Software Engineering and Applications 软件工程与应用, 2025, 14(4), 772-783
Published Online August 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.144068

Hans汉斯

PV-RAG：基于LLM的药物警戒检索增强生成模型

陈文浩[1]，朱云霞[2*]，张鸿红[1]，薛　景[1]，刘帅博[1]，查星宇[1]，翁艺航[2]，魏建香[3]

[1]南京邮电大学计算机学院，江苏　南京
[2]南京邮电大学物联网学院，江苏　南京
[3]南京邮电大学图书馆，江苏　南京

收稿日期：2025年6月11日；录用日期：2025年7月28日；发布日期：2025年8月7日

摘　要

目的/意义：随着大语言模型(LLM)在自然语言处理任务中的广泛应用，其在处理专业领域问答任务中表现出局限性。药品因其“治病且可能致病”的双重特征，须在安全警戒范围内进行使用，而传统大模型受训练数据影响难以有效表达用药安全复杂且动态的信息，易导致数据缺失引发的回答幻觉问题。方法/过程：本文提出了一种面向药物警戒的检索增强生成模型(PV-RAG)，其综合应用大语言模型LangChain框架、稀疏与稠密检索混合策略以及改进的查询改写机制，显著提升检索生成结果的准确性。结果/结论：对比实验结果表明，PV-RAG模型能够精准提取药品说明书中的关键信息，在药物警戒问答任务中表现出显著优势，为药物警戒智能化问答提供了新的技术支持和应用前景。

关键词

大语言模型，检索增强生成，药物警戒，智能问答，回答幻觉

PV-RAG: Retrieval-Augmented Generation Model for Pharmacovigilance Based on Large Language Models

Wenhao Chen[1], Yunxia Zhu[2*], Honghong Zhang[1], Jing Xue[1], Shuaibo Liu[1], Xingyu Zha[1], Yihang Weng[2], Jianxiang Wei[3]

[1]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu
[2]School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu
[3]Library of Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Jun. 11th, 2025; accepted: Jul. 28th, 2025; published: Aug. 7th, 2025

*通讯作者。

文章引用: 陈文浩, 朱云霞, 张鸿红, 薛景, 刘帅博, 查星宇, 翁艺航, 魏建香. PV-RAG: 基于 LLM 的药物警戒检索增强生成模型[J]. 软件工程与应用, 2025, 14(4): 772-783. DOI: 10.12677/sea.2025.144068

Abstract

Purpose/Significance: With the widespread application of large language models (LLMs) in natural language processing tasks, their limitations in handling domain-specific question-answering tasks have become evident. Pharmaceuticals, characterized by their dual nature of "treating diseases while potentially causing harm", must be used within strict safety monitoring frameworks. Traditional large models, constrained by training data, struggle to accurately represent the complex and dynamic nature of drug safety information, often leading to hallucinated responses due to data gaps. Methods/Process: This study proposes a Pharmacovigilance-oriented Retrieval-Augmented Generation model (PV-RAG), which integrates the LangChain framework for LLMs, a hybrid sparse-dense retrieval strategy, and an improved query rewriting mechanism to significantly enhance the accuracy of retrieval-generated responses. Results/Conclusions: Comparative experiments demonstrate that the PV-RAG model can precisely extract critical information from drug manuals and exhibits superior performance in pharmacovigilance question-answering tasks. This approach provides new technical support and application prospects for intelligent pharmacovigilance systems.

Keywords

LLM, RAG, Pharmacovigilance, Intelligent Q&A, Answer Hallucination

1. 引言

大语言模型技术的发展极大地推动了自然语言处理任务的进步，特别是在语言理解和生成方面取得了良好的表现[1]。Hadi M. U.等人[2]在关于大语言模型的概述中指出，通过使用深度学习技术和大量数据集，大语言模型能够捕捉语言中的复杂模式和长距离依赖关系，能够广泛应用于医学、教育、金融和工程等垂直领域。然而，大语言模型在理解语言的语义和上下文方面仍然存在局限性[2]，尤其是在面对专业领域知识任务时，容易因知识体系不全面或缺乏专业性而无法提供准确答案。

药品作为一种特殊的商品，具有“治病且可能致病”的双重特征。据统计 2023 年全国药品不良反应监测网络收到各类《药品不良反应/事件报告表》241.9 万份[3]，因此相关药品的使用必须在药物警戒的范围内严格规范地进行。

然而随着网络购药的普及化，民众用药安全问题更为突显，保障居民安全用药成为一项重要的民生工程。网络上药品数据名目众多内容庞杂，其中不乏过时或者错误信息，依赖于基础大语言模型易产生不准确的用药指导信息，目前大语言模型的回答幻觉现象已成为公认的问题[4]。伴随大语言模型的兴起，检索增强生成(RAG)技术受到了更多关注和研究[5]。Gao 等人[6]指出，将 LLM 的内在知识与外部数据库的庞大动态存储库协同合并，不仅可以有效应对信息更新滞后的问题，还能使大模型获得更丰富、实时的专业知识。

为此，本文在大语言模型应用 LangChain 框架基础之上，综合稀疏与稠密检索混合策略以及改进的查询改写机制，以权威的药物说明书作为数据补充，构建了一种面向药物警戒领域的检索增强生成模型(PV-RAG)，旨在解决大语言模型在药物警戒领域应用中出现的信息不准确和高冗余问题。实验对比数据

表明，PV-RAG 能够从采集整理的 13,639 条药品说明书中精准匹配出各类信息，对各类用药问题提供更精准更全面的解答，显著提高药物警戒领域问答系统的准确性和权威性，为智能医疗服务的发展提供了新的应用方向。

2. 相关工作

药物警戒的概念最早在 20 世纪 70 年代，它是指对药品不良反应及其他与药物相关问题的监测、评估、理解和预防的科学与活动[7]，包括药物不良反应监测、风险评估、风险沟通、风险最小化和持续监测等多个方面，是保障公众安全用药的重要工具。

与起步较早的发达国家相比，我国在药物警戒数据建设和民众药物知识科普方面仍有较大提升空间。例如王青宇等[8]对多个国际数据库对比研究后指出，我国药物警戒数据库在数据开放性和患者主动上报途径方面与国际数据库存在差距。另一方面，为解决药物警戒的公众教育问题，可以考虑依靠大语言模型技术构建面向公众的安全用药信息检索与问答系统。

RAG 技术为解决大语言模型易出现的幻觉问题提供了具体的解决思路，将 RAG 技术应用于药物警戒领域，使大模型不再依赖于有限的预训练知识，通过连接医学文献库、临床数据集和药物数据库等资源，弥补传统大模型在专业知识方面的不足，为用户提供更准确、更专业性的回答。

Miao J.等人[9]介绍了 NVIDIA 将 RAG 用于医疗研究的案例，通过连接电子健康记录和语义搜索功能，运用 RAG 技术支持诊断特定病情并提供治疗建议。Karan S.等人[10]提到的谷歌的 Med-PaLM，以及 Wei J.等人[11]提到的斯坦福的 BiomedLM，通过整合大规模医学数据进行微调，达到了高精度结果。其中 2023 年公布的 Med-PaLM 2 [10]在医疗问答中表现出更高专业度，回答准确率与人类医生水平接近。Yang 等人[12]的 BioASQ 系统结合了生物医学文献数据库，使模型生成医学答案时更为精准；Yagnik N.等人[13]提出的 MedLM 结合特定医学数据集，优化了模型生成效果。

综上所述，大语言模型通过大数据训练和检索增强生成机制，能够在疾病诊断、药物推荐等场景中表现出明显优势。将智能问答系统与检索增强技术相结合，能够在医学应用中实现高精度和专业化。

3. PV-RAG 模型构建

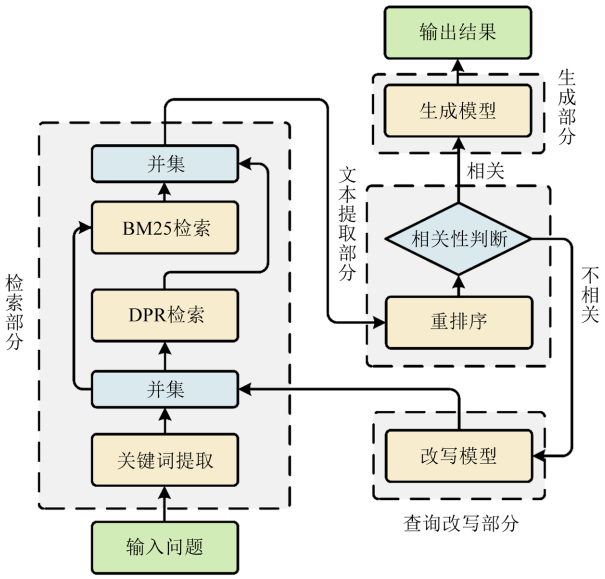

Figure 1. Model architecture diagram of PV-RAG
图 1. PV-RAG 模型结构图

本文提出的 PV-RAG (Pharmacovigilance Retrieval-Augmented Generation)模型，适用于药物警戒领域的检索增强生成任务。PV-RAG 模型总共分为四个部分，分别为检索部分、文本提取部分、查询改写部分和生成部分，其结构如图 1 所示。

3.1. 检索部分

稠密检索模型(DPR, Dense Passage Retrieval)，其通过 BERT 模型分别构造查询编码器和段落编码器，将查询与文档片段转化为稠密向量表示。具体来说，DPR 利用深度学习模型捕捉文本间的语义关联，计算查询向量与文档片段向量的 BERTScore，以此衡量二者的相关性。相比于传统的稀疏检索方法，DPR 在大多数数据集上展现了更优的性能。然而，在药物警戒领域任务中，DPR 方法的表现会因数据分布的偏移而受到影响。药物警戒领域任务通常涉及到数据大多已经过规范化处理，导致 DPR 有时难以有效捕捉药物警戒领域的语义特征[14]。此时，传统的稀疏检索方法 BM25 由于其基于词频与逆文档频率(TF-IDF)的检索策略，在某些情况下表现出更高的稳健性[14]。BM25 直接基于词匹配进行检索，能够在没有语义先验的情况下提供可靠的检索结果。

为充分利用稀疏检索和稠密检索的优势，PV-RAG 模型在检索阶段采用了 BM25 与 DPR 的混合检索方法，将二者的检索结果取并集，从而构建更全面的候选文档集合。首先，根据利用 BM25 对检索关键词在文档集合中进行词频与逆文档频率匹配，快速筛选出初步相关的文档片段。然后利用 DPR 通过查询编码器与段落编码器对查询与文档片段进行语义编码，计算出 BERTScore 相关性分数，进一步补充相关性较高但未被 BM25 检索到的片段。最终将 BM25 与 DPR 的检索结果合并，形成候选文档片段的多样化集合。

在信息检索任务中，取并集操作发生在稀疏检索(BM25)和稠密检索(DPR)的结果之间。假设查询 q 在文档集合 D 中进行检索，BM25 和 DPR 分别返回了一组候选文档集合，记为 $R_{BM25}(q)$ 和 $R_{DPR}(q)$，取并集操作如式(1)所示。

$$R_{UNION}(q) = R_{BM25}(q) \cup R_{DPR}(q) \tag{1}$$

在大模型生成任务中，输入质量直接影响生成结果的准确性。初始检索结果中可能包含与查询无关的片段，若直接输入生成模型，可能会导致模型偏离用户意图，从而影响生成的可靠性。因此，在检索阶段之后进行重排序，以确保仅将与查询密切相关的片段输入生成阶段，这也是提升任务性能的必要步骤。

相似度计算是实现这一目标的核心方法。它通过量化查询与文档片段之间的语义相似性，筛选出最相关的内容。特别是在药物警戒领域中，知识表达形式多样化，相似度计算能帮助模型更好地捕捉语义关系。通过引入多头注意力机制的引导和动态融合策略，最终形成段落的统一表征。最终基于查询和段落表征之间的关系，使用式(2)计算其相似度。

$$f(p,q) = \frac{1}{m}\sum_{i=1}^{m} E_{q,i}^{T} X_i \tag{2}$$

其中，$E_{q,i}^{T}$ 为第 i 个段落向量，X_i 为段落多向量表示。

3.2. 文本提取

在 PV-RAG 模型的框架下，文本提取部分扮演着至关重要的角色，其主要目标是确保从检索阶段获得的文档中提取出与用户查询高度相关的文本片段。这一过程包括对文档片段进行重排序、提取和利用其中的关键信息。通过这些工作，文本提取部分将为 PV-RAG 模型的生成阶段提供精炼、准确的输入，从而提高整个问答系统的性能和输出质量。

对于重排序(Re-ranking)工作，由于传统 BERT 直接用于重排序时需对查询和文档拼接输入，计算开销大，而 MR-BERT 及其变体通过双塔结构或轻量级交互模块，先独立编码查询和文档，再计算细粒度交互，显著降低计算成本[15]，因此本文使用 MRI-BERT 架构进行重排序。

对于已经给定的一个药品说明书数据集 $D=\{(q_i,P_i,Y_i)\}_{i=1}^{n}$，其中，$P_i=[p_{i1},p_{i2},\cdots,p_{ik}]$为查询文本，$Y_i=[y_{i1},y_{i2},\cdots,y_{ik}]$为第 i 个查询文本 P_i 对应的标签，每一个 y 为 0 或 1 整数变量，即 $y\in\{0,1\}$。定义重排序模型的目标为：

$$f^*=\max_f E_{\{q,D,Y\}}\vartheta(Y,F(q,P)) \tag{3}$$

其中，ϑ 为标签以及预测的评估度量。

$$F(q,P)=\{f(q,p_i)\}_{i=1}^{N} \tag{4}$$

为对应的 q 下文档的分数集合。在该架构的基础上，查询和文档片段被编码为高维嵌入向量，通过计算这些向量之间的相似度分数，可以有效判断查询与文档片段的相关性。

通过使用 MRI-BERT 架构，对药品说明书文档进行预训练，编码器可以更好地查询和正样例段落之间的语义关系，在得到与期望结果相关的信息中具有较大优势。

在检索阶段获取的文档中，可能含有与当前问题无关的信息，对大模型的处理过程产生负面影响，导致模型无法聚焦于问题的核心部分，从而干扰其对关键信息的理解和提取。因此，确保在检索阶段进行相关性判断过滤掉无关信息，提供简明、相关的文档输入，对于提升大模型的生成质量至关重要。

由于 GLM-4 模型具有更好的语义理解和高效的多语言处理能力[16]，本文基于 GLM-4-9B 模型，借助提示工程进行相关性判断。

3.3. 查询改写

对于未通过相关性判断的内容，进入查询改写部分。在这一环节，对于首次检索未能获得有效信息的查询，系统会根据输入查询中的关键术语，结合前一次关键词提取的结果，加入与这些关键词紧密相关的同义词、近义词或行业术语，以达到增强查询的多样性，确保覆盖到更多可能的相关信息的目的。通过使用文本过滤器评估内容的有效性，启动基于同义词扩展的查询改写，从而有助于减轻因不必要的查询改写而导致的用户意图偏离问题[17]。同义词拓展使用 GLM-4-9B 模型完成。对于 GLM-4-9B 模型生成的同义词，再次进入检索部分，进行新一轮更具有针对性的检索。

设置同义词拓展的提示词为：

Prompt = *f*"""

你是一名专注于医学领域的大语言模型助手。请根据以下用户查询任务，结合其关键词，生成一个用于信息检索的改写版本，要求如下：

根据{*输入文本* 1}，*对*{*输入文本* 2}*进行语义扩展*(*同义词、近义词、专业术语或缩写词*)，*并将语义拓展结果输出到*{*结果*}。*要保证改写后查询与原始语义一致，不得引入歧义或无关词。若关键词与特定人群、疾病、药品类别有关，需结合领域背景生成领域相关术语。*

{*输入文本* 1} = {*text*1}

{*输入文本* 2} = {*text*2}。

其中，{*输入文本* 2}*若含有 n 个关键词，则*{*输入文本* 2}*的格式为*{*关键词* 1}，{*关键词* 2}，{*Omit* (*n*)}，{*关键词 n*}。*其中，*{*Omit* (*n*)} = {*Omit* (*n* − 1)}，{*关键词 n* − 1}，*n 满足 n* ≥ 3；*n 满足 n* = 2 *时，*{*Omit* (*n*)} = ”*NULL*”。

{结果} = {输出文本}。其中，{输出文本}若含有 m 个关键词，则{输出文本}的格式为{新关键词 1}，{新关键词 2}，{Omit (m)}，{新关键词 m}。其中，{Omit (m)} = {Omit (m − 1)}，{新关键词 m − 1}， m 满足 m ≥ 3； m 满足 m = 2 时，{Omit (m)} = "NULL"。

要求 n + 2 ≤ m 。

"""

其中，text1 为输入的问题文本，text2 为进入检索环节前得到的关键词。通过该提示词，GLM-4-9B 可以稳定地返回同义词拓展后的新关键词。

以 text1 = “使用辰龙罗欣可能会出现什么不良反应？”，text2 = “辰龙罗欣，不良反应”为例，输出为“辰龙罗欣，副作用，不良反应，风险，NULL”。

再以 text1 = “使用药物艾克儿的不良反应有哪些？”，text2 = “艾克”为例，注意到 text2 存在对 text1 的药品名称提取不正确的现象。经查询改写过程后，输出结果为：“艾克儿，盐酸氨溴索，氨溴索，不良反应，NULL”。该结果既提取出“不良反应”，也正确地提取出了“艾克儿”药品名，以及其通用名，改写效率高。

上述提示词为优化后版本。在较早的实验中注意到，若提示词中不包含：“你是一名专注于医学领域的大语言模型助手。请根据以下用户查询任务，结合其关键词，生成一个用于信息检索的改写版本”、“要保证改写后查询与原始语义一致，不得引入歧义或无关词。若关键词与特定人群、疾病、药品类别有关，需结合领域背景生成领域相关术语。”这两处内容，则会出现改写效率低的问题。

以先前 text1 = “使用药物艾克儿的不良反应有哪些？”的问题为例，输入相同的内容，较早的低效率提示词得到的输出为：“艾克儿，艾克药物，艾克治疗，艾克的不良反应，NULL”。此组输出存在“艾克药物”、“艾克治疗”等并未准确提取“艾克儿”，为无效的改写；此组输出还存在“艾克的不良反应”这种不仅未准确提取药品名，还并未提取出“不良反应”这一关键词的现象，属于无效的改写。综上，优化后的提示词在改写效率上存在明显提升。

改写完成后，将改写后的信息输入检索控制器，进行更为精确和针对性的二次检索，从而获得更相关且有价值的搜索结果。该过程能够提高信息匹配度和检索效率，更有效地满足查询的实际需求。

3.4. 生成回答

对于通过相关性判断的内容，进入生成模型，利用 GLM-4-9B 大语言模型生成回答。

以问题“使用凯德林的注意事项有哪些？”为例，在使用 BERT-base-uncased 作为 Embedding 模型的情况下，输出的结果为：

“1. *青霉素皮肤试验：用药前需要进行皮肤试验，阳性者禁用。交叉过敏反应：对一种青霉素类抗生素过敏的人，可能对其他青霉素类抗生素也过敏，也可能对青霉胺或头孢菌素过敏。*

2. *慎用情况：过敏性疾病史：有哮喘、湿疹、枯草热、荨麻疹等过敏性疾病史的患者应慎用。*

3. *溶液稳定性与使用溶液浓度：氨苄西林溶液浓度越高，稳定性越差，且溶液的稳定性会随着温度升高而降低。致敏物质：溶液放置时间过长可能产生致敏物质，因此应尽量及时使用，不宜久置。*

4. *对诊断的干扰：用药期间，用硫酸铜法测定尿糖可能出现假阳性，但用葡萄糖酶法则不受影响。*

5. *大剂量使用：使用大剂量时，应定期检测血清钠水平。*”

4. 数据集与问题设计

4.1. 数据集

本文构建的 PV-RAG 模型在通用大语言模型的基础上综合了检索增强技术，数据直接来源于权威部

门提供的药品数据集。为确保实验结果的准确性，本实验所使用的数据集从数据来源的科学性、数据类型的多样性、实验数据分割策略等多个角度进行了考虑。

首先实验中使用的数据集来自权威部门公开的药品说明书，包括从各大药品厂商官网或国家药监局(NMPA)公开的药品说明书中提取的数据。采集的说明书数据总共涉及 13,639 种药品，确保了研究数据的科学性。

其次实验数据集中数据类型具有多样性，采集的每种药品数据包括药品的名称、成分、适应症、用法用量、禁忌症、注意事项、药物相互作用等多项内容，确保了实验设计问题的丰富性和多样性。

第三，将收集到的药物数据，转化为文档知识库，该知识库包含相关药品说明文本内容(包括该药品的中文名，组分，适应症，禁忌，注意事项和不良反应)，同时进行重新排列，并从中选择一万条有详细内容描述的数据作为初始集合；接着通过 ChineseRecrusiveTextSplitter 分词工具对数据集进行分词处理，随后，将该数据集通过随机采样的方式划分为训练集、验证集和测试集，用于训练和验证 RAG 模块的检索性能。具体数据划分如下：文档总量包含 10,000 条药品数据，验证集和测试集各包含 1000 条和 120 条药品数据信息，验证集和测试集均带有可正确回答的药品数据，用于评估检索增强模块的生成答案准确率。

考虑到问答系统中会面临大量的问答题，因此数据预处理阶段还需将数据集文本处理成问答对数据，包括问题和标准答案，用于模型的知识检索模块，帮助提升模型对具体问题的回答准确性。

4.2. 问题设计

(1) 问题类型设计

为多角度比较本文提出的 PV-RAG 模型与其他大语言模型在安全用药问答领域的特点与优势，本文实验部分设计了一套多类型的问题集合，题型包括单选、多选、判断和问答四类，涵盖了从简单选择到复杂推理的多种任务场景，旨在对比测试各模型针对不同类型问题的理解与检索能力以及生成准确回答的能力。

通过四种类型问题的组合，可以有效覆盖药品问答系统中可能遇到的多种实际应用场景，确保实验结果的全面性和代表性。实验设计的各类型问题数量占比如表 1 所示。

Table 1. Proportion of different question types
表 1. 各类型问题数量占比

问题类别	问题类型	数量(占比)
客观题	单选题	60 (40%)
	多选题	40 (26.7%)
	判断题	15 (10%)
主观题	问答题	35 (23.3%)

(2) 问题内容设计

为尽可能覆盖更多的用药安全知识，问题的内容设计主要包括以下四个方向：

① 人群适用性问题

探讨哪些人群可以安全使用特定药品，哪些人群使用药品会产生禁忌反应等等。例如：

问题 1：“哪类人群禁用头孢氨苄干混悬剂？”

问题 2：“阿司匹林适合哪些年龄段的患者使用？”

问题 3：“孕妇需要慎用下列哪些药品？”

② 药品使用禁忌问题

了解药品的禁忌症和不适合使用的情况。例如：

问题 4：“关于溴莫尼定噻吗洛尔滴眼液，对于精神抑郁，大脑功能受损，冠状动脉病变，雷诺氏症，体位性低血压，血栓闭塞性脉管炎的患者是否均应慎用？”

问题 5：“对于药品普光，尿碱化剂可减低该药品在尿中的溶解度，导致结晶尿和肾毒性的说法是否正确？”

③ 药品使用的注意事项问题

关注使用药品时需要特别注意的事项。例如：

问题 6：“茜芷胶囊和茜芷片的主要成分都是川牛膝、三七、茜草、白芷，在服用该药品时会产生哪些不良反应？”

问题 7：“使用药品金磐嗪是否会加强具有潜在肾毒性药物的毒性作用？”

④ 药品适用症状问题

询问药品适用于哪些具体症状或疾病。例如：

问题 8：“对于药品活心丸，该药品主治胸痹，心痛，用于冠心病、心绞痛的说法是否正确?”

问题 9：“科泰复用于治疗哪些症状？”

5. 实验及结果分析

5.1. 对比实验模型选取

为更好地评估构建的 PV-RAG 模型在各类问答场景中的表现，本文设计将 PV-RAG 与目前国内中文领域使用较多的智谱清言 ChatGLM-4、文心一言-3.5、讯飞星火-V4.0 Turbo、通义千问-2.5 这四种大语言模型进行对比性测试(后文仅使用模型名称，省略版本)。

5.2. 问答流程

实验问题集根据前文设计的四个问题方向的四种问题类型进行提问，并在各个大模型上依次提出相同的问题。问答流程如下：

(1) 输入问题：依次向每个大模型提出相同的问题，确保输入问题一致性。

(2) 获取模型回答：记录每个模型的回答结果，并将其与标准答案进行比对。

(3) 结果记录：每个模型的回答将按照问题的不同类型进行分类保存，用于后续分析。

5.3. 评估指标

(1) 客观题评估

对于安全用药智能问答系统而言，答案的准确性是其性能评估的最直观指标，因此在本实验中，对于客观题直接选取答案的准确率(Accuracy ACC)作为大模型性能的评估指标。

客观题中的单选题和判断题，答案是唯一的，模型的回答要么正确，要么错误，因此准确率可以通过直接计算模型的正确回答次数来获得，其计算如公式(5)所示。其中 nums 表示要评估的实例数，correct answers 为正确响应数。

$$ACC_{\text{single}} = \frac{\text{correct answers}}{\text{nums}} \tag{5}$$

对于多选题，答案可能包含多个选项，准确率使用包容性匹配的方法进行计算。设定每道题满分为100 分，若模型回答存在错误选项则该题得分为 0；若模型未选错但未完全选全，则按照公式(6)计算该题得分：

$$Score_i = \frac{\text{Number of Correct Options}_i}{\text{Total Correct Options}_i} \times 100 \tag{6}$$

根据得分，多选题的准确率计算如公式(7)所示。这种方式确保了对多选题回答的细化评估，能够反映模型对复杂问题的处理能力。

$$ACC_{multi} = \frac{\sum_{i=1}^{N} Score_i}{N \times 100} \tag{7}$$

(2) 主观题评估

对于主观题，考虑到答案中各药物名称的专业性，以及各类不良反应、用药禁忌上文字表达的多样性，采用人工方式计算平均准确率与平均冗余度两项数据作为检测大模型性能的评估指标，以确保评估的全面性。

Table 2. Subjective QA examples from major LLMs
表 2. 各大语言模型主观题问答示例

大语言模型	回答
PV-RAG	艾克儿的不良反应包括：皮疹、药物热和哮喘等过敏反应，贫血、血小板减少、嗜酸性粒细胞增多，血清氨基转移酶轻度增高，以及恶心、呕吐、腹泻及假膜性肠炎等胃肠道反应，以及由念珠菌或耐药菌引起的二重感染。
智谱清言	皮肤反应：包括皮疹、干燥、瘙痒等，有时可能会出现严重的皮肤反应，如剥脱性皮炎。 消化系统反应：如腹泻、恶心、呕吐、口腔炎、便秘等。 肝脏问题：可能会出现转氨酶升高、胆红素升高等肝功能异常。 肌肉骨骼症状：如关节痛、肌肉痛等。 骨髓抑制：可能会导致白细胞、红细胞和血小板计数降低，增加感染和出血的风险。 呼吸系统反应：如咳嗽、呼吸困难等。 神经系统症状：如头痛、感觉异常、失眠、抑郁等。 心脏问题：罕见情况下，可能会出现心律失常。 在使用艾克儿治疗期间，患者应密切关注自身症状，如有异常应及时就医。
文心一言	胃肠道反应：少数患者可能出现轻度的恶心、呕吐和腹泻等胃肠道副作用。这些副作用通常较为轻微，且用餐前服用可能会减少不适。 过敏反应：偶见荨麻疹和麻疹样皮疹。若发生荨麻疹或严重的麻疹样皮疹，应立即停止使用该药物。 肝功能异常：极少数患者可能出现暂时性的肝功能异常。因此，在使用该药物期间，患者应定期监测肝功能。
讯飞星火	1. 血液系统毒性：可能导致粒细胞减少和中度贫血。 2. 周围神经毒性反应：通常表现为深腱反射降低，麻木较少见，偶尔会有感觉异常，长期用药可能导致下肢无力。 3. 消化道自主神经系统：主要是肠麻痹引起的便秘，麻痹性肠梗阻罕见，偶尔会出现恶心呕吐。 4. 消化系统：可能引起呼吸困难或支气管痉挛，这些症状可能在注药后数分钟或数小时内发生。 5. 其他不良反应：包括进行性中度脱发、下颌痛、局部静脉炎等。
通义千问	如果您指的是特定药物的商品名或通用名，请提供更多的信息或者确认药物的正确名称，以便能够给出准确的回答。如果是处方药，通常的不良反应会在药物说明书中有详细的记载，包括可能的副作用如头痛、恶心、皮疹等，严重程度不一。对于任何药物，如果出现不良反应，应立即咨询医生或药剂师。

以常用药物阿莫西林克拉维酸钾咀嚼片(艾克尔)为例，药品说明书标注的不良反应包括：① 恶心、呕吐、腹泻及假膜性肠炎等胃肠道反应。② 皮疹、药物热和哮喘等过敏反应。③ 贫血、血小板减少、嗜酸性粒细胞增多等。④ 血清氨基转移酶可轻度增高。⑤ 由念珠菌或耐药菌引起的二重感染。⑥ 偶见兴奋、焦虑、失眠、头晕以及行为异常等中枢神经系统症状。表 2 给出各种大语言模型针对问题“使用药物艾克儿的不良反应有哪些？”生成的回答。

从表 2 的回答示例中可以看到，不同模型给出的回答差异性很大，有些模型的回答超出了药物说明书标注的范围，其真实性和可靠性存在疑问，不排除回答幻觉的可能。因此，针对主观问答题答案的准确率主要通过人工评测的方式衡量，由人工根据药品说明书的专业知识对模型生成的答案进行计算评分，包括答案的正确性、完整性和相关性。单个问答准确率计算方法为大模型生成答案与标准答案相符的条目数占标准答案总条目数的比例；所有问答题准确率取均值即为平均准确率。该指标能够科学地量化模型生成答案的精确性，反映模型在药品说明书问答任务中对关键信息的捕捉能力。其计算如公式(8)所示：

$$AverageACC = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{Number of Correct Matches}_i}{\text{Total Number of Reference Answers}_i} \tag{8}$$

数据冗余度则用于评估模型生成答案的简洁性和信息密度，通过计算答案中重复或无关信息的比例来衡量，旨在避免模型生成冗长但信息量低的回答。冗余度的计算基于模型生成答案中重复或冗余条目的数量与总生成条目数的比例。冗余度的值越低，表明模型生成的答案越简洁且信息密度越高；反之，冗余度的值越高，则表明模型生成的答案中包含较多重复或无关信息，可能导致信息冗余和用户阅读效率降低。其计算如公式(9)所示：

$$\text{Re}\,dundancy = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{Number of Irrelevant Entries}_i}{\text{Total Number of Generated Entries}_i} \tag{9}$$

5.4. 结果对比分析

基于上述评估指标，将构建的 PV-RAG 模型与目前主流的多种大语言模型进行实验对比。各模型在客观题与主观题上的问答实验对比数据见表 3 和表 4 所示。

Table 3. Accuracy data comparison in objective QA Tests
表 3. 客观题问答实验 ACC 数据对比

	PV-RAG	智谱清言	文心一言	讯飞星火	通义千问
单选题	0.7049	0.6500	0.6230	0.6721	0.5902
判断题	0.8750	0.5000	0.7500	0.6250	0.7500
多选题	0.6236	0.2966	0.5291	0.2700	0.5925

Table 4. Comparative analysis of subjective QA experiment data
表 4. 主观题问答实验数据对比

评估指标	PV-RAG	智谱清言	文心一言	讯飞星火	通义千问
平均准确率	0.8209	0.6404	0.8380	0.7153	0.6726
平均冗余度	0	0.6274	0.3460	0.3109	0.4459

表 3 展示了这五种大模型在三种客观题上的问答表现。在单选题和判断题的准确率上，PV-RAG 模型的表现较优，分别达到 0.7049 和 0.875，得益于其较强的上下文理解能力以及丰富的专业数据作为依

据，在回答准确率上表现出较好的稳定性。由此可以看到在拥有较高质量的知识库后，模型能够较为准确地识别最合适的答案，相比于其他大模型而言，检索准确率的提升较为显著。对于需要依赖多维信息进行综合的多选题，对大模型的综合检索以及辨析能力提出了更苛刻的要求，从实验结果上来看，PV-RAG 模型在多选题上的准确率为 0.6236，与单选和判断两种题型相比正确率有所下降，但对比其他大语言模型仍为最高，展现了较好的全面性。

表 4 给出了对于相同主观题在五种大模型上得到的问答平均准确率与平均冗余度对比数据。从实验结果来看，PV-RAG 和文心一言大模型在平均准确率方面表现明显优于其他模型。这表明 RAG 技术有效增强了检索能力，使其在主观问答任务中生成更准确的答案。同时，PV-RAG 的平均冗余度为 0，说明其生成答案更加精炼，避免了不必要的信息冗余，这是相比其他模型的一大优势。相比之下，智谱清言在无 RAG 的情况下准确率明显下降，且冗余度在所有模型中最高，达到了 0.6274，说明在缺乏外部知识检索增强的情况下，该模型容易生成大量与问题无关或重复的内容。其他大模型，如讯飞星火和通义千问，两方面表现居中，虽然能够较好地回答问题，但仍然存在较多的信息冗余问题。

值得一提的是，在实验过程中，除 PV-RAG 外的四种大数据模型均在问答中出现了幻觉现象，即模型会在没有相关信息支持的情况下，随意生成错误答案，例如智谱清言给出的问答题答案超过一半为不真实的数据。虽然幻觉问题已成为大语言模型的共性问题，这与大语言模型架构的固有局限有着密切关联，但如果发生在对身体健康影响重大的用药咨询问答情景中，则会给民众造成严重误导，给民众的生命安全带来一定威胁。

从实验结果来看，在拥有丰富数据集的情况下，PV-RAG 所具有的数据检索增强能力较于其他国内大模型而言具有明显的搜索准确性与优越性，一方面由于其检索动态选择不同的文档或信息片段，它能够生成较为准确的答案，另一方面在检索过程提供了信息来源的追溯功能。通过查看检索到的文档片段，用户或开发者可以清楚地看到生成内容背后的数据依据，这增强了模型的透明性和可解释性，可以有效地避免大模型幻觉而制造大量错误信息。这一实验结果进一步验证了 RAG 在主观问答任务中的适用性和优势，尤其在需要精准检索和高质量生成的医药咨询场景下具有重要应用价值。

6. 总结与展望

本文构建了面向药物警戒领域的 PV-RAG 大语言模型，结合检索增强生成技术实现安全用药智能问答系统，为有效提升安全用药问答系统的可靠性提供了有效的解决方案。实验结果表明，PV-RAG 在药品禁忌问题处理上，相较于基础的 LLM 模型，展现了更高的准确性和知识覆盖范围，特别在动态检索和查询改写机制的支持下，提升了查询效率和答案质量。系统同时表现出较强的适应性，能够处理语料多样化问题，尤其是在涉及多重因素的用药安全查询问答中，与其他大语言模型相比，展现出较高的准确率和低冗余度。为大语言模型技术在药物警戒领域的广泛应用提供了一定的参考依据。

然而，PV-RAG 仍存在一定局限性，主要体现在模型采用 RAG 架构设计，严重依赖权威机构提供的药品说明书和相关医学数据的质量，如果数据缺失或者不能及时更新药品说明书，同样都会引发回答幻觉问题。例如我国每年批准上市的新研药物多达几十上百种，部分创新药(如肿瘤靶向药)因加速审批，说明书适应症可能仅覆盖临床试验人群，未及时扩展至真实世界验证的新适应症，造成药品说明书信息不完整或滞后；而对于在我国使用面非常广泛的中药和中成药，说明书中的不良反应信息往往缺失。此外该模型在跨领域适应性、复杂医学推理能力以及系统的可解释性等方面也存在不足。

为了进一步提升系统性能，未来的工作将重点聚焦于多源数据融合和自适应更新机制的构建。首先要增强跨领域知识融合与推理能力，提高模型的可解释性，解决大规模应用中对单一数据集的严重依赖问题，例如可以嵌入知识图谱，引入结构化知识进行约束。其次要打破纯文本训练的局限，向多模态感

知、符号–神经结合和具身认知方向发展，不断缓解大语言模型的幻觉问题。随着全球化应用的需求，提升多语言和跨文化适应性和加强数据隐私保护也将是医药大模型未来的重要研究方向。

基金项目

国家社科基金重点项目“面向药物警戒的领域知识库构建与应用研究”(项目编号：23ATQ009)；江苏省社科基金一般项目“基于图谱融合的突发公共卫生事件跨领域影响机制研究”(项目编号：23TQB007)；大学生创新创业训练计划项目“基于 RAG 的药物警戒知识增强模型研究”(项目编号：202410293135Y)；大学生创新创业训练计划项目“面向中药药物警戒的知识库构建及应用研究”(项目编号：202410293134Y)。

参考文献

[1] Patil, R. and Gudivada, V. (2024) A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Applied Sciences*, **14**, Article No. 2074.

[2] Hadi, M.U., Qureshi, R., Shah, A., *et al*. (2023) Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects.

[3] 国家药品不良反应监测中心. 国家药品不良反应监测年度报告(2023 年) [EB/OL]. 国家药品监督管理局药品评价中心网站, 2024. https://www.cdr-adr.org.cn/drug_1/aqjs_1/drug_aqjs_sjbg/202403/t20240326_50614.html, 2025-04-02.

[4] 刘泽垣, 王鹏江, 宋晓斌, 等. 大语言模型的幻觉问题研究综述[J]. 软件学报, 2025(3): 1152-1185.

[5] 赵静, 汤文玉, 霍钰, 等. 大模型检索增强生成(RAG)技术浅析[J]. 中国信息化, 2024(10): 71-72+70.

[6] Gao, Y., Xiong, Y., Gao, X., *et al*. (2023) Retrieval-Augmented Generation for Large Language Models: A Survey.

[7] 祝晓雨, 张伟光, 孙树森, 等. 中国药物警戒的发展及文献计量分析[J]. 医药导报, 2019, 38(6): 820-825.

[8] 王青宇, 陈壮. 四大国际药物警戒数据库概述及其应用[J]. 药物不良反应杂志, 2023, 25(8): 497-503.

[9] Miao, J., Thongprayoon, C., Suppadungsuk, S., Garcia Valencia, O.A. and Cheungpasitporn, W. (2024) Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. *Medicina*, **60**, Article No. 445. https://doi.org/10.3390/medicina60030445

[10] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., *et al*. (2023) Large Language Models Encode Clinical Knowledge. *Nature*, **620**, 172-180. https://doi.org/10.1038/s41586-023-06291-2

[11] Wei, J., Yang, C., Song, X., *et al*. (2024) Long-Form Factuality in Large Language Models.

[12] Yang, H., Li, S. and Gonçalves, T. (2024) Enhancing Biomedical Question Answering with Large Language Models. *Information*, **15**, Article No. 494. https://doi.org/10.3390/info15080494

[13] Yagnik, N., Jhaveri, J., Sharma, V., *et al*. (2024) Medlm: Exploring Language Models for Medical Question Answering Systems.

[14] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., *et al*. (2020) Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), November 2020, 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[15] 张康, 陈明, 顾凡. 自编码器预训练和多表征交互的段落重排序模型[J]. 计算机应用研究, 2023, 40(12): 3643-3650.

[16] 李洁, 李正芳, 邹垚, 等. 大语言模型多语言词对齐能力评测方法[J]. 西南民族大学学报(自然科学版), 2024, 50(6): 681-688.

[17] Peng, W., Li, G., Jiang, Y., Wang, Z., Ou, D., Zeng, X., *et al*. (2024). Large Language Model Based Long-Tail Query Rewriting in Taobao Search. *Companion Proceedings of the ACM Web Conference* 2024, Singapore, 13-17 May 2024, 20-28. https://doi.org/10.1145/3589335.3648298