Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Management Engineering—Article

面向大语言模型驱动的医学知识检索与问答系统——框架设计与评估

刘宇炀[a,#]，李晓瑛[a,#]，罗妍[a]，杜晋华[b]，张颖[a]，吕婷钰[a]，尹浩[b]，唐小利[a,*]，刘辉[a,*]

[a] *Institute of Medical Information, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100020, China*
[b] *Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China*

ARTICLE INFO

Article history:
Received 17 June 2024
Revised 5 February 2025
Accepted 19 February 2025
Available online 26 February 2025

关键词
大语言模型
医学知识
信息检索
向量数据库

摘要

随着大语言模型（LLM）的快速发展，其处理自然语言文本的能力显著提升，为医学知识发现开辟了新路径。本文提出一种由垂域LLM驱动的医学知识检索与问答（ERQA）框架。该框架通过集成语义向量数据库与文献库，并进行医学领域的增量预训练与监督微调，以完成检索与问答任务。在新冠病毒肺炎（COVID-19）疫情与TripClick数据集上的评测显示，ERQA在多项任务中表现优异。在COVID-19数据集上，ERQA-13B的检索指标达到最优水平，具体为归一化折损累计增益NDCG@10 = 0.297、召回率Recall@10 = 0.347、平均倒数排名（MRR）= 0.370；在摘要生成任务和问答任务的性能方面也高于基线模型（ROUGE-1 = 0.434，BLEU-1 = 7.851）。在TripClick数据集上获得的结果进一步表明，ERQA对多样医学主题具有良好的适应性，是迈向高效生物医学知识检索与问答的重要尝试。

1. 引言

OpenAI公司于2022年11月发布了聊天生成预训练转换器（ChatGPT），随后于2023年3月推出了GPT-4，展示了大语言模型（LLM）的广泛应用场景[1–2]。经过在包含数十亿个token的超大规模语料上训练，LLM展现出令人瞩目的文本生成与理解能力，某些方面可与人类水平相当[3]。LLM不仅重塑了公众的创意写作与技术写作实践，也在多个科学领域取得了当前领先的表现。以“large language models”或“ChatGPT”为关键词在Web of Science中检索，至2023年11月底共返回105 722篇文献，覆盖工程学、计算机科学与医学等主要主题。

LLM本质上是统计模型，通过条件概率建模预测词序列，其在多项自然语言处理（NLP）任务中持续刷新性能指标，也进一步推动了生物医学垂域LLM的发展。代表性模型包括BioMedLM [4]、BioGPT [5]与PMC-Llama [6]等；这些模型在PubMed引文与全文等特定数据集上训练或微调，以提升其在生物医学应用中的实用性。以BioGPT为例，其依托大规模医学文献，为研究人员和医疗从业者提供信息抽取与决策辅助的工具。上述模型体现了LLM在生物医学研究中的前沿进展，一方面能够提升数据分析效率与准确性，另一方面为信息抽取[7]、医疗服务与医学教育[8]等场景提供了新的可能性。

尽管如此，LLM仍可能出现“幻觉”与“虚构”等问题。在要求高准确性与高可靠性的场景（如医疗决策辅助等）中，这些问题尤需重视。生物医学LLM产生的不

* Corresponding authors.
E-mail address: tang.xiaoli@imicams.ac.cn (X. Tang), liuhui@pumc.edu.cn (H. Liu).
# These authors contributed equally to this work.

英文原文：Engineering 2025, 50(7): 270–282
引用本文：Yuyang Liu, Xiaoying Li, Yan Luo, Jinhua Du, Ying Zhang, Tingyu Lv, Hao Yin, Xiaoli Tang, Hui Liu. Toward a Large Language Model-Driven Medical Knowledge Retrieval and QA System: Framework Design and Evaluation. *Engineering*, https://doi.org/10.1016/j.eng.2025.02.010

准确或带偏见回答可能导致延误最佳治疗时机，造成心理或生理伤害，甚至危及生命。因此，确保由生物医学LLM生成的回答经过严格验证并具备充分透明性至关重要[9]。相关可行策略包括提升训练数据质量并为模型推断提供充分的证据支持等。

鉴于LLM在NLP任务中的表现，其具有作为医学知识检索与问答（QA）系统的核心引擎的潜力。本文通过集成外部数据库，构建一种医学知识检索与QA框架，并从定性与定量两个维度开展系统评估。

2. 国内外研究现状

LLM在生物医学与医疗健康领域展现出广阔前景，涵盖临床决策支持、医学教育等多种应用。近期研究持续扩展其支撑能力，例如，将结构化数据或外部知识注入模型的数据增强型LLM可用于基于放射科报告推断肿瘤治疗反应等应用[10]。鉴于本文聚焦于医学知识检索与QA，下文主要回顾与该领域直接相关的NLP研究。

2.1. 知识抽取

知识抽取是NLP的核心任务，旨在将非结构化文本转化为结构化知识，主要包含两类子任务：命名实体识别（NER）与关系抽取（RE）。早期方法高度依赖手工特征与规则系统，随着深度学习与Transformer模型的发展，相关模型性能取得了显著提升。

Mayo诊所研发的临床文本分析与知识抽取系统（cTAKES）[11]在医学信息抽取研究中具有奠基意义。该开源框架采用模块化架构解析临床文本，将机器学习与基于规则的方法结合用于抽取任务。与此类似，知识引导的远程监督（KGDS）[12]通过引入生物医学知识，增强了系统从电子病历中抽取关系的能力。当文本中的实体难以与标准知识库对齐时，这一方法更为有效。

近期相关研究将医学知识直接编码进预训练语言模型。例如，Roy和Pan [13]将统一医学语言系统（UMLS）概念融入BERT的嵌入表示，进一步提升模型对复杂医学关系的理解与抽取能力。由美国国家医学图书馆研发的语义知识库（SemRep）同样利用基于UMLS的规则来捕获生物医学文本中的关系[14]。但美国国家医学图书馆的严格评估显示，SemRep的精确率仅为0.55、F1值为0.42，其中生物医学实体识别与规范化对其错误率贡献显著[15]。此外，也已提出面向中文医学文本的知识增强医学关系抽取（KeMRE）[16]等专用框架，通过引入源自临床指南的知识嵌入，能够增强基于BERT的RE流程。

LLM已被广泛用于评估生物医学NER与RE任务在基准数据集上的性能。例如，在BC5CDR-chemical数据集（化学-疾病RE的基准语料）上，GPT-3与GPT-4分别获得了0.73与0.82的F1分数。尽管LLM功能强大，将其在领域特定知识上进行微调仍是确保语境精准抽取的关键。

2.2. 信息检索

传统的信息检索（IR）模型（如Okapi BM25 [17]）以词频-逆文档频率（TF-IDF）为基础实现文档排序，为精准检索奠定了技术基础。这类早期模型至今仍是许多IR系统的骨干组件。

随后，面向医学场景的专用IR系统不断演进。代表性模型包括MedSearch [18]，其通过对冗长医学查询的适配，强化了传统医学网页内容的搜索能力，其中的查询重写与结果多样化技术对不熟悉复杂医学术语的用户更为友好。与此同时，采用医学主题词表（MeSH）的PubMed [19]已成为生物医学文献检索的“金标准”，帮助医学专业人员获取循证实践所需的最新进展[20]。近年的医学IR研究重点转向融合人工智能，以更好地处理复杂查询与同义词问题[21–22]。例如，提出了一种基于NLP的关键词增强与筛选方法以帮助科研人员优化检索方式[23]。该方法利用先验知识从初始检索题名与摘要中抽取有意义的候选关键词，并已在房颤主题研究中显示出有效性。Jin等[24]开发了可插拔式生物医学文献检索模块，将点击日志融入稠密检索模型，从而在相关查询的基础上获得改进的检索效果。

尽管传统方法在结构化检索方面表现出色，但在应对复杂医学查询的语境与语义细粒度方面仍显不足。LLM的引入为医学IR带来新的可能性，凭借更强的上下文理解能力，能够显著提升用户体验。通过将经典IR模型的稳健性与LLM的高级语义能力相结合，新一代IR系统在医学检索场景中正日益体现出更高的有效性与易用性。

2.3. QA系统

基于序列到序列的神经网络模型[25]能够有效提升生成具备语境感知的回答的能力。传统的检索式QA系统多依赖预定义模板[26]或搜索算法[27]，从结构化数据库或非结构化文本语料中抽取答案。近年来，将知识图谱融入神经对话系统被证明可显著提升医学对话场景中的语义理解与回答准确性[28]。类似地，常识知识感知对话生成模型（ConKADI）[29]与信道感知知识融合网络（CAKF）[30]等模型利用外部医学知识能够支持具备个性化逻辑的推理过程，从而提升医学QA的质量。

Transformer架构的发展进一步改变了医学QA领域

[31–32]。研究者构建了多种生物医学QA数据集，用于对模型进行严格评测，其中包括基于美国医师执照考试的MedQA与基于PubMed引文摘要的PubMedQA [33]。这些数据集推动了面向QA的LLM发展。值得注意的是，GPT-4与Med-PaLM 2 [34]在MedQA上分别达到86.1与86.5的准确率分数，与人类专家平均87.0的水平相当。在PubMedQA上，BioMedLM、BioGPT、Med-PaLM 2分别达到74.4、81.0、81.8的分数。尽管这些LLM在标准QA数据集上的表现已可比拟人类，但在有效应对实际生物医学问题方面，仍须开展更为全面的评估。

文本摘要被视为QA任务的一个子类，其研究可追溯至20世纪50年代，主要分为抽取式与总结式两大路径[35]。抽取式摘要通常基于TF-IDF等统计方法进行词项加权[36]或基于图模型对句子重要性进行排序[37]；总结式摘要则在理解文本主旨的基础上重述关键信息，以形成更简洁清晰的概要[38]。近年，基于Transformer的模型在总结式方法中占据主导地位。例如，GPT-4已被广泛用于医学文献综述，将冗长论文压缩为精炼摘要，便于研究人员快速获取主要结论[39–40]。此外，总结式技术也可用于临床病历笔记生成[41]与诊断报告抽取[42]等任务。

3. 数据与方法

3.1. 数据收集

本研究所使用数据以PubMed平台检索与新冠病毒肺炎疫情（COVID-19）相关的文献为主要来源。本文使用“novel coronavirus”“2019-nCoV”“COVID-19 virus”“SARS-CoV-2”和“SARS2”等关键词进行检索，过滤无关内容并确保所获文献与研究目标直接相关。为二次复核与满足模型训练需要，收集过程纳入了题名、作者、摘要、关键词、MeSH与数字对象唯一标识符（DOI）等核心元数据。所收集文献按主题可划分为七类：机制、传播、诊断、治疗、预防、病例报告与预测（表1）。为保证数据的准确性与可靠性，本文在筛选过程中充分考虑了各研究可能存在的局限与偏倚，最终得到包含426 541篇文献的数据集。

为进一步降低数据集偏倚，本文纳入公开数据集TripClick [43]作为附加基准。TripClick源自Trip Database健康检索引擎的用户交互日志，覆盖2013—2020年约520万次用户交互，并配套IR评测基准与相关元数据，用以支持深度学习IR模型的训练与评估。

在使用LLM进行语义检索与微调阶段，本文对文献进行最小化预处理，主要包含两步：①将文本统一为unicode transformation format-8-bit（UTF-8）编码，并清理乱码或非法字符；②将文章内容组织为分层JavaScript Object Notation（JSON）结构。随后由两名标注员开展复核以确认数据的准确与完整；如出现分歧，则由第三名审校者裁决。复核流程分为以下四个阶段。

• 相关性复核：标注员独立评阅各文献与研究主题的契合度，基于关键术语判定其相关性，并核对其是否匹配预设类别。

• 重复性复核：针对大规模PubMed检索中可能存在的冗余记录（如同一论文的多个版本）清除所有重复条目。

• 完整性复核：核查每篇文献是否包含必要元数据（如题名、作者、摘要、MeSH与DOI等），并检查预处理（如UTF-8编码、特殊/非法字符移除）是否正确；对缺失项通过补充来源加以完善。

• 可重复性复核：为评估标注一致性，在设定时间间隔后，随机抽取10%已复核文献由同一组标注员进行复标。

3.2. 框架设计

本文提出一种由垂域LLM驱动的医学文献检索与问答（ERQA）框架。如图1所示，ERQA将垂域LLM、文献数据库与语义向量数据库整合为一体，形成直接关联医学文献的知识获取方案。

垂域LLM以Llama2 [44]为基础，采用增量预训练与

表1 增量预训练所用数据集统计

Data source	Article type	Total	Training set	Validation set	Testing set
Novel coronavirus	Mechanism	62 170	49 739	6 217	6 214
	Transmission	22 543	18 034	2 254	2 255
	Diagnosis	65 295	52 237	6 529	6 529
	Treatment	101 943	81 553	10 194	10 196
	Prevention	109 874	87 899	10 987	10 988
	Case report	48 543	38 835	4 854	4 854
	Forecasting	16 173	12 937	1 617	1 619
TripClick	–	84 736	67 789	8 473	8 474

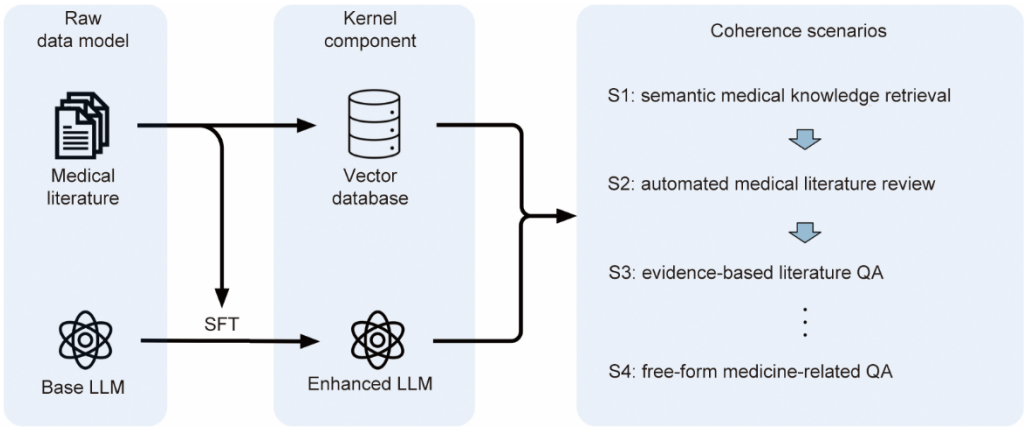

图1. 由垂域LLM驱动的ERQA框架。SFT：监督微调。

指令微调相结合的流程。Llama 2是在多样化通用语料上进行训练得到的，在医学文献检索与QA方面对细粒度语义的支撑能力有限。本文基于生物医学文本进行增量预训练，可在保持原有生成能力的同时，逐步注入领域知识[45]；随后基于提示词工程进行微调，进一步使模型能够处理问题分类、问题重构、摘要生成与基于文献的QA等任务，并约束输出质量。

文献数据库是各类学术著作的综合资料库，既保证原始文本内容的完整性与可获取性，又以篇章级为基本单元保存题名、作者、机构、摘要、关键词及结构化正文等元数据，便于在执行语义检索后回溯原文。

语义向量数据库用于支持语义检索，以段落级存储文本嵌入[46]。指定文本首先经由垂域LLM处理，取最后一层Transformer的输出生成查询向量，并利用近似最近邻检索实现高效召回[47–48]。在离线阶段，采用*K*-means将全部嵌入预聚类为若干子区域，并构建倒排文件以加速匹配；在线检索时，先在子区域中心上进行初始相似度计算，再在候选子区域内进行二次匹配，从而避免全量穷举的计算开销。通过嵌入向量与唯一文章标识之间的映射，实现由向量匹配至可读文本的有效关联。

上述组件的集成为ERQA提供了一种新的医学知识检索方案，其工作流程如图2所示。研究者首先提出问题，如“2021年之后关于新型冠状病毒（SARS-CoV-2）如何调控宿主免疫应答的研究热点是什么?”，并被ERQA模型判定为文献检索类问题；当问题包含特定书目信息约束（如发表时间、作者、机构）时，垂域LLM会识别并提取约束信息（如2021—2024年），并据此对问题进行规范化重写（如“SARS-CoV-2如何调控宿主免疫应答?”），随后将重写后的问题交由语义向量数据库处理。

系统检索出满足约束条件的前*N*个相关语义向量，并返回与之关联的唯一标识以链接到文献数据库中的完整著录信息。进一步，垂域LLM基于前*N*篇文献的题名与摘要生成精简摘要并以列表形式呈现。如研究者需要更细节的信息，可提出针对性追问（如“如何通过T细胞检测方法判定对SARS-CoV-2感染的交叉免疫应答?”）。当该追问对应某一检索结果（如DOI：10.1038/s41467-021-21856-3）时，系统将其转换为指令化问题（如“基于题为‘……’的文章信息，请回答‘……’”），据此执行检索增强生成以缓解LLM常见的“幻觉”问题。

3.3. 实现细节

垂域LLM基于基础模型Llama 2 [44]，通过增量预训练和微调完成知识增强。Llama 2采用32层解码器结构，使用均方根层归一化（RMSNorm）替代层归一化（LayerNorm），在注意力机制中具体使用的是基于分组查询注意力（GQA）的多头注意力，并通过旋转位置编码实现位置表征。在包含2万亿token、上下文窗口4096的语料上训练后，本文选择Llama-7B（70亿参数）与Llama-13B（130亿参数）作为ERQA的基础模型。

在增量预训练阶段，本文使用字节对编码对收集到的医学文本进行分词，使“angiotensin-converting enzymes”等复杂术语可被拆解为具有语义的子词单元，以便于模型高效处理医学术语。预训练以下一词预测为目标，在无监督设定下为模型注入领域知识。经过对生物医学文本的大量学习，模型学习到文献中的重要表述关系。例如，模型可以识别到“ACE抑制剂”与“高血压”的关联，或“PCR检测”与“COVID-19诊断”的关联，有助于生成更为精确的医学回答。每个epoch结束后，在从COVID-19和TripClick数据集中选取的留出验证集上评估性能。

微调阶段旨在对齐模型与医学知识检索及QA的关键任务，包括问题分类、问题重构、摘要生成与基于文献的QA（第3.2节和图2），部分微调提示词示例见表2。在问

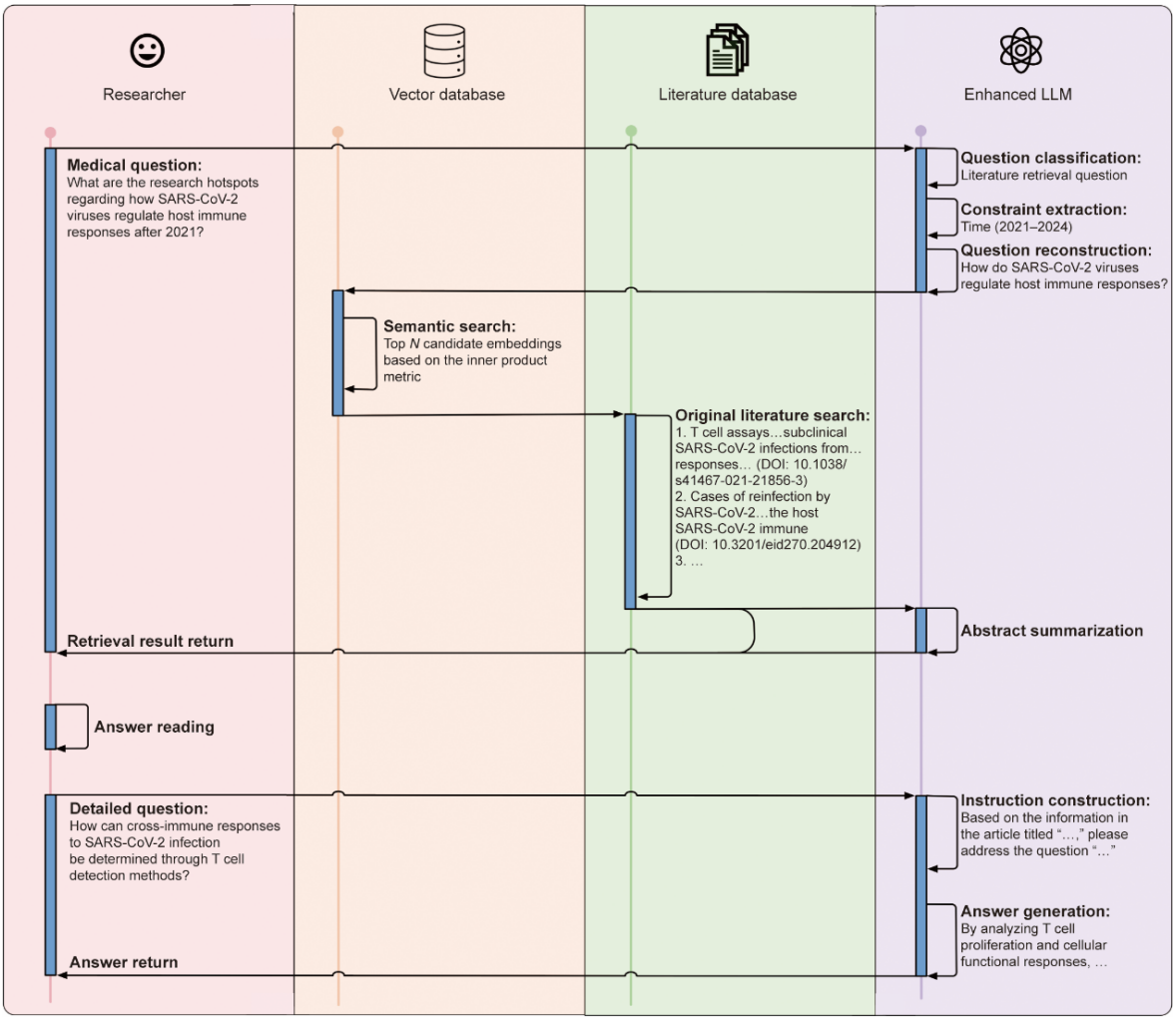

图2. ERQA框架的工作流程。

表2 不同微调任务的提示词示例

Type	Prompt examples
Question classification	1. [QUESTION] + Is the given query a literature retrieval problem or a free-format problem? 2. [QUESTION] + Identify which category the discussed query falls under: literature retrieval problem or free-format problem? 3. [QUESTION] + Determine if the provided query can be recognized as a literature retrieval problem or a free-format problem?
Question reconstruction	1. Explore the seminal contributions of the [AUTHOR] in the field of [QUESTION] with a focus on groundbreaking findings between [2010] and [2022]. → [QUESTION]; [AUTHOR]; [2010–2022] 2. Locate the primary research findings related to [QUESTION] by [AUTHOR], examining publications dated between [2017] and [2022] across esteemed academic journals including [NATURE], [CELL], and [Sciences]. → [QUESTION]; [AUTHOR]; [2017–2022]; [NATURE, CELL, SCIENCE]
Abstract summarization	1. Condense the core content of the article [TITLE] into a brief summary, drawing from the information provided in its abstract to convey the main research results as articulated in the [ABSTRACT] 2. Provide a brief encapsulation of the article [TITLE] using the information available in its abstract, distilling the main points and research outcomes articulated 3. Summarize the article [TITLE] by extracting the key details from its abstract as outlined in the [ABSTRACT]
Literature-based QA	1. Answer [QUESTION] by drawing insights from the [ABSTRACT] of the article [TITLE], extracting relevant information and addressing the specific inquiry posed 2. Respond to [QUESTION] utilizing details from the [ABSTRACT] of the article [TITLE], synthesizing key information and providing a relevant and concise answer 3. Provide an answer to [QUESTION] based on the content outlined in the [ABSTRACT] of the article [TITLE], incorporating pertinent details to address the specific query

题分类和问题重构模块中，本文以医学研究人员为目标用户，收集贴近实际的信息检索查询，并移除发表日期、作者、机构等限制以重构标准化检索问题；摘要生成模块借鉴既有研究[49–50]，以引用原文献的文献作为候选摘要来源，并进行人工复核；基于文献的QA模块主要依靠人工阅读摘要与PubMedQA中所使用的方法[33]。微调过程中采用低秩适配（LORA），冻结基础模型参数，仅更新低秩矩阵[51]。在多种秩值实验后，本文选取$r = 4$以兼顾计算开销与微调精度，并设定缩放因子$\alpha = 16$以调节低秩分量的影响。该策略在保留基础LLM通用语言能力的同时，能够高效适配领域任务。

本文按照8∶1∶1的比例将数据集划分为训练集、验证集与测试集（表1和表3）。为提升医学QA任务输出的可靠性，模型推理温度设定为0.15，因为较低温度可降低输出响应所表现出的随机性。同时结合累计概率为0.9的前P个采样，在覆盖度与“幻觉”抑制之间取得平衡。

在模型的训练与更新过程中，使用AdamW（weighted adaptive moment estimation）优化器，权重衰减为0.01，β系数为(0.9, 0.999)。学习率调度包含1500步预热，之后衰减至最大学习率的10%。学习率在预训练阶段为1×10^{-4}、微调阶段为1×10^{-5}，对应批大小分别为64与32。为防止过拟合，训练过程采用早停策略，若验证集损失连续5个epoch无改善，则停止训练。训练共使用6块NVIDIA A100图形处理单元（GPU），最长训练时长达1440 h。图3展示了微调1000次迭代过程中损失函数与困惑度的变化趋势。

4. 评估与讨论

为贴近实际应用场景，本文从文献检索、摘要生成与基于文献的QA三方面对所提框架进行性能评估。

4.1. 基线模型

为评估ERQA框架的生成性能，本文选取多种语义基线模型进行对比，包括BERT [52]、BioBERT [53]、BioClinicalBERT [54]以及一些较新的医学LLM（如BioMedLM [4]、Meditron-7B [55]和ChatDoctor [56]），以实现全面评价。

BERT是具有1.1亿参数的双向Transformer模型，通过同时考虑句子左、右侧上下文来捕获远距离依赖。BioBERT与BioClinicalBERT分别在医学文献与临床病例数据上进行微调，以提升其在医学NLP任务中的表现。

BioMedLM是专用于生物医学领域的LLM，仅在生物医学摘要与论文上训练，采用标准的Transformer堆叠架构，上下文窗口长度为1024、隐藏维度为2560，其在多类生物医学NLP应用中取得了较为优异的结果。

表3 微调数据统计

Data source	Prompt type	Total	Training set	Validation set	Testing set
Shared	Question classification	1 278	1 022	127	128
	Question reconstruction	4 932	3 945	493	495
Novel coronavirus	Abstract summarization	7 419	5 935	741	741
	Literature-based QA	13 678	10 942	1 367	1 367
TripClick	Abstract summarization	5 492	4 393	550	549
	Literature-based QA	4 328	3 462	433	433

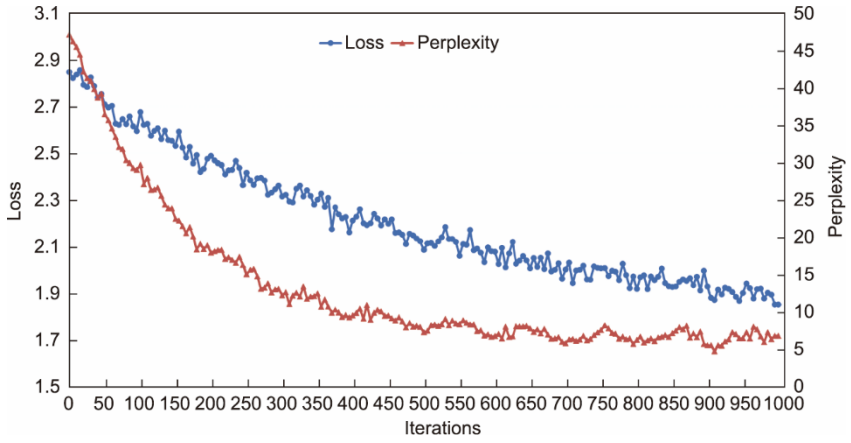

图3. ERQA模型微调过程的损失函数与困惑度变化。

Meditron-7B 则是在 Llama 2 基础上构建的更通用医学 LLM，使用 NVIDIA Megatron-LM 分布式训练器进行训练，其大规模训练过程使其能够有效处理多种医学推理任务。ChatDoctor 主要应用于医患对话场景，基于 Llama 架构在大规模医患交互数据上完成微调。ChatDoctor 同时结合了经过人工整理的离线知识库与外部资源（如 Wikipedia）进行实时检索增强，以更好地应对真实临床查询。为确保在所收集数据集上的公平对比，以上对比模型的所有超参数均在相关数据集上通过网格搜索进行调优。

4.2. 文献检索

如 3.2 节所述，向量数据库为语义检索提供了可行路径，其中向量嵌入对最大化检索性能至关重要。本节评估不同文献检索模型的嵌入表示所带来的影响。对于 COVID-19 数据集，本文采用包括文章类别与人工反馈在内的两类金标准；对于 TripClick 数据集，采用来自点击日志的记录作为金标准。所选模型均以归一化折损累计增益（NDCG）、召回率（Recall）、平均倒数排名（MRR）等指标进行评价。

在 COVID-19 数据集上使用文章类别作为金标准评估检索表现时，ERQA-7B 与 Meditron 表现相当，且均显著优于 ChatDoctor 与基于 BERT 的模型。如图 4 所示，ERQA-7B 的 NDCG@10 指标（0.897）略低于 Meditron（0.899），但高于 ChatDoctor（0.893）和 BioMedLM（0.885）。Recall@10 指标呈现相同趋势，ERQA-7B 的 Recall@10 指标为 0.906，略低于 Meditron（0.907），并优于 ChatDoctor（0.902）和 BioMedLM（0.894）。上述结果表明，在 COVID-19 数据集上，ERQA-7B 与 Meditron 的检索准确性更高，较 BERT 系模型具有明显优势。此外，本文在该数据集上以人工反馈为真值进行额外评估。如图 5（a）所示，ERQA-7B（NDCG@10 = 0.264、Recall@10 = 0.289）再次优于 ChatDoctor（NDCG@10 = 0.257、Recall@10 = 0.279）。Meditron 的数值（NDCG@10 = 0.261、Recall@10 = 0.287）略低于 ERQA-7B，两者整体表现接近。BERT 系模型（如 BioClinicalBERT）结果（NDCG@10 = 0.221、Recall@10 = 0.237）明显偏弱。上述结果表明，当以人工反馈为评价准则时，小规模模型存在局限。

在 TripClick 数据集上[图 5（b）]，ERQA-7B（NDGG@10 = 0.337, Recall@10 = 0.279）表现稳定，与 Meditron（NDCG@10 = 0.341、Recall@10 = 0.276）基本相当，且领先于 ChatDoctor（NDCG@10 = 0.332、Recall@10 = 0.272）。ERQA-13B 在 TripClick 数据集上取得最显著提升，其 NDCG@20 指标为 0.428、Recall@50 指标为 0.391，明显高于次优模型 ERQA-7B（NDCG@20 = 0.348、Recall@50 = 0.376）。该结果表明，更大规模模型与更精细的微调在应对多样且噪声较高的真实检索数据时更具优势。

4.3. 摘要生成

摘要生成作为文本摘要的一个细分方向，在科学与医学研究语境下对内容准确性要求较高。为评估本模型及对比基线的性能，本文采用 ROUGE 指标（包括 ROUGE-1、ROUGE-2、ROUGE-L）这些指标通过与参考摘要的重叠程度进行度量，其中 ROUGE-1 关注一元重叠，ROUGE-2 关注二元重叠，ROUGE-L 关注最长公共子序列[35]。

在 COVID-19 数据集上的结果[图 6（a）]显示，LLM 系模型整体优于 BERT 系。其中，ERQA-13B 在各项指标上均为最优，其 ROUGE-1 指标为 0.434，而 ERQA-7B 的 ROUGE-1 指标为 0.420。相较传统模型，ERQA-13B 在 ROUGE-1 上较 BERT、BioBERT、BioClinicalBERT 分别

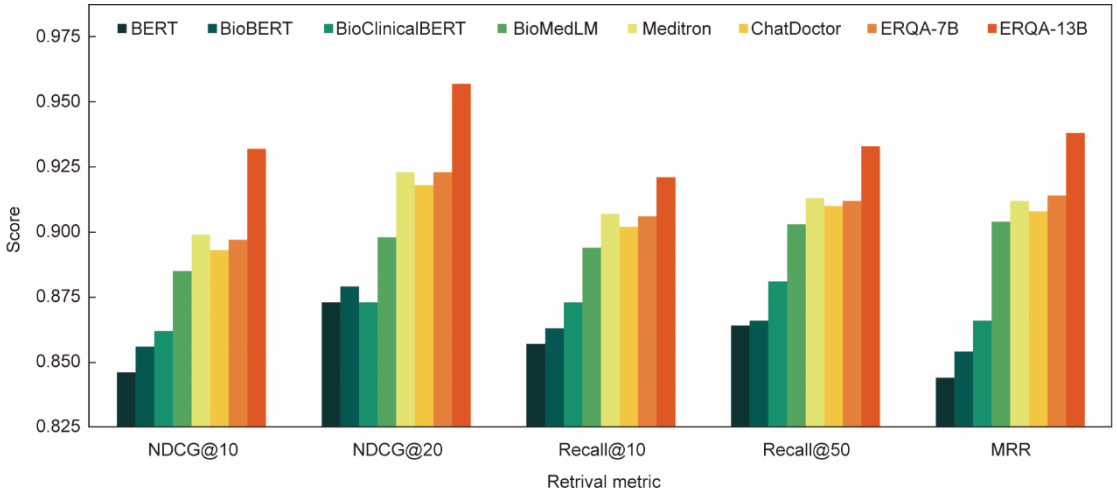

图 4. COVID-19 数据集上以类别为真值的检索性能。

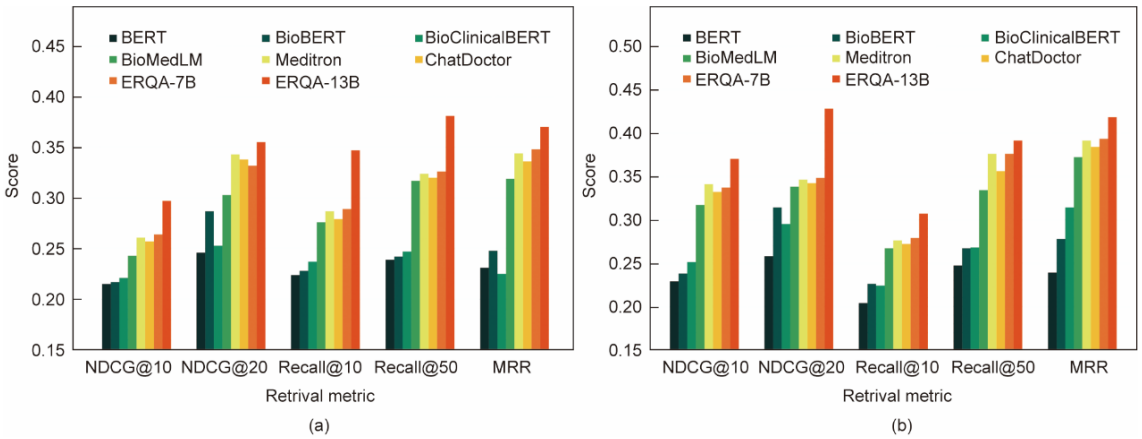

图5. COVID-19与TripClick数据集上以人工反馈为真值的检索性能。（a）COVID-19数据集；（b）TripClick数据集。

提升28.4%、33.95%、19.89%。BioMedLM、Meditron、ChatDoctor的ROUGE-1指标分别为0.409、0.413、0.411，未超过ERQA模型。对于ROUGE-2，ERQA-13B的指标为0.203，较BioMedLM提升16.67%；ERQA-7B与Meditron的指标分别为0.184、0.181。在ROUGE-L指标上，ERQA-13B的结果为0.345，ERQA-7B的结果为0.329，Meditron的结果为0.320。

在TripClick数据集上[图6（b）]，ERQA-13B的各项指标均处于最优，其ROUGE-1指标为0.421。Meditron与ChatDoctor的ROUGE-1指标分别为0.400、0.387，ERQA-7B的ROUGE-1指标为0.403。在ROUGE-2上，ERQA-7B的指标为0.294、ERQA-13B的指标为0.303，均高于Meditron（0.286）与ChatDoctor（0.275）。ROUGE-L指标中，ERQA-13B的指标为0.367，次之分别是ERQA-7B（0.331）、Meditron（0.327）、ChatDoctor（0.316）。表4给出了ERQA在摘要任务中的若干示例，ERQA系模型生成的摘要更为准确且上下文相关性更强，进一步证明了该框架在医学知识抽取场景中的优势。

4.4. 基于文献的QA

最新提出的生物医学大模型Med-PaLM 2基于MedQA与MedMCQA数据集训练，在美国执业医师资格考试（USMLE）上取得了与专业医生相当的成绩。这类数据集以多项选择形式组织，因此模型性能主要以准确率衡量。与此不同，ERQA面向医学知识检索与QA场景，任务以基于上下文的阅读理解为主，传统准确率并不适用。为此，本文采用BLEU指标[57]评估生成内容的流畅性与语境相关性，用于衡量其在上下文理解任务中生成回答的质量。

如表5所示，在COVID-19与TripClick两个数据集上，LLM系模型整体显著优于BERT系模型。BERT、BioBERT、BioClinicalBERT的BLEU得分相对较低，表明其在生成高质量、与语境一致的QA回答方面存在局限。

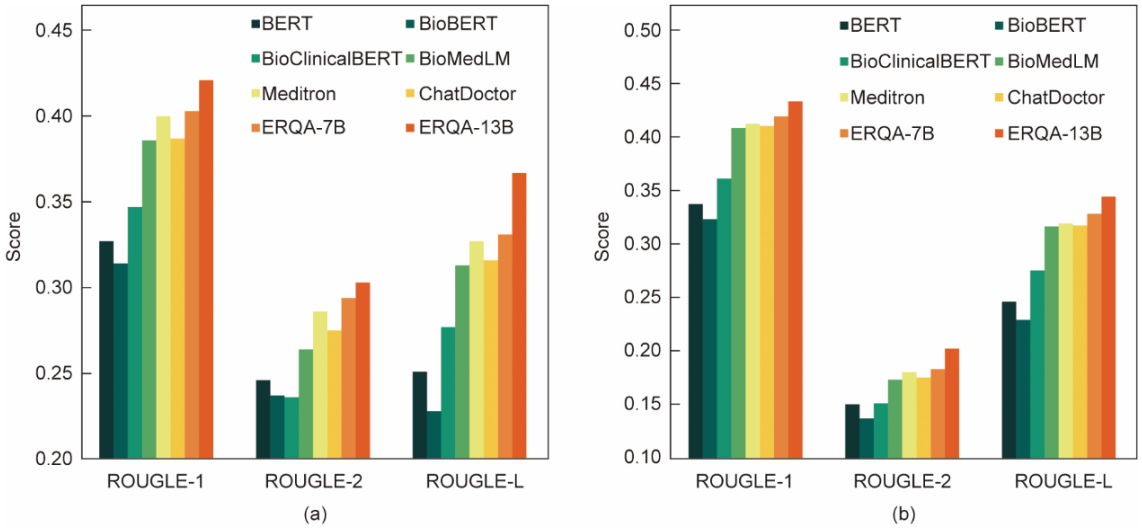

图6. COVID-19与TripClick数据集上的摘要生成性能比较。（a）COVID-19数据集；（b）TripClick数据集。

表4 摘要生成案例

Title	Abstract summarization
SARS-CoV-2 inactivation in aerosol by means of radiated microwaves (PMID: 37515131)	Coronaviruses cause respiratory infections in mammals and birds, including humans. Air transmission is the main route of SARS-CoV-2 infection. We developed a method to inactivate the virus using electromagnetic waves. In the laboratory, we consistently achieved a 90% reduction in infectivity with a safe Radio Frequency (RF) wave emission. This is the first demonstration of SARS-CoV-2 inactivation using RF waves in conditions compatible with human presence. Further research is needed to apply this technology in different environments and explore its potential against other viruses
Array-based multiplex and high-throughput serology assays (PMID: 36781805)	Serological tests are crucial for diagnosing infections, monitoring vaccination responses, and tracking disease transmission. The COVID-19 pandemic emphasized the need for rapid development of reliable serological tests. This study presents a multiplex serological assay based on bead array technology, capable of detecting anti-SARS-CoV-2 antibodies using minimal sample volume and high-throughput analysis. The method is adaptable to include antigens representing new variants, and it is being expanded to enable parallel detection of antibodies against multiple infectious agents
Brief report: Declining rates of SARS-CoV-2 vaccine uptake among patients with thoracic malignancies (PMID: 36792425)	The study analyzed vaccine uptake among 242 patients with thoracic malignancies. It revealed a decline in uptake of subsequent vaccine doses, with 75% receiving the recommended third dose, 39% receiving the recommended fourth dose, and 5% receiving the recommended fifth dose. Additional vaccinations were found to increase humoral immunity. The findings emphasize the need to understand the reasons for decreased vaccine uptake and highlight the importance of counseling patients with lung cancer on public health recommendations

PMID: PubMed identifier.

表5 COVID-19与TripClick数据集上基于文献的QA性能

Model	COVID-19		TripClick	
	BLEU-1	BLEU-4	BLEU-1	BLEU-4
BERT	0.541	0.013	0.568	0.004
BioBERT	0.632	0.011	0.454	0.003
BioClinicalBERT	0.790	0.018	0.573	0.004
BioMedLM	5.843	0.672	5.703	0.417
Meditron	6.278	0.725	6.004	0.432
ChatDoctor	6.083	0.705	5.842	0.426
ERQA-7B	6.467	0.722	6.284	0.447
ERQA-13B	7.851	0.873	6.543	0.536

BioMedLM在COVID-19数据集上的BLEU-1为5.843、BLEU-4为0.672，在TripClick上的BLEU-1和BLEU-4分别为5.703、0.417。Meditron与ChatDoctor进一步凸显了LLM的优势。Meditron在COVID-19数据集上的BLEU-1和BLEU-4分别为6.278和0.725，在TripClick数据集上分别为6.004、0.432。经专门面向医学检索与QA任务微调的ERQA模型在两数据集上均超越上述基线，ERQA-7B在COVID-19数据集上的BLEU-1和BLEU-4分别为6.467、0.722，在TripClick数据集上分别为6.284、0.447，更大参数量的ERQA-13B在此基础上取得进一步提升。

为进一步理解LLM的QA能力，本文对ERQA进行了人工评测，以连贯性、一致性与满意度为评分维度[34]。各维度采用百分制并划分为四级，以细化不同维度的性能判断：1~25分表示性能较差；26~50分表示性能一般；51~75分表示性能良好；76~100分表示性能优秀。

• 连贯性衡量回答的逻辑衔接与论证支撑。高得分（76~100）表示生成的语句概念准确、逻辑严谨；低得分（1~25）表示内容逻辑松散或难以理解。

• 一致性衡量回答与来源文献的一致程度，防止“幻觉”或事实性错误。高得分代表与来源高度一致，低得分反映偏离或不准确。

• 满意度衡量回答对用户信息需求的满足度，包括完整性与信息量。高分表示回答充分且相关，低分表示未满足预期。

在人工评价过程中，三位领域专家独立对ERQA生成的QA对进行评分，根据连贯性、一致性和满意度对其回答进行0~100分的评分。为确保评测可靠性，本文对每个QA对分别计算Krippendorff's α以评估三项指标的评审者一致性，并将0.75设为阈值：当$\alpha \geq 0.75$时，视为评分一致，取三位评审者的均值作为该QA对的最终得分；若$\alpha < 0.75$，则将该QA对标记为复评，由评审者复核并讨论分歧，直至α超过阈值。完成全部QA对的评审与分歧消解后，本文对三项指标（连贯性、一致性和满意度）的最终得分取全体样本平均，以保证整体评估的一致可靠。

如图7所示，将模型规模由70亿参数提升至130亿参数，可在连贯性、一致性和满意度三项指标上提升ERQA框架的表现。其中，较大模型在满意度上提升更为明显，

显示其从文献中提取准确且相关信息方面的能力增强。然而，这些性能提升并未完全抵消额外计算成本带来的代价。在医学知识检索与QA应用的实际部署中，ERQA-7B可能是更为均衡的选择。

模型能够给出结构清晰、表述明确的回答，但在部分情况下会引入与来源细节不一致之处，这可能源于LLM的固有“幻觉”。例如，针对COVID-19免疫的一个问题，生成文本对抗体保护持续时间表述有误，导致一致性评分下降。除此之外，也有部分回答能够满足信息需求但在涉及复杂医学概念时逻辑衔接不足，影响连贯性评分。为评估ERQA的医学文本理解、相关信息抽取与语境化回答能力，本文采用来自COVID-19文献的多样化摘要作为输入查询。表6汇总了ERQA在严格推理需求下的典型输出案例，并据此提出后续优化方向。

医学知识检索与QA通常涉及多轮依赖上下文的交互。尽管ERQA支持处理较长上下文，其最大上下文长度仍受所用模型结构限制。例如，本研究使用的Llama-7B的上下文窗口上限为4096 tokens，足以覆盖多数医学查询。同时，可检索的上下文长度还受向量数据库中已预处理的摘要与关键文献段落所限。所提出的ERQA框架能在token预算内提供较为详尽的回答，并通过上下文约束保持回答的完整性。

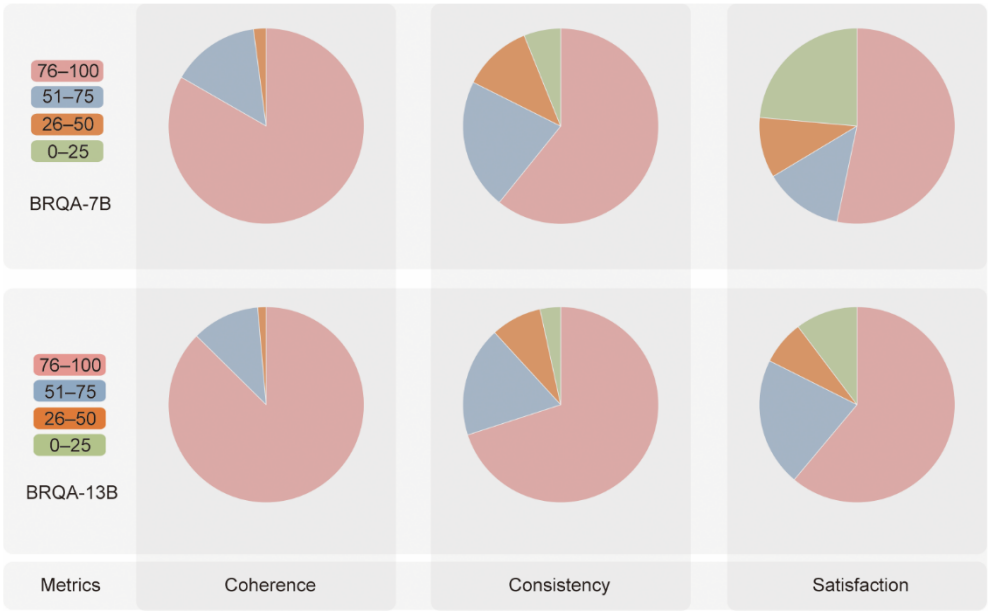

图7. 基于文献的QA性能评估。

表6 基于文献的QA生成案例

PMID	Question	Answer
37069722	What are the challenges faced in controlling the COVID-19 pandemic despite the administration of vaccines?	Despite the administration of vaccines, challenges in controlling the COVID-19 pandemic include the emergence of SARS-CoV-2 variants, breakthrough and reinfection cases in vaccinated individuals, new cases in low-to-middle income countries, and concerns regarding screening for asymptomatic infections and managing diagnosed cases.
36420736	What are some of the potential long-term contributors to the symptoms of COVID-19?	Perturbations in T cells, B cells, mast cells, autoantibodies, and metabolic reprogramming are potential contributors to the long-term symptoms of COVID-19.
36369746	How did the COVID-19 pandemic impact the circulation of common seasonal respiratory viruses (SRV) in the Ottawa region of Canada?	The COVID-19 pandemic led to a significant reduction in the incidence of common SRV, including influenza, in the Ottawa region of Canada. The introduction of nonpharmaceutical measures, such as masking, to prevent the transmission of SARS-CoV-2 likely contributed to the near absence of SRV cases. However, rhinoviruses and enteroviruses were relatively unaffected during this period.
37544384	How long were the levels of protective antibodies maintained after the second dose of vaccine in SARS survivors?	The levels of protective antibodies (anti-SARS-CoV-2 spike IgG, anti-SARS-CoV-2 spike receptor-binding domain IgG, and anti-SARS-CoV-2 neutralizing antibodies) were maintained at very high levels even 166 days after the second dose of the COVID-19 vaccine in SARS survivors.

IgG: immunoglobulin G.

4.5. 消融实验

本文在COVID-19与TripClick数据集上开展消融实验，以评估ERQA框架各组件的贡献。消融实验包括四个比较对象：原始Llama 2（用作基准模型）；变体Llama2 w/VD（即在QA任务中引入文献向量数据库增强的Llama2模型）；ERQA（含向量数据库）；ERQA w/o VD（表示去除向量数据库的ERQA）。

在依赖嵌入库和向量数据库的文章检索任务中，本文比较了Llama2 w/VD与ERQA的表现。表7显示，ERQA相比Llama2 w/VD有明显提升，体现出ERQA的微调流程对检索任务的积极作用。在摘要生成任务中，本文同时关注Llama2与ERQA w/o VD的比较，以检验模型能否有效提炼文章主题。图8显示，ERQA w/o VD的ROUGE分数显著提升，证明增量预训练与微调分别在医学知识掌握与任务对齐两方面改进了模型性能，使其优于原始基线。

表7 不同模型在COVID-19和TripClick数据集上的文章检索消融对比结果

Metric	COVID-19		TripClick	
	Llama2 w/VD	ERQA	Llama2 w/VD	ERQA w/o VD
NDCG@10	0.239	0.264	0.247	0.337
NDCG@20	0.318	0.332	0.326	0.348
Recall@10	0.245	0.289	0.251	0.279
Recall@50	0.287	0.326	0.31	0.376
MRR	0.292	0.348	0.348	0.393

在文献检索和摘要生成的下游任务中，基于文献的QA最能直接体现LLM驱动的医学知识挖掘能力。尽管微调与数据库均为ERQA的QA任务服务，垂域LLM（ERQA w/o VD）与Llama2 w/VD仍可完成QA，但性能不及完整ERQA框架。针对BLEU指标（表8），ERQA w/o VD与Llama2 w/VD表现相近，均低于ERQA。以表6中PMID: 37069722的问题为例，ERQA w/o VD的回答为：“尽管已接种疫苗，仍有若干挑战阻碍疫情控制，包括疫苗犹豫、资源匮乏地区的疫苗可及性不足、难以覆盖脆弱人群以及公共卫生基础设施与资源不足。”该回答忽略了“病毒变异”等关键信息，并引入与问题不符的“疫苗犹豫”。与此相比，完整ERQA在两数据集的BLEU指标均优于其他版本。消融结果表明微调与向量数据库集成的联合效应显著增强了其在QA任务上的性能。

表8 不同模型在COVID-19和TripClick数据集上的基于文献的QA消融对比结果

Model	COVID-19		TripClick	
	BLEU-1	BLEU-4	BLEU-1	BLEU-4
Llama2	4.626	0.527	4.332	0.274
Llama2 w/VD	5.942	0.713	0.871	0.415
ERQA w/o VD	5.883	0.692	5.764	0.424
ERQA	6.467	0.772	6.284	0.447

4.6. 局限性

在文献检索、摘要生成与基于文献的QA三类任务的评估中，本文结果显示LLM相较传统语言模型在理解与运用医学知识方面具有显著优势，这为基于文献的医学知识挖掘带来新的可能。但面向实际应用的LLM驱动医学检索与QA系统仍面临若干挑战。

其一，检索粒度与嵌入层选择直接影响结果质量。向量数据库中的文本切分粒度若过粗，易导致信息过载或上下文缺失；过细则可能割裂语境、遗漏要点。另一方面，LLM的不同隐层捕获不同抽象层级的信息，不同生成嵌入的模型层会显著影响检索质量。向量检索提升语义相关性的同时，基于元数据的匹配往往在精确性与可解释性上更具优势；如何在向量嵌入与元数据特征之间取得平衡，是系统设计的关键。

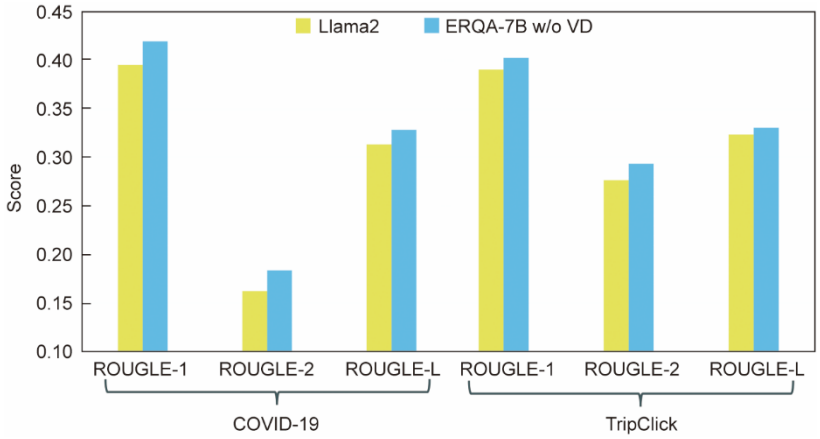

图8. 不同模型在COVID-19和TripClick数据集上的摘要生成消融对比结果。

其二，在基于文献的QA评估中，ERQA的连贯性得分高于一致性与满意度，主要原因是训练语料中的偏倚或不一致会导致生成内容偏离事实。为缓解这一问题，未来可引入外部知识源与核验机制，对生成回答进行证据校对与事实验证，缓解生成“幻觉”问题。

5. 结论

本文提出一种由垂域LLM驱动的医学知识检索与QA框架ERQA，主要将三大核心组件整合为统一流程，并在多种场景（文献检索、摘要生成、基于文献的QA）下进行定量与定性评估。实验结果表明，该方法在推进生物医学知识发现方面具有潜力。未来工作将纳入更大规模的生物医学文献数据集，并引入更多评价指标，持续提升模型性能。

CRediT authorship contribution statement

Yuyang Liu: Writing-review & editing, Writing-original draft, Validation, Methodology, Data curation, Conceptualization. **Xiaoying Li:** Writing-original draft, Methodology, Conceptualization. **Yan Luo:** Investigation, Formal analysis, Data curation. **Jinhua Du:** Software, Methodology. **Ying Zhang:** Formal analysis, Data curation. **Tingyu Lv:** Formal analysis, Data curation. **Hao Yin:** Writing-review & editing, Supervision. **Xiaoli Tang:** Writing-review & editing, Supervision, Funding acquisition. **Hui Liu:** Writing-review & editing, Supervision, Funding acquisition.

致谢

本研究得到中国医学科学院医学与健康科技创新工程（2021-I2M-1-033）与国家重点研发计划（2022YFF0711900）的资助。

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. 2023. arXiv:230308774.
[2] OPENAI. ChatGPT [Internet]. San Francisco: OPENAI; undated [cited 2024 May 5]. Available from: https://openai.com/blog/chatgpt/.
[3] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930–40.
[4] Bolton E, Hall D, Yasunaga M, Lee T, Manning C, Liang P. BioMedLM: a domain-specific large language model for biomedical text. Stanford: Stanford Center for Research on Foundation Models; 2022.
[5] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief bioinform 2022;23(6):bbac409.
[6] Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inform Assoc 2024;31(9):1833–43.
[7] Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2024;25(1):bbad493.
[8] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9(1):e45312.
[9] Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature? J Am Soc Nephrol 2023; 34(8):1302–4.
[10] Tan RSYC, Lin Q, Low GH, Lin R, Goh TC, Chang CCE, et al. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. J Am Med Inform Assoc 2023;30(10): 1657–64.
[11] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.
[12] Zhao Q, Xu D, Li J, Zhao L, Akhtar RF. Knowledge guided distance supervision for biomedical relation extraction in Chinese electronic medical records. Expert Syst Appl 2022;204:117606.
[13] Roy A, Pan S. Incorporating medical knowledge in BERT for clinical relation extraction. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021 Nov 7–11; Online. Stroudsburg: Association for Computational Linguistics; 2021.
[14] Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics 2020;21(1):188.
[15] Hristovski D, Dinevski D, Kastrin A, Rindflesch TC. Biomedical question answering using semantic relations. BMC Bioinformatics 2015;16(1):6.
[16] Qi T, Qiu S, Shen X, Chen H, Yang S, Wen H, et al. KeMRE: Knowledge-enhanced medical relation extraction for Chinese medicine instructions. J Biomed Inform 2021;120:103834.
[17] Whissell JS, Clarke CL. Improving document clustering using Okapi BM25 feature weighting. Inf Retrieval 2011;14(5):466–87.
[18] Luo G, Tang C, Yang H, Wei X. MedSearch: a specialized search engine for medical information retrieval. Proceedings of the 17th ACM Conference on Information and Knowledge Management; 2008 Oct 26–30; Napa Valley, CA, USA. New York City: Association for Computing Machinery; 2008.
[19] Canese K, Weis S. PubMed: the bibliographic database. In: The NCBI handbook. 2nd ed. Bethesda: National Center for Biotechnology Information (US); 2013.
[20] Vanopstal K, Buysschaert J, Laureys G, Vander SR. Lost in PubMed. Factors influencing the success of medical information retrieval. Expert Syst Appl 2013; 40(10):4106–14.
[21] Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. EBioMedicine 2024;100:104988.
[22] Mourão A, Martins F, Magalhães J. Multimodal medical information retrieval with unsupervised rank fusion. Comput Med Imaging Graph 2015;39:35–45.
[23] Ma J, Wu X, Huang L. The use of artificial intelligence in literature search and selection of the PubMed database. Sci Program 2022;16(1):1–9.
[24] Jin Q, Shin A, Lu Z. Lader: log-augmented dense retrieval for biomedical literature search. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2023 July 23–27; Taipei, China. New York City: Association for Computing Machinery; 2023.

[25] Zeng H, Liu J, Wang M, Wei B. A sequence to sequence model for dialogue generation with gated mixture of topics. Neurocomputing 2021;437:282–8.
[26] Gomes Jr J, de Mello RC, Ströele V, de Souza JF. A hereditary attentive template-based approach for complex knowledge base question answering systems. Expert Syst Appl 2022;205:117725.
[27] Guu K, Lee K, Tung Z, Pasupat P, Chang M. REALM: retrieval augmented language model pre-training. In: Proceedings of International Conference on Machine Learning; 2020 Feb 15–17; Shenzhen, China. New York City: Association for Computing Machinery; 2020.
[28] Varshney D, Zafar A, Behera NK, Ekbal A. Knowledge graph assisted end-to-end medical dialog generation. Artif Intell Med 2023;139:102535.
[29] Wu S, Li Y, Zhang D, Zhou Y, Wu Z, eds. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5 – 10; Online. Stroudsburg: Association for Computational Linguistics; 2020.
[30] Wu S, Li Y, Zhang D, Wu Z. Generating rational commonsense knowledge-aware dialogue responses with channel-aware knowledge fusing network. IEEE/ACM Trans Audio Speech Lang Process 2022;30:3230–9.
[31] Pereira J, Fidalgo R, Lotufo R, Nogueira R. Visconde: multi-document QA with GPT-3 and neural reranking. In: Proceedings of the European Conference on Information Retrieval; 2023 Apr 2–6; Dublin, Ireland. Berlin: Springer; 2023.
[32] Huang D, Wei Z, Yue A, Zhao X, Chen Z, Li R, et al. DSQA-LLM: domain-specific intelligent question answering based on large language model. In: Proceedings of the International Conference on AI-generated Content; 2023 Aug 25–26, Shanghai, China. Berlin: Springer; 2023.
[33] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. 2019. arXiv:190906146.
[34] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31:943–50.
[35] El-Kassas WS, Salama CR, Rafea AA, Mohamed HK. Automatic text summarization: a comprehensive survey. Expert Syst Appl 2021;165:113679.
[36] Khan R, Qian Y, Naeem S. Extractive based text summarization using *K*-means and TF-IDF. Int J Electron Bus 2019;12(3):33–44.
[37] Parveen D, Ramsl HM, Strube M. Topical coherence for graph-based extractive summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015 Sep 17–21; Lisbon, Portugal. Stroudsburg: Association for Computational Linguistics; 2015. p. 1949–54
[38] Gehrmann S, Deng Y, Rush AM. Bottom-up abstractive summarization. 2018. arXiv:180810792.
[39] Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology 2023;1(2):100017.
[40] Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. npj Digit Med 2023; 6(1):158.
[41] Abacha AB, Yim WW, Adams G, Snider N, Yetisgen-Yildiz M. Overview of the MEDIQA-Chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In: Proceedings of the 5th Clinical Natural Language Processing Workshop; 2023 Jul 9 – 14; Toronto, ON, Canada. Stroudsburg: Association for Computational Linguistics; 2023.
[42] Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024;34(5):2817–25.
[43] Rekabsaz N, Lesota O, Schedl M, Brassey J, Eickhoff C. TripClick: the log files of a large health web search engine. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021 Jul 11–15; Virtual Event. New York City: Association for Computing Machinery; 2021. p. 2507–13.
[44] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. 2023. arXiv:230709288.
[45] Ma G, Wu X, Wang P, Lin Z, Hu S. Pre-training with large language model-based document expansion for dense passage retrieval. 2023. arXiv:230808285.
[46] Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. npj Digit Med 2021;4(1):68.
[47] Cunningham P, Delany SJ. K-nearest neighbour classifiers-a tutorial. ACM Comput Surv 2021;54(6):1–25.
[48] Hezel N, Barthel KU, Schall K, Jung K. Fast approximate nearest neighbor search with a dynamic exploration graph using continuous refinement. 2023. arXiv:230710479.
[49] Yasunaga M, Kasai J, Zhang R, Fabbri AR, Li I, Friedman D, et al. ScisummNet: a large annotated corpus and content-impact models for scientific paper summarization with citation networks. In: Proceedings of the AAAI conference on artificial intelligence; 2019 Jan 27 – February 1; Honolulu, HI, USA. Washington, DC: Association for the Advancement of Artificial Intelligence (AAAI) Press; 2019.
[50] Chen Y, Polajnar T, Batchelor C, Teufel S. A corpus of very short scientific summaries. In: Proceedings of the 24th Conference on Computational Natural Language Learning; 2020 Nov 19 – 20; Online. Stroudsburg: The Association for Computational Linguistics; 2020.
[51] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: low-rank adaptation of large language models. 2021. arXiv:210609685.
[52] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. arXiv:181004805.
[53] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
[54] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019. arXiv:190403323.
[55] Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. Meditron-70b: scaling medical pretraining for large language models. 2023. arXiv:231116079.
[56] Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. Chatdoctor: a medical chat model fine-tuned on a large language model meta-AI (Llama) using medical domain knowledge. Cureus 2023;15(6):e40895.
[57] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics; 2002 Jul 7–12; Philadelphia, PA, USA. Stroudsburg: Association for Computational Linguistics; 2002.