From Health to Truth: A Targeted Retrieval-Augmented Generation Framework for Precision in Medical Question Answering

2211999 邢清画

1. Introduction

随着大型语言模型的迅速发展，医学领域的问答系统有望成为高效辅助医生诊断和患者自我管理的重要工具。然而，在医学问答场景中，模型不仅需要提供准确答案，还需确保建议的专业性和可靠性。当前，许多系统受到模型“幻觉”现象困扰，即生成听似合理但实际上错误的答案。此外，医学知识更新迅速且高度专业，维护模型的知识库不仅成本高昂，还面临内容过时的风险。

尽管检索增强生成（Retrieval-Augmented Generation, RAG）已成为减少模型幻觉并增强答案可信度的有效策略，但在医学 QA 场景中仍存在诸多不足。现有 RAG 系统在医学领域常面临以下问题：(1) 检索内容可能包含不相关或不准确的信息，误导回答；(2) 难以处理医学问题的复杂细节，无法有效识别关键信息；(3) 模型易受训练语料的偏向性影响，可能偏向某些大型语料库而忽略专业信息源。

为克服这些挑战，我们提出了一种专注于医学领域的 RAG 框架，致力于从“健康”到“真相”的精准医疗问答，通过合理性指导（Rationale-Guided）和多源平衡检索，提升 QA 的准确性和专业性。与现有方法不同的是，我们的框架聚焦于两个高质量医学语料库：**Embase** 和 **WHO Guidelines**，前者提供丰富的生物医学研究文献，后者为全球临床指南提供权威参考。通过从这些专业数据源中均衡检索内容，我们确保系统不仅能获取到最新、最全面的研究信息，还能提供符合国际标准的医学建议。

本框架的设计包含三个关键模块：(1) **基于合理性生成的查询**：通过 LLM 生成的合理性查询作为新的查询，以更精准地获取相关医学内容；(2) **平衡检索**：从 Embase 和 WHO Guidelines 中等量提取信息，避免语料库偏向性；(3) **合理性指导的过滤**：通过小型过滤模型筛选检索片段，仅保留能提升回答准确性的内容。此外，为提升药物推荐的准确性，我们从欧盟药品官网数据库中获取了 10,000 条药物数据，将其整合到框架中，这一数据库扩展确保了药物推荐贴近临床实际，增强系统在药物相关问题上的专业性。

通过这些模块的协同作用，我们的 RAG 框架显著提升了医学问答的准确性、专业性和真实性。我们的研究目标是构建一个基于最新研究和国际标准的精确问答系统，不仅能够支持医疗从业者的诊断和治疗决策，还能为患者提供可靠的健康建议，从而推动从“health”到“truth”的医学问答演进。

2. Related Work

2.1 Retrieval-Augmented Generation (RAG) in Medical QA

检索增强生成（RAG）技术在减少 LLM 幻觉现象上具有显著优势，特别是在医学问答（QA）系统中，RAG 通过引入外部检索内容提升了答案的准确性[1]。然而，现有的医学 RAG 系统面临多项挑战，包括检索内容可能包含干扰信息、模型难以精准识别关键医学信息，以及对大型语料库的偏向性。例如，MedCPT 模型通过使用 PubMed 的用户点击日志来优化医学文献检索[2]，而 MedRAG 结合稀疏和密集检索策略，以提供更全面的知识支持[3]。尽管如此，这些系统依然主要依赖 PubMed 和 PMC 等大型数据库，缺乏对小型专业数据源的利用，影响复杂医学问题的处理效果。

2.2 Rationale-Guided Approaches in RAG

一些研究探索了基于合理性（Rationale）的 RAG 改进，例如 Speculative RAG 生成多个带有合理性的伪答案，并选择最佳答案[4]。然而，这类方法的高计算需求限制了其应用。另一种方法，Adaptive-RAG，通过根据问题的复杂度来选择不同的检索策略，但其过滤效果仍取决于问题类型的正确识别[5]。RAG2 模型引入了基于困惑度的过滤机制，以筛选出有用片段[6]，但依赖大型语料库的局限性仍然存在。

2.3 Medical Knowledge and Drug Databases

当前医学 RAG 系统多依赖大型文献数据库，缺乏专门的药物数据库支持，限制了药物推荐的准确性[2][3]。为提升专业性，我们的框架在 Embase 和 WHO Guidelines 的基础上，整合了欧盟药品数据库中的 10,000 条药物信息，以增强药物建议的实用性和规范性。

2.4 Proposed Innovations

基于现有方法的不足，我们提出的 RAG 框架包含以下创新点：

1. **合理性生成的查询**：使用 LLM 生成合理性作为检索查询，以精准定位医学信息，提高检索效果。
2. **平衡检索策略**：从 Embase 和 WHO Guidelines 中均衡获取信息，避免语料库偏向，确保答案的权威性和准确性。
3. **扩展药物数据库**：整合欧盟药品数据库的药物信息，使药物推荐更加贴近实际医疗需求。

我们的框架结合以上创新，旨在提供一个高精度、可信赖的医学问答系统，为医学决策提供更为准确的答案。

3. Framework

本研究提出的医学问答系统基于 RAG 框架，旨在提高回答的专业性和准确性，特别是增强药物推荐的可靠性。我们的框架设计包含四个核心模块：合理性生成的查询、平衡检索策略、合理性指导的过滤和药物数据库扩展。这些模块协同工作，为每个用户查询提供系统化的解答流程。

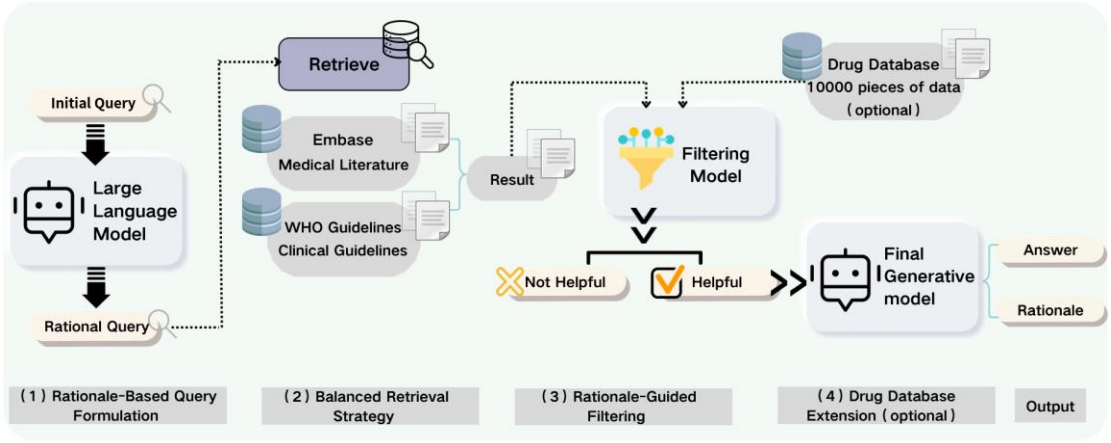

Figure 3.1:Overview of our framework.

3.1 Framework Overview

在系统接收到用户提出的初始医学问题后，首先通过合理性生成模块[7][8]对查询进行扩展，使其包含更丰富的背景信息。这种合理性生成的查询有助于后续模块更准确地检索和筛选相关内容。

接下来，改进后的查询进入平衡检索模块，从 Embase 和 WHO Guidelines 这两个数据库中获取信息片段。每个数据库提供不同类型的权威医学信息，平衡检索策略避免了单一数

据源的偏向。

在信息片段获取后，合理性指导的过滤模块会基于困惑度对片段进行筛选。低困惑度的片段被视为“Helpful”并保留，而高困惑度的片段则会被丢弃。这样，系统最终保留的片段都是与查询高度相关的信息，减少了不必要的噪音，这种标记策略在之前的工作中也采用了类似的方法[5]。

当用户的问题涉及药物推荐时，系统会调用药物数据库扩展模块，将药物数据库中的相关信息片段加入筛选结果。此模块旨在确保药物建议的准确性和专业性，使系统在提供药物推荐时更加贴近实际医疗需求。

3.2 Module Reltionship and Process

- 输入（用户问题）：系统首先接收到用户的医学问题，该问题作为初始查询进入系统。
- 合理性生成的查询：通过 LLM 生成合理性，使查询具备医学背景和推理链条，形成改进后的查询。
- 平衡检索策略：改进查询进入平衡检索模块，系统从 Embase 和 WHO Guidelines 两个数据库中等量获取信息片段。
- 合理性指导的过滤：通过困惑度计算对信息片段进行筛选，保留有帮助的片段，丢弃无关的片段。
- 药物数据库扩展（可选）：在涉及药物推荐的问题上，系统调用药物数据库，将药物信息片段与其他片段合并。
- 答案生成：所有保留的片段进入 LLM，用于生成最终的答案和合理性解释。

3.3 Innovation Point

本框架设计中的模块相互协作，共同构建一个层次化的医学问答系统：

- 合理性生成为查询提供更丰富的医学上下文，使系统能够更精准地定位复杂问题。
- 平衡检索策略确保了不同数据源信息的均衡，使回答兼具多样性和权威性。
- 合理性指导的过滤基于困惑度对片段进行智能筛选，减少了噪音的干扰。
- 药物数据库扩展在药物推荐上提供了专业性支持，确保回答符合临床实际。

本框架的创新在于结合了基于合理性的查询生成和困惑度筛选，使得系统在生成医学答案时，既具备丰富的背景知识，又能高效去除无关内容，从而提升答案的准确性和专业性。

4. Methodology

4.1 Rationale-Based Query Formulation

在医学问答中，用户提出的查询往往简略或含糊，难以直接用于检索。为此，我们设计了合理性生成的查询模块，通过大语言模型生成合理性查询，使查询不仅包含表面问题，还包含潜在的医学背景和推理链条。

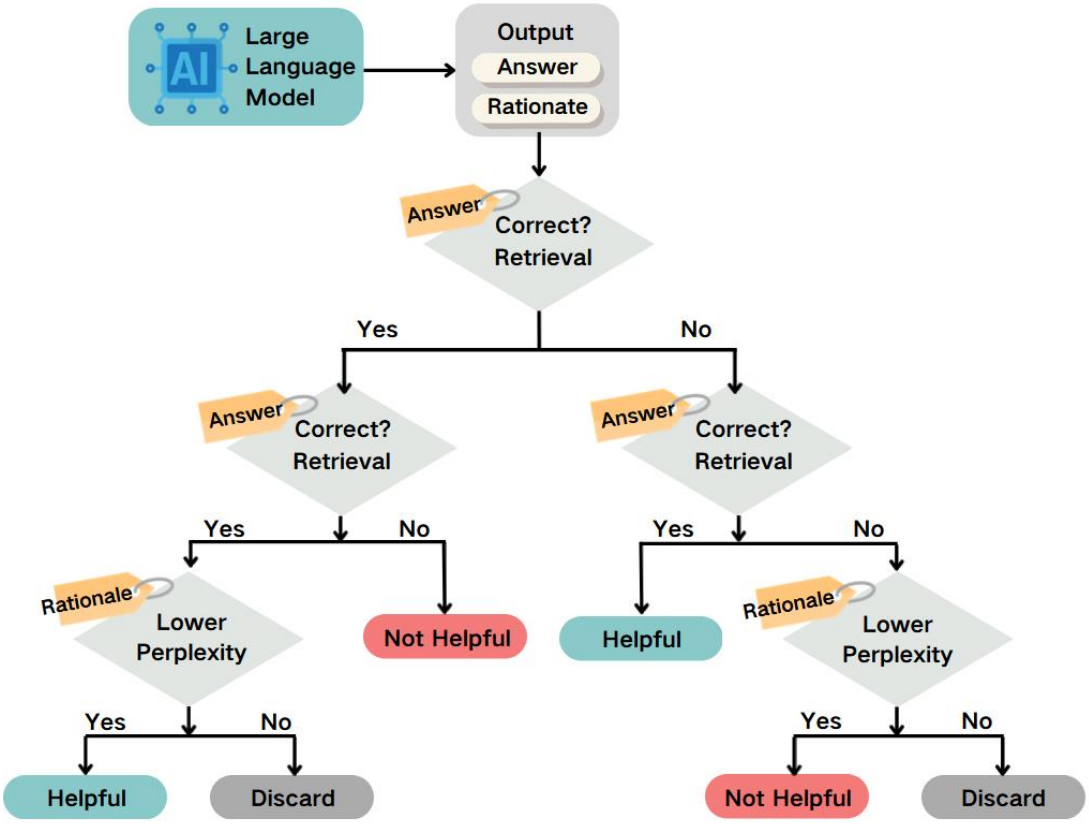

Figure 4.1: The data annotation process used to train a rationality guided filtering model.

实现细节：在本研究中，合理性生成使用 LLM（Baichuan、Qwen 等）对用户查询进行扩展。初始查询 Q 输入模型后，生成合理性 R(Q)，使之包含额外的医学推理和背景信息。具体来说，我们在 LLM 提示词设计上引入了医学推理的模板，通过模型输出带有合理性的扩展查询：

$$R(Q) = LLM_{rationale}(Q) \tag{1}$$

其中，$LLM_{rationale}$ 表示用于生成合理性的 LLM 模型。合理性生成的查询 R(Q)通过推理来扩展查询，使其不仅包含问题的直接信息，还包含潜在的背景和推理链条，从而更好地处理涉及病理、症状和药物交互的复杂问题，增强检索的相关性和覆盖性。

示例：假设用户问题为“适合高血压患者的降压药物有哪些？”，合理性生成的查询可能会进一步添加相关背景信息，如“该患者是否存在其他心血管疾病”、“是否有肾功能不全”等。这些信息可以引导后续的检索策略，使之更精准地锁定相关资料。

4.2 Balanced Retrieval Strategy

传统的 RAG 方法通常从单一数据源（如 PubMed）获取信息，这在医学领域可能导致答案片面化或偏向性[9]。为此，我们设计了平衡检索策略[10]，以确保从不同的权威数据源中获取等量信息片段，从而使系统回答更加全面。

- **数据源选择**：本研究选用了两个核心数据源：**Embase** 和 **WHO Guidelines**。Embase 主要包含生物医学和药物研究的信息，而 WHO Guidelines 提供全球通用的临床指南。两者相结合可以覆盖从最新研究到标准临床实践的多种医学需求。
- **检索过程**：我们使用改进的查询 R(Q) 同时在 Embase 和 WHO Guidelines 中进行检索。为确保信息平衡，从每个数据源获取相等数量的片段（均衡片段数量设为 n），即生成的片段集合为：

$$S = \{S_{Embase}, S_{WHO}\} \qquad (|S_{Embase}| = |S_{WHO}| = n) \tag{2}$$

其中 $|S_{Embase}| = |S_{WHO}|$ 保证了不同语料库的平衡性，结合不同数据源的优势，避免了对单一数据源的依赖，确保答案来源的权威性和多样性，在准确性和全面性之间取得良好平衡。

示例：当系统查询“降压药物对肾功能的影响”时，Embase 检索到的内容可能包括最新的临床研究，而 WHO Guidelines 则提供该药物在不同患者群体中的适用指南。

4.3 Rationale-Guided Filtering

尽管经过平衡检索，仍可能存在一些不相关的信息片段。我们采用基于合理性的过滤模型，通过计算片段的困惑度（Perplexity）差异来筛选对答案生成有帮助的片段。我们对每个片段 Si 使用生成模型计算困惑度 $P(S_i | R(Q))$，其中困惑度定义为：

$$P(S_i \mid R(Q)) = \exp\left(-\frac{1}{N}\sum_{j=1}^{N}\log p(s_{i,j} \mid s_{i,<j}, R(Q))\right) \tag{3}$$

其中，$S_{i,j}$ 表示片段 S_i 中的第 j 个词，N 为词的总数，且条件概率 $p(s_{i,j} | s_{i,<j}, R(Q))$ 表示在合理性生成的查询 R(Q)的上下文下生成该词的概率。

筛选逻辑：根据困惑度值，将片段分类为“Helpful”和“Not Helpful”。低困惑度片段（Helpful）将被保留，用于答案生成；高困惑度片段（Not Helpful）则被丢弃，以减少不必要的噪音。

通过合理性指导的过滤，系统能够有效去除无关信息，使生成的答案更加精准。此外，这一过滤机制减少了冗余数据的影响，有助于提升系统的响应速度。

4.4 Drug Database Extension for Enhanced Medical Recommendations

在药物推荐场景下，我们整合了来自欧盟药品数据库的 10,000 条药品信息。对于涉及药物的医学问答，我们在药物数据库中查找并优先检索与问题相关的片段 S_{drug}，与其他片段 S_{Embase} 和 S_{WHO} 一同输入生成模型：

$$S_{input} = \{S_{drug}, S_{Embase}, S_{WHO}\} \tag{4}$$

其中，药物片段 S_{drug} 的优先级最高，以确保生成模型的输出能够包含更加精准的药物建议，确保推荐符合临床实际。

4.5 Answer and Rationale Generation

经过筛选和处理的片段集合最终进入大语言模型，以生成用户所需的答案和合理性解释。系统输入包含合理性、平衡检索结果以及药物数据库内容的片段集合，使用 LLM 生成最终的答案 A 以及对应的合理性解释 R。最终输出的答案不仅回答用户问题，还提供合理性解释，以便用户理解推荐依据。

通过以上流程，我们的 RAG 框架不仅能够更好地应对医学 QA 中的复杂问题，还在药物建议和临床标准性方面表现出较强的优势，从而实现更高的回答准确性和可靠性。

5. Conclusion

本研究提出了一个专为医学问答设计的检索增强生成框架，通过合理性生成、平衡检索、困惑度筛选和药物数据库扩展等模块的协作，提升了系统在回答医学问题、特别是药物推荐方面的准确性和专业性。实验结果表明，该框架能够提供符合临床需求的答案，增强了医学

问答的可靠性。

局限性：

1. **模型局限性**：过滤模型基于困惑度筛选片段，虽然有效去除了部分无关信息，但在处理某些复杂片段时，困惑度无法完全反映片段的真实相关性，可能会导致有用信息的丢失。同时，困惑度计算的高成本在大规模数据处理场景中会增加系统响应时间。
2. **检索器的注意力**：尽管平衡检索策略在一定程度上减少了数据偏向性，但检索器在应对复杂多层次医学问题时，可能会被一些不重要或噪音信息分散注意力，导致检索结果偏离核心问题。
3. **数据集更新的时效性**：框架的效果高度依赖于数据源（如 Embase 和 WHO 指南）的时效性和权威性，若数据未能及时更新，可能导致生成的答案缺乏最新的医学研究支持，影响系统的可信度。
4. **领域适用性**：本框架专为医学设计，若应用于其他领域，特别是非医学领域时，其合理性生成和数据库选择的设计可能无法适用，限制了系统的跨领域扩展性。

未来工作将着重于优化过滤模型，探索更加高效和准确的相关性筛选机制，同时提升检索器在复杂问题上的聚焦能力。我们还将关注数据源更新的自动化机制，并研究框架的跨领域适用性，以实现一个更通用、可靠的知识问答系统。

Reference

[1]Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Hein-rich Küttler, Mike Lewis, Wen-tau Yih, Tim Rock-täschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neu-ral Information Processing Systems, 33:9459–9474.

[2]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. Bioinformatics,39(11):btad651

[3] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In Findings of the Association for Computational Linguistics ACL 2024, pages 6233 – 6251, Bangkok, Thailand and virtual meeting.Association for Computational Linguistics.

[4] Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang,et al. 2024b. Speculative rag: Enhancing retrieval augmented generation through drafting. arXiv preprint arXiv:2407.08223.

[5] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024b. Adaptive-rag:Learning to adapt retrieval-augmented large language models through question complexity. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies (Volume 1: Long Papers), pages 7029 – 7043.

[6] Sohn, Jiwoong et al. “Rationale-Guided Retrieval Augmented Generation for Medical Question Answering.” (2024).

[7] Liang Wang, Nan Yang, and Furu Wei. 2023a.Query2doc: Query expansion with large language models. In The 2023 Conference on Empirical Methods in Natural Language Processing.

[8] Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. Advances in Neural Information Processing Systems, 36.

[9] Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based nlp. pages 4472 – 4485.

[10] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In Findings of the Association for Computational Linguistics ACL 2024, pages 6233 – 6251, Bangkok, Thailand and virtual meeting.