自适应混合检索增强大模型的农作物病虫害智能问答方法

杨 俊[1], 杨婉霞[1*], 杨 森[1], 何 亮[2,3], 张 娣[1]

(1. 甘肃农业大学 机电工程学院,甘肃兰州 730070,中国; 2. 新疆大学 计算机科学与技术学院,新疆乌鲁木齐 830046,中国; 3. 清华大学 电子工程系/北京信息科学与技术国家研究中心,北京 100084,中国)

摘 要: **[目的/意义]** 充分发挥隐含在农业大数据中的分散、异构和无关联农业知识的潜在应用价值,通过构建知识库,结合检索技术用于增强大模型输出专业的农业知识,为促进农业知识快速服务于生产实践提供有效手段。**[方法]** 提出了检索增强大模型的农作物病虫害智能问答方法,该方法通过自建知识库并协同优化分块策略、自适应检索机制与结构化提示工程,实现了农业病虫害领域知识有效增强大模型的精准专业问答。具体提出了自适应混合检索增强生成方法(Retrieval-Augmented Generation, RAG),首先在固定长度分块时引入重叠机制缓解语义割裂,同时,采用向量语义相似度匹配与主题高度相关的文本分块加以存储。根据问题复杂度设计了动态路由的单跳(BM25算法)检索与多跳检索。然后将文本方法与多种基线方法在不同查询类型和不同复杂度查询等多方面进行了对比实验。**[结果和讨论]** 本研究方法在Qwen1.5-7B-Chat模型上的效果最佳,准确率达到89.6%,对单跳与多跳查询的准确率分别达到0.921和0.748,较Self-RAG与Adaptive-RAG多跳查询的准确率分别提升0.082和0.059,说明本研究方法能更好地推理多跳等复杂查询。**[结论]** 本研究方法在生成答案的准确性、相关性和全面性方面具有显著优势。未来的工作将探索融合多模态知识库。

关键词: 自适应混合检索;文本分块;病虫害;智能问答

中图分类号: TP391.1 **文献标识码:** A **文章编号:** SA202506026

引用格式:杨俊, 杨婉霞, 杨森, 何亮, 张娣. 自适应混合检索增强大模型的农作物病虫害智能问答方法[J]. 智慧农业(中英文), 2026, 8(1): 52-61. DOI: 10.12133/j.smartag.SA202506026

YANG Jun, YANG Wanxia, YANG Sen, HE Liang, ZHANG Di. Intelligent Q&A Method for Crop Diseases and Pests Using LLM Augmented by Adaptive Hybrid Retrieval[J]. Smart Agriculture, 2026, 8(1): 52-61. DOI: 10.12133/j.smartag.SA202506026 (in Chinese with English abstract)

0 引 言

全球气候变暖和极端天气促进了农作物病虫害的发生,扩大了影响范围,使得传统依赖经验的防治方式面临严峻挑战[1]。在以大数据、互联网为基座的信息化时代[2],伴随着人工智能的再次崛起,知识成为推动大数据-大模型和知识服务协调发展的交合力。为此,从海量分散、异构和无序的农业大数据中提取潜在的极具价值的知识,将其关联和结构化,并结合大模型智能问答使其快速应用于农业生产实践,服务于农业各主体,快速推进农业生产的科学化和精准化是现代农业发展的必然趋势[3]。现有农业问答系统多基于关键词匹配或简单规则库[4],其缺陷是对复杂症状的语义理解不足,且通用大语言模型在垂直领域应用中也暴露出明显缺陷,如训练数据覆盖不全引发事实性幻觉、多跳推理能力薄弱导致复杂查询结果不好。为此,本研究首先自建了专业且具有高质量的典型农作物病虫害知识库,采用检索技术拓展了大模型的外源知识,确保大模型生成内容的时效性、领域性和准确性。

事实证明,检索增强生成(Retrieval-Augmented Generation, RAG)极大地拓宽了大模型的知识边界,提升了大模型生成输出内容的可靠性与可溯源性[5]。自2020年LEWIS等[6]首次提出RAG框架,研究者们通过多种方法优化RAG,并在问答

收稿日期:2025-06-13

基金项目:新一代人工智能国家科技重大专项(2022ZD0115801)

作者简介:杨 俊,硕士研究生,研究方向为自然语言处理。E-mail: 2412399856@qq.com

*通信作者:杨婉霞,博士,教授,研究方向为农业信息化。E-mail: yangwanxia@163.com

copyright©2026 by the authors (登录www.smartag.net.cn免费获取电子版全文)

系统、推荐系统，以及教育咨询等诸多领域落地应用。以问答系统为例，Google Research团队[7]构建了跨模态联合检索框架，通过同步检索文本与图像知识，实现多模态场景知识推理，该方法计算资源需求高，标注成本制约了落地泛化性。CHEN等[8]通过从数据库中检索相关代码段及摘要，为生成模型提供丰富的补充信息，显著提升了代码与摘要生成模型的性能表现。随着研究深入，GraphRAG创新性地引入知识图谱（Knowledge Graph, KG），利用图的连接性来扩展检索的语义信息，其核心是将知识图谱的结构化关系嵌入到模型的生成过程中，利用知识增强大语言模型（Large Language Model, LLM）的推理能力，解决现有RAG应对复杂查询和多跳推理的不足[9]。相对而言，该方法资源消耗较大，很难应用于资源受限的场景。为了实现知识整合任务，受海马体记忆索引理论的启发，HippoRAG模拟人脑的长期记忆检索过程，基于LLM从语料中构建KG作为人工海马体索引[10]。对于新的查询，HippoRAG识别其中的关键词后在KG上运行个性化PageRank（PR）算法，以实现跨段落的信息检索，本质上是在单次检索步骤中执行多跳推理。该方法的难点在于记忆系统识别信息的完整性，实现过程复杂，应用存在一定难度。SelfRAG引入可训练的检索控制器，通过自省机制评估检索结果与生成内容质量，该方法通过生成后再修正抑制幻觉，提升事实一致性，但其系统复杂度较高[11]。AdaptiveRAG对于查询复杂度分类，自适应选择分层检索策略与资源分配模式，实现了高效响应和灵活的数据处理能力[12-14]。

由上述分析可知，检索增强生成技术通过增加知识库，优化检索算法，提高了LLM生成的效果和效率。然而，现有RAG方法在农业领域的研究不足主要体现在：农业领域知识库构建的规模不够，质量不高，关键为多源实体对齐准确率不高；在检索策略层面，固定分块策略破坏了农业文本语义的连贯性，单一检索算法难以兼顾单跳查询的精确匹配与多跳查询的语义泛化需求；生成层面，专业领域知识与LLM的协同效果不佳[15-17]。为此，本研究提出基于知识检索增强大模型的农作物病虫害知识智能问答方法，通过自建高质量专业知识库，设计自适应混合检索（Adaptive Hybrid Retrieval, AHR）方法和结构化提示工程引导模型分步推理，增强了模型生成答案的专业性与逻辑性，解决了现有方法无法处理复杂查询的问题。

1　材料与方法

1.1　数据收集

本研究主要采集了小麦、水稻、玉米、马铃薯和棉花共5种典型农作物病虫害及防治数据，其来源涵盖科研文献库、农业政务平台、技术论坛及网页等。主要通过优化网络爬虫工具中的XPath或CSS（Cascading Style Sheets）选择器[18]，精确锁定并提取网页等与病虫害相关的文本信息。然后运用人工校验方法精准定位与剔除异常值，确保数据的有效性，最终形成了87 901条非结构化数据，为后续数据分块存储、检索，以及从数据中抽取知识奠定了基础。

1.2　知识库构建

知识库构建流程如图1所示。其中最关键环节为半结构化和非结构化数据的知识抽取和实体对齐。为了规范化、标准化组织知识，首先基于本体论构建了病虫害知识的模式层（如图2所示），其中主要定义了病虫害领域概念、实体、属性，以及它们之间的关系，为结构复杂、动态变化和多维度关联的农业病虫害领域知识提供专业的表示框架。数据标注与知识抽取是知识库构建中获取结构化知识的关键步骤，采用操作简便、功能强大的doccano标注工具完成少量样本标注。知识抽取使用具有强大泛化能力的小样本学习模型通用信息抽取技术UIE（Universal Information Extraction）[19]，无需大量标注数据，只需40个标注样本微调模型便可抽取模式层定义的实体等。然后结合Bert向量语义与大模型提示工程，实现了知识库中实体的精准对齐。首先使用Bert将实体向量化后，基于余弦相似度计算实体向量之间的语义相似度，初步筛选相似度高的实体。然后基于“千问”大模型（https://www.tongyi.com/）设计少样本提示词（few-shot prompts），借助提示词使模型深入挖掘实体属性与类别标签，全面理解实体之间的语义关联，实现了知识库中实体的精准对齐，最终获取了916 239条三元组知识库（如“小麦锈病，发病症状，叶片上出现锈红色斑点”），部分抽取结果如表1所示。从数据的采集、数据预处理的人工校验、监督学习的知识抽取和粗细粒度的实体对齐等多环节保证了构建数据库的质量。最后分别采用MySQL[20]与Elasticsearch[21]存储了关系数据和非结构化语义数据，以满足自适应检索需求。

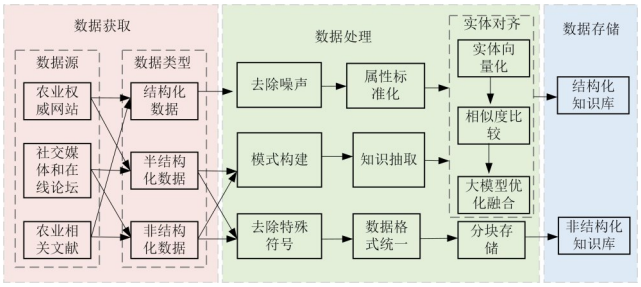


图1 农作物病虫害知识库构建流程

Fig. 1 Construction process of crop diseases and pests knowledge base

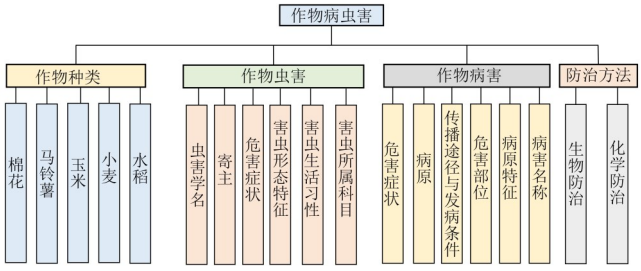


图2 农作物病虫害知识库模式层

Fig. 2 Mode layer of crop diseases and pests knowledge base

表1 农作物病虫害知识部分抽取结果展示

Table 1 Display of extraction results for crop diseases and pests knowledge

实体	属性	属性值
水稻稻瘟病	寄主	水稻
玉米黑束病	危害部位	叶片
棉花叶烧病	寄生方式	菌丝体
水稻稻瘟病	病害名称	稻瘟病
小麦锈病	危害部位	叶片、叶鞘
小麦蚜虫	危害部位	叶片、茎部
玉米矮花叶病	传播途径	蚜虫的扩散
棉花枯萎病	危害症状	叶脉褪绿变黄
水稻稻瘟病	发病条件	高温高湿环境

1.3 检索增强生成方法

本研究提出的检索增强问答生成方法的总体架构如图3所示，主要由分块存储与索引、知识库检索，以及答案生成三大核心模块构成，该方法的创新思路详述如下。

1.3.1 混合分块

由于传统单一分块方法无法应对农业病虫害领域文本语义的专业性和复杂性，本研究提出一种融合固定长度分块与语义优化的混合分块方法。该方法分为两个阶段：首先，结合重叠机制，对病虫害文本进行固定长度分块，以保证分块语义的完整性。然而分块后的文本中仍存在一些与主题无关的

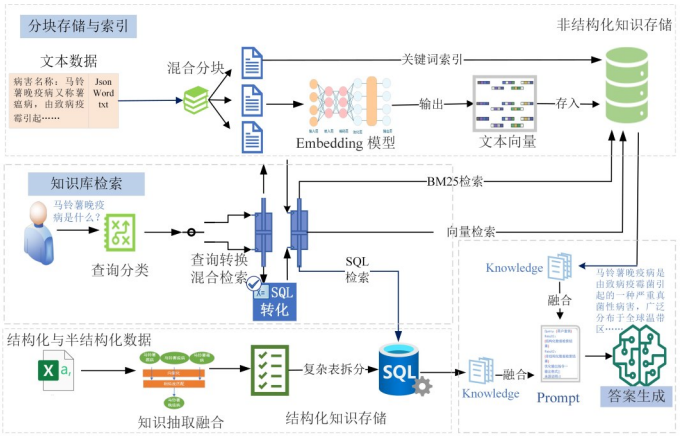


图3 检索增强问答生成方法的总体架构

Fig. 3 The overall structure of the retrieval enhanced Q&A generation method

文本块，为消除此类文本块的干扰，同时缓解存储的压力，故引入向量语义相似度匹配法对分块结果进一步优化，即精确计算分块片段向量 $\boldsymbol{V}_i=\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ 与知识库中参考向量集合 $\boldsymbol{V}_{\text{库}}=\{\boldsymbol{v}_{\text{ref}1}, \boldsymbol{v}_{\text{ref}2}, \ldots, \boldsymbol{v}_{\text{ref}n}\}$ 的最大余弦相似度，确保分块文本与专业知识的相关性，其计算如公式（1）所示。

$$\text{sim}_{\max}=\max_{\boldsymbol{v}_{\text{ref}}\in \boldsymbol{V}_{\text{库}}}\left(\frac{\boldsymbol{v}_i\cdot\boldsymbol{v}_{\text{ref}i}}{\|\boldsymbol{v}_i\|\cdot\|\boldsymbol{v}_{\text{ref}i}\|}\right) \tag{1}$$

式中：$\|\boldsymbol{v}_i\|$ 表示第 i 个向量的模值，设定阈值 $\tau\in[0, 1]$，若最大相似度低于阈值即 $\text{sim}_i\leqslant\tau$，则判定为不相关并剔除。

1.3.2 检索推理

本研究知识推理主要包含查询分类等三个关键环节，具体详述如下：

1）查询分类。为提高查询效率，依据农作物病害症状词典和逻辑表达模板，将查询问题划分为单跳查询与多跳查询。其主要是通过对问题语义的精准提取与分类实现。即首先将口语化问题映射至预定义的农业专业术语（如“叶子长毛”：“菌丝体滋生”），以消除方言差异对语义理解的干扰。为此，本研究设置了120组这样的语义映射。其次基于三元组实体关系特点，构建了312条单跳查询触发词规则库和相应的模板（如表2），将能从模板中匹配到的查询定义为单跳查询，快速定位问题类型。最后通过轻量级规则树终裁：优先匹配触发词，无触发词时使用UIE模型判断实体数量，根据实体数量划分为单一实体归单跳和多实体/关系归多跳。

2）单跳查询的检索过程。本研究提出的自适

表2　农作物病虫害研究单跳查询模板

Table 2　Single-hop query template for crop diseases and pests research

模板类型	表达式示例	用户输入示例
症状解决	{作物}{症状}怎么办?	小麦叶锈病怎么办?
病虫害	{症状}是什么病?	水稻叶子上的白点是什么病?
方法询问	{病害}用什么药?	小麦赤霉病用什么药?
传播途径	{病虫害}是如何传播的?	稻瘟病是如何传播的?
环境影响	{病虫害}与哪些环境因素有关?	稻瘟病与哪些环境因素有关?

应混合检索根据问题的复杂程度实现单跳和多跳分类查询。单跳查询的检索过程主要分为三个步骤。首先，从问题中提取关键信息。然后，运用BM25算法将提取的信息与Elasticsearch索引所构建的文档内容进行匹配。最后，将匹配结果输入大模型用于生成答案。BM25算法针对查询中的每个词项，独立计算其在各文档中的得分并累加，生成问题与查询结果的综合相关性得分。得分越高，则文档与查询的相关性越强。BM25算法对查询Q与文档D的相关性评分计算如公式（2）所示。

$$\sum Score(D,Q)=\sum_{i=1}^{n} IDF(r_i)\cdot \frac{(k_1+1)\cdot f(r_i,D)}{f(r_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{avgdl}\right)} \tag{2}$$

式中：给定问题$\boldsymbol{R}$中包含词r_1，r_2，r_3，…，r_n；k_1和b是调节参数；k_1是一个正系数，用于控制词频的饱和度，较高的k_1值意味着词频对评分的影响更大；b是用于控制文档长度对评分的影响的参数，取值在0到1之间，当$b=1$时，文档长度的影响最大，当$b=0$时，文档长度不影响评分；$IDF(r_i)$是每个词r_i的逆文档频率（计算如公式（3）所示），$f(r_i, D)$是词r_i在文档D中的频率；$|D|$是文档D的长度；$avgdl$是所有文档片段的平均长度[22]。

$$IDF(r_i)=ln\left(\frac{N-n(r_i)+0.5}{n(r_i)+0.5}+1\right) \tag{3}$$

式中：N表示文本片段集合的总数量；$n(r_i)$表示r_i的文档数。

3）多跳查询的检索过程。本研究的多跳查询检索策略如图4所示，首先进行查询解析转换，即将用户输入的查询转换为结构化条件（用于结构化查询语言（Structured Query Language, SQL）检索）与语义向量表示（用于向量检索（Dense Vector Retrieval, DVR））， 然后在不同的知识库中进行检索，最后构建最佳检索结果列表用于增强大模型的理解。其中，SQL检索的核心机制在于将数据请求转化为逻辑严密的指令，例如，若需查询2023年华北地区小麦锈病的发生记录，可构建查询语句：SELECT pest_name，location，date_reported FROM pest_outbreaks WHERE crop_type='小麦' AND pest_name='锈病' AND location LIKE '%华北%' AND YEAR（date_reported）=2023。数据库引擎通过B+Tree索引快速定位满足条件的记录，执行过程为：首先解析WHERE条件为布尔表达式树，其次利用索引筛选候选文档，然后对中间结果进行连接或聚合运算，实现在关系型数据库中数据的精确定位。DVR[23]是通过一个预训练的语言模型将查询文本转换成固定长度的向量表示，然后使用余弦相似度判断查询文本与知识库文本的相似性。其向量化过程的数学表征可形式化定义为：给定一段文本T，通过编码器将其映射到一个稠密向量如公式（4）所示，其中$\boldsymbol{d}$是向量维度。具体来说，对于输入序列，编码器输出每个token对应的隐藏状态向量。通常取［CLS］标记对应的向量作为整个序列的表示。

$$\boldsymbol{v}_T=\text{Encoder}(x)_{[\text{CLS}]} \tag{4}$$

式中：Encoder表示编码器函数，代表输入文本的token序列；$\boldsymbol{v}_T$为文本编码后的语义向量。这些向量捕捉了文本的语义信息，使得具有相似意义但不完全相同的词语或句子能够在向量空间中靠近。

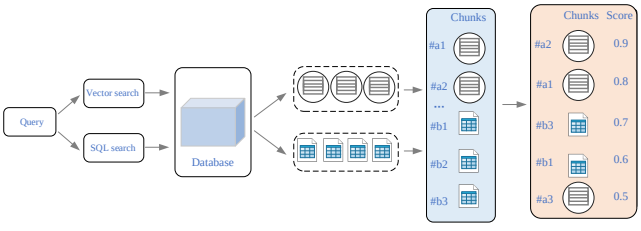


图4　多跳查询检索方式

Fig. 4　Multi-hop query retrieval method

本研究根据查询问题的解析转换自适应的选择SQL或DVR检索通道，这种混合检索充分利用了SQL检索的准确定位性和DVR有效捕捉语义关联性，以更好地处理复杂逻辑的嵌套关系。例如，在查询关于马铃薯早疫病与晚疫病的症状区别及防治意见时，可以通过SQL语句直接访问包含疾病名称和症状描述等字段，提供一些精确简短的关键信

息。DVR可检索与“马铃薯早疫病与晚疫病的症状区别”相关片段，在提供关键信息的基础上增加了症状的相关背景信息。

对于多跳查询的检索结果，若将存在差异性的SQL检索与DVR结果直接合并会导致信息冗余与逻辑断裂。在此，本研究采用倒数排名融合RRF（Reciprocal Rank Fusion）[24]算法对SQL检索与DVR结果进行排序融合。该算法无需预先训练或调整权重参数即可实现异构检索结果的协同优化，其计算如公式（5）所示。

$$score(d) = \sum_{q \in Q} \frac{1}{k + rank_q(d)} \tag{5}$$

式中：Q表示检索方式类型；$rank_q(d)$为文档d在第q种检索结果中的排名；k为调节参数，常称其为平滑因子。RRF算法的排序结果是排名靠前的文档贡献显著权重，越靠后的排名，权重贡献度越低。

1.4 大模型生成

大模型答案生成过程分上下文组织和模型的推理引导。上下文组织的核心是将检索到的相关文档片段按其与问题的相关性从高到低进行二次排序，从而识别并选择能够最大程度满足问题需求的相关文档。然后进行关键片段精炼，即在选定的相关文档片段中识别最具信息量和指导性的片段，这些片段能够直接回答问题或能提供重要的背景信息，从而提升信息的有效性和针对性。模型推理引导则包含多步推理与答案规范化两个重要环节。在多步推理过程中，通过链式[25]思维模板（prompt，如表3所示）强制模型显式分解推理过程，这种层层递进式的逻辑引导促使大语言模型逐步深入分析输入信息，深度挖掘复杂语义关联，规避思维跳跃。答案规范化环节则对模型生成的初步答案进行格式、语义等方面的规范处理，使其符合既定的输出标准，提升答案的可读性与可用性。

表3 答案生成提示词设计模板

Table 3 Answer generation prompt word design template

环节	步骤	内容
上下文组织(为模型推理提供背景等信息)	P1	你是一名农业病虫害防治专家,需基于知识库检索到的内容,依据解读的用户问题生成专业、可信的答案。 请严格遵循以下步骤生成回答:用户查询{query} **上下文组织** 读取定位相关片段并排序: - 文档1:{检索片段1} - 文档2:{检索片段2} - ……(最多保留Top K个相关片段) 根据这些片段与{query}的相关性,从高到低对文档片段重新排序,结果存入数组中如[文档2编号,文档1编号……]
	P2	**关键信息提炼** - 从上述文档中提取与问题直接相关的信息;(如症状描述、病原特征、防治方案等)
模型多步推理与答案生成	P3	**模型多步推理** 根据以下链式模板分步分析问题,确保答案逻辑严谨: 1. *症状匹配*: - 用户描述:“{用户输入症状}” - 匹配知识库症状:“{检索片段中的症状描述}” - 关联病害:“{病害名称}”(引用文档) - 关联特征:{斑点/断裂/疱状病斑/表皮破裂}(引用文档) - 关联部位:{叶片/茎/穗}(引用文档) 2. *病因推断*: - 病原类型:{真菌/细菌/病毒/害虫}(引用文档) - 诱发因素:{环境条件/种植习惯}(引用文档) - 传播途径:{昆虫/真菌/风雨}(引用文档) 3. *防治建议*: - 农业防治:{轮作/土壤处理}(引用文档) - 化学防治:{农药}推荐2~3种农药,注明用量(引用文档) - 生物防治:{天敌/微生物制剂}(引用文档)
	P4	**答案生成** 请从病虫害名称、病原、病原特征、危害部位、危害症状、发病条件、传播方式、防治方法这些方面对{query}全面分析,结果控制在200字左右;逻辑清晰,确保内容简练、无冗余。

2　结果与分析

实验数据集主要使用了自建数据集和公共数据集两类，其中自建的典型农作物病虫害问答数据集（包含200条单跳查询、200条多跳查询，以及用于测试多样化问答能力的150条比较型、150条判断型、150条选择型查询）。公共数据集为生物医学领域PubMedQA。

2.1 实验环境与评价指标

为验证本研究提出的AHR-RAG（Adaptive Hybrid Retrieval-RAG）方法性能，根据实验需求配置了相应的实验环境及参数，详细配置信息如表4所示。

表4　实验环境及参数配置信息

Table 4　Experimental environment and parameter configuration information

实验环境及参数	配置信息
操作系统	Linux Ubuntu
Python	3.10
CUDA	13.4
GPU	4*NVIDIA GeForce GTX 1080Ti
Torch	1.21.3
显存	48 G
BM25算法参数	k_1= 1.5；b=0.75
RRF算法参数	k = 60
模型回答超参数	Temperature= 0.3

为了定量化评估模型输出的回答效果，本研究选择了问答系统中常用的量化评价指标：召回率、准确率、F_1值，以及响应时间。相关计算如公式（6）～公式（8）所示。

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{6}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

式中：*Precision*为准确率；*Recall*为召回率；TP（True positive）为真正例，指模型正确识别出的病虫害信息，如准确判断病症并推荐对应防治措施；FP（False positive）为假正例，是模型误判产生的错误信息，如将一种病害误诊为另一种并提出不当防治建议；FN（False nagative）为假负例，指模型未能检索或回答本应包含的关键专业知识，例如遗漏某种重要病原或有效的环境调控措施。

2.2 检索召回量Top K确定实验

为了探究召回数量（Top K）对各检索算法性能的影响，采用单跳和多跳混合数据集，在保持其他参数不变的前提下，调整Top K值，观察各算法检索结果的准确率与召回率的变化，其实验结果如表5与图5所示。

表5　召回率与准确率随Top K的变化

Table 5　Recall and precision vary with Top K

Top K	准确率				召回率			
	AHR	DVR	SQL	BM25	AHR	SQL	DVR	BM25
1	0.607	0.554	0.452	0.407	0.523	0.487	0.412	0.376
2	0.705	0.654	0.559	0.501	0.634	0.598	0.521	0.463
3	0.756	0.703	0.608	0.554	0.692	0.649	0.584	0.531
4	0.782	0.758	0.659	0.626	0.745	0.708	0.642	0.604
5	0.822	0.774	0.706	0.679	0.791	0.743	0.726	0.657
6	0.785	0.743	0.653	0.627	0.762	0.721	0.668	0.613
7	0.764	0.698	0.621	0.582	0.733	0.683	0.637	0.561
8	0.735	0.662	0.578	0.547	0.705	0.647	0.602	0.535

分析表5可知，当Top K从1增至5时，各检索算法的性能基本都在提升，但AHR检索算法表现更佳，其召回率与准确率分别从0.523、0.607提升至0.791、0.822，平均每增加1个召回结果，召回率提升约6.7个百分点。这说明适量扩大候选集可覆盖更全面的相关文档。当Top K>5后，四种检索方法的召回率与准确率均出现下降，当Top K达到8时，AHR召回率由0.791降至0.705，准确率由0.822降至0.735。说明随着Top K的增加，引入了无关文档导致冗余数据的积累，会对LLM产生负面影响。在Top K=5的峰值点，AHR检索的召回率较DVR（最优单一检索方式）高8.95%。因此，本研究最终选择Top K=5作为检索的默认配置。

2.3 对比实验

2.3.1　基线模型对比

为验证AHR-RAG方法在垂直领域问答方面的性能，基于本研究构建的农作物病虫害数据集，选择了较为出色的Self-RAG、Adaptive-RAG等基线模型与本研究方法进行对比，结果如表6所示。

分析表6可知，AHR-RAG方法在垂直领域的问答任务中优势显著。相较于无检索策略的基线模型，AHR-RAG在Qwen1.5-7B-Chat模型上将准确率和F_1值分别提升了0.193和0.170。与主流改进方法相比，AHR-RAG在保持较低响应时间的同时，

表6 农作物病虫害智能问答研究基线模型对比

Table 6 Comparison of baseline models for Intelligent Q&A research on crop diseases and pests

RAG	Model	*Recall*	*Precision*	F_1	Time/s
No RAG	Qwen1.5-7B-Chat	0.725	0.703	0.714	1.02
Naive RAG	Qwen1.5-7B-Chat	0.751	0.735	0.743	1.13
Self-RAG	Qwen1.5-7B-Chat	0.823	0.845	0.834	3.42
Adaptive-RAG	Qwen1.5-7B-Chat	0.848	0.878	0.863	2.63
AHR-RAG	Qwen1.5-7B-Chat	0.872	0.896	0.884	2.43
AHR-RAG	GLM	0.865	0.868	0.866	2.71
AHR-RAG	Baichuan	0.867	0.857	0.862	2.68

其F_1值较Self-RAG和Adaptive-RAG分别提高0.05和0.021，准确率高达0.896，Self-RAG检索在面对复杂多变的输入时，对检索结果和生成内容的评估可能出现偏差，导致模型在调整输出时出现误判，从而影响整体性能。Adaptive-RAG通过预训练分类器动态选择检索策略，与AHR-RAG检索策略相似，但AHR-RAG增加了混合分块法与检索融合机制，提升分块质量的同时增强了向量检索与SQL检索的协同，即融合了向量检索的深层语义关联与SQL检索的结构化知识精准匹配性，有效克服了单一检索的模糊性、机械性，显著提升了对复杂农业问题回答的准确性和知识覆盖度，使得检索结果更为精准高效。另外，AHR-RAG在GLM和Baichuan等不同基座模型上依然表现稳定，验证了本方法的普适性。

为进一步评估本方法的领域迁移性与泛化能力，使用生物医学领域公共数据集PubMedQA进行对比实验，结果如表7所示。

表7 农作物病虫害智能问答研究公共数据集测试

Table 7 Test on public dataset for intelligent Q&A research on crop diseases and pests

Method	*Recall*	*Precision*	F_1	Time/s
无(RAG)	0.352	0.437	0.389	0.56
Naive-RAG	0.448	0.451	0.449	1.36
Self-RAG	0.528	0.496	0.511	3.54
Adaptive-RAG	0.601	0.562	0.581	2.46
AHR-RAG	0.709	0.583	0.640	3.23

由表7可知，AHR-RAG方法有较好的通用性，不管是农业领域还是医疗领域，AHR-RAG相较于Self-RAG与Adaptive-RAG性能均有提升。可能的原因是AHR-RAG的混合分块方法结合重叠机制充分保留上下文关联性，更好地增强模型推理的语义信息。

2.3.2 不同查询类型实验

本研究全面评估了多种检索增强生成方法对多类型查询问题的回答效果。根据农业知识问答任务的特征，将查询模式划分为比较型、判断型和选择型三类。实验采用分层抽样方法构建测试集，通过控制变量法确保各模型在相同评估条件下进行对比，实验结果如表8所示，针对选项位置偏差问题，采用均衡选项分布策略以消除选择偏好对评估结果的影响。

表8 农作物病虫害智能问答研究不同查询类型准确率测试

Table 8 Accuracy testing for different query types of intelligent Q&A research on crop diseases and pests

模型	比较型	判断型	选择型
Qwen	0.652	0.803	0.741
Naive-RAG	0.684	0.847	0.795
Self-RAG	0.725	0.882	0.836
Adaptive-RAG	0.754	0.899	0.852
AHR-RAG	0.883	0.925	0.898

分析表8可知，AHR-RAG方法在应对多样化农业问答需求方面展现出卓越的适应性，尤其在复杂知识对比与精准决策场景下优势显著。相较于模型仅利用先验知识的方法，本研究AHR-RAG方法对比较、判断、选择三种查询类型的准确率分别提升了0.231、0.123、0.157。Naive-RAG在处理复杂语义及多样化查询时能力有限。Self-RAG引入自反馈机制，能够较好地适应多变的农作物病虫害知识体系及多样性查询场景。Adaptive-RAG能依据查询复杂度动态调整检索策略，在一定程度上提高了检索适配性。而AHR-RAG之所以在三类查询问题上的性能均有领先，其原因是本方法设计了一套针对农业垂直领域问答特性的闭环（知识库构建→检索增强→答案生成）优化方案。高质量知识库为答案生成提供了专业、准确的知识原料；自适应混合检索实现了精准检索与语义关联的最优组合供给；结构化提示工程严格引导模型按领域知识逻辑推理并忠实于检索结果生成答案。

2.3.3 不同复杂度查询实验

为了全面评估AHR-RAG方法在农作物病虫害问答生成领域中的表现，本研究构建了包含不同复杂度的问答数据集，依据实体嵌套层级与逻辑连接词密度将测试集划分为单跳查询与多跳查询两类，实验结果如表9所示。

表 9　单跳查询与多跳查询准确率测试

Table 9　Accuracy test of single-hop and multi-hop queries

模型	单跳查询			多跳查询		
	Recall	*Precision*	F_1	*Recall*	*Precision*	F_1
Qwen	0.681	0.722	0.701	0.469	0.461	0.465
Naive-RAG	0.750	0.797	0.773	0.603	0.593	0.598
Self-RAG	0.843	0.882	0.862	0.678	0.666	0.672
Adaptive-RAG	0.861	0.892	0.876	0.701	0.689	0.695
AHR-RAG	0.895	0.921	0.908	0.723	0.748	0.735

由表 9 分析可得，Naive-RAG 框架受限于其端到端简单检索范式，在多跳查询中因语义解析不完整导致准确率相较于单跳查询时下降了 0.204，Self-RAG 应对多跳查询的回答有局限性。Adaptive-RAG 采用多知识源迭代检索策略，虽在单跳查询中达到 0.892 的准确率，但知识融合阶段的注意力稀释效应导致其从单跳查询时的 0.892 的准确率降到多跳查询时的 0.689。而 AHR-RAG 有效克服了复杂语义解析与知识融合难题，在单跳查询时准确率为 0.921，相较于 Adaptive-RAG 提升了 0.029；在多跳查询场景中，其准确率为 0.748。相较于 Self-RAG 和 Adaptive-RAG，准确率分别提升 0.082 和 0.059。

2.3.4　消融实验

为验证分块和检索等对本研究方法整体性能的具体贡献，在本研究构建数据集上进行了消融实验。具体设置如表 10 所示，其中 NO chunk 表示去掉混合分块，采用固定长度块分块方法，长度为 512 token，其他不变；NO Retrieval 表示去掉自适应混合检索方法，采用传统向量相似度检索，其他不变；NO Prompt 表示去掉 prompt，使用默认提示模板，其他不变；实验结果如表 10 所示。

表 10　农作物病虫害智能问答研究消融实验结果

Table 10　Results of ablation experiments of intelligent Q&A research on crop diseases and pests

模型对比组	*Recall*	*Precision*	F_1	Time/s
AHR-RAG	0.872	0.896	0.884	2.6
NO chunk	0.851	0.878	0.864	3.1
NO Retrieval	0.835	0.823	0.829	3.4
NO Prompt	0.867	0.883	0.875	2.5

根据表 10 分析可得，相较于固定长度分块方法，AHR-RAG 方法的准确率提升 0.018，F_1 值提升 0.020，响应时间缩短了 0.5 s。在检索增强生成方面，AHR-RAG 通过优化提示词策略，相比直接传递检索结果给模型输出，准确率提升了 0.013，F_1 值提升 0.009，尽管响应时间稍有增加，但生成质量得到了显著提升，生成逻辑性更强的回答、内容更丰富的答案。

2.3.5　生成问答实例对比

将本方法与千问大模型在简单和复杂问题下的回答生成示例进行了展示，具体如图 5 和图 6 所示。



图 5　AHR-RAG 方法生成农作物病虫害的答案示例

Fig. 5　Example of crop pests and diseases answers generated by the AHR-RAG method

通过对比分析发现，相较于千问大模型，本方法生成的答案在内容上更为精确与完善。即使用户提供的症状描述较为模糊，本方法的检索模块仍能通过对症状信息进行深度挖掘，筛选出高度契合的病害信息，为用户提供全面且有价值的参考。

3　结　论

本研究提出的自适应混合检索增强大模型智能问答方法的研究成果总结为以下 3 点。

1）构建了含 87 901 条非结构化文本与 916 239 条结构化三元组的规模较大的高质量多类型病虫害

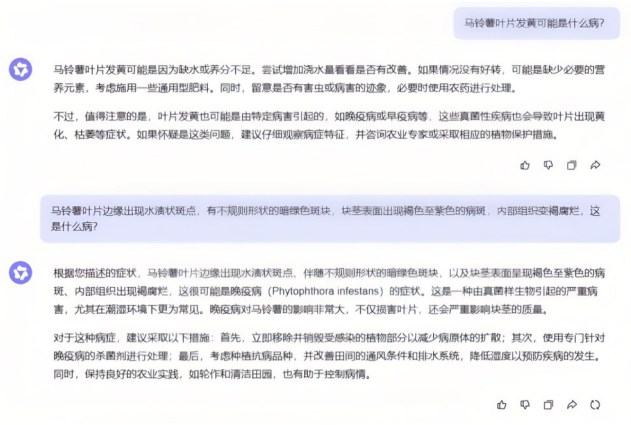


图6 千问基线模型生成农作物病虫害的答案示例

Fig. 6 Example of answers to crop diseases and pests generated by the Qianwen baseline model

知识库，为农业知识服务和模型微调提供可靠的领域知识。

2）设计的混合分块策略突破了传统分块方法在农业文本处理中的性能瓶颈，相较于常规固定长度分块法，其准确率提升0.018，F_1值提升0.020。

3）实现了动态路由检索结果和提示信息的有机融合机制，显著提升复杂查询的语义解析能力。在多跳查询时，本研究方法相较于Self-RAG和Adaptive-RAG，准确率分别提升0.082和0.059，并基于本研究方法设计实现了智能问答系统。

未来工作将围绕多模态知识融合、低资源场景优化与边缘端部署开展。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] 赵春江. 农业知识智能服务技术综述[J]. 智慧农业(中英文), 2023, 5(2): 126-148.
ZHAO C J. Agricultural knowledge intelligent service technology: a review[J]. Smart Agriculture, 2023, 5(2): 126-148.

[2] 赵泽行, 吴晓鹏, 王怡馨, 等. 基于知识图谱的农作物病虫害问答系统研究[J]. 智能化农业装备学报(中英文), 2024(4): 39-50.
ZHAO Z X, WU X P, WANG Y X, et al. Research on question answering system for crop diseases and pests based on knowledge graph[J]. Journal of Intelligent Agricultural Mechanization, 2024 (4): 39-50.

[3] YONG D, HAIFENG L, YA-NAN C, et al. Similar cluster frequency entropy: A novel uncertainty estimator for detecting large language models confabulations[J]. Chinese Journal of Electronics, 2025, 34(5):1-14.

[4] NIRANJAN P Y, RAJPUROHIT V S, SANNAKKI S S. Question answering system for agriculture domain using machine learning techniques: Literature survey and challenges[J]. International Journal of Computational Systems Engineering, 2020, 6(2): 91.

[5] MALIK S, KHAREL H, DAHIYA D S, et al. Assessing ChatGPT4 with and without retrieval-augmented generation in anticoagulation management for gastrointestinal procedures[J]. Annals of Gastroenterology, 2024, 37(5): 514-526.

[6] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[EB/OL]. arXiv: 2005.11401, 2020.

[7] ZHANG Z, WEN L, ZHAO W. Rule-KBQA: Rule-guided reasoning for complex knowledge base question answering with large language models[C]// Proceedings of the 31st International Conference on Computational Linguistics. Bangkok, Thailand: International Committee on Computational Linguistics, 2025: 8399-8399.

[8] CHEN J K, HU X, LI Z H, et al. Code search is all you need? improving code suggestions with code search[C]// 2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE). Piscataway, New Jersey, USA: IEEE, 2024: 880-892.

[9] HAN Y, YANG T, YUAN M, et al. Construction of a maritime knowledge graph using GraphRAG for entity and relationship extraction from maritime documents[J]. Journal of Computer and Communications, 2025, 13(2): 68-93.

[10] GUTIÉRREZ B J, SHU Y H, GU Y, et al. HippoRAG: Neurobiologically inspired long-term memory for large language models [EB/OL]. arXiv: 2405.14831, 2024.

[11] JEONG M, SOHN J, SUNG M, et al. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models[J]. Bioinformatics, 2024, 40(Suppl 1): i119-i129.

[12] ZHANG Y F, ZHAO D, LI H Y, et al. AdamRAG: Adaptive algorithm with ravine method for training deep neural networks[J]. Neural Processing Letters, 2025, 57(3): 53.

[13] LI Y. A dynamic knowledge base updating mechanism-based retrieval-augmented generation framework for intelligent question-and-answer systems[J]. Journal of Computer and Communications, 2025, 13(1): 41-58.

[14] TU Q S, GUO J, LI N, et al. Mitigating grand challenges in life cycle inventory modeling through the applications of large language models[J]. Environmental Science & Technology, 2024, 58 (44): 19595-19603.

[15] TILTON Z, LAVELLE J M, FORD T, et al. Artificial intelligence and the future of evaluation education: Possibilities and prototypes[J]. New Directions for Evaluation, 2023, 2023(178/179): 97-109.

[16] PERKINS G, ANDERSON N W, SPIES N C. Retrieval-augmented generation salvages poor performance from large language models in answering microbiology-specific multiple-choice questions[J]. Journal of clinical microbiology, 2025, 63(3): e01624-24.

[17] SIDDHARTH L, LUO J X. Retrieval augmented generation using engineering design knowledge[J]. Knowledge-Based Systems, 2024, 303: 112410.

[18] 于平. 基于大数据的深度学习网络爬虫算法在信息搜集与处理中的应用[J]. 科技资讯, 2024, 22(16): 55-57.
YU P. Application of deep learning web crawler algorithm based on big data in information collection and processing[J]. Science & Technology Information, 2024, 22(16): 55-57.

[19] YUAN W W, YANG W X, HE L, et al. Research on entity and relationship extraction with small training samples for cotton pests and diseases[J]. Agriculture, 2024, 14(3): 457.

[20] 张朵朵, 张雪茹, 王永杰, 等. 基于C/S架构和MySQL数据库的肉类加工信息管理系统的设计与构建[J]. 食品与生物技术学报, 2024, 43(9): 99-106.
ZHANG D D, ZHANG X R, WANG Y J, et al. Design and construction of meat processing information management system based on C/S architecture and MySQL database[J]. Journal of Food Science and Biotechnology, 2024, 43(9): 99-106.

[21] 范晓磊, 陈钊, 高金萍. Elasticsearch在林业数据领域的应用[J]. 世界林业研究, 2025, 38(1): 60-66.
FAN X L, CHEN Z, GAO J P. Application of elasticsearch in the field of forestry data[J]. World Forestry Research, 2025, 38(1): 60-66.

[22] 郑少帅, 翁境鸿, 蒋小洋. 基于BM25、文本Embeddings与交叉编码器的民航客服知识库检索研究[J]. 无线互联科技, 2023, 20 (24): 122-125.
ZHENG S S, WENG J H, JIANG X Y. Research on civil aviation

customer service knowledge base retrieval based on BM25, text vector method and cross encoder[J]. Wireless Internet Technology, 2023, 20(24): 122-125.
[23] 王苑铮, 范意兴, 陈薇, 等. 稠密向量实体检索模型的二值化提速压缩[J]. 模式识别与人工智能, 2023, 36(1): 60-69.
WANG Y Z, FAN Y X, CHEN W, et al. Binarization for speed-boosting compression of dense vector-based entity retrieval models[J]. Pattern Recognition and Artificial Intelligence, 2023, 36 (1): 60-69.
[24] 赵征宇, 罗景, 涂新辉. 基于多粒度语义融合的信息检索方法[J]. 计算机应用, 2024, 44(6): 1775-1780.
ZHAO Z Y, LUO J, TU X H. Information retrieval method based on multi-granularity semantic fusion[J]. Journal of Computer Applications, 2024, 44(6): 1775-1780.
[25] MAHMOUDI E, ERICKSON B, VAHDATI S, et al. Prompt optimization and chain of thought reasoning for automated classification of echocardiography reports using privacy-preserving open-source language models[J]. Journal of The American College of Cardiology, 2025, 85(12): 2139.

Intelligent Q&A Method for Crop Diseases and Pests Using LLM Augmented by Adaptive Hybrid Retrieval

YANG Jun[1], YANG Wanxia[1*], YANG Sen[1], HE Liang[2,3], ZHANG Di[1]

（1. *College of Electrical and Mechanical Engineering*, *Gansu Agricultural University*, *Lanzhou* 730070, *China*; 2. *School of Computer Science and Technology*, *Xinjiang University*, *Urumqi* 830046, *China*; 3. *Beijing National Research Center for Information Science and Technology/ Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*）

Abstract: [Objective] Extracting valuable knowledge from vast amounts of dispersed, heterogeneous, and unstructured agricultural big data, correlating and structuring it, and enhancing large models to form intelligent question-answering systems enables the effective delivery of services to all in agriculture. This approach can rapidly advance the scientific and precision-based development of agricultural production. Existing agricultural Q&A systems lack enough semantic understanding of complex symptoms, while general-purpose large language models (LLMs) produce factual hallucinations due to incomplete training data coverage. The research aims to address the issues of insufficient scale and low quality in the construction of knowledge bases in the agricultural field. **[Methods]** First, disease and pest data were collected along for five typical crops: wheat, rice, corn, potatoes, and cotton. Using manual verification methods, outliers were precisely identified and removed, ultimately yielding 87 901 unstructured data entries. Then, a few-shot learning model was employed to extract entities defined in the pattern layer, and these entities were aligned with the semantic vectors of Bert and LLM prompt engineering, ultimately yielding a triplet knowledge base of 916 239 entries for knowledge retrieval. A knowledge retrieval-augmented LLM approach for intelligent Q&A on crop diseases and pests was proposed, specifically the adaptive hybrid retrieval-augmented generation (AHR-RAG) approach. Firstly, an overlapping mechanism was introduced during fixed-length segmentation to mitigate semantic fragmentation. Simultaneously, vector semantic similarity was used to match highly related text blocks with the topic for optimization and storage. Then, single-hop and multi-hop retrieval were designed based on the complexity of the problem. Single-hop retrieval used the BM25 algorithm to match information extracted from the query with document content in the Elasticsearch index, feeding the results into the LLM to enhance answer generation. Multi-hop retrieval first converted user queries into structured conditions and semantic vector representations. Results retrieved from different knowledge bases were then fused using reciprocal rank fusion (RRF) and fed into the LLM. **[Results and Discussions]** The proposed method was experimentally compared with multiple baseline approaches, including different query types and complexity queries. The results demonstrated that the proposed method achieved accuracy and F_1 improvements of 0.193 and 0.170, respectively, on the Qwen1.5-7B-Chat model. Compared to the improved methods Self-RAG and Adaptive-RAG, AHR-RAG maintained low response times while achieving F_1 improvements of 0.05 and 0.021, respectively, with an accuracy as high as 0.896. For multi-type question-answering tasks, compared to the Naive-RAG method that relied solely on prior knowledge, the AHR-RAG approach achieved accuracy improvements of 0.231, 0.123, and 0.157 for comparison, judgment and selection query types, respectively. For parsing complex semantics, AHR-RAG also demonstrated significant advantages. In single-hop queries, its accuracy reached 0.921, representing a 0.029 improvement over Adaptive-RAG. In multi-hop query scenarios, its accuracy reached 0.748, achieving gains of 0.082 and 0.059 over Self-RAG and Adaptive-RAG respectively. In retrieval-augmented generation, AHR-RAG achieved a 0.013 increase in accuracy and a 0.009 improvement in F_1 by optimizing prompt strategies, compared to directly feeding retrieval results to the model for output. **[Conclusions]** This research demonstrates strong adaptability to diverse query types and excels at reasoning complex queries such as multi-hop searches. It delivers significant advantages in answer generation accuracy, relevance, and comprehensiveness, producing responses with enhanced logical coherence and richer content. Future work will explore the integration of multimodal knowledge bases.

Key words: adaptive hybrid retrieval; text blocking; diseases and pests; intelligent Q&A

Foundation items: The National Key R&D Program of China (2022ZD0115801)

First author: YANG Jun, E-mail: 2412399856@qq.com

***Corresponding author:** YANG Wanxia, E-mail: yangwanxia@163.com