

基于人工智能大模型技术的果蔬农技知识智能问答系统

王 婷^{1,2}, 王 娜³, 崔运鹏^{1,2*}, 刘 娟^{1,2}

(1. 中国农业科学院农业信息研究所, 北京 100081, 中国; 2. 农业农村部农业大数据重点实验室, 北京 100081, 中国; 3. 96962 部队, 北京 102206, 中国)

摘 要: [目的/意义] 乡村振兴战略给农业技术推广提出新的要求, 使农业推广知识的供给形式有待进一步创新。以果蔬农技知识服务为需求导向, 基于前沿大语言模型技术, 面向新型农业知识导读和知识问答等农技推广服务, 构建果蔬农技知识智能问答系统。[方法] 基于草莓种植户需求分析, 把草莓栽培农技知识划分为不同主题, 形成知识对象识别和知识问答两种大模型下游任务, 结合机器自动标注和人工标注的方法构建小样本高质量训练语料; 通过对比已有的4种大语言模型: Baichuan2-13B-Chat、ChatGLM2-6B、Llama-2-13B-Chat、ChatGPT 的性能表现, 选择性能最优的模型作为基础模型, 按照“优质语料+预训练大模型+微调”的研究思路, 训练具有语义分析、上下文关联和生成能力, 能够适应多种下游任务的深度神经网络, 构建农业知识问答大模型; 采用数据优化、检索增强生成技术等多种策略缓解大模型幻觉问题; 研发果蔬农技知识智能问答系统, 生成高精度、无歧义的农业知识答案, 同时支持用户多轮问答。[结果和讨论] 以精准率和召回率为命名实体识别任务的性能表现指标, 参与测评的国内主流模型在微调后不同知识主题下的平均精准率均超过85%, 平均召回率表现各异, 其中知识实体类型的数量、标注语料数量等因素都会影响大模型性能; 以幻觉率和语义相似度为知识问答任务的性能表现指标, 数据优化、采用检索增强生成技术等策略以10%~40%的幅度有效降低大模型幻觉率, 并有效提高大模型的语义相似度。[结论] 在农业领域的命名实体识别和知识问答任务中, 预训练大模型ChatGLM的表现性能最优。针对预训练大模型下游任务的微调和基于检索增强生成(Retrieval-Augmented Generation, RAG)技术的模型优化可以缓解大模型幻觉问题, 显著提升大模型性能。大模型技术具有创新农技知识服务模式、优化农业知识推广的潜力, 能够有效降低种植户获取高质量有效知识的时间成本, 引导更多的种植户实现农业技术创新和转型。但是由于性能不稳定等诸多问题, 大模型的优化方法和具体场景应用仍需进一步深入研究。

关键词: 大模型; 生成式预训练变换器; 农技知识; 智能问答; 命名实体识别

中图分类号: TP399; S126

文献标志码: A

文章编号: SA202311005

引用格式: 王婷, 王娜, 崔运鹏, 刘娟. 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. DOI: 10.12133/j.smartag.SA202311005

WANG Ting, WANG Na, CUI Yunpeng, LIU Juan. Agricultural technology knowledge intelligent question-answering system based on large language model[J]. Smart Agriculture, 2023, 5(4): 105-116. DOI: 10.12133/j.smartag.SA202311005 (in Chinese with English abstract)

0 引 言

现代信息技术的飞速发展促进了中国农业技术的不断进步。如何推广普及先进实用的农业技术成

为中国农业现代化亟须解决的问题^[1]。传统农业技术推广存在“供需矛盾突出”的问题: 首先, 很多农业科研成果没有得到有效的推广和使用, 导致农业生产力和市场需求不匹配。其次, “大水漫灌式”

收稿日期: 2023-11-01

基金项目: 北京市数字农业创新团队项目(BAIC10-2023); 中国农业科学院基本科研业务费项目(JBYW-AII-2023-31); 国家重点研发计划项目(2022YFF0711902)

作者简介: 王 婷, 研究方向为深度学习方法的理论研究与应、生信分析。E-mail: wangting01@caas.cn

*通信作者: 崔运鹏, 博士, 研究员, 研究方向为农业大数据挖掘分析、自然语言处理、生信分析。E-mail: cuiyunpeng@caas.cn

copyright©2023 by the authors

的农技推广无法满足农业生产者的个性化、易接受、好吸收的需求。在乡村振兴战略给农业技术推广提出新要求的背景下,农业推广知识的供给形式有待进一步创新^[2]。如何在果蔬栽培过程中,针对种植户面临的实际生产问题,随时随地提供高质量的答疑解惑,并辅助种植户快速有效地消化和吸收农业知识,都是农技知识推广亟须解决的问题,实现技术分别对应自然语言处理领域中的命名实体识别和知识问答。

深度学习、大规模神经网络等人工智能技术的快速突破性进展,尤其是大语言模型(Large Language Models, LLM)的出现,给命名实体识别和知识问题提供了新的手段和方法,从而也使得拟人化、智能化的农业技术推广成为可能^[3]。LLM在大规模数据训练过程中可以自动学习一些高级复杂的功能,拥有更准确的逻辑推理学习能力,在很多方面都拥有了接近人类认知的表现^[4]。聊天生成式预训练变换器(Chat Generative Pre-trained Transformer, ChatGPT)的出现将大语言模型技术推向了爆发阶段,引发了大模型技术产业前所未有的发展契机。OpenAI、微软、谷歌、Facebook、百度等科技巨头企业纷纷布局了相关业务,在大语言模型领域持续不断地进行探索创新,推出ChatGPT、GPT-4、文心一言、通义千问、ChatGLM-6B、Baichuan2-13B-Chat等人工智能大模型产品。

同时,中国政府也推出多项与人工智能领域相关的利好政策。其中,《北京市促进通用人工智能创新发展的若干措施(2023—2025年)》提出“开展大模型创新算法及关键技术研究”“加强大模型训练数据采集及治理工具研发”等,面向政务服务、医疗、科学研究、金融等领域拓展应用场景,抢抓大模型发展机遇,推动通用人工智能领域实现创新引领。

因此,本研究在把握农业需求、跟踪前沿人工智能技术的基础上,以草莓栽培技术为例,利用预训练大语言模型,按照“优质语料+预训练大模型+微调”的大模型技术研究思路,训练具有语义分析、上下文关联和生成能力,能够适应知识对象实体识别、关系抽取、知识问答等多种下游任务的深度神经网络模型,构建新一代果蔬农技知识问答系

统,能够为农业生产者提供高精度、无歧义的有效知识,全面助力农业生产者解决实际问题。

1 智能问答技术相关研究

1.1 命名实体识别

根据命名实体识别领域的研究进展,主要可以划分为以下3个阶段。

(1) 基于词典和规则的方法。在专家制定好规则和词典后,通过文本匹配的方式实现命名实体识别。这种方法过于依赖专家知识,人工成本和时间成本较高,且无法面向新领域、新实体类型或新数据集进行迁移或扩展^[5]。

(2) 基于传统机器学习的方法。命名实体识别被转化为序列标注问题,当前的预测标签不仅与当前的输入特征相关,还与之前的预测标签相关,利用序列之间的强相互依赖关系来实现命名实体识别。常见方法包括隐马尔可夫模型^[6]、支持向量机^[7]和条件随机场^[8]等。

(3) 基于深度学习的方法。在命名实体识别任务适用于非线性转化的特性条件下,由于深度学习以端到端的训练方式,通过梯度传播方法构建更复杂的网络结构,从而能够提取自然语言中更有效的特征,更好地挖掘命名实体和实体类型之间的关系,所以注意力机制、图神经网络、迁移学习、远监督学习等热门研究技术成为命名实体识别目前的主流研究方向。常见方法包括BiLSTM-CRF^[9]、IDCNN-CRF^[10]、CAN-NER^[11]、LatticeLSTM^[12]、BERT^[13]和ERNIE^[14]等。

1.2 知识问答

根据知识问答领域的研究进展,主要可以划分为以下4个阶段。

(1) 传统基于规则的方法。通常包括问题分类、答案检索和答案生成。这种方法人工成本和时间成本较高,且无法有效处理未知问题和答案^[15]。

(2) 基于知识图谱的方法。首先通过结构化数据、文本语料库和半结构化数据构建领域知识图谱,然后基于此提取准确和详细的答案。由于知识图谱存在过于依赖专家知识、知识不完备性、缺乏

语言理解等缺点，所以这种方法具有很大的局限性^[16]。

(3) 基于传统深度学习的方法。基于深度学习方法中小型神经网络，如循环神经网络、长短期记忆网络和注意力机制等，将自然语言转化为语义表示，并将问题和答案表示为向量形式，通过问题向量和答案向量之间的相似匹配得分确定最优答案^[17]。

(4) 基于大语言模型的方法。把深度学习方法中大型神经网络作为预训练模型，如BERT、RoBERTa^[18]、BART^[19]、GPT^[20]和ChatGLM等，根据下游任务进行微调，以实现迁移学习。这些模型的参数规模巨大，网络结构十分复杂，在设计的预训练任务下从大规模无标注文本中学习自然语言上下文相关的意义和结构，可以捕捉到更丰富的语言特征，从而能够更好地应对各种自然语言处理任务，知识问答便是其中一个。但是由于GPT、ChatGLM等大模型生成回复时，在“事实性”“实时性”等方面存在缺陷，在需要精准回答的领域知识问答场景无法满足需求，所以必须借助外部知识库生成高质量的准确回复。

2 果蔬农技知识需求分析

以针对草莓的果蔬农技知识需求分析为例，本研究分别对北京市12个区县的草莓种植户进行随机抽样问卷调查，基本涵盖了北京市郊区主要的草莓生产区域，涉及种植户基本情况、技术需求及获得途径等。本次调研共计发放193份调查问卷，回收165份，问卷有效率为85.5%。

(1) 果蔬农技知识需求调研种植户的基本特征，包括性别、年龄、学历、家庭背景、家庭收入、草莓单产水平等，如表1所示。

(2) 从调研种植户的区域分布来看，基本上涵盖了北京市郊区主要的草莓生产区域，包括昌平、通州、顺义等，如图1所示。

(3) 种植户的技术需求类型。种植户对草莓种植农业技术的整体需求如表2所示。93.3%的种植户需要增加产量的良种技术；91.3%的种植户需求是病虫害防控技术；排在第3位的是节本高效栽培技术；种植户需求较低的技术是省工机械技术、贮

表1 果蔬农技知识需求调研种植户的基本特征

Table 1 The characteristics of farmers related with the requirement study on agricultural knowledge

样本种植户	描述	百分比/%
性别	男	95.5
	女	4.5
年龄	40岁及以下	29.1
	40~50	35.3
	50岁及以上	35.6
学历	初中及以下	81.3
	高中及以上	18.7
家庭务农人口比例	30%及以下	14.2
	30%~90%	72.1
	90%及以上	13.7
年家庭收入	5万及以下	24.4
	5万~8万	63.9
	8万及以上	11.7
草莓单产水平	20 t/hm ² 及以上	49.3
	5~20 t/hm ²	47.2
	5 t/hm ² 及以下	3.5

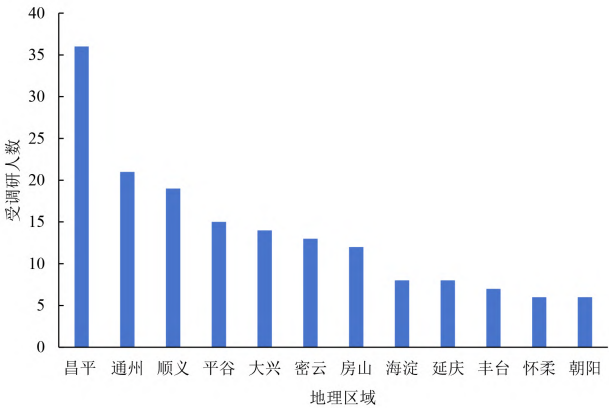


图1 果蔬农技知识需求调研种植户的区域分布

Fig. 1 The district distribution of farmers related with the requirement study on agricultural knowledge

运及加工技术。总体上，北京市郊设施草莓种植户对各种类型技术都表现出强烈的需求欲望，各类技术都有七成以上的种植户选择。设施草莓生产大多是大棚保护地生产，属于劳动密集型农业，机械作业的替代性较小，所以种植户对省工机械技术的需求较少。

(4) 种植户的技术需求。通过种植户按照急需性和重要性对设施草莓栽培技术需求的排序发现，种植户普遍对良种技术需求迫切。首先，27.9%的种植户将提高品质良种技术的急需程度排在第1

表2 草莓种植户的农技需求分布

Table 2 Agricultural technology demand distribution of strawberry farmers

技术类型	需求 比例/%	种植户对现有技术水平的评价			
		满足/%	基本 满足/%	有待 提高/%	亟待 提高/%
种苗培育技术	75.5	13.85	23.08	40.00	23.08
水肥及管理技术	81.2	7.69	32.31	38.46	21.54
病虫害防治技术	91.3	7.58	16.67	51.52	24.24
贮运及加工技术	85.6	13.85	29.23	47.69	9.23
优质生产技术	72.1	10.77	21.54	41.54	26.15
增加产量良种技术	93.3	21.54	33.85	23.08	21.54
节本高效栽培技术	90.7	7.69	26.15	53.85	12.31
省工机械技术	77.7	10.77	38.46	33.85	16.92
新品种应用	69.3	25.76	37.88	16.67	19.70
新农药和肥料应用技术	67.9	32.31	26.15	30.77	10.77
新农具应用技术	73.7	29.23	30.77	26.15	13.85
品牌经营技术	49.2	33.85	15.38	30.77	20.00

位；其次有24.5%的种植户将提高产量良种技术排在第2位，说明随着生活水平的提高，种植户顺应人们消费需求的变化，在高产稳产的基础上愈发重视草莓的品质；急需性排名第3位的是病虫害防控技术，19.8%的种植户认为最需要此项技术；种植户对水肥及管理技术、省工机械技术、节本高效栽培技术的重视程度相当，分别排在第4、5、6位；在对各种类型技术进行排序的过程中，41.1%的种植户把贮运及加工技术排在最后，可见目前大部分种植户不太关注草莓的产后技术。

(5) 种植户获取技术的来源途径。关于种植户技术获得途径的调查显示，近九成种植户没有参加过任何和草莓种植相关的技术培训，基本上还是以自己摸索、凭经验生产为主；其次是通过向其他种植户学习获得技术。48.97%的种植户生产主要依靠自己；39.65%的种植户则是模仿其他种植户生产。在其他获得技术途径方面，4.91%的种植户从政府各级农技推广站的农技人员那里学到相关生产技术；2.65%和2.26%的种植户分别从媒体或乡村干部那里获得生产技术知识，农业合作组织或龙头企业 在草莓生产技术供给方面的作用则微乎其微。

基于以上需求分析，本研究通过构建基于大模型技术的果蔬农技知识智能问答系统创新农业技术推广模式，以此加强对农业技术知识服务的管控，降低种植户优化农技知识储备的时间成本，从而引导更多的种植户实现农业技术创新和转型。

3 果蔬农技知识智能问答大模型构建

本研究基于农技推广现状和草莓农技知识需求分析，针对农技知识导读和农技知识问答等农业领域行业应用，形成知识对象识别和知识问答两种大模型下游任务，获取相关农技知识语料，构建小样本高质量标注语料，训练农业领域知识大模型，并构建果蔬农技知识智能问答系统，总体框架如图2所示。

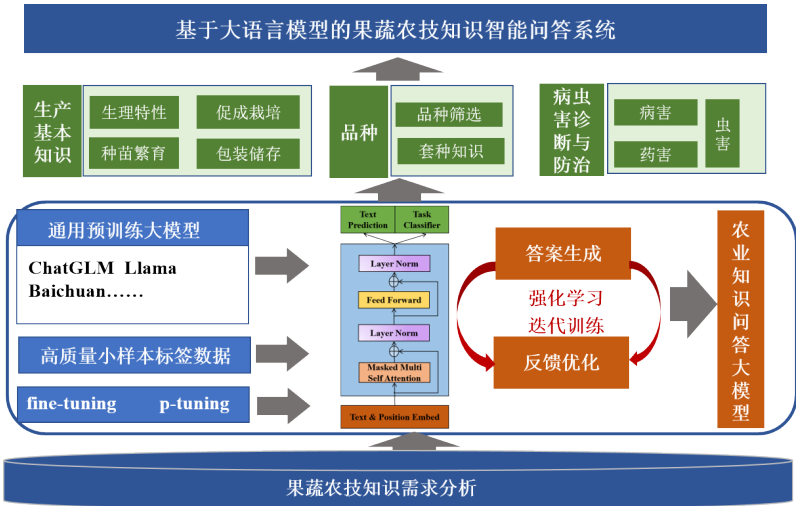


图2 果蔬农技知识智能问答系统总体框架图

Fig. 2 The general framework of intelligent question answering system oriented to agricultural technology

3.1 获取农技知识语料

基于本研究第3节中种植户对草莓农技知识的需求分析,本研究首先对草莓栽培技术知识进行了梳理,划分了草莓生产基本知识、品种筛选、套种知识、虫害诊断与防治、病害诊断与防治、药害诊断与防治6个专题的知识对象。然后,分别根据果蔬农技知识知识对象确定实体类型,形成关键词,并构建关键词检索式。举例说明如下:专题病害的诊断与防治的知识对象包括农作物、作物部位、病害诊断与防治;知识对象“病害诊断与防治”的实体类型包括病害、症状、防治措施,关键词为病害、灰霉病、霜霉病等,对应的关键词检索式为:Summary: (“病害” or “灰霉病” or “霜霉病” or……)。

面向草莓农技知识的6个专题,通过关键词检索式检索知网、维普、万方、国家科技图书文献中心(National Science and Technology Library, NSTL)、中国科学引文数据库等权威学术网站,下载相关摘要或全文,经过清洗、内容重新组织等数据预处理过程,共获取869条草莓农技知识语料。

3.2 形成小样本高质量标注语料

分别针对果蔬农技知识对象识别和果蔬农技知识问答两种下游任务形成预训练大模型微调需要的小样本高质量标注语料。

(1) 标注知识对象实体。利用机器自动标注和人工标注结合的方式对869条知识语料进行标注。首先针对每种知识对象实体类型随机抽取40条知识语料,使用标注工具doccano进行人工标注^[21]。然后把其他标注语料放到大模型ChatGPT里生成伪标签,最后通过人工校对形成大模型标注语料。其中,数据属性包含知识实体、知识实体类型及其位置。

(2) 构建问答对。根据知识语料的内容,通过领域专家转换成高质量的问答对。

(3) 构建训练集和测试集。构建的标注语料中,不同知识主题下,标注语料、知识对象、知识对象类型、问答对数量的具体分布如表3所示。其中,70%的标注语料和问答被用作大模型微调;30%作为测试集检验模型性能。

表3 大模型训练标注语料数据统计

Table 3 The statistical analysis of labeled database for the LLM training

知识专题	生产基 本知识	品种 筛选	套种 知识	虫害诊断 与防治	病害诊断 与防治	药害诊断 与防治	总量
标注语料 数量	151	107	61	212	232	106	869
知识对象 数量	5	1	3	4	4	4	21
知识实体类 型数量	9	2	4	6	6	5	32
问答对数量	63	47	22	44	48	36	260

3.3 构建农业知识实体识别和农业知识问答大模型

农业知识实体识别和农业知识问答大模型的构建过程如下所述:

(1) 初期基础模型候选集构建。基于以下3个方面的综合考虑构建初期基础预训练大模型集:① SuperCLUE 针对中文通用大模型的逻辑推理、知识百科、语言理解、生成创作、对话等各种能力的排名;② github 平台上相关应用获取的星数;③ 是否开源可商用,最终选择 Baichuan2-13B-Chat (Baichuan)^[22]、ChatGLM2-6B (ChatGLM)、Llama-2-13B-Chat (Llama)^[23] 作为初期基础预训练大模型集,并同时和业界热度比较高的大模型 ChatGPT^[24] 进行性能对比。

(2) 预训练大模型微调。由于 Lora 微调^[25] 是一种通过将模型参数矩阵分解为低秩形式,只对部分重要参数进行微调,从而加速模型训练和提高模型性能的方法,具有以下优点:① 参数量小,对算力和存储的要求较低;② 属于小型微调,减少模型参数优化的时间,效率较高;③ 性能与全参微调方法相差不大;④ 生成单独模块,可以和其他微调方法结合使用,所以针对知识实体识别任务,本研究采用 Lora 微调方法。

由于知识问答任务需要根据大模型的生成内容进行多次调整,所以本研究采用 Prompt-tuning 方法^[26] 进行微调。Prompt-tuning 微调通过设计和优化输入提示或模板,来指导预训练模型生成所需要的输出内容,以适应下游任务,其优点包括:① 不需要修改模型参数,计算成本低,需要的训练时间

少,效率高;②属于手动试错过程,允许创建特定任务的提示,可以更灵活地适应各种下游任务;③对于超过10亿参数量的模型来说,小样本甚至是零样本的微调性能也能够极大地被激发出来。

(3)模型优化。随着大模型在各个领域的应用不断增加,很多研究表明大模型生成的文本存在不遵循原文或者不符合事实的现象,称之为大模型幻觉问题^[27]。针对本研究的内容,根据大模型幻觉问题产生内容的性质,主要分为3种类型:信息冲突,大模型的回复是错误的信息;无中生有,对大模型的回复无法验证其真假;信息不匹配,大模型的回复和问题不相关。

针对大模型幻觉问题,分别从数据和模型两个方面进行了微调优化。在数据方面,分别采用标注语料去重和人工剔除可能导致幻觉的数据^[28];在模型方面,采用检索增强生成(Retrieval-Augmented Generation, RAG)方法^[29]增强大模型对领域知识的理解和生成能力。

RAG技术基于检索模型,将大模型和外部知识源结合在一起,不仅可以通过知识库的补充,提高专业知识的广度和深度,加深大模型对专业知识的理解和分析,还为大模型赋予能够随时查询最新知识的能力,保持大模型的知识更新,保证知识的时效性。因此,RAG技术能够有效缓解大模型“幻觉”导致的大模型生成结果不准确、不相关或虚构的问题,提高大模型生成答案内容的质量和有效性。同时,在非开源大模型的应用中,还能够有效解决数据安全的问题,保证敏感数据的安全性。RAG技术具体的实现过程如下:①构建外部知识库。基于维基百科、科技文献、技术教程等数据源,通过网络爬取的方法获取专业领域知识,形成针对知识问答的外部知识源。②构建知识向量库。将知识库中的文档转换成可以处理的文本块,使用向量模型将文本块转换成向量形式,并创建文本块向量的索引,以实现知识库内容的快速搜索。③向量检索。对用户查询内容进行处理,提取关键词转化为向量形式,利用知识向量索引,通过近似最近邻检索等方法在知识向量库中查找最相近的文本块向量,获取与用户查询内容最相关的知识片段。④答案生成。基于Prompt,使得大模型结合用户查询内

容和检索到的知识片段给出问题答案,既包含了用户查询内容的上下文信息,也融合了知识库中检索到的信息。

4 结果与讨论

4.1 知识实体识别结果分析

4.1.1 性能评价指标

本研究采用精准率(Precision, Pre)和召回率(Recall)作为评价知识实体识别的性能指标^[30]。

(1)精准率:Pre为预测正确的实体个数与预测的实体总个数的商,表示衡量检测知识实体识别的整体有效性。

(2)召回率:Recall为预测正确的实体个数与标注的实体总个数的商,表示正确识别知识实体的数量占有所有知识实体的比值。

4.1.2 精准率分析

本研究对比了微调前后不同大模型在不同知识主题下的精准率,如表4所示,可以看出:①微调对不同模型的作用不同。微调后Llama的精准率由76.6%提升到86.52%,得到明显的提升;Baichuan、ChatGPT和ChatGLM的提升幅度为10%~15%;国内业界主流模型ChatGLM、Baichuan、Llama,微调后在不同知识主题下的平均精准率均达到85%以上。②不同模型的精准度不同。在微调前,精准率最高的模型是ChatGPT;微调后,精准度最高的模型是ChatGLM。③知识对象实体类型的数量对精准度的影响不同。相比其他知识主题,定义了9种实体类型的生产基本知识主题的精准率最低;定义了2种实体类型的品种筛选主题的精准率最高。④微调使用的标注语料规模影响模型的精准度。在实体类型数量相同的条件下,标注语料越多,精准率越高。例如,在知识主题虫害诊断和防治、病害诊断与防治、药害诊断与防治的知识实体类型数量都是4的条件下,药害诊断与防治的标注语料为232条,知识主题虫害诊断和防治、病害诊断与防治的语料条数分别为212和106条,药害诊断与防治的模型精准率最高。对于命名实体识别任务,如本研究中的知识实体识别,大语言模型的表现性能和模型架构、标注语料规模、实体类型数量等因素

有密不可分的关系。

表4 初期预训练大模型精准率分析

Table 4 Accuracy analysis of initial pre-trained LLMs

知识主题	Baichuan		Llama		ChatGPT		ChatGLM	
	B/%	A/%	B/%	A/%	B/%	A/%	B/%	A/%
生产基本知识	62.7	82.9	66.7	82.9	62.7	81.9	79.7	86.5
品种筛选	76.3	87.4	80.3	87.4	76.3	82.4	80.3	89.4
套种知识	71.8	85.9	77.8	85.9	73.8	79.9	77.8	87.9
虫害诊断与防治	72.3	85.9	78.3	88.5	74.7	80.9	78.3	91.6
病害诊断与防治	75.0	86.5	80.3	89.2	74.1	81.7	78.3	92.5
药害诊断与防治	72.2	85.2	76.2	85.2	72.2	80.2	76.2	87.2
平均值	71.7	85.6	76.6	86.5	72.3	81.2	78.4	89.2

注：A代表模型微调后召回率；B代表模型微调前召回率。

4.1.3 召回率分析

对比了微调前后各个模型的召回率，如表5所示，可以看出：①不同模型的召回率不同。微调前，召回率最高的模型是Llama；微调后，召回率最高的模型是ChatGPT。②微调前后，精准率和召回率的趋势不一致。微调后，所有模型的精准率都有所增加，但只有模型ChatGLM和Llama的召回率是增加的，其他模型都稍微有些降低。③不同模型的精准率和召回率的表现不一致。微调后，ChatGLM的平均精准率最高，而Llama的平均召回率最高。

表5 初期预训练大模型召回率对比分析

Table 5 Recall analysis of initial pre-trained LLMs

知识主题	Baichuan		Llama		ChatGPT		ChatGLM	
	B/%	A/%	B/%	A/%	B/%	A/%	B/%	A/%
生产基本知识	56.1	47.5	57.3	59.5	51.8	46.7	69.7	70.4
品种筛选	70.7	69.6	70.1	72.9	70.4	62.4	70.2	72.1
套种知识	66.6	67.5	69.2	71.5	68.9	60.1	68.0	70.6
虫害诊断与防治	68.5	63.2	68.2	76.7	69.4	63.5	67.5	75.9
病害诊断与防治	67.2	65.4	68.1	77.9	68.2	65.9	69.3	76.3
药害诊断与防治	65.3	67.1	67.5	77.2	67.3	61.4	65.9	70.1
平均值	65.7	63.4	66.7	72.6	66.0	60.0	68.4	72.6

注：A代表模型微调后召回率；B代表模型微调前召回率。

4.1.4 微调效果分析

由本研究对Lora微调前后的模型性能分析可得，Lora微调后，所有测评模型在所有知识主题下的精准率都有明显提高，大部分召回率都有所提高。Lora微调在预训练大模型已有的神经网络结构和参数的基础上，根据下游任务——命名实体识别更改输出层，并通过输入带有实体类型标签的语料数据来更新模型的部分参数，有效地利用了深度神

经网络强大的泛化能力，所以相比微调之前能够更好地适应下游任务，从而提升模型的精准率和召回率，优化了模型针对命名实体识别任务的表现性能。精准率提升的幅度和微调语料的数量有关，所以会出现相同模型在不同知识主题下的精准率得到不同程度提升的情况。由于模型参数的优化，在模型识别的命名实体中正确识别的比例更大，识别的实体数量相对会有所减少，于是会造成召回率升高幅度小于精准率升高幅度，甚至相比微调前降低的现象。

4.2 农技知识问答结果分析

4.2.1 大模型性能定量分析

本研究通过指标幻觉率和语义相似度定量评估大模型提供答案的质量和幻觉问题的严重程度，并对比分析了大模型在“微调+优化”前后的表现性能。

(1) 性能评价指标。包括幻觉率（Hallucination Rate, HR）和语义相似度（Semantic Similarity, SS）。

幻觉率。模型提供的答案中存在大模型幻觉现象的答案占所回答问题总量的比例，作为衡量大模型提供的答案是否符合事实及大模型幻觉问题严重程度的评价指标。本研究采用“投票决定制”决定大模型提供的答案是否存在大模型幻觉问题，邀请相关专业领域的5名研究人员针对“是”或“否”进行投票，票数多的选项作为最终结果。

语义相似度。衡量模型预测答案和语料库中提供答案之间的相似程度^[31]。具体流程如下：对预测答案和语料答案进行分词、去除停用词；使用词频-逆文本频率指数（Term frequency-Inverse Document Frequency, TF-IDF）方法计算两条答案的词频向量；计算两个词频向量的余弦相似度，值越大表示预测答案和语料答案的相似度越大。

(2) 幻觉率分析。以幻觉率为指标对比分析不同知识主题下ChatGLM、Baichuan、Llama、ChatGPT这4种大模型在“微调+优化”前后出现大模型幻觉现象的严重程度，如图3所示。其中，图3中蓝色条形代表“微调+优化”前的幻觉率；橙色条形代表“微调+优化”后的幻觉率。由图3可得：

① 不同知识主题范围下,大模型幻觉问题的缓解程度不同。对于药害诊断与防治专题,药害的种类相对病虫害较少,且标注语料比较充分,所以大模型幻觉问题的改善更显著。② 相同“微调+优化”方式下,不同大模型对大模型幻觉问题的缓解程度不同。例如,在病害诊断与防治主题下,ChatGPT

的幻觉率降低 10.9%, ChatGLM 降低 43.5%, 其他主题的对比可以得出类似结论。“微调+优化”后,幻觉率以 10%~40% 的幅度降低,大模型幻觉现象得到有效缓解,大模型问答回复的质量有了明显提升。

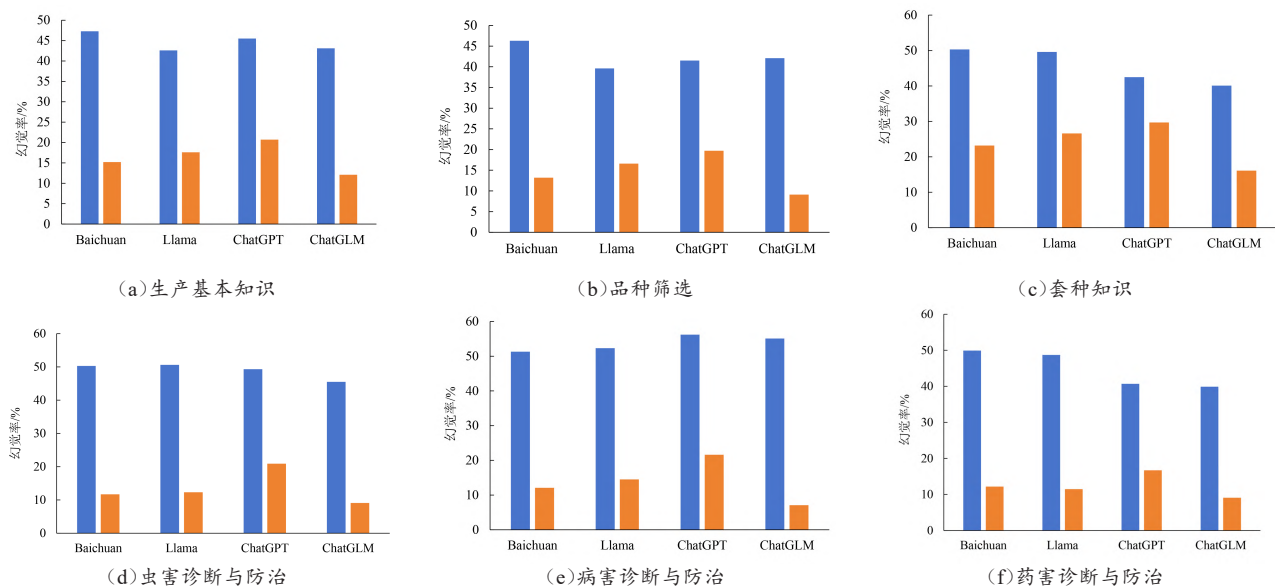


图3 不同知识主题下大模型“微调+优化”前后的幻觉率对比分析

Fig. 3 The hallucination rate comparative analysis of LLMs before and after fine-tuning and optimization in different knowledge topics

(3) 语义相似度分析。由于知识主题病害诊断与防治的数据量最大,本研究以此主题为例,通过指标语义相似度对比分析4种大模型的预测答案和标注答案的相似性,如图4所示。其中,图4中蓝色条形代表“微调+优化”前的语义相似度;橙色条形代表“微调+优化”后的语义相似度。由图4可得,在模型“微调+优化”后,所有测评模型的语义相似度都得到提升,增长幅度均大于15%;在模型“微调+优化”后,ChatGLM的表现最好,其次是Llama,最后是ChatGPT和Baichuan。

(4) “微调+优化”效果分析。由本研究对“微调+优化”前后的模型性能分析可得, prompt-tuning 微调和 RAG 技术优化后,所有测评模型的幻觉率都大幅度降低,语义相似度都明显提高。由于模型在 prompt-tuning 的微调过程中,根据答案生成的质量不断调整模型的输入提示,迭代优化模型性能,从而可以更灵活地适应知识问答任务。同时,由于 RAG 技术给模型增加了外部语料库,使得模型能够更深入地理解知识问答任务的领域知识。因此,

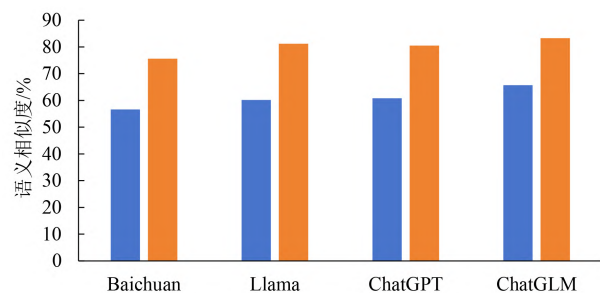


图4 大模型“微调+优化”前后的语义相似度对比分析

Fig. 4 The semantic similarity comparative analysis of LLMs before and after fine-tuning and optimization

所有测评模型在“微调+优化”后,大模型生成的答案质量得到显著提升。

4.2.2 具体问答案例分析

本节以 ChatGLM 对问题“大拱棚促成草莓栽培有什么特点”的回复为例,说明大模型在微调前后性能表现的变化,如表6所示。ChatGLM 在微调前提供的答案不仅分别介绍了“大拱棚”和“半促成”栽培的特点,还针对“草莓品种选择”“土

壤选择”和“病虫害控制”等栽培技术给出建议。虽然这些扩展知识都是在大拱棚半促成栽培方式下草莓栽培需要注意的问题，但却导致答案中存在“避重就轻，重点丢失”的问题。在微调后，Chat-

GLM 提供的答案首先概括介绍“大拱棚半促成草莓栽培”及其优点，然后对比分析“露地促成栽培”和“半促成栽培”的不同，相比之前的回复，对应问题更具有针对性，质量得到明显提升。

表6 大模型“微调+优化”前后的回复内容对比

Table 6 The comparative analysis of LLMs' reply before and after fine-tuning and optimization

问题	微调前	微调后
大拱棚半促成草莓栽培有什么特点	<p>大拱棚半促成草莓栽培是一种草莓种植技术,其特点是:</p> <p>1)大拱棚:该栽培结构较大,通常采用60~70 cm的高拱棚,顶部有通风口和排水口,可以保持土壤湿度和通风良好</p> <p>2)半促成:在这种栽培结构中,草莓生长期较短,通常为20~30 d,而果实成熟期较长,通常为50~60 d。因此,草莓的生长期和果实成熟期比较接近,可以满足消费者对快速生长和大量果实的需求</p> <p>3)草莓品种选择:半促成草莓栽培适合选择生长期较短、果实大小适中、口感鲜美的草莓品种,如红颜、赤星、巨峰等</p> <p>4)土壤选择:半促成草莓栽培通常需要在土壤肥沃、排水良好的地区进行种植,同时需要合理施肥和浇水,以保持土壤湿度和营养供给</p> <p>5)病虫害控制:半促成草莓栽培需要加强对病虫害的控制,可以使用农药进行预防和治理,以保证草莓的安全和丰收</p>	<p>大拱棚半促成草莓栽培是一种高效、经济、方便的草莓种植技术,适合在不同地区进行推广和应用。大拱棚半促成栽培是指草莓在露地促成栽培与大拱棚保护下促成栽培的混合栽培形式。相对于露地促成栽培,半促成栽培具有以下几个特点:</p> <p>1)充分利用白天高温资源,加快草莓生长,缩短上市期,经济效益高</p> <p>2)节省劳动力,减少人工辅助次数,节约劳动力成本</p> <p>3)在半促成栽培期间,温度更为灵活,可以根据实际情况适时调整温度,使植株生长更为适宜</p> <p>4)露地促成栽培草莓在低温下容易“休眠”,而半促成栽培可以在露地促成栽培与保护地促成栽培之间实现“中间状态”,顺利实现“休眠”与“清醒”的转变</p> <p>5)大拱棚半促成栽培相对于露地促成栽培更为灵活,可提前采果,减少空袋率,提高果实产量</p>

4.3 果蔬农技知识智能问答系统

基于以上分析可得，ChatGLM 在农业领域命名实体识别任务和知识问答任务中的表现都是最优的，且是开源可商用的，所以本研究选择 Chat-GLM 作为构建果蔬农技知识智能问答系统的基础预训练模型。果蔬农技知识智能问答系统的功能包括以下两个部分：

(1) 智能问答。针对用户的问题需求给出相应

的答案，如图5所示。其中，ChatGLM 模型的命名实体识别任务在智能问答中的表现形式为：输入一段农业科技文献、农业新闻等农业领域知识，同时以自然语言的形式要求系统给出农业实体类型。

(2) 问答信息源查看。在系统给出问题答案的同时，会把模型生成答案参考的信息源提供给用户。用户可以通过点击信息源链接获取信息源的详细信息，包含文本、图片等多模态的内容，如图6所示。



图5 果蔬农技知识智能问答系统问答界面

Fig. 5 The QA interface of agriculture technology question answering system



图6 检索信息源详细内容展示

Fig. 6 The details presentation of indexed information sources

5 结论

在农业知识服务有待进一步优化的背景下,本研究在分析大模型业界发展态势的基础上,面向农技知识导读和农技知识问答等农业领域行业应用,利用 Baichuan2-13B-Chat、ChatGLM2-6B、Llama-2-13B-Chat、ChatGPT 这4种已有的预训练大模型,构建知识实体识别和知识问答两种农业领域大模型,形成果蔬农技知识智能问答系统。在知识实体识别任务中,通过精准率和召回率对比分析了4种大模型的性能表现,可以得出:

(1) 针对农业领域的命名实体识别任务和知识问答任务,预训练大模型 ChatGLM 的表现性能最优。

(2) 针对预训练大模型下游任务的微调和基于 RAG 技术的模型优化可以显著提升大模型的性能,且对不同大模型性能的影响程度不同。

(3) 知识实体类型的数量、标注语料数量等因素都会影响大模型性能。在知识问答任务中,采用了数据优化、检索增强生成技术等大模型幻觉缓解策略,通过幻觉率和语义相似度两种指标定量分析了大模型幻觉现象和缓解策略使用的有效性,并通过具体问答案例对比分析了微调前后大模型回复质量的变化。

通过以上分析,可以得出大模型技术在知识对象实体识别和知识问答任务中表现优异。随着大语言模型在涌现能力中的不断升级,在多模态知识处理和决策制定方面都将有突破性的表现。但是大模型依然存在性能不稳定等各种问题,在行业领域内的应用有待进一步深入研究。

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] 李秋生,郑凯欣,刘小春. 新时代基层农技推广体系改革创新实践探索、制约因素及深化路径[J]. 世界农业, 2022(2): 80-89.
- [2] 刘娜. 浅析我国农业推广的现状与策略[J]. 河北农业, 2023, 10: 27-28.
- [3] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. Computer science, 2019: ID 160025533.
- [4] ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models[EB/OL]. arXiv: 2303.18223, 2023.
- [5] 李冬梅, 罗斯斯, 张小平, 等. 命名实体识别方法研究综述[J]. 计算机科学与探索, 2022, 16(9): 1954-1968.
- [6] MORWAL S. Named entity recognition using hidden markov model (HMM)[J]. International journal on natural language computing, 2012, 1(4): 15-23.
- [7] EKBAL A, BANDYOPADHYAY S. Named entity recognition using support vector machine: A language independent approach[J]. International journal of electrical and computer engineering, 2010, 4(3): 589-604.
- [8] SONG S L, ZHANG N, HUANG H T. Named entity recognition based on conditional random fields[J]. Cluster computing, 2019, 22(3): 5195-5206.
- [9] LUO L, YANG Z H, YANG P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics, 2018, 34(8): 1381-1388.
- [10] CHANG Y A, KONG L, JIA K J, et al. Chinese named entity recognition method based on BERT[C]// 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA). Piscataway, New Jersey, USA: IEEE, 2021: 294-299.
- [11] ZHU Y Y, WANG G X, KARLSSON B F. CAN-NER: Convolutional attention network for Chinese named entity recognition[EB/OL]. arXiv: 1904.02141, 2019.
- [12] ZHANG Y, YANG J. Chinese NER using lattice LSTM[EB/OL]. arXiv: 1805.02023, 2018.
- [13] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. arXiv: 1810.04805, 2018.
- [14] SUN Y, WANG S H, LI Y K, et al. ERNIE: Enhanced representation through knowledge integration[EB/OL]. arX-

- iv: 1904.09223, 2019.
- [15] RILOFF E, THELEN M. A rule-based question answering system for reading comprehension tests[C]// Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems-Volume 6. New York, USA: ACM, 2000: 13-19.
- [16] YANI M, KRISNADHI A A. Challenges, techniques, and trends of simple knowledge graph question answering: A survey[J]. Information, 2021, 12(7): ID 271.
- [17] SHARMA Y, GUPTA S. Deep learning approaches for question answering system[J]. Procedia computer science, 2018, 132: 785-794.
- [18] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. arXiv: 1907.11692, 2019.
- [19] CHIPMAN H A, GEORGE E I, MCCULLOCH R E. BART: Bayesian additive regression trees[J]. The annals of applied statistics, 2010, 4(1): 266-298.
- [20] PEREIRA J, FIDALGO R, LOTUFO R, et al. Visconde: Multi-document QA with GPT-3 and Neural Reranking[C]// European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2023: 534-543.
- [21] DAUDERT T. A web-based collaborative annotation and consolidation tool[J]. International conference on language resources and evaluation, 2020: 7053-7059.
- [22] YANG A Y, XIAO B, WANG B N, et al. Baichuan 2: Open large-scale language models[EB/OL]. arXiv: 2309.10305, 2023.
- [23] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models[EB/OL]. arXiv: 2302.13971, 2023.
- [24] FLORIDI L, CHIRIATTI M. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and machines, 2020, 30(4): 681-694.
- [25] DING N, QIN Y J, YANG G A, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature machine intelligence, 2023, 5(3): 220-235.
- [26] LIU X A, JI K X, FU Y C, et al. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 61-68.
- [27] BANG Y J, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[EB/OL]. arXiv: 2302.04023, 2023.
- [28] LEE K, IPPOLITO D, NYSTROM A, et al. Deduplicating training data makes language models better[EB/OL]. arXiv: 2107.06499, 2021.
- [29] PENG B L, GALLEY M, HE P C, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback[EB/OL]. arXiv: 2302.12813, 2023.
- [30] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models[EB/OL]. arXiv: 2307.03109, 2023.
- [31] CORLEY C, MIHALCEA R. Measuring the semantic similarity of texts[C]// Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment-EMSEE '05. Morristown, New Jersey, USA: Association for Computational Linguistics, 2005: 13-18.

Agricultural Technology Knowledge Intelligent Question-Answering System Based on Large Language Model

WANG Ting^{1,2}, WANG Na³, CUI Yunpeng^{1,2}, LIU Juan^{1,2}

(1. Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; 2. Key Laboratory of Big Agri-data, Ministry of agriculture and rural areas, Beijing 100081, China; 3. Unit 96962, Beijing 102206, China)

Abstract:

[Objective] The rural revitalization strategy presents novel requisites for the extension of agricultural technology. However, the conventional method encounters the issue of a contradiction between supply and demand. Therefore, there is a need for further innovation in the supply form of agricultural knowledge. Recent advancements in artificial intelligence technologies, such as deep learning and large-scale neural networks, particularly the advent of large language models (LLMs), render anthropomorphic and intelligent agricultural technology extension feasible. With the agricultural technology knowledge service of fruit and vegetable as the demand orientation, the intelligent agricultural technology question answering system was built in this research based on LLM, providing agricultural technology extension services, including guidance on new agricultural knowledge and question-and-answer sessions. This facilitates farmers in accessing high-quality agricultural knowledge at their convenience.

[Methods] Through an analysis of the demands of strawberry farmers, the agricultural technology knowledge related to strawberry cultivation was categorized into six themes: basic production knowledge, variety screening, interplanting knowledge, pest diagnosis

and control, disease diagnosis and control, and drug damage diagnosis and control. Considering the current situation of agricultural technology, two primary tasks were formulated: named entity recognition and question answering related to agricultural knowledge. A training corpus comprising entity type annotations and question-answer pairs was constructed using a combination of automatic machine annotation and manual annotation, ensuring a small yet high-quality sample. After comparing four existing Large Language Models (Baichuan2-13B-Chat, ChatGLM2-6B, Llama 2-13B-Chat, and ChatGPT), the model exhibiting the best performance was chosen as the base LLM to develop the intelligent question-answering system for agricultural technology knowledge. Utilizing a high-quality corpus, pre-training of a Large Language Model and the fine-tuning method, a deep neural network with semantic analysis, context association, and content generation capabilities was trained. This model served as a Large Language Model for named entity recognition and question answering of agricultural knowledge, adaptable to various downstream tasks. For the task of named entity recognition, the fine-tuning method of Lora was employed, fine-tuning only essential parameters to expedite model training and enhance performance. Regarding the question-answering task, the Prompt-tuning method was used to fine-tune the Large Language Model, where adjustments were made based on the generated content of the model, achieving iterative optimization. Model performance optimization was conducted from two perspectives: data and model design. In terms of data, redundant or unclear data was manually removed from the labeled corpus. In terms of the model, a strategy based on retrieval enhancement generation technology was employed to deepen the understanding of agricultural knowledge in the Large Language Model and maintain real-time synchronization of knowledge, alleviating the problem of LLM hallucination. Drawing upon the constructed Large Language Model, an intelligent question-answering system was developed for agricultural technology knowledge. This system demonstrates the capability to generate high-precision and unambiguous answers, while also supporting the functionalities of multi-round question answering and retrieval of information sources.

[Results and Discussions] Accuracy rate and recall rate served as indicators to evaluate the named entity recognition task performance of the Large Language Models. The results indicated that the performance of Large Language Models was closely related to factors such as model structure, the scale of the labeled corpus, and the number of entity types. After fine-tuning, the ChatGLM Large Language Model demonstrated the highest accuracy and recall rate. With the same number of entity types, a higher number of annotated corpora resulted in a higher accuracy rate. Fine-tuning had different effects on different models, and overall, it improved the average accuracy of all models under different knowledge topics, with ChatGLM, Llama, and Baichuan values all surpassing 85%. The average recall rate saw limited increase, and in some cases, it was even lower than the values before fine-tuning. Assessing the question-answering task of Large Language Models using hallucination rate and semantic similarity as indicators, data optimization and retrieval enhancement generation techniques effectively reduced the hallucination rate by 10% to 40% and improved semantic similarity by more than 15%. These optimizations significantly enhanced the generated content of the models in terms of correctness, logic, and comprehensiveness.

[Conclusion] The pre-trained Large Language Model of ChatGLM exhibited superior performance in named entity recognition and question answering tasks in the agricultural field. Fine-tuning pre-trained Large Language Models for downstream tasks and optimizing based on retrieval enhancement generation technology mitigated the problem of language hallucination, markedly improving model performance. Large Language Model technology has the potential to innovate agricultural technology knowledge service modes and optimize agricultural knowledge extension. This can effectively reduce the time cost for farmers to obtain high-quality and effective knowledge, guiding more farmers towards agricultural technology innovation and transformation. However, due to challenges such as unstable performance, further research is needed to explore optimization methods for Large Language Models and their application in specific scenarios.

Key words: LLM; generative pre-trained transformer; agricultural technology knowledge; intelligent question answering; name entity identity

Foundation items: Beijing Digital Agriculture Innovation Consortium Project (BAIC10-2023); Fundamental Research Funds of Chinese Academy of Agricultural Sciences (JBYW-AII-2023-31); National Key Research and Development Program of China (2022YFF0711902)

Biography: WANG Ting, E-mail: wangting01@caas.cn

***Corresponding author:** CUI Yunpeng, E-mail: cuiyunpeng@caas.cn

(登陆 www.smartag.net.cn 免费获取电子版全文)