刘琪,王皓冉,方婉茹,等.FruitQA-Sys:面向农业果蔬领域的专家问答系统[J].华中农业大学学报,2026,45(3):68-76.
DOI:10.13300/j.cnki.hnlkxb.2026.03.006

FruitQA-Sys:面向农业果蔬领域的专家问答系统

刘琪[1],王皓冉[1],方婉茹[1],徐娟[2],魏小梅[1,3]

1.华中农业大学信息学院,武汉 430070;2.果蔬园艺作物种质创新与利用全国重点实验室,武汉 430070;
3.农业智能技术教育部工程研究中心,武汉 430070

摘要 为解决现有大语言模型在农业果蔬种植等专业细分场景中知识覆盖不足、难以满足实际智能化应用需求的问题,采用领域数据构建、参数高效微调与检索增强生成相结合的方法,提出了一种面向农业果蔬领域的专家问答系统FruitQA-Sys。该系统通过生成器-打分器模块构建大规模高质量果蔬问答数据集,基于LoRA策略对领域大模型FruitBOT进行微调,并结合引导式检索增强生成技术IntroRAG,实现果蔬专业知识的精准问答。结果显示,在自动评估中,FruitQA-Sys的BERTScore为0.76,较DeepSeek-V3提升1.3%,与Qwen-3.0持平;GLEU为0.37,较DeepSeek-V3和Qwen-3.0分别提升2.7%;在人工评估中,系统综合得分达到8.56,较DeepSeek-V3和Qwen-3.0分别提升11.5%和15.7%。结果表明,领域微调大模型与IntroRAG的协同机制能够有效提升果蔬种植场景下智能问答系统的专业性与准确性。

关键词 LoRA微调;检索增强生成(RAG);垂直大模型;果蔬知识问答系统;数据集构建
中图分类号 S126 **文献标识码** A **文章编号** 1000-2421(2026)03-0068-09

农业果蔬种植在保障粮食安全、维护生态平衡与推动乡村振兴中具有核心地位。随着气候变化与病虫害谱系演替,农业生产管理对精细化知识、实时决策与可操作化建议的需求日益增强,因此对高准确性与可追溯性的智能辅助系统提出了更高要求[1]。近年来,大语言模型(large language models,LLMs)在智能问答、知识检索、对话系统等通用语言理解与生成任务上成效显著[2]。为满足农业领域的智能化应用需求,研究人员已开展LLM的垂直领域探索,并结合检索增强技术提升其输出可靠性。Zhang等[3]提出了基于LLM的IPM-AgriGPT病虫害预测管理系统,该系统通过图增强注意力框架结合农业知识图谱进行特定场景推理,有效提升了模型在病虫害防治问题上的事实一致性和场景适应性。王婷等[4]提出了基于大模型技术的果蔬农技知识智能问答系统,该系统结合果蔬栽培及病虫害防治知识库,实现农业技术问题的精准问答和智能咨询服务,有效提升了农技服务的智能化水平。Yang等[5]提出ShizishanGPT,在RAG框架下集成搜索引擎、农业知识图谱、通用GPT-4及作物表型预测代理等模块,通过整合设计实现更全面的农业问答功能。这些系统通过LLMs的自然语言理解能力与知识库的专业性相结合,使系统能够在提供知识指导的同时兼顾信息的科学性和操作可行性。但是这些研究直接接入通用领域大模型,缺乏对领域数据的适配,因而模型输出易出现细节缺失、不够专业和可靠的问题。

Yang等[6]构建了一个基于微调LLM结合检索增强生成(RAG)技术的农业问答系统AgriGPT,该系统采用复杂检索和知识图谱多跳推理技术,使之在处理复杂农业问答时能够同时利用多种信息源,从而显著提升回答的准确性和专业性。Zhao等[7]构建了一个基于LLM微调技术的植物病害检测的高效诊断与决策支持系统,该系统结合知识图谱知识推理在提升病虫害识别准确率的同时还能输出可解释的防治建议。这些研究通过领域数据的微调使LLM获得了专业领域的知识,因而在应用中与特定领域场景更加适配,输出更加精准。但是此类研究所采用的微调数据集完全来自LLM合成(如AgriGPT),缺乏人工专家的指导,使得微调模型的问答准

收稿日期:2025-09-30
基金项目:国家重点研发计划项目(2023YFD2300600);中央高校基本科研费专项(2662025PY017)
刘琪,E-mail:2024317120031@webmail.hzau.edu.cn
通信作者:魏小梅,E-mail:may@mail.hzau.edu.cn;徐娟,E-mail:xujuan@mail.hzau.edu.cn

确性和领域适应性有限[8]。另外，由于农业领域专用数据集的匮乏和垂直大模型技术探索的不足[9]，现有农业垂直大模型在果蔬种植这一细分领域的研究和应用存在明显空白[10]。

为解决上述问题，本研究提出FruitQA-Sys，一个面向果蔬种植应用场景的专家知识问答系统，旨在采用文本问答的形式实现领域专业知识问答。

1　材料与方法

1.1　数据来源

本研究收集的文本数据来源于农业相关教材和已发表的论文，确保了数据的专业性和正确性。经过筛选仅保留了果蔬种植管理、病虫害防治及作物采收等相关的内容。具体做法是，提取论文的标题、摘要、教材的书名等字段，调用DeepSeek API通过这些字段判断是否属于果蔬领域相关论文，其中大模型提示词设计如图1所示。筛选完成后得到2 390本教材与108 127篇相关的论文。

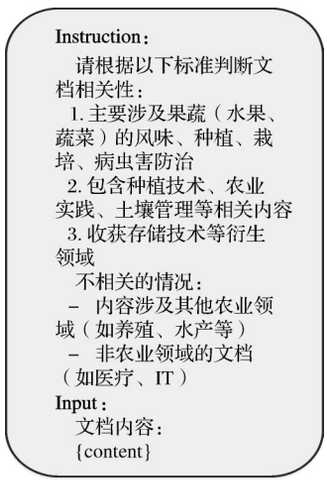


图1　大模型文献筛选提示

Fig. 1　LLM prompt for screening literature

1.2　FruitQA-Sys 系统概述

从图2可以看出，FruitQA-Sys专家问答系统的实现过程包括3个步骤：首先，构建了用于微调垂直模型的数据集；接着，采用微调技术实现了在LLM基座上的微调；最后，通过构建的文本向量库基于IntroRAG检索技术生成大模型的输入提示。

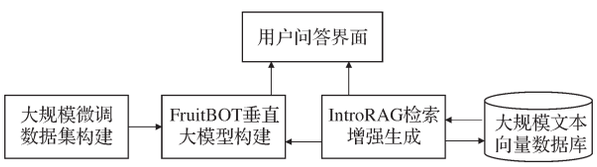
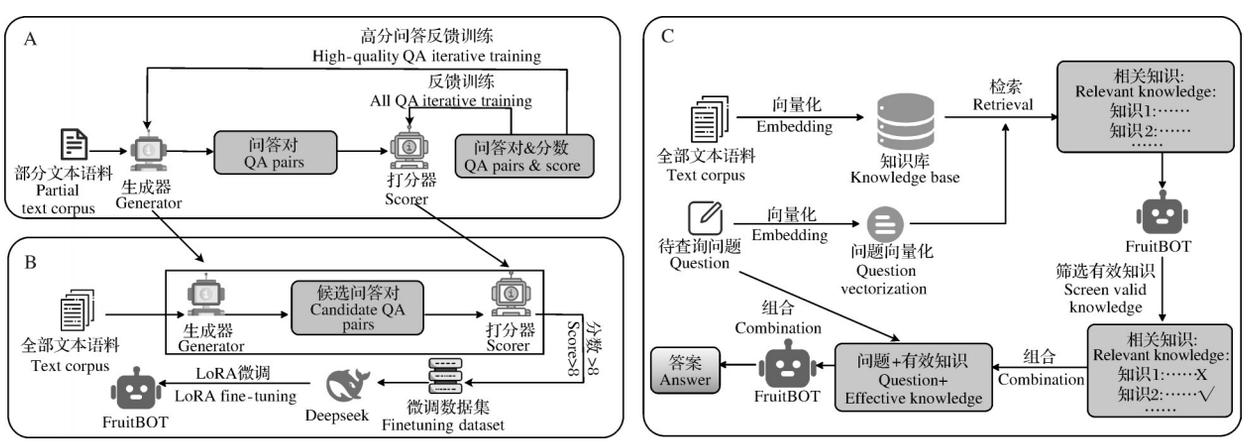


图2　FruitQA-Sys总体设计

Fig. 2　FruitQA-Sys overview

基于上述整体实现流程，系统的核心功能可进一步划分为3个关键模块，其内部结构与功能设计如图3所示，具体而言，FruitQA-Sys分为3个模块：(1)生成器-打分器训练模块。该模块通过循环迭代的方式训练了一个“生成器-打分器”模型，用来自动生成微调问答对。(2)FruitBOT微调模块。采用LoRA技术基于上一步生成的大规模问答语料对开源大模型基座进行微调，以增强模型对果蔬专业知识的理解与生成能力。(3)引导式检索增强生成(IntroRAG)模块。该模块基于收集的文本语料构建大规模的文本向量库，并在问答过程中采用IntroRAG技术实现FruitBOT的可信输出。



A. 生成器与打分器训练 Generator and scorer training；B. LoRA微调 LoRA fine-tuning；C. 引导式检索增强生成框架 IntroRAG framework.

图3　FruitQA-Sys构建详细流程图

Fig. 3　FruitQA-Sys construction detailed workflow

1.3 FruitBOT微调数据集构建

FruitBOT大模型的微调数据集是通过一个生成器和一个打分器模型来生成的。生成器和打分器的模型基座也是LLM，通过人工标注的种子数据集引导，进行迭代训练得到。训练完成后，该模型可以通过原始文本语料生成FruitBOT的微调数据集。

1）生成器-打分器种子数据集构建。本研究人工标注了种子数据集，包括500条生成器数据和500条打分器数据。构建种子数据集采用的原始文本为果蔬种植领域的专业论文，涉及的内容覆盖果蔬种植、果蔬风味、病虫害防治、采收与存储等果蔬种植全产业链数据。

种子数据集的原始语料以JSON文件格式保存，每条样本包含“原文”“问题”“答案”3个字段。其中，“原文”对应原始文本片段，“问题”与“答案”则基于原文内容设计，确保逻辑一致性与可溯源性。为了适配生成器微调的输入格式，原始JSON格式被转换为图4中的格式。转换后的数据结构包含instruction、input与output 3个部分。其中，instruction用于明确模型在生成问答对时的具体要求，即通过规范性说明指导模型“该做什么、如何做”，以确保生成方向与任务目标一致；input对应模型训练所需的原始输入，即原始文本片段；output则由3个子字段构成，分别为Q：基于输入文本生成的问题；C：生成原文中与该问答对相关的知识；A：与问题对应的答案。

Instruction：
基于给定文本生成多样化的问题和对应的答案，需满足以下标准：质量规范：每个问题必须对应文本中的明确知识点；避免假设性提问（如“如果……”）；禁止使用模糊指代（如“这个”“上述”等，应使用具体试验或方法名称）；生成的问答对在脱离文章背景后也能自洽、准确地解释相关知识。输出为严格的JSON格式，包含以下键，“question”：生成的问题；“cot”：答案的相关背景信息或段落；“answer”：答案，只生成农业相关知识问答对。

Input：
打蜡对柑橘果实的主要作用包括：增强果品表面光泽，改善外观品质，提高商品价值；堵塞果品表面气孔和皮孔，降低失水，在一定时期内有利于保持果品新鲜度；具有气调作用，限制果蔬与外界氧气和二氧化碳等气体交换，形成微型气调环境；减轻果实所受机械损伤。这些作用使得打蜡成为柑橘商品化处理的重要环节。

Output：
Q：打蜡对柑橘果实有哪些主要作用？
C：打蜡对柑橘果实的主要作用包括：增强果品表面光泽，改善外观品质，提高商品价值；堵塞果品表面气孔和皮孔，降低失水，在一定时期内有利于保持果品新鲜度；具有气调作用，限制果蔬与外界氧气和二氧化碳等气体交换，形成微型气调环境；减轻果实所受机械损伤。
A：打蜡主要作用有以下几个方面：(1) 增强果品表面光泽，改善外观品质，提高商品价值；(2) 堵塞果品表面气孔和皮孔，降低失水，在一定时期内有利于保持果品新鲜度；(3) 具有气调作用，限制果蔬与外界O_2和CO_2等气体交换，形成微型气调环境；(4) 减轻果实所受机械损伤。

图4 生成器种子数据格式

Fig. 4 Seed data format for generator

打分器的训练目标在于使打分器具备区分不同质量问答对的能力，以此筛选出高质量问答对。为了标注打分器的500条种子数据集，首先设计了一个总分10分制的评分体系，该体系从正确性、专业性、自洽性、正交性以及领域相关性5个维度对标注数据质量进行量化评价，具体分值分配与评分细则见图5。

打分器种子数据集标注过程如下：针对一段原始文本，人工标注高、中、低3个质量梯度的问答对，3个梯度的打分区间为：低质量0～4分、中质量5～7分、高质量8～10分。且每个打分由图5中的5项评价维度得分求和所得。打分器的训练集样本格式如图6所示，其中instruction字段用于规定10分制评分规则及各维度细则，input包含待评分的问答对及其对应的原文文本片段，output则记录该问答对的最终得分，并给出基于5项评价维度的评分理由，例如“正确性得3分，因答案基本符合原文，但存在1处细节偏差”。

2）生成器-打分器训练。通过迭代循环的方式来训练生成器-打分器模型，训练过程分为2步：首先各采用500个人工标注的问答对生成器和打分器进行初代指令微调；然后逐轮使用100篇文本语料生成问答对并进行循环增量微调[11]。生成器-打分器模型的训练基座为DeepSeek-R1-Distill-Qwen-14B，是DeepSeek官方基于DeepSeek-R1通过知识蒸馏技术优化而来，继承了原模型的核心推理能力[12]。

在初始训练阶段，利用种子数据集对模型进行第1次训练，获得初代版本的生成器-打分器模型，该模型将作为第二阶段迭代训练的初始模型。第二阶段的增量式迭代过程如图3A所示，具体来说，每一轮迭代开始时，随机抽取100篇论文，由生成器自动生成问答对，经过打分器筛选后新增的高分问答对直接用于下一轮生成器微调，使每次微调都是新的数据。增量式学习既能确保模型持续学习新的知

1. 正确性：
4分：答案与原文完全一致，无任何歧义或错误。
2分：核心事实正确，但存在细节缺失或次要信息不完整。
0分：存在事实性错误或与原文严重不符。
2. 专业性：
2分：使用规范农业术语，逻辑严密，表述专业。
1分：术语基本正确，但逻辑松散或表述不够严谨。
0分：存在专业知识错误或术语使用不当。
3. 自洽性：
2分：问题可独立理解，满足以下全部条件：①明确涉及农业细分领域（如作物、土壤、农机等）；②关键术语为通用术语或已有明确定义；③逻辑完整，不依赖隐含信息或上下文。
1分：问题需基础农业知识辅助理解，或存在轻微模糊。
0分：问题依赖原文信息，存在未定义概念或模糊指代。
4. 正交性：
1分：问题/答案组合在同批次数据中唯一，无实质重复。
0分：与同批次中其他问题/答案组合实质重复。
5. 领域相关性：
1分：明确涉及农业实体与技术（如作物品种、土壤管理、农机操作等）。
0分：与农业无关或仅涉及边缘概念。

图5　打分器种子数据集评分规则

Fig. 5　Scoring criteria for the scorer seed dataset

Instruction：
作为农业果蔬种植领域数据质量审核员，请按以下规则严格打分：正确性(0~4分)。4: 答案与原文完全一致且无歧义；2: 核心事实正确但细节缺失；0: 存在事实性错误。
（其他打分规则略，参考图5）
Input：
审核以下问答对质量：原文：……，问题：……，答案：……。
Output：
分数:
部分得分:{“正确性”：“专业性”：“自洽性”：“正交性”:“领域相关性”:}，
原因:[“正确性:……,得…分”,
“专业性：……，得…分”，
“自洽性：……，得…分”，
“正交性：……，得…分”，
“领域相关性：……，得…分”]

图6　打分器种子数据格式

Fig. 6　Seed dataset format for scorer

识，还能防止生成器-打分器的训练出现过拟合，因而能保持模型的稳定性与泛化能力。

跟上述种子数据集的格式一样，生成器生成的数据也采用“原文”“问题”“答案”3个字段进行存储。同时，考虑到部分论文篇幅较长，可能导致大模型的理解与生成质量下降，论文在输入打分器前做了分块处理，以减轻长文本带来的干扰。生成的问答对由打分器进行质量评估，评估方式采用本文“1.3”提出的10分制评分标准。对于得分为满分（10分）的问答，将其进行格式转换后加入生成器的训练集。同时，将所有问答对的评分结果、打分理由及其对应样本整理为打分器训练数据，以提升打分器对数据质量的判别能力。

生成器-打分器训练的迭代终止条件由人工检验与自动评分联合决定。具体而言，人工检验是在每轮生成问答对后，随机抽取100条包含果蔬各方面知识问答对进行人工评估，评价标准采用图5中10分制评分规则，人工评分分数超过8分即为合格，当合格率达到90%时即为满足条件；自动评分则通过打分器打分，当得分超过8分的问答对达到90%时满足条件。仅在2项标准均被满足时，迭代过程方可终止。此双重决策不仅保证了数据质量，而且确保了自动判别与人工评价之间的一致性。

综上所述，该迭代循环框架通过“生成→判别→筛选→回写”的闭环机制，在低人工依赖度下实现了高质量训练样本的持续扩充与模型能力的渐进优化。Puri等[13]的研究表明，生成器-打分器模型所获取的微调数据集比人工标注问答对效率更高，比完全依赖大模型生成的数据集质量更专业、准确。因此该方法在保证数据质量的同时，兼顾了可扩展性与效率，为农业等垂直领域大模型的高效微调提供了一种可行的数据合成方法。

3）微调数据生成。生成器-打分器模型训练完成后，将收集的全部文本语料输入模型，用来生成Fruit-BOT的微调数据集。如图3B所示，首先将获得的原始文本输入生成器，自动生成候选问答对；随后，这些问答对被输入打分器中进行质量评估，得分在8分以上的被保留。最终获得94 344条高质量问答对，总计16 095 922个Tokens，平均Token数为170.61。

1.4　基于LoRA的FruitBOT微调

LoRA[14]作为一种针对大模型的参数高效微调方法，有效解决了传统微调需更新模型全部参数的弊端。其核心设计思路是冻结原始大模型的权重参数，仅在部分权重矩阵上引入可训练的增量矩阵以实现下游任务适配。具体而言，假设大模型某一层

的原始权重矩阵为 $W\in R^{d\times k}$(d为输入维度,k为输出维度),LoRA在该矩阵基础上引入可训练的增量矩阵$\triangle W$,使得模型实际使用的权重矩阵W'为:

$$W'=W+\Delta W \tag{1}$$

为进一步压缩可训练参数量,LoRA利用低秩矩阵的性质对$\triangle W$进行低秩分解,将其表示为2个小维度矩阵的乘积,如公式(2)所示:

$$\Delta W=B\cdot A \tag{2}$$

式(2)中,$A\in R^{r\times k}$为降维矩阵,$B\in R^{d\times r}$为升维矩阵,关键超参数r(低维空间维度)满足$r\ll\min(d,k)$。由此可得LoRA的完整权重更新公式(3):

$$W'=W+B\cdot A \tag{3}$$

该设计使得可训练参数仅为$r\times(d+k)$,相较于传统微调需更新的$d\times k$个参数大幅减少,显著降低了训练成本与显存占用[15],其效果与全参数微调相比,性能损失较小。有研究结果表明,LoRA在多数下游任务中的性能与全参数微调相比,精度损失极小[16]。因此,本研究在模型微调过程中选择采用LoRA方案,以在保证效果的同时兼顾训练效率与资源消耗。

如图3B所示,微调数据集准备好以后,送入基座模型DeepSeek-R1-Distill-Qwen-7B。微调过程中超参数设置如下:学习率为1×10^{-4},训练轮次(epoch)为10,低秩维度r取16,LoRA缩放因子α=16,批大小(batch size)为2,权重衰减(weight decay)为0.01,并采用bf16精度进行训练。其中,学习率决定参数更新的速度与稳定性;epoch控制模型在训练集上的迭代次数;r与α共同影响LoRA模块的表达能力和有效容量;batch size影响梯度估计的稳定性与显存占用;weight decay用于防止过拟合;而bf16精度则在保证数值稳定的同时有效降低显存开销与计算成本。

1.5 引导式检索增强生成(IntroRAG)问答

1)文本向量库构建。经过领域数据集的微调,已经将领域知识内化到FruitBOT的参数中,使其可以较为准确地完成专业知识问答。但大模型的固有局限在实际场景中逐渐显现[17]:模型参数一旦固化,在遇到训练语料未覆盖或已过时的信息时容易产生知识幻觉,从而给出不可靠或与事实不符的回答;同时农业知识时效性强,若要纳入新信息通常必须重新微调,既耗时又难以实现即时更新[18]。为此本研究在FruitQA-Sys中引入检索增强生成RAG技术[19]以缓解知识动态更新和模型幻觉问题。

当前的RAG技术主要使用人工构建的知识库来增强大模型输出可靠性,知识库包括文本向量知识库(vector DB)、知识图谱等。其中向量知识库简单高效、技术成熟、易于实现和扩展。本研究构建的文本向量库以农业果蔬种植领域的教材与已发表论文为数据源[20],从源头上保证了知识的专业性与可靠性,并确保知识体系的系统性与全面性。原始数据包括2 390本教材与108 127篇论文,构建的向量数据库采用了1 024维向量来存储文本,这种高维表征可以更容易捕捉语义空间中细微的文本差异[21]。例如,该表示方式能够有效地区分“水稻的抗旱性”与“小麦的抗旱性”,而非笼统地归为“作物的抗旱性”。最终,本研究构建了一个规模达到69 GB的大规模农业领域文本向量数据库,为后续检索与问答生成提供了坚实的知识支撑。

2)引导式检索增强生成(IntroRAG)。在基于垂直大模型FruitBOT的知识检索和问答应用中,本研究设计了一种引导式的检索增强生成框架(IntroRAG),可以显著提高答案的准确性,实现过程如图3C所示。与传统RAG方法不同,该框架在检索过程中加入FruitBOT作为引导模块,以提升检索结果的有效性。具体而言,首先将用户问题向量化,并在知识向量库中进行相似性检索[22]。传统方法通常会将检索到的知识直接输入至生成模型,而IntroRAG则先利用FruitBOT对检索结果进行筛选和判断,以确认其是否包含回答问题所需的关键信息[23],从而避免无关知识干扰生成过程。筛选后的高相关性知识最终作为补充信息输入至FruitBOT,以支持模型生成更加准确、专业的回答。结果表明,结合了IntroRAG的问答在保证检索广度的同时提高了知识利用的精度,有效增强了检索与生成之间的协同。

1.6 模型评测方法

为系统评估FruitQA-Sys的问答性能,本研究设计了自动评估与人工评估相结合的综合评价方案。自动评估用于量化模型输出质量,人工评估用于验证回答的专业性与实用性等。

参考Wang等[24]的研究方法,本研究引入了人工评估,以进一步从专家的角度检验模型在实际应用场景中的表现。评估专家由3名具有农业学科背景的专业人员组成,评估过程采用正确性、专业性、完整性、流畅度等4个维度进行打分,每个维度满分10分,最后按照得分进行加权平均获得模型的总的得分。不同维度的权重分配为:(1)正确性40%,用于衡量模型输出内容是否事实准确,是否与农业知识

和常识一致；(2)专业性25%，用于评估回答是否展现出足够的专业领域知识深度和准确性，包括使用专业术语、提供翔实的论据或数据；(3)完整性20%，关注回答是否覆盖问题的核心要点，能否形成较为系统的说明；(4)流畅度15%，从语言表达层面考察答案是否语法正确、逻辑连贯以及可读性良好[25-26]。

打分过程中，每条回答均由2位专家在4个维度上独立打分，然后按照加权平均计算各自的总得分，若2位专家的评分差异不超过3分，则认为评分一致，取两人平均分作为该条回答的最终分数；否则，则交由第3位专家进行仲裁并给出最终分数(最终分数应该跟前面任意一个差异不超过3分)，用以作为人工评估的综合指标。

2　结果与分析

2.1　测试集和评估标准

目前未见有公开发表的农业果蔬领域垂直大模型的评测基准，参考Kaur等[27]的做法，本研究构建了一个专用评测数据集。数据集包括100条人工编写的高质量问答对，涵盖的主题覆盖果蔬种植的全产业链，基本满足了本领域的核心需求。数据标注人员来自农业领域的专家，标注过程经过多轮专家评估和验证，确保答案内容正确。评估指标则采用大模型的通用指标BERTScore[28]、GLEU[29]、ROUGE-L[30]。BERTScore是基于预训练语言模型的上下文嵌入表示，用来计算候选答案与参考答案之间的语义相似度，该指标能够捕捉同义表达和语义层面的对应关系。GLEU是通过同时考虑候选答案与参考答案之间的*n*-gram精度与召回率来衡量二者一致性。ROUGE-L是衡量生成答案对参考答案的覆盖情况。

2.2　与通用大模型性能比较

本研究选择了3个通用大模型DeepSeek-R1-Distill-Qwen-7B、DeepSeek-V3、Qwen-3.0进行性能比较。第一个模型为FruitBOT微调的基座模型(简称FB_Base)，后两者是目前有代表性的主流通用大模型。三者均由国内机构研发，在中文处理方面相较于其他通用大模型有明显优势。由表1可知，FruitQA-Sys在3项指标中均取得最优成绩，大幅超越未微调的基准模型FB_base，说明经过领域数据微调和RAG检索后，大模型对特定领域的知识理解能力更强，输出结果更准确。与DeepSeek-V3和Qwen-3.0模型对比发现，FruitQA-Sys在3个指标上均呈现小幅优势，具体而言，FruitQA-Sys的BERTScore达到0.76，与Qwen-3.0持平，但相比DeepSeek-V3提升约1.3%；GLEU达到0.37，相较DeepSeek-V3和Qwen-3.0均提升约2.7%；ROUGE-L达到0.27，相比DeepSeek-V3提升约8.0%，相比Qwen-3.0提升约3.8%。但在模型参数量方面，FruitQA-Sys远远小于2个基准模型，这在农业应用场景中非常重要，模型轻量化正是适配农业实际需求的关键技术指标。总的来说，FruitQA-Sys系统在农业问答任务中，既能在语言表达的准确性、流畅性上面可以跟目前超大参数的模型媲美，又由于领域微调和IntroRAG技术使其在专业知识的关键信息把握、精准匹配上面优于未微调基座模型。

表1　模型输出评分

Table 1　Model output scores

模型 Model	参数量 Parameters	BERTScore	GLEU	ROUGE-L
FB_Base	7×10^9	0.71	0.23	0.14
DeepSeek-V3	671×10^9	0.75	0.36	0.25
Qwen-3.0	235×10^9	0.76	0.36	0.26
FruitQA-Sys	7×10^9	**0.76**	**0.37**	**0.27**

为直观呈现不同模型针对专业问题答案的准确性，本研究借助图7进行案例展示。由图7可以看出，DeepSeek-V3在回答中表现为较多假设性与不确定性，Qwen-3.0直接说明知识库中无相关信息，无法回答，而FruitQA-Sys的回答直接给出结论并附带量化结果。由此可以看出，FruitQA-Sys由于包含大量的领域专业知识，因此可以直接输出确证的领域事实，其专业性、针对性和可操作性明显优于通用模型。

2.3　人工评估结果

由图8可以看出，FruitQA-Sys在四大维度上均取得了最高分，尤其在正确性、专业性和完整性方面明显领先于其他模型，分别以9.0、8.5和8.0分位列第一，证明提出的系统在领域知识储备和专业的输出上表现突出。在输出流畅度方面，所有模型表现相当，这个符合通用大模型的特点，也说明FruitQA-Sys在提升专业性的同时并没有损失语言特性。综上所述，FruitQA-Sys在综合性能尤其在农业任务的专业性方面展现出明显优势。

2.4　消融实验

为评估不同模块对模型性能的影响，本研究利用3种配置：(1)FB_Base＋IntroRAG：未微调的基础模型结合IntroRAG技术；(2)FruitBOT：使用在基础模型上微调的垂直模型；(3)FruitBOT＋In-



图7 问答结果案例分析

Fig. 7 Case study of Q&A results

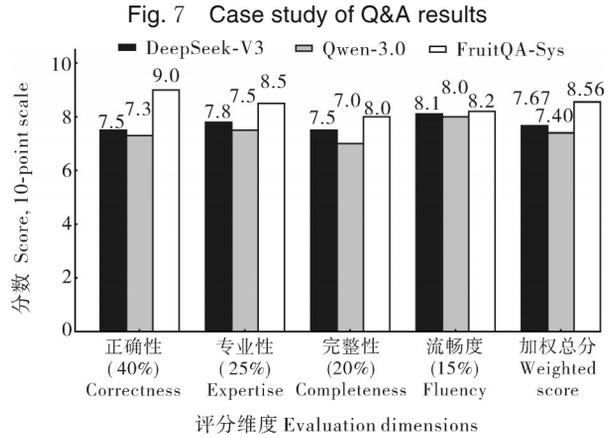


图8 模型人工评价结果

Fig. 8 Manual evaluation results of the model

troRAG：微调模型结合IntroRAG技术，即FruitQA-Sys系统，在测试集上进行自动评估。由表2可以看出，FruitBOT＋IntroRAG显著高于仅使用IntroRAG的基础模型和仅进行微调无IntroRAG组件的FruitBOT。说明领域数据弥补了LLM在垂直领域知识、表达范式和任务能力上的不足，垂直大模型结合RAG的协同作用，可有效提升模型语义匹配、细节一致性与信息覆盖能力。

表2 消融实验

Table 2 Ablation study

模块 Module	BERTScore	GLEU	ROUGE-L
FB_Base＋IntroRAG	0.72	0.28	0.20
FruitBOT	0.72	0.27	0.20
FruitBOT＋IntroRAG	**0.75**	**0.37**	**0.27**

3 讨 论

针对LLMs在农业果蔬种植领域智能化应用的空白，本研究构建了面向该场景的专家问答系统FruitQA-Sys。系统通过“生成器-打分器迭代数据集构建”“LoRA微调大模型”“引导式检索增强”三大核心模块，形成“数据生成-模型优化-知识补充”完整LLM应用解决方案。研究表明，该系统在正确性、专业性、流畅性上显著优于通用LLMs。从设计与实践看，FruitQA-Sys优势有：其一，生成器-打分器迭代框架突破数据瓶颈，相较其他研究完全依赖自动合成语料模型来说，通过人工种子数据初始化、多维度筛选（正确性、专业性等）及增量整合，能够以低成本生成高质量问答对；其二，LoRA微调可以适配资源约束，相较全参数微调的高耗特性，借助低秩矩阵分解仅更新少量参数，在保证微调模型领域适配性的同时，降低训练成本与显存占用；其三，IntroRAG可以通过补充知识弥补静态模型的时效性，同时借FruitBOT筛选检索结果提升回答的相关性。

LLM应用属于前沿技术，目前仍然有很多问题有待研究，未来将围绕该任务展开持续优化和场景适配方面的研究，研究计划包括：(1)提升文本向量库的检索速度。本研究中69 GB的文本向量数据库规模过大，问答需经过“向量化-检索-筛选-生成”全流程，检索与筛选环节耗时较长，需要研究相关检索技术提升效率；(2)模型的场景验证。提出的模型需要在真实场景中反复验证和优化，包括在真实生产环境中收集多样化作物与区域数据，提升模型的泛化性与跨区域适用性。

参考文献 References

[1] 赵瑞雪，杨晨雪，郑建华，等. 农业智能知识服务研究现状及展望[J]. 智慧农业（中英文），2022，4(4)：105-125. ZHAO R X, YANG C X, ZHENG J H, et al. Agricultural intelligent knowledge service: overview and future perspectives[J]. Smart agriculture, 2022, 4(4): 105-125 (in Chinese with English abstract).

[2] QIU X P, SUN T X, XU Y G, et al. Pre-trained models for natural language processing: a survey[J]. Science China technological sciences, 2020, 63(10): 1872-1897.

[3] ZHANG Y Q, FAN Q J, CHEN X, et al. IPM-AgriGPT: a large language model for pest and disease management with a G-EA framework and agricultural contextual reasoning [J/OL]. Mathematics, 2025, 13(4): 566[2025-09-30]. https://doi.org/10.3390/math13040566.

[4] 王婷,王娜,崔运鹏,等.基于人工智能大模型技术的果蔬农技知识智能问答系统[J].智慧农业(中英文),2023,5(4):105-116. WANG T, WANG N, CUI Y P, et al. Agricultural technology knowledge intelligent question-answering system based on large language model[J]. Smart agriculture, 2023, 5(4): 105-116 (in Chinese with English abstract).

[5] YANG S T, LIU Z H, MAYER W, et al. ShizishanGPT: an agricultural large language model integrating tools and resources[C]//Web Information Systems Engineering (WISE 2024). December 2-5, 2024, Doha, Qatar. Cham: Springer, 2024: 284-298.

[6] YANG B, ZHANG Y, FENG L, et al. AgriGPT: a large language model ecosystem for agriculture [PP/OL]. arXiv, 2023: 2308.15085 [2025-09-30]. https://doi. org/10.48550/arXiv.2508.08632.

[7] ZHAO X Y, CHEN B Y, JI M X, et al. Implementation of large language models and agricultural knowledge graphs for efficient plant disease detection [J/OL]. Agriculture, 2024, 14(8): 1359 [2025-09-30]. https://doi. org/10.3390/agriculture14081359.

[8] NIRANJAN P Y, RAJPUROHIT V S, SANNAKKI S S. Question answering system for agriculture domain using machine learning techniques: literature survey and challenges [J/OL]. International journal of computational systems engineering, 2020, 6(2): 91[2025-09-30]. https://doi.org/10.1504/IJCSYSE.2020.111208.

[9] 郭旺,杨雨森,吴华瑞,等.农业大模型:关键技术、应用分析与发展方向[J].智慧农业(中英文),2024,6(2):1-13. GUO W, YANG Y S, WU H R, et al. Big models in agriculture: key technologies, application and future directions [J]. Smart agriculture, 2024, 6(2): 1-13 (in Chinese with English abstract).

[10] 张云慧,张智,刘杰.病虫测报智能化研究进展[J].现代农药,2023,22(5):9-16. ZHANG Y H, ZHANG Z, LIU J. Research progress of intelligent in forecasting for diseases and insect pests [J]. Modern agrochemicals, 2023, 22(5): 9-16 (in Chinese with English abstract).

[11] WANG Y Z, KORDI Y, MISHRA S, et al. Self-instruct: aligning language models with self-generated instructions[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. July 7-9, 2023, Toronto, Canada. Toronto: ACL, 2023: 13484-13508.

[12] GUO D, YANG D, ZHANG H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning [PP/OL]. arXiv, 2025: 2501.12948[2025-09-30]. https://doi.org/10.1038/s41586-025-09422-z.

[13] PURI R, SPRING R, SHOEYBI M, et al. Training question answering models from synthetic data [PP/OL]. arXiv: 2002.09599 [2025-09-30]. https://doi. org/10.48550/arXiv.2002.095996.

[14] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models [PP/OL]. arXiv, 2021: 2106.09685 [2025-09-30]. https://doi. org/10.48550/arXiv.2106.09685.

[15] 刘铮,朱轶.农业大语言模型的部署及应用开发探索[J].中国农业文摘-农业工程,2025(6): 60-64. LIU Z, ZHU Y. Exploration on deployment and application development of agricultural large language models[J]. China agricultural abstracts-agricultural engineering, 2025(6): 60-64 (in Chinese).

[16] BIDERMAN D, PORTES J, ORTIZ J J G, et al. LoRA learns less and forgets less[PP/OL]. arXiv, 2024: 2405.09673 [2025-09-30]. https://doi.org/10.48550/arXiv.2405.09673.

[17] 张钦彤,王昱超,王鹤羲,等.大语言模型微调技术的研究综述[J].计算机工程与应用,2024,60(17):17-33. ZHANG Q T, WANG Y C, WANG H X, et al. Comprehensive review of large language model fine-tuning [J]. Computer engineering and applications, 2024, 60(17): 17-33 (in Chinese with English abstract).

[18] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation [J]. ACM computing surveys, 2023, 55(12): 1-38.

[19] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[PP/OL]. arXiv, 2005.11401 [2025-09-30]. https://doi.org/10.48550/arXiv.2005.11401.

[20] JIANG J C, YAN L, LIU H F, et al. Knowledge assimilation: Implementing knowledge-guided agricultural large language model [J/OL]. Knowledge-based systems, 2025, 314: 113197 [2025-09-30]. https://doi.org/10.1016/j.knosys.2025.113197.

[21] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). November 3-7, 2019, Hong Kong, China. Hong Kong: ACL, 2019: 3980-3990.

[22] KARPUKHIN V, OGUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). November 16-20, 2020, Stroudsburg, PA, USA. Stroudsburg: ACL, 2020: 6769-6781.

[23] 于合龙,沈金梦,毕春光,等.基于知识图谱的水稻病虫害智能诊断系统[J].华南农业大学学报,2021,42(5):105-116. YU H L, SHEN J M, BI C G, et al. Intelligent diagnostic system for rice diseases and pests based on knowledge graph[J]. Journal of South China Agricultural University, 2021, 42(5):

105-116 (in Chinese with English abstract).
[24] WANG C X, CHENG S R, GUO Q P, et al. Evaluating open-QA evaluation [PP/OL]. arXiv, 2023: 2305.12421 [2025-09-30]. https://doi.org/10.48550/arXiv.2305.12421.
[25] ZHAO J M, XU L X, MA Z Z, et al. A review of plant leaf disease identification by deep learning algorithms [J/OL]. Frontiers in plant science, 2025, 16: 1637241 [2025-09-30]. https://doi.org/10.3389/fpls.2025.1637241.
[26] 符鹏,杨海平. 大语言模型能力评价体系构建及问题研究[J]. 情报探索, 2024(11): 34-40. FU P, YANG H P. Construction and investigation of evaluation system for large language model capabilities [J]. Information research, 2024(11): 34-40 (in Chinese with English abstract)
[27] KAUR R, BHANKHAR A S, SALH J S, et al. Leveraging synthetic data for question answering with multilingual LLMs in the agricultural domain [PP/OL]. arXiv, 2025: 2507.16974 [2025-09-30]. https://doi.org/10.48550/arXiv.2507.16974.
[28] ZHANG T Y, KISHORE V, WU F, et al. BERTScore: evaluating text generation with BERT [PP/OL]. arXiv, 2019: 1904.09675 [2025-09-30]. https://doi. org/10.48550/arXiv.1904.09675.
[29] MUTTON A, DRAS M, WAN S, et al. GLEU: automatic evaluation of sentence-level fluency [C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. June 25-27, 2007. Prague, Czech Republic. Prague: ACL, 2007: 344-351.
[30] LIN C Y. Rouge: a package for automatic evaluation of summaries [C]//Proceedings of the Text Summarization Branches Out. July 25-26, 2004, Barcelona, Spain. Barcelona: ACL, 2004: 74-81.

FruitQA-Sys: an expert question-answering system for fruit and vegetable domain

LIU Qi[1], WANG Haoran[1], FANG Wanru[1], XU Juan[2], WEI Xiaomei [1,3]

1. *College of Informatics, Huazhong Agricultural University, Wuhan 430070, China*;
2. *National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, Wuhan 430070, China*;
3. *Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China*

Abstract A domain-specific expert question-answering system, FruitQA-Sys, was proposed by integrating domain data construction, parameter-efficient fine-tuning, and retrieval-augmented generation to solve the problems of insufficient knowledge coverage and difficulty in meeting the practical requirements of intelligent applications in specialized sub-scenarios including the planting of fruit and vegetable with existing large language models. The proposed system constructed a large-scale and high-quality fruit and vegetable question-answer dataset through a generator-scorer module, finely tuned the domain-specific large model FruitBOT based on the LoRA strategy, and incorporated the guided retrieval-augmented generation technique IntroRAG to achieve accurate question and answer of professional knowledge of fruit and vegetable. The results show that in automatic evaluation, the FruitQA-Sys achieves a BERTScore of 0.76, which is a 1.3% improvement over the DeepSeek-V3, and is on par with another mainstream general-purpose large model Qwen-3.0; its GLEU score is 0.37, representing a 2.7% improvement over both DeepSeek-V3 and Qwen-3.0. The overall score in human evaluation reached 8.56, representing improvements of 11.5% and 15.7% over DeepSeek-V3 and Qwen-3.0, respectively. It is indicated that the collaborative mechanism of domain-specific fine-tuning and IntroRAG can effectively improve the professionalism and accuracy of intelligent question-answering systems in scenarios of planting fruit and vegetable.

Keywords LoRA finetuning; retrieval-augmented generation (RAG); domain-specific large language model; fruit and vegetable QA system; dataset construction

(责任编辑:葛晓霞)