevant documents.

Here are the relevant documents:
{{context}}

Here is the statement:
{{statement}}

Output only the document IDs with which supports the statement.
Do not output other things.

Figure 8: 针对引用查找者的提示模板。

Prompts template for Attribute Judge

Recall:
You are a helpful medical expert.
Based on the document, determine whether the statement is fully supported or not.
Options:
- Fully Supported: The statement is fully supported by the document.
- Not Fully Supported: The statement is not fully supported by the document.
Provide only your chosen option.
Document: {{premise}}
Statement: {{hypothesis}}

Precision:
You are a helpful medical expert.
Based on the document, determine whether it supports the statement.
Options:
- Fully Support: The document fully supports the statement.
- Partial Support: The document supports part of the statement, but some parts are missing.
- Cannot Support: The document cannot support the statement.
Provide only the chosen option.
Document: {{premise}}
Statement: {{hypothesis}}

Figure 9: 属性判断提示模板。

E.1 注释指南

引文召回衡量所有引文的组合支持陈述的程度。

- 对于每一个陈述，将所有提供的引用（例如，PubMed 文章）作为一个整体进行审查。
- 确定这些引用的综合信息是否完全支持或无法完全支持该陈述。

引用精确度衡量每个单独的引用支持陈述的效果。

- 对于每个引用，评估其是否能够完全支持、部分支持或不支持该陈述。
- 对每个引用独立重复此评估。

注意：请仅使用 PubMed 文章的摘要作为引用，而不是整篇文章（只查看摘要，而不是整篇文章）。

关于“完全支持”的澄清：该判断取决于陈述和引文内容之间的关系。

- 文章中未提及的词语：如果陈述中的词语代表的事物与文章描述的完全不同（例如，不同的医学术语没有重叠），那么陈述不能被视为“完全支持”。在这种情况下，支持可能是“未支持”或“部分支持”，这取决于信息之间的相关程度。如果词语描述的是文章中提到的某个更广泛概念的子类或具体实例（例如，文章讨论一种治疗类别，而陈述中提到该类别中的一种治疗），那么引用可以被认为是“部分支持”。
- 完全支持标准：只有当声明中的所有关键术语和概念都被引用中的信息直接涉及并明确支持时，声明才能被视为“完全支持”。

E.2 例子

- 声明：“像苹果这样的水果是维生素 C 的丰富来源。”
- 引用 1：文章提到“橙子、草莓和猕猴桃等水果是维生素 C 的极好来源”。
- 引用 2：这篇文章讨论了“苹果很有营养，但专注于它们的纤维含量”，并未提及维生素 C。

回顾（文献组合）：如果你把这两个引文放在一起看，它们并不能完全支持这个陈述。虽然引文 1 提到了富含维生素 C 的水果，但并未明确提到苹果；而引文 2 并没有提供关于苹果中维生素 C 的相关信息。

召回分数：不支持。

精准度（每个引用分别计算）：

- 引用 1: 提供“部分支持”，因为它提到富含维生素 C 的水果，但没有具体说明苹果。
- 引文 2：提供了“没有支持”，因为它根本没有提到维生素 C。

F 案例研究：MedCite 与开放领域基线的比较

为了说明现有基准的局限性和 MedCite 的优点，我们在 BioASQ 数据集中对问题进行了案例研究：“avelumab 对尿路上皮癌有效吗?”

我们比较了第 6.1 节中描述的四种配置：以 Llama-3-8B-I 作为基础大型语言模型的 MedRAG、PRG、PGC、MedCite。

下面我们展示每个系统输出的摘录，接着是语句级别的引用评估。完整的输出和支持文件见附录 F.2 。

F.1 分析

这个案例凸显了现有方法的几个关键限制以及 MedCite 的优势。MedRAG 和 PRG 都依赖于单轮的问题级检索。因此，重要的论点可能完全缺少支持文档（MedRAG），或者由于检索粒度过于粗略而被错误引用（PRG）。相比之下，PGC 在支持覆盖范围方面表现更好：所有声明都通过事后引文检索得到了支持。然而，它的生成完全基于参数模型知识，使其难以更新，并容易随着医学知识的发展而出现事实漂移。

MedCite 弥合了这些差距。通过在检索到的文档中奠定初始答案生成的基础（如 PRG），然后通过语句级检索和重新排序改进引用支持（如 PGC），它实现了两者的最佳结合：答案的正确性、完整性和适应性。在像医学这样高风险的领域中，知识更新频繁且可信度至关重要，这种混合设计比纯粹的参数化或粗略检索方法提供了更稳健和可验证的解决方案。

F.2 案例研究的完整输出和引用

Question: Is obesity related to cognitive decline?
Llama-3-8B-I. answer: Obesity is a complex condition that is associated with various health risks, including cardiovascular disease, type 2 diabetes, and certain types of cancer [1]. Research has also suggested that obesity may be linked to cognitive decline, as it can lead to chronic inflammation, oxidative stress, and changes in brain structure and function [2]. Studies have shown that obese individuals are at a higher risk of developing dementia and Alzheimer's disease, and that weight loss may help to improve cognitive function [3]. Therefore, it is likely that obesity is related to cognitive decline. References : [1] World Health Organization. (2018). Obesity and overweight. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>; [2] Yaffe, K., et al. (2013). The association between obesity and cognitive decline in older adults. Journal of the American Medical Association, 310(11), 1155-1162; [3] Luchsinger, J. A., et al. (2013). Obesity and cognitive decline in older adults. Journal of Alzheimer's Disease, 35(2), 247-255.
UltraMedical answer: Obesity is a condition characterized by excessive body fat that poses a risk to health. It is associated with various health issues, including cardiovascular diseases, diabetes, and certain types of cancer. To determine if obesity is related to cognitive decline, we need to consider the evidence linking obesity to brain health. 1. Research has shown that obesity can lead to changes in the brain's structure and function. For instance, studies have indicated that obesity is associated with reduced brain volume and alterations in brain activity, which could potentially affect cognitive functions. 2. Additionally, obesity is known to increase the risk of metabolic syndrome, which includes conditions such as hypertension, high blood sugar, and high cholesterol. These conditions have been linked to an increased risk of cognitive decline and dementia. 3. Inflammation is another factor to consider. Obesity is a state of chronic low-grade inflammation, and inflammation is a known risk factor for cognitive decline. Adipose tissue, particularly visceral fat, secretes pro-inflammatory cytokines that can affect brain health. 4. Furthermore, obesity can lead to sleep apnea and other sleep disturbances, which are associated with cognitive deficits. Poor sleep quality has been linked to a higher risk of cognitive decline and Alzheimer's disease. 5. Based on the evidence, it is reasonable to conclude that obesity is related to cognitive decline. The mechanisms include direct effects on brain structure and function, metabolic disturbances, inflammation, and sleep disturbances. Therefore, the answer is A. Yes, obesity is related to cognitive decline. References : [1] Hall, K. D., & Guo, J. (2017). Obesity: portable operating instructions. JAMA, 318(17), 1685-1686; [2] Yaffe, K., et al. (2013). The metabolic syndrome, inflammation, and risk of cognitive decline and dementia: the Health, Aging, and Body Composition Study. JAMA, 310(3), 197-208; [3] Spiegel, K., et al. (2009). Sleep, cognition, and health: emerging research directions. Dialogues in Clinical Neuroscience, 11(4), 667-679.
GPT-4o answer: Obesity has been linked to cognitive decline through various mechanisms. Excess body fat can lead to inflammation and insulin resistance, both of which are associated with negative effects on brain health [1][2]. Additionally, obesity is a risk factor for cardiovascular diseases, which can impair blood flow to the brain and contribute to cognitive decline [2][3]. Studies have shown that individuals with obesity may have a higher risk of developing conditions like Alzheimer's disease and other forms of dementia [1][3]. Therefore, there is substantial evidence supporting the relationship between obesity and cognitive decline. References : [1] Smith, E., Hay, P., Campbell, L., & Trollor, J. N. (2011). A review of the association between obesity and cognitive function across the lifespan: implications for novel approaches to prevention and treatment. Obesity Reviews, 12(9), 740-755; [2] Kivipelto, M., Ngandu, T., Fratiglioni, L., Viitanen, M., Kreholt, I., Winblad, B., & Nissinen, A. (2005). Obesity and vascular risk factors at midlife and the risk of dementia and Alzheimer disease. Archives of Neurology, 62(10), 1556-1560; [3] Whitmer, R. A., Gunderson, E. P., Quesenberry, C. P., Zhou, J., & Yaffe, K. (2007). Obesity in middle age and future risk of dementia: a 27 year longitudinal population based study. BMJ, 335(7630), 705.

Table 10: 使用参数化引用方法生成的医学参考示例。这些引用通常不正确且格式不佳。最重要的是，由于缺乏与来源的 API 访问，它们很难被评估。

Question: Is avelumab effective for bladder cancer?		
Method	Generated Answer	Statement Evaluation
MedRAG	Avelumab is a PD-L1 inhibitor that has been approved by the FDA for the treatment of urothelial carcinoma (UC) in patients who have received platinum-based chemotherapy. The JAVELIN Bladder 100 trial showed that avelumab maintenance treatment after first-line chemotherapy significantly prolonged overall survival...	• 陈述 1：没有引用—无法验证 • 声明 2：不支持—文件覆盖范围不明确
PRG	Avelumab is a PD-L1 inhibitor that has been shown to be effective in treating advanced urothelial carcinoma. In the JAVELIN Bladder 100 trial, avelumab was found to significantly prolong overall survival compared to best supportive care alone [17]. Additionally, avelumab has been approved by the FDA for the treatment of patients with locally advanced or metastatic urothelial carcinoma who have disease progression during or following platinum-containing chemotherapy [17].	• 声明 1：不完全支持——未提供引用 • 陈述 2：完全支持 • 声明 3：未完全支持——引用的文件未提及 FDA 批准
PGC	Avelumab is a monoclonal antibody that targets PD-L1, a protein that helps cancer cells evade the immune system [1][2]. It has been studied in various clinical trials for the treatment of bladder cancer, particularly in patients with locally advanced or metastatic urothelial carcinoma [3][4]. The results of these trials have shown that avelumab can improve overall survival and response rates [5][6]. Therefore, based on the available evidence, I conclude that avelumab is effective for bladder cancer [7][8].	• 所有陈述均得到充分支持
MedCite	Avelumab is a PD-L1 inhibitor that has been shown to be effective in treating advanced urothelial carcinoma [2][3]. In the JAVELIN Bladder 100 trial, avelumab was found to significantly prolong overall survival compared to best supportive care alone [1][4]. Additionally, avelumab has been approved by the FDA for the treatment of patients with locally advanced or metastatic urothelial carcinoma who have disease progression during or following platinum-containing chemotherapy [1][5][6].	• 所有陈述均得到充分支持

Table 11: 模型生成的答案与示例的陈述评估的比较。

Question: Is avelumab effective for bladder cancer?
MedRAG Answer: After reviewing the documents, I found that avelumab is a PD-L1 inhibitor that has been approved by the FDA for the treatment of urothelial carcinoma (UC) in patients who have received platinum-based chemotherapy. The JAVELIN Bladder 100 trial showed that avelumab maintenance treatment after first-line chemotherapy significantly prolonged overall survival compared to best supportive care alone. I also found that avelumab has been shown to be effective in treating patients with locally advanced or metastatic UC who have responded or stabilized after first-line platinum-containing chemotherapy. The National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology recommend avelumab as a category 1 preferred checkpoint inhibitor maintenance option for patients with cisplatin-eligible and cisplatin-ineligible locally advanced or metastatic UC. Based on the evidence, I conclude that avelumab is effective for bladder cancer.
PRG Answer: Avelumab is a PD-L1 inhibitor that has been shown to be effective in treating advanced urothelial carcinoma. In the JAVELIN Bladder 100 trial, avelumab was found to significantly prolong overall survival compared to best supportive care alone [17]. Additionally, avelumab has been approved by the FDA for the treatment of patients with locally advanced or metastatic urothelial carcinoma who have disease progression during or following platinum-containing chemotherapy, or have disease progression within 12 months of neoadjuvant or adjuvant platinum-containing chemotherapy [17]. Cited Documents : Document [17] (Title: Avelumab Maintenance Therapy for Advanced or Metastatic Urothelial Carcinoma, PMID: 32945632): "Platinum-based chemotherapy is standard-of-care first-line treatment for advanced urothelial carcinoma. However, progression-free survival and overall survival are limited by chemotherapy resistance. ..."
PGC Answer: Avelumab is a monoclonal antibody that targets PD-L1, a protein that helps cancer cells evade the immune system [1][2]. It has been studied in various clinical trials for the treatment of bladder cancer, particularly in patients with locally advanced or metastatic urothelial carcinoma [3][4]. The results of these trials have shown that avelumab can improve overall survival and response rates [5][6]. Therefore, based on the available evidence, I conclude that avelumab is effective for bladder cancer [7][8]. Cited Documents : Document [1] (Title: Immune checkpoint inhibitors of PD-L1 as cancer therapeutics, PMID: 31488176): "Since the discovery of immune checkpoint proteins, there has been a special interest in developing antibodies that block programmed cell death 1 receptor (PD-1) and programmed cell death receptor ligand 1 (PD-L1) ..." Document [2] (Title: Avelumab: combining immune checkpoint inhibition and antibody-dependent cytotoxicity, PMID: 28274143): "Immune checkpoint inhibition holds great promise for selected tumors. ..." Document [3] (Title: Immunotherapy in bladder cancer-quo vadis? Update on current trials and developments, PMID: 32468092): "Great advances have been made for the treatment of urothelial carcinoma by the introduction of checkpoint inhibitors (CPI). ..." Document [4] (Title: The biology and rationale of targeting nectin-4 in urothelial carcinoma, PMID: 33239713): "Bladder cancer is the tenth most common cancer type worldwide. Urothelial carcinoma is the most common type of bladder cancer and accounts for 90 % of bladder cancer cases in the USA and Europe. ..." Document [5] (Title: Avelumab for advanced Merkel cell carcinoma in the Netherlands: a real-world cohort, PMID: 32948651): "Merkel cell carcinoma (MCC) is associated with high recurrence rates and poor survival when metastatic disease is present. ..." Document [6] (Title: Clinical Evaluation of Avelumab in the Treatment of Advanced Urothelial Carcinoma: Focus on Patient Selection and Outcomes, PMID: 35237074): "First-line therapy for treatment of advanced urothelial carcinoma includes combination platinum-based chemotherapies, ..." Document [7] (Title: Plain language summary of results from the JAVELIN Bladder 100 study: avelumab maintenance treatment for advanced urothelial cancer, PMID: 35416053): "This is a plain language summary of an article originally published in <iThe New England Journal of Medicine</i. ..." Document [8] (Title: Cost-Effectiveness of Avelumab Maintenance Therapy Plus Best Supportive Care vs. Best Supportive Care Alone for Advanced or Metastatic Urothelial Carcinoma, PMID: 35570929): "Avelumab (MSB0010718C) is a fully human anti-programmed cell death ligand 1(PD-L1) antibody against PD-L1 interactions and enhances immune activation against tumor cells in the meantime. ..."

Table 12: 案例研究答案和引用的文献。

MedCite:
Avelumab is a PD-L1 inhibitor that has been shown to be effective in treating advanced urothelial carcinoma [2][3]. In the JAVELIN Bladder 100 trial, avelumab was found to significantly prolong overall survival compared to best supportive care alone [1][4]. Additionally, avelumab has been approved by the FDA for the treatment of patients with locally advanced or metastatic urothelial carcinoma who have disease progression during or following platinum-containing chemotherapy, or have disease progression within 12 months of neoadjuvant or adjuvant platinum-containing chemotherapy [1][5][6].
Cited Documents :
Document [1] (Title: Avelumab Maintenance Therapy for Advanced or Metastatic Urothelial Carcinoma, PMID: 32945632): "Platinum-based chemotherapy is standard-of-care first-line treatment for advanced urothelial carcinoma. ..."
Document [2] (Title: Clinical Evaluation of Avelumab in the Treatment of Advanced Urothelial Carcinoma: Focus on Patient Selection and Outcomes, PMID: 35237074): First-line therapy for treatment of advanced urothelial carcinoma includes combination platinum-based chemotherapies, though resistance and long-term toxicity concerns to these regimens cause limitations in progression-free survival and overall survival. ..."
Document [3] (Title: Which place for avelumab in the management of urothelial carcinoma?, PMID: 31286802): "<bIntroduction</b: Urothelial carcinoma (UC) has a poor prognosis, with the only standard first-line metastatic treatment being platinum-based chemotherapy. ..."
Document [4] (Title: Patient-reported Outcomes from JAVELIN Bladder 100: Avelumab First-line Maintenance Plus Best Supportive Care Versus Best Supportive Care Alone for Advanced Urothelial Carcinoma, PMID: 35654659): In JAVELIN Bladder 100, avelumab first-line maintenance plus best supportive care (BSC) significantly prolonged overall survival (OS; primary endpoint) versus BSC alone in patients with advanced urothelial carcinoma (aUC) without disease progression with first-line platinum-containing chemotherapy. ..."
Document [5] (Title: FDA Approval Summary: Atezolizumab for the Treatment of Patients with Progressive Advanced Urothelial Carcinoma after Platinum-Containing Chemotherapy, PMID: 28424325): "Until recently in the United States, no products were approved for second-line treatment of advanced urothelial carcinoma. ..."
Document [6] (Title: Avelumab in metastatic urothelial carcinoma after platinum failure (JAVELIN Solid Tumor): pooled results from two expansion cohorts of an open-label, phase 1 trial, PMID: 29217288): "The approval of anti-programmed death ligand 1 (PD-L1) and anti-programmed death 1 agents has expanded treatment options for patients with locally advanced or metastatic urothelial carcinoma. ..."

Table 13: 案例研究答案和引用的文献。