SELF-MEDRAG：一种用于可靠医学问答的自反思混合检索增强生成框架

JESSICA RYAN[1,*]，ALEXANDER I. GUMILANG[1], ROBERT WILIAM[1],DERWIN SUHARTONO[1]

[1]计算机科学学院，Bina Nusantara University，Anggrek 校区，Jl. Raya Kb. Jeruk No. 27，11530，印度尼西亚雅加达 *通讯作者：jessica.ryan@binus.ac.id

摘要

大语言模型（LLM）在医学问答（QA）方面展现出巨大潜力，但仍容易产生幻觉和无依据推理，因此在高风险临床场景中的可靠性受到限制。检索增强生成（RAG）通过整合外部知识缓解了这些问题，但传统的单轮检索往往无法解决需要多步推理的复杂生物医学查询。为此，我们提出 Self-MedRAG，这是一种自反思混合框架，旨在模拟临床推理中的迭代式假设验证过程。Self-MedRAG 集成了混合检索策略，通过倒数排名融合（RRF）结合稀疏检索器（BM25）和稠密检索器（Contriever），以最大限度地提高证据覆盖率。该框架利用生成器生成答案及其支持性理由，再由轻量级自反思模块通过自然语言推理（NLI）或基于大语言模型的验证进行评估。如果推理依据缺乏充分的证据支持，系统便会自动改写查询并迭代优化上下文。我们在 MedQA 和 PubMedQA 基准测试集上评估了 Self-MedRAG。结果表明，我们的混合检索方法显著优于单检索器基线。此外，引入自反思循环带来了显著提升：MedQA 的准确率从 80.00% 提高到 83.33%，PubMedQA 的准确率从 69.10% 提高到 79.82%。这些发现证实，将混合检索与基于证据的迭代自我反思相结合，能够有效减少无依据主张，并提升基于大语言模型系统的临床可靠性。

关键词：检索增强生成（RAG）；医学问答；大语言模型；混合检索。

1. 引言

大语言模型（LLM）通过综合复杂的生物医学知识，为医学问答（QA）展现出巨大潜力，并为临床医生和患者提供支持 [1,2]。然而，尽管大语言模型具备诸多优势，仍受幻觉、过度自信的推理以及对静态预训练语料库的依赖所限制，而这些语料库无法跟上医学证据的快速发展。在高风险临床场景中，这些缺陷会带来严重风险 [3]。

检索增强生成（RAG）通过以外部证据为大语言模型的输出提供依据，已成为一种颇具前景的解决方案 [4]。然而，传统检索增强生成通常采用单轮流程：检索一次，作答一次。这种静态工作流程在一般领域有效，但在医学语境中往往难以胜任，因为相关问题需要反复澄清、修订假设，并逐步收集证据，这与临床医生进行鉴别诊断时的推理过程十分相似。单次检索往往不足以捕捉细微的生物医学知识或解决含糊不清的问题 [5,6]。

近期研究通过自我反思和迭代式检索增强生成方法来应对这些局限。Self-RAG 使模型能够审视自己的答案并请求更多证据，但其完整实现复杂、计算开销大，且并未针对临床推理领域进行专门优化 [7]。相反，轻量级迭代式检索增强生成系统会执行多轮推理—检索循环，但往往缺乏领域感知机制，无法评估检索到的证据是否真正支持模型的推理依据。因此，这些方法要么引入不必要的复杂性，要么未能满足医学问答对领域感知证据的要求 [8,9]。

为应对这一挑战，我们提出 **Self-MedRAG**，一种用于可靠医学问答的自我反思式混合检索增强生成框架。Self-MedRAG 通过结合混合检索、利用现有自然语言推理和大语言模型的轻量级自我反思模块，以及迭代式查询优化，模拟临床医生循序渐进的推理过程。这一工作流使该框架能够逐步增强事实依据，同时保持临床一致性和透明度。

我们的贡献如下：

1. 我们提出了**Self-MedRAG**，这是一个迭代式检索增强生成框架，将检索、生成、验证和查询优化整合到一个统一循环中，专为医学问答设计。 2. 我们利用现成的自然语言推理和大语言模型，构建了一种轻量级自我反思机制，用于指导额外检索并提升证据对齐，同时避免大量计算。 3. 实证结果表明，Self-MedRAG 在多个医学问答基准上提升了事实准确性、证据支撑和临床安全性。

2. 相关工作

2.1. 医学问答

医学问答基准测试集用于评估模型生成临床可靠、以证据为依据的回答的能力。广泛使用的数据集

包括用于基于医学考试进行诊断推理的 MedQA、用于基于研究摘要开展循证生物医学推理的 PubMedQA，以及用于广泛医学知识测试的 MMLU-Medical 和 MedMCQA [10–13]。MIRAGE基准进一步整合了这些资源，提供标准化的检索增强生成评估设置，将来自多个医学问答数据集的问题与精选检索语料库配对，以评估依据支撑质量、检索依赖性和事实准确性，从而推动检索增强方法在医学问答中的应用 [14]。

2.2. 医学领域的大语言模型

随着大规模大语言模型的兴起，许多研究探索了将预训练语言模型应用于医学问答。BioBERT、PubMedBERT、Med-PaLM 和 PMC-LLaMA 等模型通过对非结构化临床文本进行推理并处理复杂的多步骤问题，在生物医学领域展现出强大的性能 [15–17]。然而，尽管取得了这些进展，大语言模型在临床情境中仍面临显著局限，包括产生幻觉、依赖静态或不完整的预训练数据、对人口统计特征和领域偏移敏感，以及偏见、可解释性和公平决策方面的挑战；这些问题限制了其在临床决策支持中的可靠性 [18]。先前的研究还在检索增强生成和监督微调条件下评估了 GPT-4.1 和 DeepSeek 等 API模型以及开源大语言模型，为混合检索和基于证据的医学问答提供了基线 [19]。

2.3. 检索增强生成与知识图谱

检索增强生成（RAG）通过以外部证据为基础生成输出，提高了大语言模型的可靠性。经典RAG系统依赖单一检索步骤，使用BM25等稀疏方法，通过基于关键词的词项重叠匹配查询与文档；或使用E5、BGE和Contriever等稠密模型，将文本编码为语义向量表示，以捕捉超越精确措辞的深层含义 [20–22]。混合检索通过倒数排名融合（RRF）或加权融合等方法组合多个检索器的输出，并进一步应用交叉编码器重排序器，为大语言模型识别最相关的段落，从而增强覆盖范围 [23,24]。在医学场景中，BioBERT、PubMedBERT和MedCPT等领域适配检索器利用生物医学语料库和PubMed搜索日志提高精确率，构成医学RAG系统的领域专用骨干 [25,26]。

近期的基于图的RAG系统将文档检索扩展至知识图谱和生物医学本体，使模型能够访问实体和关系层面的信息，从而围绕疾病、症状和治疗开展结构化推理。在医学领域，MedRAG和MedGraphRAG等框架将文本检索与图遍历或图导航相结合，以增强证据支撑和多跳推理。这些方法提升了事实性和可解释性，但也增加了图构建、计算负载、索引和维护方面的复杂性 [27–29]。

2.4. RAG 工作流和批评模型

RAG 工作流已从传统的单次检索发展到先进架构，这些架构结合迭代检索、模块化任务分解和基于批评器的优化，逐步提升答案质量。在这些进展的基础上，智能体式 RAG 进一步赋予大语言模型自主规划和决策的能力

决定何时检索或重新表述查询、核验证据，以及调用外部工具进行复杂的多步推理 [30–33]。这些先进 RAG 范式的核心是批评模块，负责确保事实性和上下文效用 [34]。例如，Self-RAG 学习到的反思标记可用于自我批评其生成的答案、识别无依据主张，并判断是否需要额外检索；CRAG（纠正式检索增强生成）的检索充分性评估器则用于检查上下文相关性 [7,35]。

在医学领域，若干框架将这些工作流理念应用于临床推理。MRD-RAG通过多轮检索和生成，模拟临床医生的诊断过程；每一轮都基于先前的答案和检索到的证据，逐步完善结论 [36]。i-Med-RAG在医学领域采用轻量级迭代循环，在不增加大量计算开销的情况下，逐步提升答案质量和检索覆盖率 [5]。SIM-RAG引入了充分性批评器，在证据充分时提前停止推理，从而提高效率；但它依赖通用领域充分性信号，这些信号可能无法反映临床要求 [6]。这些系统凸显了迭代推理、评判和证据充分性在医学决策支持场景中的重要性，但它们都未能将强大的领域基础与高效的自我反思充分结合起来。

3. 方法论

3.1. Self-MedRAG 概述

Self-MedRAG 是一种用于可靠医学问答的迭代检索增强生成框架。如图1所示，该系统遵循一个循环，其过程与临床医生收集证据、生成临时解释并重新评估推理是否得到充分支持的方式相似。在每次迭代中，模型都会检索上下文段落，生成答案及其推理依据，然后通过评分模块评估该推理依据。如果推理依据的部分内容缺乏支持或相互矛盾，系统就会重新表述查询，以针对缺失的信息，并重复检索和生成。这个反思循环使 Self-MedRAG 能够逐步增强事实依据，同时确保最终答案及其推理依据在临床上连贯且以证据为基础。

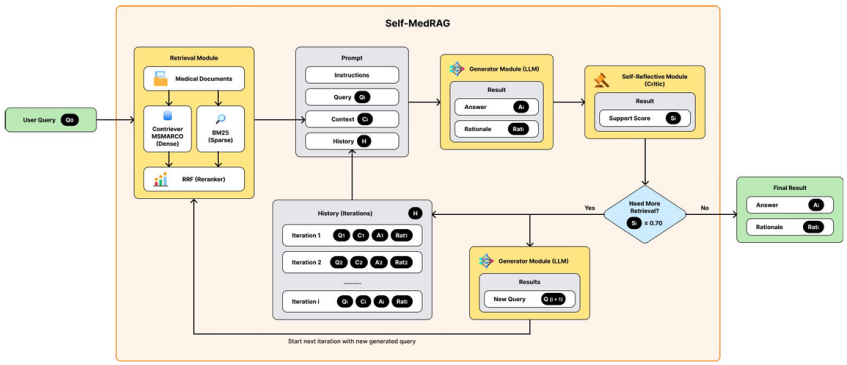

图 1. Self-MedRAG流程

Self-MedRAG 的完整推理工作流程总结于算法1中，该算法展示了检索、生成、反思和细化之间的迭代交互。

算法1 Self-MedRAG推理

要求： 初始问题 Q_0, 历史记录 H，检索器 R，生成器 G，Self-反思 SR
确保： 最终答案 A 及推理依据 Rat

```
i ← 0
当 true 时执行
  输入： Q_i
  检索上下文 C_i ← R(Q_i)
  生成答案和推理依据 (A_i, Rat_i) ← G(Q_i, C_i, H)
  计算推理依据支持分数 S_i ← 自我反思(Rat_i, C_i)
  如果 S_i ≥ 0.70 则
    返回 (A_i, Rat_i)
  else
    更新历史记录 H ← H ∪ {(Q_i, C_i, A_i, Rat_i)}
    提取不受支持的论据 U_i ← {r ∈ Rat_i | r 不受C_i支持}
    更新查询： Q_{i+1} ← Q_i ∪ U_i
    i ← i + 1
  结束 if
结束 while
```

3.2. 评估数据集

我们在两个广泛使用的医学问答基准测试集 MedQA 和 PubMedQA 的测试集中随机抽取了共 1,000 个问题，用于评估 Self-MedRAG。这些数据集为严格评估我们框架的事实准确性、证据支撑和临床推理能力提供了测试平台。

- **MedQA** 是一个医学选择题考试数据集，涵盖美国医师执照考试第一步、第二步和第三步中的临床场景。每个样例均包含一个问题、一组答案选项和正确答案 [37]。

- **PubMedQA** 是一个源自 PubMed摘要的生物医学问答数据集。每个问题都对应一篇研究摘要，并有一个“是/否/可能”的答案。我们使用测试集的人工子集 [38]，并移除了答案为“可能”的所有问题。

3.3. 检索方法

检索模块提供 Self-MedRAG 在每次迭代中使用的上下文。对于输入查询 Q_i，系统同时采用稀疏检索和稠密检索路径，以最大限度地覆盖临床相关信息。

BM25 被用作稀疏检索器，利用词汇匹配和基于词频-逆文档频率的评分，捕捉高精确率的生物医学术语和精确短语匹配。作为补充，系统集成了 Contriever-MSMARCO 作为稠密检索器，将查询和文档编码为嵌入向量，并通过点积相似度对其排序。这使系统能够检索语义上

相关、但表面措辞可能与查询没有明显共同之处的内容。稠密嵌入通过 FAISS 索引存储并高效检索。

为综合利用两种检索信号的优势，采用倒数排名融合（RRF）合并 BM25 和 Contriever 返回的候选结果。

$$RRF(d) = \sum_{i=1}^{N} \frac{1}{K+rank_i(d)} \quad (1)$$

对于每个段落d，其 RRF 融合分数按照式 (1) 计算，其中$K = 60$ 遵循标准做法。由此得到的融合排序生成检索上下文C_i，为后续的生成和验证步骤提供证据基础。

3.4. 生成器模块

生成器模块采用 DeepSeek 实现。DeepSeek 是一个大语言模型（LLM），作为 Self-MedRAG 的主要推理和答案生成组件。在每次迭代 i 中，LLM 都会接收一个结构化提示，其中包含：(1) 当前查询 Q_i；(2) 检索到的上下文 C_i；(3) 鼓励以证据为依据、保持临床谨慎性并避免无依据主张的系统指令；以及 (4) 可用时提供的推理历史 H，以保持多步推理的连贯性。

随后，DeepSeek 会生成两项输出。它会生成符合任务格式的答案 A_i，例如根据问题类型给出二选一或多选题的选项，并生成推理依据 Rat_i，解释其决策并引用检索到的上下文中的证据。该推理依据必须与C_i紧密关联，以确保模型的推理过程透明、可追溯且有临床依据。

大语言模型的可靠性和迭代稳定性使其能够在多个周期中细化推理、整合新检索到的证据，并在必要时纠正先前的假设。随后，答案与推理依据组成的配对(A_i, Rat_i)会被传递给自反思模块，由该模块判断在系统确定最终回复之前，是否需要进一步检索、查询改写或进一步细化。

3.5. 自反思模块

自反思模块评估大语言模型生成的推理依据 Rat_i 是否得到检索到的证据上下文 C_i 的充分支持。此验证决定系统能否敲定答案，还是必须启动新一轮迭代。通过将每条推理依据陈述与 C_i 中的所有段落进行比较，评估其是否得到证据支持。实验采用了两种验证机制：

- **基于 NLI 的验证：**使用 RoBERTa-large-MNLI 模型执行自然语言推理。对于每一对证据上下文和推理依据，证据段落作为前提，推理依据陈述作为假设。模型输出蕴含标签和置信度分数。对于每条陈述，系统选择蕴含置信度最高的证据段落作为其最佳支持依据。• **基于大语言模型的验证**：提示 Llama~3.1-8B 充当自然语言推理分类器，并输出蕴含、中立或矛盾。其置信度分数采用与 NLI 模型相同的方法，从概率形式的输出中解析得出。

根据某条推理依据陈述的最佳蕴含置信度是否超过验证阈值 $\tau = 0.5$，将其标记为受支持或不受支持。随后，该模块计算推理依据支持分数S_i，其定义为Rat_i中受支持陈述所占的比例。如果$S_i \geq \theta = 0.7$，系统将(A_i, Rat_i)作为最终输出返回。否则，系统会收集不受支持的推理依据元素，并用其构造优化后的查询Q_{i+1}，从而触发额外一次迭代。阈值 τ 和 θ 是根据初步验证实验选定的，以平衡检索深度和答案准确率。

4. 结果与讨论

表1：PubMedQA 和 MedQA 上的检索与评判器配置对比。

方法	检索	评判器	PubMedQA 准确率/精确匹配 F1		MedQA 准确率/精	F1
无评判器						
基础检索增强生成	MedCPT	-	65.97	62.60	39.90	40.49
基础检索增强生成	BM25	-	66.80	60.67	41.74	41.92
基础检索增强生成	Contriever	-	67.90	64.41	43.30	41.15
基础检索增强生成	BM25 + Contriever + RRF	-	69.10	64.45	80.00	79.93
使用评判器						
自我反思式 **MedRAG**	BM25 + Contriever + RRF	自然语言推理 (roberta-large-mnli)	79.82	78.40	83.30	83.30
自我-MedRAG	BM25 + Contriever + RRF	Llama3.1–8B	78.76	77.31	82.90	82.90

表1所呈现的结果展示了不同检索策略和评判器配置下的性能趋势。对于基础检索增强生成方法，在 PubMedQA 和 MedQA 数据集上，结合 BM25 与 Contriever 并使用倒数排名融合（ RRF ）的混合检索，性能均显著优于任何单一检索器。在 PubMedQA 上，BM25 和 Contriever 的准确率分别为 66.80% 和 67.90%，通过 RRF 融合后，准确率略微提升至 69.10%。在 MedQA 上，这一效果更为显著：该方法将准确率从单独使用 BM25 时的 41.74% 和单独使用 Contriever 时的 43.30%，大幅提升至 80.00%。这一显著提升表明，利用 RRF 融合检索能够整合 BM25 提供的高精确率词法信号与 Contriever 检索到的语义相关段落，从而更广泛地覆盖临床相关证据。

MedCPT检索的表现较低（在 PubMedQA 上为 65.97%，在 MedQA 上为 39.90%），表明依赖单一基于嵌入的检索器存在局限。尽管 MedCPT [39] 是在生物医学语料库上训练的，但其训练采用了对比学习目标，这往往会生成较为平滑的嵌入空间，使句子趋于聚集到宽泛的主题簇中 [40]。因此，MedCPT 擅长检索一般主题相关的段落，但可能不太擅长区分细粒度差异，例如具体的诊断标准、报告的效应量或临床结局的方向。

与 MedCPT 不同，BM25 能够进行精准的词汇匹配，而 Contriever 则通过通用领域对比学习着重实现语义对齐。将二者与倒数排名融合结合后，这些互补信号能够带来更多样的证据。因此，融合检索生成的上下文既能捕捉精确术语，也能涵盖经改述的生物医学推理，从而提升性能。

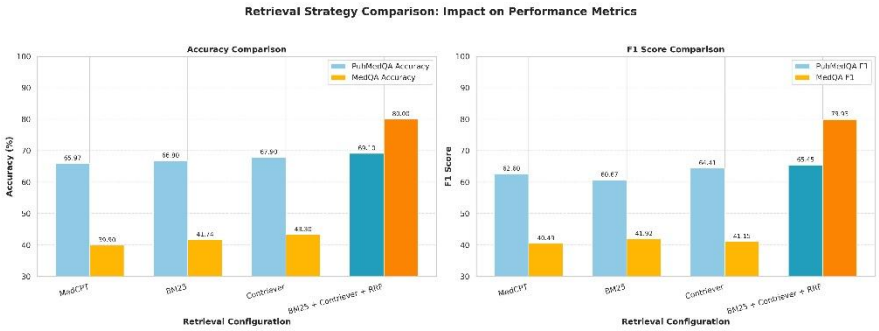

图 2. PubMedQA 和 MedQA 数据集上不同检索策略的准确率和 F1 分数比较。

为了直观地证实混合检索方法带来的性能提升，图 2 展示了 PubMedQA 和 MedQA 基准测试集上准确率和 F1 分数的比较分析。柱状图凸显了仅依赖单一检索器的局限性，例如稀疏检索器（BM25）或密集检索器（Contriever、MedCPT），尤其是在 MedQA 数据集上，单一方法的性能在准确率和 F1 分数方面均仅徘徊于 40%–43% 左右。值得注意的是，利用倒数排名融合（RRF）整合 BM25 和 Contriever 后，这一性能差距得以弥合，使 MedQA 的准确率提升至约 80%。

除了简单检索之外，我们提出的 Self-MedRAG 框架还取得了显著提升。如表1所示，加入迭代验证后，PubMedQA 数据集的准确率从 69.10%（采用 RRF 的基础检索增强生成）提升至 79.82%；使用基于自然语言推理的评判器时，MedQA 数据集的准确率则从 80.00% 提升至 83.33%。PubMedQA 上近十个百分点、MedQA 上超过三个百分点的提升，凸显了第 3.5 节所述自反思模块的作用。在多轮迭代过程中，生成器模块会生成答案和推理依据，并根据更新后的检索上下文对其进行优化，使系统能够纠正缺乏支持的假设并改进其推理。

自反思模块会评估生成器模块生成的每条推理依据是否得到检索到的证据支持，并判断是否需要进一步迭代优化。本文评估了两种验证机制：基于自然语言推理的评判器（roberta-large-mnli）和基于大语言模型的评判器（Llama 3.1–8B）。自然语言推理评判器的表现略胜于大语言模型评判器，在 PubMedQA 和 MedQA 上的准确率分别达到 79.82% 和 83.33%，而后者分别为 78.76% 和 82.90%。考虑到 roberta-large-mnli 模型经过了以蕴含为重点的

训练，其表现略优符合预期。尽管如此，这两种评判器的表现都优于无评判器、非迭代基线，表明性能提升源于迭代机制本身，而非具体选择了哪种评判器。

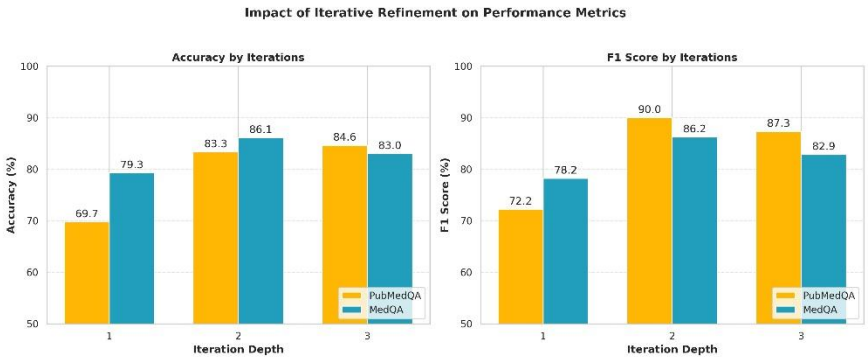

图 3. PubMedQA 和 MedQA 上迭代过程中准确率与 F1 分数的累积提升。

图 3 展示了在准确率和 F1 分数方面，针对两个数据集进行自反思模块迭代过程所产生的累积影响。我们观察到，在两个数据集上，第一次迭代到第二次迭代之间的性能均显著提升：MedQA 的准确率从 79.3% 升至 86.1%，PubMedQA 则从 69.8% 升至 83.3%。这一上升趋势证实，自反思模块在识别并纠正不受支持的论据方面具有提升性能的潜力。然而，将该过程扩展到第三次迭代似乎会导致收益递减：PubMedQA 的性能趋于平稳，而 MedQA 的性能则略有下降。

总体而言，研究结果证实了 Self-MedRAG 集成设计的有效性。混合检索能够生成更丰富的证据集，迭代推理能够增强事实依据，而基于证据的自我验证则能减少无依据主张。这些组成部分协同作用，使两个数据集上的表现均得到显著提升，尤其是在需要多步推理并综合多份生物医学文档的任务中。研究结果凸显了检索多样性、迭代优化以及推理依据层面验证的重要性，这些方法有助于克服传统检索增强生成流程在生物医学领域中的局限性。

5. 结论

我们提出了 Self-MedRAG，这是一种面向可靠生物医学问答的自反思混合检索增强生成框架。在 PubMedQA 和 MedQA 上进行的实验表明，结合 BM25 和 Contriever 并采用倒数排名融合的混合检索，性能显著优于单一检索器；与单独使用各方法相比，MedQA 准确率几乎翻倍。加入自反思模块后，性能进一步提升：PubMedQA 准确率从 69.10% 提高到 79.82%，MedQA 准确率从 80.00% 提高到 83.33%，这证明了迭代推理依据

验证在纠正无依据主张方面的有效性。基于自然语言推理的评判器和基于大语言模型的评判器均能提升结果，其中前者在两个数据集上的表现略优于后者，不过，大部分收益仍来自迭代优化本身。

总体而言，这些发现验证了将多样化检索策略与迭代式推理依据级验证相结合，能够显著增强生物医学问答能力，尤其是对于需要多步推理并综合多个来源信息的任务。它们凸显了检索多样性、自我反思式推理和基于证据的验证的重要性，这些方法有助于克服传统生物医学领域检索增强生成流程的局限性。

术语表

Q_0	初始问题/查询
Q_i	第i次迭代时的查询
H	推理历史
R	检索器模块
G	生成器模块
SR	自反思模块
A	最终答案
A_i	第i次迭代的答案
Rat	推理依据
Rat_i	第 i 次迭代的推理依据
C_i	第 i 次迭代检索到的上下文
S_i	第 i 次迭代的推理依据支持分数
U_i	第 i 次迭代的不受支持的推理依据元素
d	段落/文档
K	倒数排名融合常数（K = 60）
i	迭代索引

希腊符号

τ	蕴含置信度的验证阈值（τ = 0.5）
θ	推理依据支持分数阈值（θ = 0.7）

缩写

LLM	大语言模型
RAG	检索增强生成
QA	问答
NLI	自然语言推理
RRF	倒数排名融合
BM25	最佳匹配25（稀疏检索算法）
词频-逆文档频率	词频-逆文档频率
FAISS	Facebook AI 相似性搜索
美国医师执照考试	美国医师执照考试
MNLI	多体裁自然语言推理
准确率/精确匹配	准确率/精确匹配
F1	F1 分数（精确率和召回率的调和平均数）
MedCPT	医学对比式预训练变换器
CRAG	纠正式检索增强生成
SIM-RAG	具备充分性感知的迭代式医学检索增强生成
MRD-RAG	多轮诊断式检索增强生成

参考文献

1. Zheng, Y.; Gan, W.; Chen, Z.; Qi, Z.; Liang, Q.; and Yu, P.S.（2024年）。医学领域的大语言模型：综述。*arXiv*预印本 *arXiv:2405.13055*。检索自 https://arxiv.org/abs/2405.13055 2. Yang, X.; Li, T.; Su, Q.; Liu, Y.; Kang, C.; Lyu, Y.; 等（2024年）。大语言模型在疾病诊断与治疗中的应用。《中华医学杂志》，138(2)，130–142。 3. Yang, R.; Ning, Y.; Keppo, E.; Liu, M.; Hong, C.; Bitterman, D.S.; 等（2025年）。检索增强生成在医疗保健生成式人工智能中的应用。*Npj Health Systems*，2(1)。 4. Gupta, S.; Ranjan, R.; and Singh, S.N.（2024年）。检索增强生成（检索增强生成）的全面综述：演进、当前格局与未来方向。*arXiv*预印本 *arXiv:2410.12837*。检索自 https://arxiv.org/abs/2410.12837 5. Xiong, G.; Jin, Q.; Wang, X.; Zhang, M.; Lu, Z.; and Zhang, A.（2024年）。通过迭代式后续问题改进医学领域的检索增强生成。*arXiv*预印本 *arXiv:2408.00727*。检索自 https://arxiv.org/abs/2408.00727 6. Yang, D.; Zeng, L.; Rao, J.; and Zhang, Y.（2025年）。知之为知之：通过自我练习学习多轮检索增强生成中何时继续搜索。第*48*届国际*ACM SIGIR*信息检索研究与发展会议论文集（*SIGIR '25*），1305–1315。 7. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H.（2023年）。Self-RAG：通过自我反思学习检索、生成与批判。*arXiv*预印本 *arXiv:2310.11511*。检索自 https://arxiv.org/abs/2310.11511 8. Bora, A.; and Cuayáhuitl, H.（2024年）。基于检索增强生成的大语言模型在医疗聊天机器人应用中的系统分析。《机器学习与知识提取》，6(4)，116。 9. Dong, X.; Zhu, W.; Wang, H.; Chen, X.; Qiu, P.; Yin, R.; 等（2025年）。检索之前先讨论：智能体主导的讨论如何改进医学问答中的检索增强生成。*arXiv*预印本 *arXiv:2504.21252*。检索自 https://arxiv.org/abs/2504.21252 10. Jin, D.; Pan, E.; Oufattole, N.; Weng, W.H.; Fang, H.; and Szolovits, P.（2020年）。该患者患有什么疾病？一个来自医学考试的大规模开放领域问答数据集。*arXiv*预印本 *arXiv:2009.13081*。检索自 https://arxiv.org/abs/2009.13081 11. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X.（2019年）。PubMedQA：一个生物医学研究问答数据集。*2019*年自然语言处理经验方法会议与第*9*届国际自然语言处理联合会议论文集（*EMNLP-IJCNLP*），中国香港，2567–2577。 12. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; 等（2020年）。衡量大规模多任务语言理解能力。*arXiv*预印本 *arXiv:2009.03300*。检索自 https://arxiv.org/abs/2009.03300 13. Pal, A.; Umapathi, L.K.; and Sankarasubbu, M.（2022年）。MedMCQA：一个面向医学领域问答的大规模多学科多项选择数据集。机器学习研究论文集（*PMLR*），174。

14. Xiong, G.; Jin, Q.; Lu, Z.; and Zhang, A.（2024年）。医学领域检索增强生成的基准测试。计算语言学协会会议论文集：*ACL 2024*，泰国曼谷。 15. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; and Kang, J.（2020年）。BioBERT：用于生物医学文本挖掘的预训练生物医学语言表征模型。《生物信息学》，36(4)，1234。 16. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; 等（2021年）。面向生物医学自然语言处理的领域特定语言模型预训练。《*ACM*医疗保健计算汇刊》，3(1)，1–23。 17. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; 等（2022年）。大语言模型编码临床知识。*arXiv*预印本 *arXiv:2212.13138*。检索自 https://arxiv.org/abs/2212.13138 18. Maity, S.; and Saikia, M.J.（2025年）。医疗保健与医疗应用中的大语言模型：综述。《生物工程》，12(6)，631。 19. Wiliam, R.; Jessica, R.; Gumilang, A.I.; and Suhartono, D.（2025年）。使用基于症状的数据集进行疾病诊断的大语言模型：探索推理与非推理。*2025*年*ICORIS*国际会议论文集。 20. Li, X.; Lipp, J.; Shakir, A.; Huang, R.; and Li, J.（2024年）。BMX：熵加权相似度与语义增强词法搜索。*arXiv*预印本 *arXiv:2408.06643*。检索自 https://arxiv.org/abs/2408.06643 21. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z.（2024年）。BGE M3嵌入：通过自知识蒸馏实现多语言、多功能与多粒度文本嵌入。*arXiv*预印本 *arXiv:2402.03216*。检索自 https://arxiv.org/abs/2402.03216 22. Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; 等（2021年）。通过对比学习进行无监督稠密信息检索。*arXiv*预印本 *arXiv:2112.09118*。检索自 https://arxiv.org/abs/2112.09118 23. Wang, M.; Tan, B.; Gao, Y.; Jin, H.; Zhang, Y.; Ke, X.; 等（2025年）。平衡混合：混合搜索中权衡取舍的实验分析。*arXiv*预印本 *arXiv:2508.01405*。检索自 https://arxiv.org/abs/2508.01405 24. Verma, S.; Jiang, F.; and Xue, X.（2025年）。超越检索：结合交叉编码器、GPT重排序器与大语言模型开展生物医学问答集成。*arXiv*预印本 *arXiv:2507.05577*。检索自 https://arxiv.org/abs/2507.05577 25. Han, H.; Wang, Y.; Shomer, H.; Guo, K.; Ding, J.; Lei, Y.; 等（2024年）。结合图的检索增强生成（GraphRAG）。*arXiv*预印本 *arXiv:2501.00309*。检索自 https://arxiv.org/abs/2501.00309 26. Neha, F.; Bhati, D.; and Shukla, D.K.（2025年）。医疗保健中的检索增强生成（检索增强生成）：全面综述。人工智能，6(9)，226。 27. Zhao, X.; Liu, S.; Yang, S.Y.; and Miao, C.（2025年）。MedRAG：利用知识图谱引导的推理增强医疗保健副驾驶中的检索增强生成。*arXiv*预印本 *arXiv:2502.04413*。检索自 https://arxiv.org/abs/2502.04413 28. Wu, J.; Zhu, J.; Qi, Y.; Chen, J.; Xu, M.; Menolascina, F.; 等（2024年）。医疗图检索增强生成：通过图实现安全的医疗大语言模型

检索增强生成。*arXiv* 预印本 *arXiv:2408.04187*。获取地址：https://arxiv.org/abs/2408.04187 29. Yang, H.; Chen, H.; Guo, H.; Chen, Y.; Lin, C.S.; Hu, S.; 等（2024年）。LLM-MedQA：通过大语言模型案例研究增强医学问答。*arXiv* 预印本 *arXiv:2501.05464*。获取地址：https://arxiv.org/abs/2501.05464 30. Singh, A.; Ehtesham, A.; Kumar, S.; 和 Khoei, T.T. （ 2025 年）。智能体式检索增强生成：智能体式 RAG 综述。*arXiv* 预印本 *arXiv:2501.09136*。获取地址：https://arxiv.org/abs/2501.09136 31. Jiang, P.; Ouyang, S.; Jiao, Y.; Zhong, M.; Tian, R.; 和 Han, J.（2025年）。基于大语言模型的检索与结构化增强生成。*arXiv* 预印本 *arXiv:2509.10697*。获取地址：https://arxiv.org/abs/2509.10697 32. Li, Y.; Zhang, W.; Yang, Y.; Huang, W.C.; Wu, Y.; Luo, J.; 等（2025年）。迈向具备深度推理能力的智能体式 RAG：大语言模型中 RAG 推理系统综述。*arXiv* 预印本 *arXiv:2507.09477*。获取地址：https://arxiv.org/abs/2507.09477 33. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; 等（2023年）。大语言模型的检索增强生成：综述。*arXiv* 预印本 *arXiv:2312.10997*。获取地址：https://arxiv.org/abs/2312.10997 34. Dong, G.; Jin, J.; Li, X.; Zhu, Y.; Dou, Z.; 和 Wen, J.R.（2025年）。RAG-Critic：利用自动化批评者引导的智能体工作流实现检索增强生成。计算语言学协会第*63*届年会论文集，3551–3578。35. Yan, S.Q.; Gu, J.C.; Zhu, Y.; 和 Ling, Z.H.（2024年）。纠正性检索增强生成。*arXiv* 预印本 *arXiv:2401.15884*。获取地址：https://arxiv.org/abs/2401.15884 36. Sun, P.; Chen, Y.; Li, X.; 和 Chu, X.（2025年）。用于模拟临床推理的多轮诊断式 RAG 框架。*arXiv* 预印本 *arXiv:2504.07724*。获取地址：https://arxiv.org/abs/2504.07724 37. openlifescienceai。（2024年）。MedQA [数据集]。获取地址：https://huggingface.co/datasets/openlifescienceai/medqa/ 38. qiaojin。（2024年）。PubMedQA [数据集]。获取地址：https://huggingface.co/datasets/qiaojin/PubMedQA/ 39. Jin, Q.; Kim, W.; Chen, Q.; Comeau, D.C.; Yeganova, L.; Wilbur, W.J.; 等（2023年）。MedCPT：利用大规模 PubMed 搜索日志进行零样本生物医学信息检索的对比预训练变换器。《生物信息学》，39(11)。40. Gao, T.; Yao, X.; 和 Chen, D.（2021年）。SIMCSE：句子嵌入的简单对比学习。*arXiv* 预印本 *arXiv:2104.08821*。获取地址：https://arxiv.org/abs/2104.08821