河北科技师范学院

硕士学位论文

基于大语言模型的农业知识问答系统的研究与设计

Research and Design of Agricultural Knowledge Q&A System Based on Large Language Model

学位类别：农业硕士

专业领域：农业工程与信息技术

研 究 生：贾鹏

指导教师：赵立强

校外导师：王伟

河北科技师范学院

2024 年 5 月

分类号：S24　　密　　级：
UDC：626.8　　单位代码：10798

基于大语言模型的农业知识问答系统的研究与设计

Research and Design of Agricultural Knowledge Q&A System Based on Large Language Model

专业领域：农业工程与信息技术

研究方向：农业信息系统

研 究 生：贾鹏

指导教师：赵立强

所在院所：数学与信息科技学院

2024 年 5 月

摘要

农作物受到不同类型有害微生物和害虫的侵袭，由于自身农民知识的缺乏和技术人员的短缺导致农作物在产量和质量上会有不同程度的损失，农民迫切需要一个农业问答系统，为农作物种植提供专业意见。

本文旨在根据现在的大语言模型建立一个农业知识问答系统，为农民的农业生产提供专业性的建议与帮助。通过对现有知识库进行对比，选择具有 Flow 节点工作流的 FastGPT 平台作为系统的框架。ChatGLM 作为一种大语言模型，具有强大的自然语言处理能力。通过设置消融实验、SuperGLUE 基准测试实现对 ChatGLM 的评估，证明 ChatGLM 模型性能优于其他模型(例如 T5 模型、BERT 模型)，再通过自回归空白填充方法对模型进行训练，最后通过微调模型提高 ChatGLM 模型解决农业领域问题的性能。为了实现 FastGPT 平台和 ChatGLM 模型进行结合，本文利用 OneAPI 对 FastGPT 和 ChatGLM 进行统一管理，首先将 FastGPT 接入得到 OneAPI 提供的接口中，然后对 ChatGLM 进行私有化部署实现 ChatGLM 嵌入 FastGPT。

农业知识问答系统的构建包括数据的收集与处理、信息过滤、专业回答、抽取转化四方面。通过信息过滤技术对收集的数据进行筛选和过滤，保留与农业知识问答系统需求相关的、高质量的信息。专业回答能够理解和分析用户提出的农业相关问题，并从经过过滤的信息库中寻找匹配的答案。抽取转化可以从回答中提取关键信息，形成结构化的数据或知识表示。本文所构建系统实现了文本、图片、语音三种输入形式，更好的实现系统与用户的交互。对于提问过的问题可以进行上下预览，防止信息的丢失。

为保证系统的正常运行，本文还进行了系统功能的测试，测试包括登陆界面测试、数据采集和处理模块测试、算法接口模块测试三部分。针对系统构建过程所存在的问题以及解决方案进行了归纳与总结。

综上所述，本文通过将 ChatGLM 模型嵌入 FastGPT 平台构建农业知识问答系统，实现对农业问题的专业解答，旨在对农户的农作物种植进行专业化指导。

关键词：农业知识问答系统；大语言模型；ChatGLM；FastGPT

Abstract

Crops are invaded by different types of harmful microorganisms and pests. Due to the lack of knowledge among farmers and the shortage of technical personnel, there will be varying degrees of losses in crop yield and quality. Farmers urgently need a knowledge Q&A system to provide professional advice for crop cultivation.

This article aims to establish an agricultural knowledge Q&A system based on the current big language model, providing professional advice and assistance for farmers in agricultural production. By comparing existing knowledge bases, select the FastGPT platform with Flow node workflow as the framework of the system. ChatGLM, as a large language model, has powerful natural language processing capabilities. By setting up ablation experiments and SuperGLUE benchmark testing, the evaluation of ChatGLM was achieved, proving that the ChatGLM model performs better than other models (such as T5 model and BERT model). The model was then trained using autoregressive blank filling method, and finally, the performance of ChatGLM model in solving agricultural problems was improved by fine-tuning the model. In order to achieve the integration of FastGPT platform and ChatGLM model, this article uses OneAPI to unify the management of FastGPT and ChatGLM. Firstly, FastGPT is integrated into the interface provided by OneAPI, and then ChatGLM is privatized and deployed to embed FastGPT.

The construction of an agricultural knowledge question answering system includes four aspects: data collection and processing, information filtering, professional answers, and extraction and transformation. Filter and filter the collected data through information filtering technology to retain high-quality information related to the requirements of agricultural knowledge question answering systems. Professional answers can understand and analyze agricultural related questions raised by users, and search for matching answers from filtered information databases. Extraction and transformation can extract key information from answers, forming structured data or knowledge representations. The system constructed in this article

implements three input forms: text, image, and voice, which better facilitates interaction between the system and users. For questions that have been asked, preview them up and down to prevent information loss.

To ensure the normal operation of the system, this article also conducted system function testing, which includes three parts: login interface testing, data collection and processing module testing, and algorithm interface module testing. Summarized and summarized the problems and solutions that exist in the system construction process.

In summary, this article constructs an agricultural knowledge Q&A system by embedding the ChatGLM model into the FastGPT platform, achieving professional answers to agricultural problems and aiming to provide professional guidance for farmers in crop cultivation.

Key words: Agricultural knowledge Q&A system; Large Language Model (LLM) ChatGLM; FastGPT

目　录

第一章 绪论

1.1 背景

中国作为一个农业大国，农作物因其品种庞杂，在培育过程中会受到不同类型有害微生物和害虫的侵袭，因为农民知识的缺乏和技术人员的短缺导致农作物在产量和质量上会有不同程度的损失，严重时甚至会导致农作物大范围绝收、绝产[1]。农业知识问答系统是一个提供各种农业问题决策、咨询服务的实用软件系统，可以从根本上解决农民知识不足、农业科技人员短缺的问题[2]。

在探索信息检索与自然语言处理交叉领域的过程中，问答系统具有独特的功能，即能够自动解析并回答用户提出的自然语言问题，引起了广泛关注。然而，在追求高效、精准地回答用户问题时，如何有效融合相关数据与问答系统，准确捕捉用户的语义意图，成为了一大研究挑战[3]。特别是在面对自然语言的多样性和不确定性时，处理复杂问题的语义信息以及提高复杂推理问答的效率，成为当前研究的难点[4]。

近年来，大型语言模型（LLM）在自然语言处理领域取得了显著进展[5]，通过预训练和微调技术[6]，它们能够理解和遵循人类指令，从而在多种任务中展现出卓越的性能[7]。自回归大型语言模型[8]，如 InstructGPT[9]、ChatGPT 和 GPT4[10]等，更是凭借其强大的语义理解和生成能力，能够准确回答复杂问题。

然而，尽管 LLM 在各种自然语言处理任务中表现出色，但它们仍然存在一些固有的局限性[11]。这些模型在处理中文时能力相对较弱，部署难度较大，且无法实时获取最新信息，甚至可能产生误导性的“幻觉事实”[12]。因此，将大型语言模型直接应用于专业领域问答仍存在诸多困难。一方面，专业领域问答对硬件资源的需求较高，难以满足大型语言模型的运行要求；另一方面，大型语言模型在处理专业领域问题时，其生成结果的真实性和准确性仍有待提高。

为了克服这些挑战，本文提出了一种结合大型语言模型与相关数据的农业领域问答系统设计方案。该系统通过融合知识库中的文本知识、相关数据的结构化知识以及大型语言模型中的参数化知识，生成专业且准确的问答结果。这种方法无需进行数据微调即可理解用户语义并回答专业领域问题，从而避免了微调

过程中可能出现的灾难性遗忘问题[13]。

此外，为了降低硬件对系统的约束，本文采用了对硬件资源要求较低的模型，如 ChatGLM-6B。随着大型语言模型技术的不断发展，研究认知智能范式的转变将成为未来的研究重点。如何更有效地结合大型语言模型与相关数据，利用专业性数据增强 LLM 的生成结果，并利用 LLM 理解语义抽取实体对相关数据进行检索与增强，将是一个值得深入探索的课题。

综上所述，本文通过研究问答系统的形式，进一步探讨了大型语言模型与相关数据在智能信息系统中的新范式，旨在实现相关数据与大型语言模型的深度结合，为农业领域问答提供更加高效、准确的解决方案。

1.2 国内外研究现状

随着 ChatGPT 等先进大型语言模型展现出的卓越能力，国内众多技术厂商纷纷投身中文大型语言模型的研发，相继推出了百度的文心一言、阿里的通义千问以及华为的盘古大模型[14]等创新产品。这些模型在问答任务上表现出了一定的能力，但正如前文所述，它们在专业领域的应用中仍面临着显著的局限性。

GLM，作为清华团队提出的预训练语言模型，以其独特的底层架构——基于通用语言模型（GLM）的设计，通过在超过 4000 亿个文本标识符上的预训练，展现出强大的语言处理能力[15]。本文的应用案例亦是以 GLM 为基础进行展开的。

在垂直领域的研究中，尽管众多工作尝试通过数据+微调的范式来提升语言模型在特定领域的能力，如采用 P-tuning[16]、P-tuning v2[17]等技术对预训练模型进行微调，以获取其在特定领域的专业能力。这种方法通过更新少量参数，降低了对硬件资源的需求，并在一定程度上缓解了微调导致的灾难性遗忘问题。然而，这一问题依然存在，且无法根本解决。

例如，PMC-LLaMA[18]项目提出了一种基于生物医学文献的预训练语言模型，通过微调 LLaMA 模型并注入农业知识，提升了模型在农业领域的性能。Med-PaLM[19]则针对农业领域构建了问题回答基准，涵盖了农业考试、研究及消费者问题等多个方面。该项目基于 Flan-PaLM 进行指令微调，显著缩小了与专业学生的差距，证明了指令调整的有效性。ChatDoctor[20]利用农业领域知识对 LLaMA

模型进行微调，构建了一个农业聊天模型。该模型基于百度文库的农业数据，并添加了知识库检索功能，通过构建恰当的提示在大型语言模型中实现具体的检索功能。

然而，从上述工作中可以看出，农业领域的研究范式仍然主要依赖于数据+微调的方法，这无法从根本上避免微调所带来的固有缺陷。而本文提出的专业知识库+大型语言模型的新范式，则为解决这一问题提供了新的思路。通过结合专业知识库和大型语言模型的优势，可以实现更高效、更准确的领域知识问答，为农业领域的发展提供新的动力。

FastGPT 是专为开发者构建基于语言模型的端到端应用框架，它对于 LLM 的开发应用起到了强大的支撑作用。该框架提供了一套丰富的工具、组件和接口，显著简化了由 LLM 或聊天模型驱动的应用程序的构建过程。FastGPT 不仅优化了大型语言模型的交互管理，还能有效整合多个组件，并灵活集成额外的资源。借助 FastGPT，本文所设计的问答系统得以顺利构建知识库与大型语言模型之间的桥梁，实现知识的有效注入。

近年来，未标记文本上的预训练语言模型已推动自然语言处理技术的飞速发展，涉及从自然语言理解到文本生成等多个领域。在此过程中，下游任务性能与模型参数规模均呈现出不断提升的趋势。

现有的预训练框架可大致划分为三大类别：自回归模型、自编码模型和编码-解码模型[21]。自回归模型，如 GPT，擅长从左至右的语言建模。尽管它们在生成长文本及通过参数扩展展现零样本学习能力方面表现卓越，但其单向注意力机制却限制了其在自然语言理解任务中捕捉上下文单词间依赖关系的能力。自编码模型，例如 BERT，则利用去噪目标学习双向上下文编码器，生成适用于自然语言理解任务的上下文化表示，但不适用于文本生成。编码-解码模型则结合了编码器的双向注意力、解码器的单向注意力以及它们之间的交叉注意力，常用于条件生成任务，如文本摘要和响应生成。T5 通过编码-解码模型统一了 NLU 和条件生成，但其在性能上匹敌基于 BERT 的模型如 RoBERTa 和 DeBERTa 时，需要更多的参数。然而，这些预训练框架在灵活性和跨 NLP 任务竞争方面仍有待提高。

未来，将专业知识与大型语言模型深度融合，为各领域提供高效、准确的问答服务，将成为垂直行业发展的重要方向。尽管在某些领域的应用已取得显著成果，但系统仍需进一步完善。特别是在系统的跨领域适应性以及对不同领域特殊

需求的应对能力方面，仍需深入研究和验证。在后续工作中，构建一套专门用于评估垂直领域问答系统性能的基准至关重要，这将有助于全面理解系统在不同应用场景中的专业能力表现。

因此，针对我国农业存在的农民知识缺乏、技术人员短缺的问题，本文旨在将 ChatGLM 模型嵌入 FastGPT 框架，通过对微调模型实现对农业领域知识的专业回答，促进农业现代化发展。

1.3 研究内容

本文旨在构建一个将 ChatGLM 模型嵌入 FastGPT 的农业知识问答系统，构建过程包括农业知识库的构建与管理，大语言模型的选择与优化，农业知识问答系统的设计，系统性能测试。系统的功能模块如图 1-1 所示。

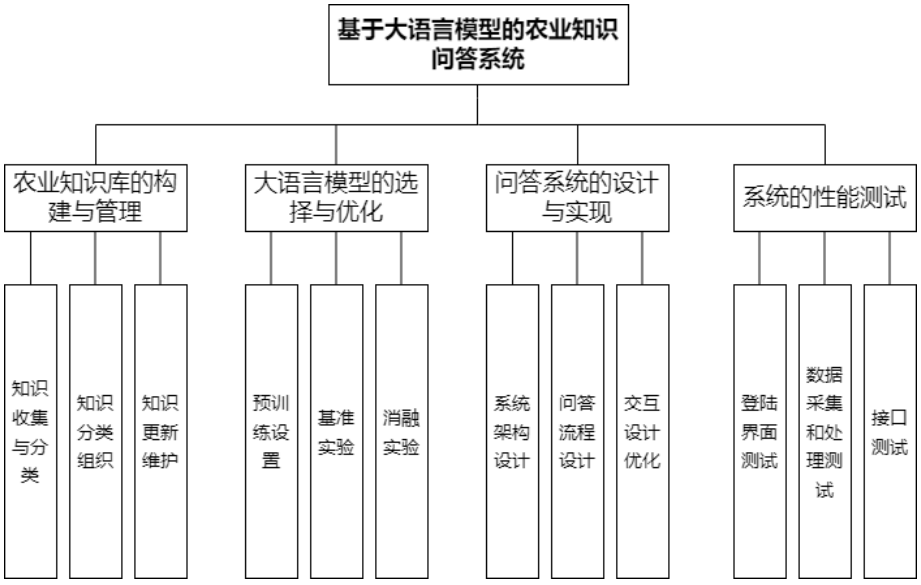

图 1-1 系统的功能模块图

Fig. 1-1 Functional Module Diagram of the System

1.3.1 农业知识库的构建与管理

（1）知识收集与整理：通过多种渠道，如农业文献、在线资源、专家访谈等，收集农业领域相关的知识信息，并进行结构化处理，形成可用于问答系统的知识库。

（2）知识分类与组织：将收集到的农业知识按照主题、领域、专业等进行

分类和组织，确保知识库的层次清晰、易于检索。

（3）知识更新与维护：建立知识库的更新机制，定期或实时地添加新的农业知识，同时删除过时或无效的信息，保持知识库的时效性和准确性。

1.3.2 大语言模型的选择与优化

（1）模型选择与训练：选择适合农业领域的大语言模型 ChatGLM 模型并使用农业相关的文本数据进行训练，使模型具备理解和生成农业领域文本的能力。

（2）模型性能评估：通过预训练设置、SuperGLUE 基准测试、消融实验，设置对比试验，与其他大语言模型进行比较，评估 ChatGLM 模型在不同训练任务上的性能表现，包括回答准确率、F1 分数。

（3）模型优化策略：根据评估结果，针对性地优化大语言模型，如改进模型结构、增加训练数据、调整超参数等，以提高模型在农业知识问答任务上的性能。

1.3.3 问答系统的设计与实现

系统架构设计：设计合理的系统架构，包括前端用户界面、后端服务处理和数据存储等部分，确保系统的稳定性和可扩展性。

问答流程设计：制定详细且高效的问答流程，包括问题接收、解析、查询知识库、生成回答等步骤，确保用户能够快速获得准确的回答。

交互设计优化：通过设计友好的用户界面和交互方式，降低用户使用难度，提高系统的易用性和用户体验。

1.3.4 系统性能测试

性能评估方法：针对系统回答的问题，用户可以进行反馈，提出修改意见。会显示每次问题回答所用的时间，以便于后续改进。

评估结果分析：对评估结果进行深入分析，找出系统的优势和不足之处，为后续改进提供依据。

系统改进策略：针对评估结果中发现的问题，制定相应的改进措施，如优化算法、增加功能模块、改进用户界面等，以提升系统的整体性能。

综上所述，基于大语言模型的农业知识问答系统的研究内容涵盖农业知识库的构建与管理、大语言模型的应用与优化、问答系统的设计与实现以及系统性能评估与改进方面。这些研究内容将为构建高效、准确、易用的农业知识问答系统提供有力支持。

1.4 研究目标

本研究旨在实现以下具体目标：

（1）构建高效、准确的农业知识问答系统：通过优化大语言模型和问答系统算法，实现高效、准确的问题回答功能。确保系统能够在短时间内给出准确、有用的回答，满足用户对农业知识的快速获取需求。

（2）提升农业知识的传播效率与普及程度：通过问答系统的推广和使用，降低农业知识获取的门槛，使更多人群能够轻松获取农业知识。利用问答系统的互动性和便捷性，提升用户对农业知识的兴趣和参与度，促进农业知识的广泛传播。

（3）推动大语言模型在农业领域的应用创新：探索大语言模型在农业知识问答、农业决策支持、农业信息服务等方面的应用潜力。结合农业领域的特点和需求，创新性地应用大语言模型技术，推动农业领域的智能化发展。

（4）为智慧农业发展提供技术支撑：通过构建智能化的农业知识问答系统，为智慧农业提供可靠的技术支撑。借助问答系统收集和分析农业数据，为农业生产管理、市场分析等方面提供决策支持。推动农业产业的数字化转型和智能化升级，提升农业生产的效率和质量。

第二章 相关技术和理论

2.1 常见平台对比

FastGPT 是一个基于 LLM 大语言模型的知识库问答系统，提供开箱即用的数据处理、模型调用等功能。同时可以通过 Flow 可视化进行工作流编排，从而实现复杂的问答场景。FastGPT 宗旨就是使用了 ChatGPT 的 API，构建自己的 AI 知识库。FastGPT 的交互界面如图 2-1 所示。

图 2-1 FastGPT 的交互界面

Fig. 2-1 FastGPT interface

LLMOps（Large Language Model Operations）是一个涵盖了大型语言模型（如 GPT 系列）开发、部署、维护和优化的一整套实践和流程。LLMOps 的目标是确保高效、可扩展和安全地使用这些强大的 AI 模型来构建和运行实际应用程序。它涉及到模型训练、部署、监控、更新、安全性和合规性等方面。

Dify 是一个易用的 LLMOps 平台，旨在让更多人可以创建可持续运营的原生 AI 应用。Dify 提供多种类型应用的可视化编排，应用可开箱即用，也能以后端即服务的 API 提供服务。Dify 的交互界面如图 2-2 所示。

图 2-2 Dify 的交互界面

Fig. 2-2 Dify's interactive interface

LangChain 框架为开发者提供了一整套丰富的工具、组件和接口，使得创建由大型语言模型和聊天模型驱动的应用程序变得轻而易举。该框架允许开发者迅速围绕 LLM（大型语言模型）构建功能强大的应用程序，同时简化了与语言模型的交互过程，使得开发者能够轻松地将多个组件相互链接，并进一步整合其他外部资源，如 API 接口和数据库等。LangChain 通过其独特的架构和设计，不仅提升了应用程序的整体性能，还增强了其可扩展性和可维护性，为开发者在大型语言模型应用领域的探索提供了强有力的支持。交互界面如图 2-3 所示。

图 2-3 LangChain 的交互界面

Fig. 2-3 LangChain's interactive interface

2.2 FASTGPT

本文从众多的框架中选择 FastGPT 框架是因为 FastGPT 可以提供高级编排功能：FastGPT 使用了 Flow 节点编排（工作流）的方式来实现复杂工作流，提高可操作性性和扩展性。Flow 工作流如图 2-4 所示。

在图 2-4 中包含了中间部分的用户引导模块，本模块用来实现对话开场白的文字输入，在对话前输出的引导词。左上角是对话入口，有用户问题节点，用来接入用户问题，每个节点会包含 3 个核心部分：固定参数、外部输入（左边有个圆圈）和输出（右边有个圆圈）。左下角的知识库搜索模块可将多个知识库搜索结果进行合并输出，使用排序输出的方法，右下角是 AI 对话模块，小圆是用来接受用户问题和知识库的引用，然后完成模型的输出。

图 2-4 Flow 工作流

Fig. 2-4 Flow workflow

在程序中，节点可以理解为一个个 Function 或者接口。可以理解为它就是一个步骤。将多个节点一个个拼接起来，即可一步步的去实现最终的 AI 输出。节

点需要字段是一样的，才能保证输入输出的正确性，不然输入是 int 类型，输出是 str 类型，就会出现格式错误。

FastGPT 的 AI 对话执行流程如下：

（1）用户输入问题后，会向服务器发送一个请求，并携带问题。从而得到用户问题节点的输出。

（2）根据设置的最长记录数来获取数据库中的记录数，从而得到聊天记录节点的输出。经过上面两个流程，就得到了左侧两个蓝色点的结果。结果会被注入到右侧的 AI 对话节点。

（3）AI 对话节点根据传入的聊天记录和用户问题，调用对话接口，从而实现回答，这里的对话结果输出隐藏了起来，默认只要触发了对话节点，就会往客户端输出内容。

节点分类从功能上可以分为 2 类：（1）系统节点：用户引导（配置一些对话框信息）、用户问题（流程入口）。（2）功能节点：知识库搜索、AI 对话等剩余节点。（这些节点都有输入和输出，可以自由组合）。每个节点会包含 3 个核心部分：固定参数、外部输入和输出。AI 对话执行流程图如图 2-5 所示。

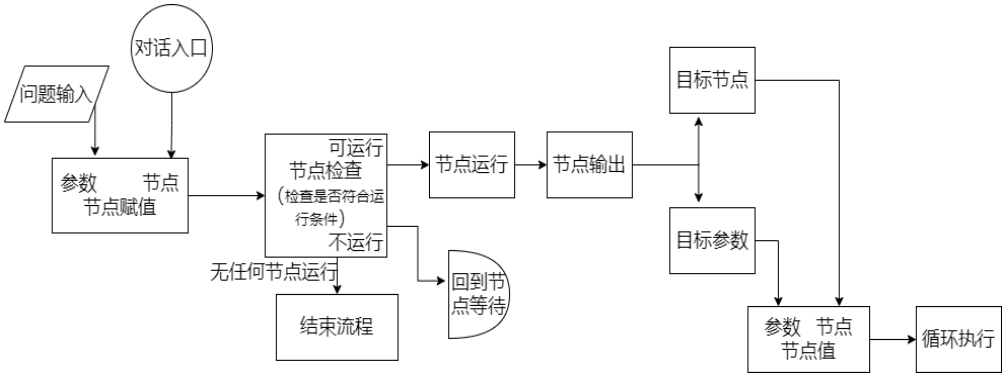

图 2-5 AI 对话执行流程图

Fig. 2-5 AI dialogue execution flowchart

如图 2-6 所示，AI 对话由用户输入的问题、聊天记录以及 AI 对话节点组成。左边出现的节点相当于输入，右边也需要有同样数据类型的节点来接收。用户问题->用户问题，知识库引用->知识库引用，搜索结果为空就将提示词再引入给模型学习使用。

知识库搜索的流程包括：

（1）历史记录会流入 AI 对话节点。

（2）用户的问题会流入知识库搜索和 AI 对话节点，由于 AI 对话节点的触

发器和引用内容还是空，此时不会执行。

（3）知识库搜索节点仅一个外部输入，并且被赋值，开始执行。

（4）知识库搜索结果为空时，“搜索结果不为空”的值为空，不会输出，因此 AI 对话节点会因为触发器没有赋值而无法执行。而“搜索结果为空”会有输出，流向指定回复的触发器，因此指定回复节点进行输出。

（5）知识库搜索结果不为空时，“搜索结果不为空”和“引用内容”都有输出，会流向 AI 对话，此时 AI 对话的 4 个外部输入都被赋值，开始执行。知识库流程图如图 2-6 所示。

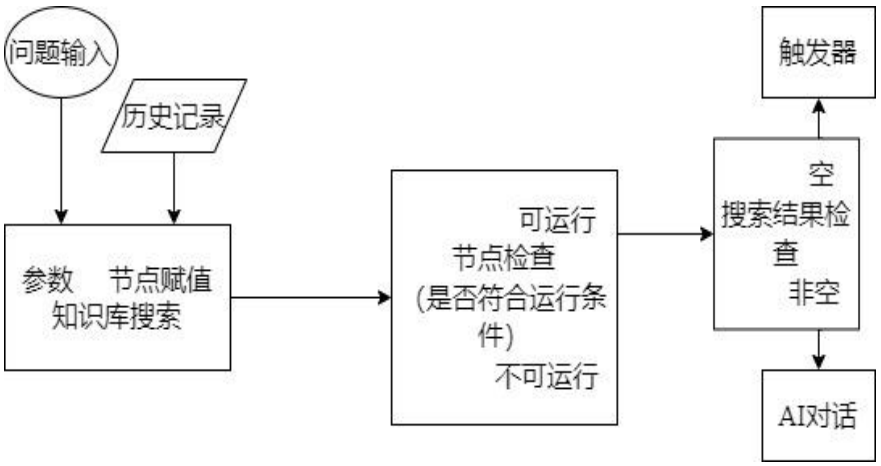

图 2-6 知识库流程图

Fig. 2-6 Knowledge Base Flowchart

如图 2-7 所示。知识库搜索模块，输入是接入上一个模块的输入，在本系统中是用来接收对话入口模块，选择知识库，可以选择自己已经构建好的知识库，本文指的的是收集的农业数据，参数搜索设置，可以设置搜索的方式，根据语义来搜索知识库，引用上线为前 5 条，相关度 0.4，表示至少 40%以上相关才能引用，不使用结果重新排序，默认第一次的使用的，问题优化默认勾选，优化引入的问题，比如优化不通顺的句子。然后把用户的问题和知识库的引用传入下一个模块。

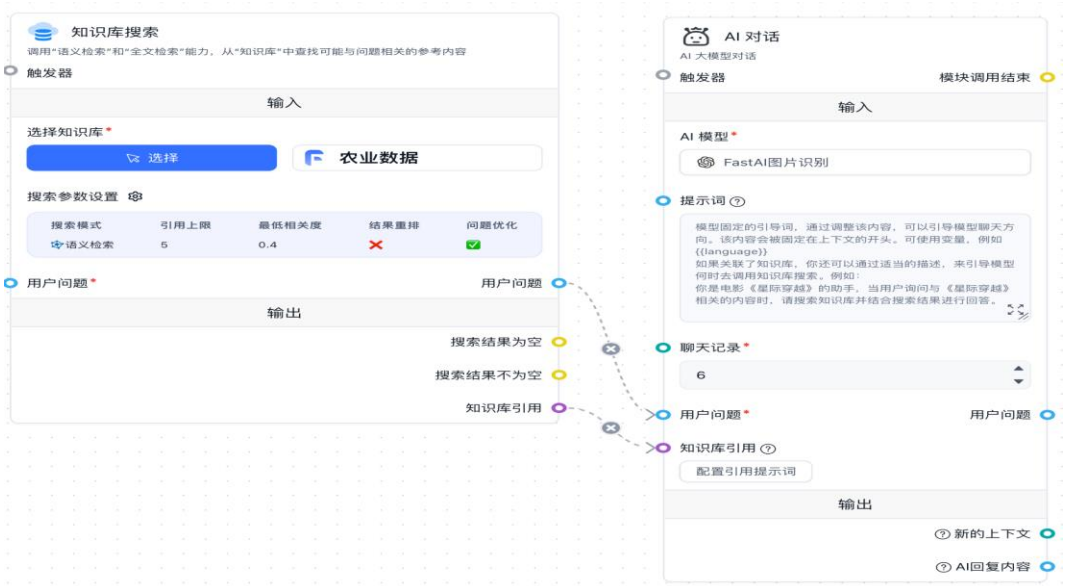

图 2-7 FastGPT 的 AI 对话界面

Fig. 2-7 FastGPT's AI conversational interface

对话模型、温度、回复上限、系统提示词和限定词为固定参数，同时系统提示词和限定词也可以作为外部输入，意味着如果有输入流向了系统提示词，那么原本填写的内容就会被覆盖。

触发器、引用内容、聊天记录和用户问题则为外部输入，需要从其他节点的输出流入。回复结束则为该节点的输出。

FastGPT 工作流的运行与单出入口的工作流不同，FastGPT 的工作流可以指定不同的入口，并且没有固定的出口，而是以节点运行结束作为出口，如果在一个轮调用中，所有节点都不再允许，则工作流结束。

工作流的执行原则包括：

（1） 仅关心已连接的外部输入，即左边的圆圈被连接了参数。

（2） 当已连接的内容都被赋值的时候触发。（这个地方经常会遇到，连接了很多根输入线，但是只要有一个输入没有值，这个节点也不会执行）

（3） 可以多个输出连接到一个输入，后续的值会覆盖前面的值。

2.3 ChatGLM 介绍

ChatGLM 系列模型（如 ChatGLM3-130B），经过大量标识符的中英双语训

练，辅以监督微调、反馈自助、人类反馈强化学习等技术对齐人类意图，具备问答、多轮对话、代码生成等能力[22]。在逻辑推理、内容创作、代码生成和信息提取等各业务场景上有着出色的表现。具体如下：

（1） 注入知识及代码预训练，初具推理能力。ChatGLM 系列模型具备广博的知识面以及灵活的知识关联能力，并通过注入代码预训练加强了推理能力；可以根据输入的指令提示，迅速联想出相关的大量知识及概念，并找出最适合的推理链条。

（2） 海量数据预训练，极具潜力的创作能力。ChatGLM 大模型基于海量数据预训练，获得了关于语言、知识和创作技巧的深入理解；使得大模型能够释放出极具潜力的创作能力，能够持续源源不断的产生丰富、广泛、新颖的高质量原创内容。

（3） 根据指令生成代码，并给出代码解释。ChatGLM 系列模型经过代码数据的预训练，支持根据自然语言提示快速生成代码，并给出代码解释；同时支持多种编程语言，支持续写、翻译、注释、bugfix 等能力。

（4） 强大的语言理解能力，智能信息提取。ChatGLM 系列模型具备强大的语言理解能力，可以深入理解文本信息之间的逻辑关系，从非结构化的文本信息中抽取所需的结构化信息。

第三章 ChatGLM 模型训练与优化

ChatGLM 模型的预训练设置、SuperGLUE 基准测试、消融实验、自回归空白填空以及模型微调步骤，都是为了提高模型的准确性和性能，使其更好地适应各种自然语言处理任务。

预训练设置是模型训练的基础。通过对大量无标注数据进行学习，模型可以捕捉到语言的规律和结构，为后续的特定任务训练提供有力的支持。ChatGLM 模型的预训练设置会考虑到数据的清洗和标注，以及选择合适的生成式语言模型算法等，从而确保模型能够充分学习到语言的内在规律。

SuperGLUE 基准测试是衡量模型性能的重要工具。SuperGLUE 针对的是已经达到挑战上限的会话式 AI 深度学习模型，为其提供更难的挑战，旨在构建能处理更加复杂和掌握更细微差别的语言模型。ChatGLM 模型在 SuperGLUE 上进行测试，可以全面评估模型在自然语言理解方面的能力，为模型的进一步优化提供指导。

消融实验是一种理解和评估神经网络模型的技术。通过对模型的各个组件或功能进行逐步删除，观察模型性能的变化，可以深入了解模型的工作原理，识别出关键的组件和特征。对于 ChatGLM 模型来说，消融实验有助于发现模型中的冗余部分或性能瓶颈，为模型的改进提供方向。

自回归空白填空是模型训练中的一种技术。通过优化自回归空白填空目标，模型可以更好地处理语言中的上下文信息，提高生成文本的连贯性和准确性。这对于 ChatGLM 模型来说，有助于提升其在对话生成等任务上的性能。

模型微调是将预训练模型用于特定任务的技术。通过对已经训练好的模型进行微小的调整，可以使其更好地适应新的数据集和任务。对于 ChatGLM 模型来说，模型微调有助于提升模型在特定场景下的性能，如情感分析、问答系统等。

3.1 预训练设置

本文为了公平比较 BERT 模型和 ChatGLM 模型，使用 BooksCorpus 和英文维基百科作为预训练数据，使用 BERT 的小写单词片段分词器，词汇表大小为

30,000。使用与 BERTBase 和 BERTLarge 相同的架构分别训练 GLMBase 和 GLMLarge，参数分别为 110M 和 340M。

本文通过训练两个混合填空目标以及文档级、句子级目标的大型模型 GLMDoc 和 GLMSent 来实现多任务预训练。此外，还增加训练了两个参数为 410M 和 515M 的 ChatGLM 模型。

通过训练与 RoBERTa 模型具有相同数据、分词和超参数的大型模型 GLMRoBERTa 模型与 SOTA 模型进行比较。

3.2 SuperGLUE 基准测试

为了评估预训练的 GLM 模型，本文在 SuperGLUE 基准测试上进行了实验，并报告了标准度量指标。SuperGLUE 包含 7 个具有挑战性的 NLU 任务，并将分类任务重新制定为使用人工设计的填空问题，采用 PET 的方法对每个任务进行了预训练的 GLM 模型微调。

通过选择 BERTBase 模型和 BERTLarge 模型作为基线，对 GLMBase 模型和 GLMLarge 模型在相同的语料库上进行相似数量的预训练设置，实现 GLMBase 模型与 GLMLarge 模型的公平比较。选择 T5 模型、BERTLarge 模型和 RoBERTaLarge 模型作为基线，对 ChatGLM 模型与 GLMRoBERTa 模型进行比较，但由于 T5 模型和 BERTLarge 模型在参数数量上没有匹配，所以用参数为 220M 的 T5Base 模型和参数为 710M 的 T5Large 模型替代测试。

通过表 3-1 可以得出在相同数量的训练数据的情况下，GLM 在 ReCoRD 数据集阅读理解任务、CommonsenseQA 任务、Winograd Schema Challenge 任务、Recognizing Textual Entailment 任务、Boolean Questions 任务、Word-in-Context、COPA Balanced 任务的准确率都胜过 BERT，无论是使用基础架构还是大型架构。从平均数值来看，GLMBase 模型比 BERTBase 模型准确率高 4.6%，且 GLMLarge 模型比 BERTLarge 模型准确率高 5.0%，这清楚地证明了 ChatGLM 模型在 NLU 任务中的由于 BERT 模型。在 RoBERTaLarge 基准设置中，GLMRoBERTa 在超过一半任务测试中取得了基线上的进步，平均准确率略有上涨。GLMRoBERTa 模型在大多数任务测试中的准确率高于 T5Large 模型，从平均准确率来看，GLMRoBERTa 模型比 T5Large 模型高 1.7%。

表 3-1 SuperGLUE 基准测试

Table 3-1 SuperGLUE benchmarks

模型	ReCoRD F1/Acc.	COA Acc.	WSC Acc.	RTE Acc.	BoolQ Acc.	WiC Acc.	CB F1/Acc.	Avg
BERTBase	65.4/64.9	66.0	65.4	70.0	74.9	68.8	70.9/76.8	66.1
GLMBase	73.5/72.8	71.0	72.1	71.2	77.0	64.7	89.5/85.7	70.7
BERTLarge	76.3/75.6	69.0	64.4	73.6	80.1	71.0	94.8/92.9	72.0
UniLMLarge	80.0/791	72.0	65.4	76.5	80.5	69.7	91.0/91.1	74.1
GLMLarge	81.7/81.1	76.0	81.7	74.0	82.1	68.5	96.1/94.6	77.0
GLMDoc	80.2/79.6	77.0	78.8	76.2	79.8	63.6	97.3/96.4	75.7
GLMSent	80.7/80.2	77.0	79.8	79.1	80.8	70.4	94.6/93.7	76.8
GLM410M	81.5/80.9	80.0	81.7	79.4	81.9	69.0	93.2/96.4	78.0
GLM515M	82.3/81.7	85.0	81.7	79.1	81.3	69.4	95.0/96.4	78.8
T5Base	76.2/75.4	73.0	79.8	78.3	80.8	67.9	94.8/92.9	76.0
T5Large	85.7/85.0	78.0	84.6	84.8	84.3	71.6	96.4/98.2	81.2
BARTLarge	88.3/87.8	60.0	65.4	84.5	84.3	69.0	90.5/92.9	76.0
RoBERTaLarge	89.0/88.4	90.0	63.5	87.0	86.1	72.6	96.1/94.6	81.5
GLMRoBERTa	89.6/89.0	82.0	83.7	87.2	84.7	71.2	98.7/98.2	82.9

本文通过在同一训练批次中，以相等的机会抽取短跨度样本和长跨度样本（文档级或句子级）来评估多任务设置中的性能。关于多任务模型评估任务包括 NLU 任务，零样本语言建模。

对于多任务模型评估任务中的 NLU 任务，本文进行了基于 Yahoo Answers 数据集的 SuperGLUE 基准测试评估，通过对 BERT 模型、BLM 模型以及

ChatGLM 模型不同掩码比率的测试来评估各种模型的性能。如表 3-2 结果所示，ChatGLM 模型在不同掩码比率下的准确率均高于 BERT 模型和 BLM 模型。GLMLarge 模型的性能高于 GLMDoc 模型。

表 3-2 基于 Yahoo Answers 数据集的模型评估

Table 3-2 Model evaluation based on the Yahoo Answers dataset

掩码比率	10%	20%	30%	40%	50%
BERT	82.8	66.3	50.3	37.4	26.2
BLM	86.5	73.2	59.6	46.8	34.8
GLM_{Large}	87.8	76.7	64.2	48.9	38.7
GLM_{Doc}	87.5	76.0	63.2	47.9	37.6

本文通过在 LAMBADA 数据集上预测一段文本的最后一个单词来实现评估系统模拟文本中的长距离依赖关系的能力。为此，使用相同的数据和分词方式训练了一个 GPTLarge 模型作为基线，方便实验观察，GPTLarge 模型长文本生成能力的分值为 50。

GLM 模型的结果显示在图 3-1 中。在相同数量的参数下，使用单向注意力编码上下文 GLMDoc 模型的长文本生成能力数值为 48 低于 GPTLarge 模型。如果将 GLM 模型的参数增加到 410M，则 GLM410M 模型性能接近 GPTLarge 模型。当 GLM 模型的参数进一步增加到 515M 时，长文本生成能力数值增长到 51，性能优于 GPTLarge 模型。在相同数量的参数下，使用双向注意力编码上下文可以提高语言建模的性能，在这种设置下，GLMDoc 模型、GLM410M 模型、GLM515M 模型长文本生成能力数值为 49、53、55 均高于单向注意力编码上下文形式。通过研究 2D 位置编码对长文本生成的贡献。可以发现去除 2D 位置编码会导致语言建模的准确性降低，困惑度增加。

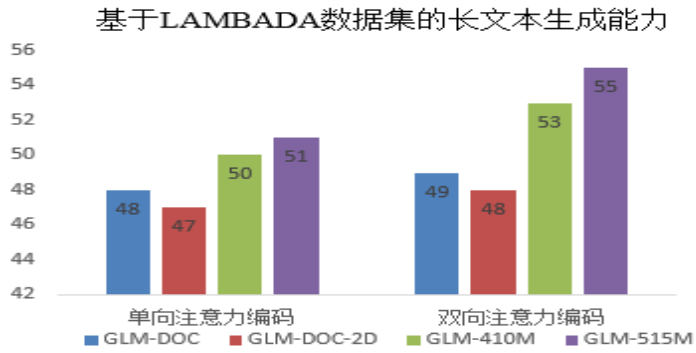

图 3-1 长文本生成能力模型评估

Fig. 3-1 Evaluation of Long Text Generation Ability Models

3.3 消融实验

本文通过设立 7 个测试任务，ReCoRD 数据集阅读理解任务、CommonsenseQA 任务、Winograd Schema Challenge 任务、Recognizing Textual Entailment 任务、Boolean Questions 任务、Word-in-Context、COPA Balanced 任务对 ChatGLM 模型进行消融分析。首先，为了与 BERT 模型进行比较，使用的与预训练数据和超参数训练了一个 BERTLarge 模型即表 3-3 中的 BERTLarge（reproduced）。BERTLarge（reproduced）模型平均准确率比官方 BERTLarge 模型即表 3-3 中的 BERTLarge 低 0.8%，而比于官方 GLMLarge 模型即表 3-1 中的 GLMLarge 平均准确率低 5.5%。这证实了 GLM 在 NLU 任务上优于 BERT 模型。其次，本文对作为序列分类器微调的 GLM 模型即表 3-3 中的 finetune 模型，以及使用填空样式微调的 BERT 模型即表 3-3 中的 BERTLarge(cloze)模型进行任务测试，由表 3-3 可知，与使用填空样式微调的 BERT 模型相比，在 ReCoRD 数据集阅读理解任务和 Winograd Schema Challenge 任务中，作为序列分类器微调的 GLM 模型无论是 ReCoRD 数据集阅读理解任务的 F1 分数和准确率还是 Winograd Schema Challenge 任务中的准确率都远远高于使用填空样式微调的 BERT 模型，实验证明了 GLM 模型在处理可变长度空白方面的能力高于 BERT 模型。

对于 ChatGLM 在 NLU 任务上的性能的提升，填空样式微调至关重要。如表 3-3 所示，使用填空样式微调的 GLM 模型即 GLMLarge（cloze）比序列分类器微调的 GLM 模型即 finetune 模型平均数值高 7 个点的性能。

最后，本文比较了具有不同预训练设计的 GLM 变体，以了解它们的重要性。GLM 模型去除跨度混洗即-shuffle spans 模型，模型始终从左到右预测蒙面跨度，导致模型的性能严重下降。GLM 模型使用不同的 sentinel 标记代替单个[MASK]标记来表示不同的蒙面跨度即表 3-3 中的+sentinel tokens 模型，该模型的性能也低于的常规 GLM 模型。本文通过对 GLM 模型于 BERT 模型进行类似的填空目标预训练，得出 GLM 模型在三个方面与 BERT 模型不同：（1）GLM 由单个编码器组成。（2）GLM 对蒙面跨度进行混洗。（3）GLM 使用单个[MASK]而不是多个 sentinel 标记。

表 3-3 GLM 的消融实验

Table 3-3 Ablation experiments of GLM

模型	ReCoRD F1/Acc.	COPA Acc.	WSC Acc.	RTE Acc.	Bool Q ACC.	WiC Acc.	CB F1/EM	Avg
BERTLarge	76.3/75.6	69.0	64.4	73.6	80.1	71.0	94.8/92.9	72.0
BERTLarge (reproduced)	82.1/81.5	63.0	63.5	72.2	80.8	68.7	80.9/85.7	71.2
BERTLarge (cloze)	70.0/69.4	80.0	76.0	72.6	78.1	70.5	93.5/91.1	73.2
GLMLarge	81.7/81.1	76.0	81.7	74.0	82.1	68.5	96.1/94.6	77.0
（cloze） finetune	81.3/80.6	62.0	63.5	66.8	80.5	65.0	89.2/91.1	70.0
-shuffle spans	82.0/81.4	61.0	79.8	54.5	65.8	56.3	90.5/92.9	68.5
+sentinel tokens	81.8/81.3	69.0	78.8	77.3	81.2	68.0	93.7/94.6	76.0

3.4 自回归空白填充

ChatGLM 是通过优化自回归留空填充目标进行训练的。给定一个输入文本$x = [x_1, \cdots, x_n]$，多个文本跨度$\{s_i, \cdots, s_m\}$被抽样，其中每个跨度s_i对应于 x 中一系列连续的标记$[s_{i,1}, \cdots, s_{i,\ l_i}]$。

本文将每个跨度都用一个“[MASK]”标记替换，从而创建一个被破坏的文本$x_{corrupt}$。模型会以自回归的方式从这个被破坏的文本中预测缺失的标记。这意味着当模型预测缺失的标记时，它可以查看破坏的文本以及先前预测的跨度。为了充分考虑不同跨度之间的相互关系，会随机重新排列跨度的顺序，类似于排列语言模型。

形式上，令Z_m为长度为 m 的索引序列[1,2, ⋯ , m]的所有可能排列的集合，$s_{z_i} < s_{z_{i'}}$为$[s_{z_i}, \cdots, s_{z_{i-1}}]$，本文定义预训练目标为

$$\sum_{z_1=1}^{m} \cdots \sum_{z_1=1}^{m} P\left(s_{z1}, \ldots, s_{z_m} | x_{corrupt}\right) \tag{0-1}$$

按照从左到右的顺序生成每个填空中的标记，即生成跨度s_i的概率被分解为

$$P(s_i | x_{corrupt}, s_1, \ldots, s_{i-1}) \tag{0-2}$$

本文使用以下技术实现自回归留空填充目标。输入 x 被分为两部分：Part A 是破坏的文本$x_{corrupt}$，Part B 包含被掩盖的跨度。Part A 的标记可以相互关注，但不能关注 B 中的任何标记。Part B 的标记可以关注 Part A 和 B 中的先行标记，但不能关注 B 中的任何后续标记。每个跨度都用特殊标记“[START]”和“[END]”进行填充，用于输入和输出。

3.5 微调 GLM 模型

对于 ChatGLM 模型来说，模型微调有助于提升模型在特定场景下的性能，对于下游自然语言理解（NLU）任务，线性分类器将由预训练模型生成的序列或标记的表示作为输入，并预测正确的标签。这些实践与生成式预训练任务不同，导致了预训练和微调之间的不一致性。将所有任务都变成 GLM 填补空白的生成任务，分为分类任务和文本生成任务两种。

分类任务具体为给定一个标准样本（x，y）将输入文本 x 通过模版转化为有一个[MASK]字符填空问题 c(x)。标签 y 也映射到了填空问题的答案 v(y) 。模型预测不同答案的概率对应了预测不同类别的概率。文本生成任务流程如图 3-2 所示。生成概率为

$$p(y|x)=\frac{p\left(v(y)\middle|c(x)\right)}{\Sigma_{y'\in y}p\left(v(y')\middle|c(x)\right)} \tag{3-3}$$

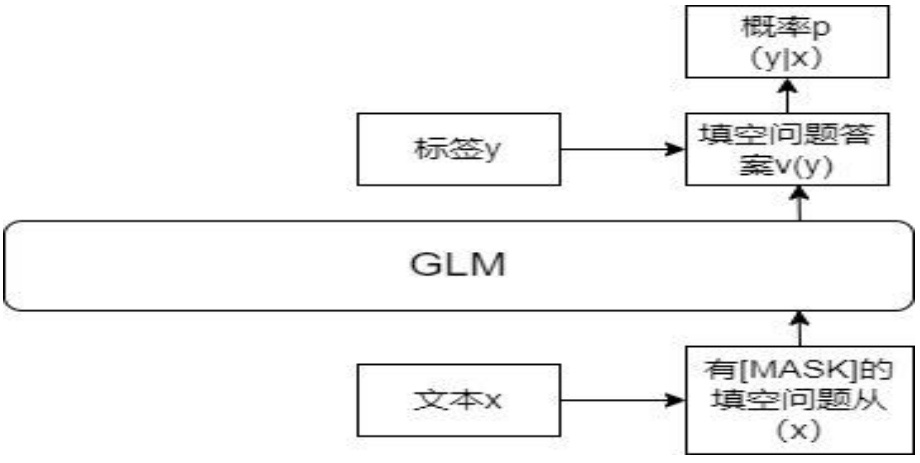

图 3-2 分类任务流程图

Fig. 3-2 Classification task flowchart

文本生成任务直接将 GLM 作为一个自回归模型的应用。比如：给定的上下文构成输入的部分的 A，在结尾附上一个[MASK]字符，模型用自回归的方式去生成 B 部分的文本。文本生成任务的流程图如图 3-3 所示。

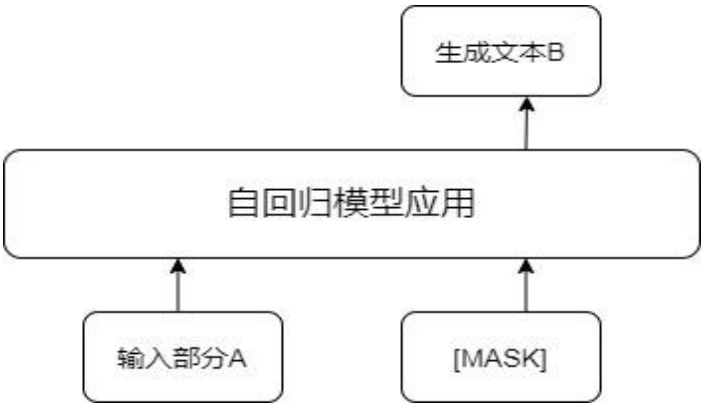

图 3-3 文本生成任务流程图
Fig. 3-3 Text generation task flowchart

3.6 本章小结

GLM 是一个用于自然语言理解和生成的通用预训练框架。本文证明了 NLU 任务可以被构建为条件生成任务，因此可以通过自回归模型来解决。GLM 将不同任务的预训练目标统一为自回归的空白填充，采用混合注意力掩码和新颖的二维位置编码。实证研究表明，GLM 在 NLU 任务上超越了先前的方法，并且可以有效地为不同任务共享参数。

第四章 模型嵌入

本文实现了 FastGPT 平台和 ChatGLM 模型通过 Intel 的 OneAPI 进行统一管理和交互。首先在本地服务器上对 FastGPT 平台和 ChatGLM 模型成功部署并提供 API 服务，将两者配置成能够接受和响应 HTTP 请求的形式。通过明确服务端口和认证机制，实现在后续 OneAPI 平台进行接入配置。在 OneAPI 平台上，需要分别将 FastGPT 和 ChatGLM 注册为独立的服务端点，要详细设定它们在 OneAPI 中的接口信息，包括但不限于 API 调用的 URL、请求方法（POST、GET 等）、必要的认证令牌和请求头等参数。依据 OneAPI 的集成规范，需要编写相应的代理逻辑或者路由规则，使得来自 OneAPI 的请求可以根据预设的路径映射，准确地被转发到对应的 FastGPT 和 ChatGLM 服务上。完成对接后进行全面的集成测试，验证从 OneAPI 发送请求时，是否能够顺利触发 FastGPT 和 ChatGLM 模型进行推理计算，并将结果回传给客户端。这一过程中可能需要对模型服务进行微调以满足 OneAPI 接口的要求，特别是当 FastGPT 和 ChatGLM 之间需要协同工作时，确保二者间的通信和数据格式转换得以无缝衔接。将 FastGPT 和 ChatGLM 与 OneAPI 进行整合的关键在于合理部署模型服务、精细配置 API 接口以及恰当实现请求转发与响应处理，从而实现统一管理、高效互动的目标。模型嵌入流程如图 4-1 所示。

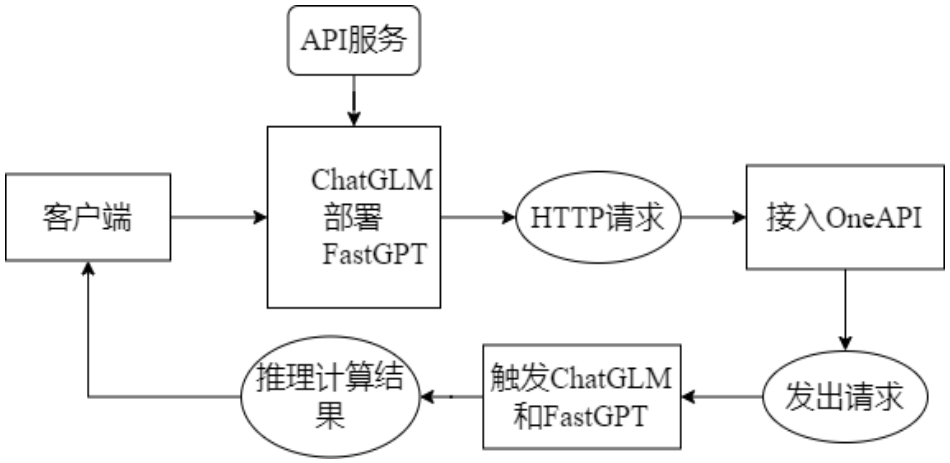

图 4-1 模型嵌入流程

Fig. 4-1 Model Embedding Process

4.1 FastGPT 接入 OneAPI

根据默认的用户名 root 以及密码 password，将 base url 填入 OneAPI 提供的 API 接口。假设 OneAPI 地址是 https://xxx.cloud.sealos.top，那么 base url 就是 https://xxx.cloud.sealos.top/v1。如果你的 OneAPI 和 FastGPT 都部署在 Sealos 中，这里的 base url 可以填入 OneAPI 的内网地址，例如内网地址是：http://one-api-wkskpejy.ns-sbjre322.svc.cluster.local:3000/v1。api key 填入由 OneAPI 提供的令牌。接入界面如图 4-2 所示。填好参数之后，点击部署应用。部署完成后，点击确认，跳转到应用详情。等待应用的状态变成 running 之后，点击外网地址即可通过外网域名直接打开 FastGPT 的 Web 界面。接下来把 ChatGLM3-6B 模型部署好，然后再回来接入 FastGPT。

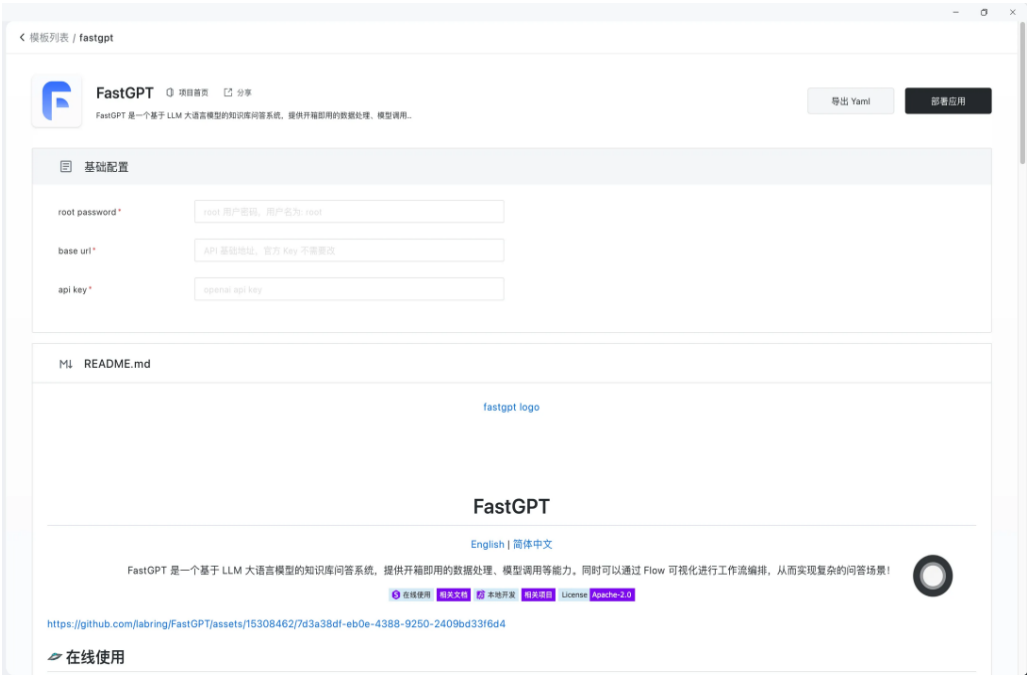

图 4-2 FastGPT 的接入界面

Fig. 4-2 FastGPT access interface

4.2 私有化部署 ChatGLM

ChatGLM3-6B 是 ChatGLM3 系列中的开源模型，在保留了前两代模型对话

流畅、部署门槛低等众多优秀特性的基础上， ChatGLM3-6B 的基础模型 ChatGLM3-6B-Base 采用了更多样的训练数据、更充分的训练步数和更合理的训练策略。在语义、数学、推理、代码、知识等不同角度的数据集上测评显示，ChatGLM3-6B-Base 具有在 10B 以下的基础模型中最强的性能。更完整的功能支持：ChatGLM3-6B 采用了全新设计的 Prompt 格式，除正常的多轮对话外。同时原生支持工具调用、代码执行和 Agent 任务等复杂场景。

硬件要求：

显卡：Nvidia，需要支持 cuda。不量化默认 FP16 精度大约需要 13GB 显存，INT8 大约需要 8GB 显存，INT4 大约需要 5GB 显存。

内存：32GB，内存不足，可开起虚拟内存。

安装依赖包需要输入命令：cd ChatGLM3，进入到克隆下来的程序所在目录和输入命令：pip install -r requirements.txt，开始安装依赖，这个过程也需要等待一会。等待安装完成后，继续执行命令安装 jupyter 内核，要使用使用 code interpreter 功能的就需要安装。安装过程包括输入命令：pip install jupyter，回车，等待安装完成。输入命令：ipython kernel install --name chatglm3-demo --user,注：#name 为项目 demo 中指定的，使用 demo 的话默认内核名称为 chatglm3-6b-demo。要使用程序包中的 composite_demo,这个包含的功能比较齐全，所以接下来要继续安装该目录下所需要的依赖。输入命令：cd composite_demo,回车，进入 composite_demo 目录。在执行接下来的命令行，需要先修改下 D:\ChatGlm3\ChatGLM3\composite_demo 中的 requirements.txt 文件，在电脑中打开这个文件。如图 4-3 所示。

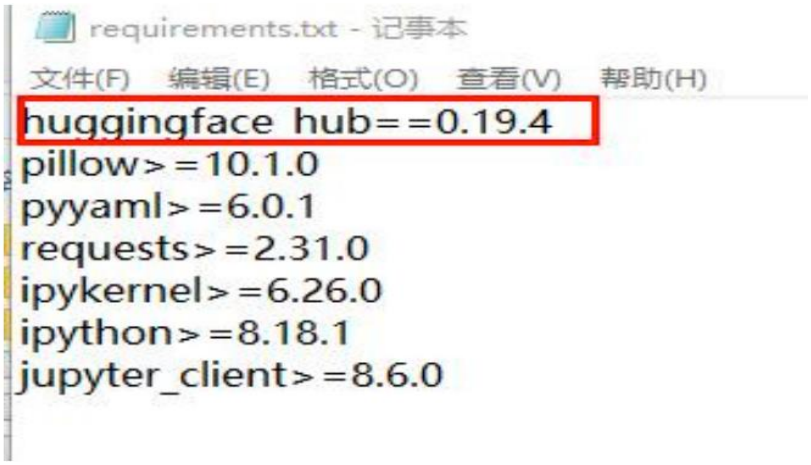

图 4-3 requirements.txt 文件信息

Fig. 4-3 requirements.txt file information

将第一行的>=改为==。因为>=会安装最新版的，最新版的跟程序中使用的不匹配，会导致运行出错。修改完后，回到命令行界面，输入：pip install -r requirements.txt,回车，开始安装依赖，这个过程也需要等待一会。

运行程序：

修改模型路径，如图 4-4 所示，由于将模型下载到了：D:\chatglm3\chatglm3-130b,这个程序中默认的路径不同，所以运行前先修改下程序中的模型路径。用记事本或者其他 python 编辑器，打开 D:\chatglm3\ChatGLM3\composite_demo 中的 client.py。

```
from __future__ import annotations

import os
import streamlit as st
import torch

from collections.abc import Iterable
from typing import Any, Protocol
from huggingface_hub.inference._text_generation import TextGenerationStreamResp
from transformers import AutoModel, AutoTokenizer, AutoConfig
from transformers.generation.logits_process import LogitsProcessor
from transformers.generation.utils import LogitsProcessorList

from conversation import Conversation

TOOL_PROMPT = 'Answer the following questions as best as you can. You have acce

#MODEL_PATH = os.environ.get('MODEL_PATH', 'THUDM/chatglm3-6b')
MODEL_PATH = os.environ.get('MODEL_PATH', 'D:\\chatglm3\\chatglm3-6b')
PT_PATH = os.environ.get('PT_PATH', None)
PRE_SEQ_LEN = int(os.environ.get("PRE_SEQ_LEN", 128))
TOKENIZER_PATH = os.environ.get("TOKENIZER_PATH", MODEL_PATH)
```

图 4-4 模型修改路径

Fig. 4-4 Modify the path of the model

4.3 GLM 嵌入 FastGPT

实现 ChatGLM 模型嵌入 Fastgpt 框架首先需要登录 Sealos 国内版集群，网址为：https://cloud.sealos.top/然后打开应用管理：应用名称随便填，镜像为：registry.cn-hangzhou.aliyuncs.com/ryyan/chatglm.cpp:chatglm3-q5_1，GPU 和内存拉到最大值，不然跑不起来。容器暴露端口设置为 8000。然后点击右上角的部署。操作如图 4-5 所示。

图 4-5 网络基础配置

Fig. 4-5 Basic network configuration

部署完成后，点击查看运行日志：打开 OneAPI 的 Web 界面，添加新的渠道：OpenAI，模型名称可以通过自定义模型名称来设置。代理地址填入 ChatGLM3-130B 的 API 地址。如果把 OneAPI 和 ChatGLM3-130B 全部部署在 Sealos 中，那就可以直接填 ChatGLM3-6B 的内网地址。具体操作如图 4-6 所示。

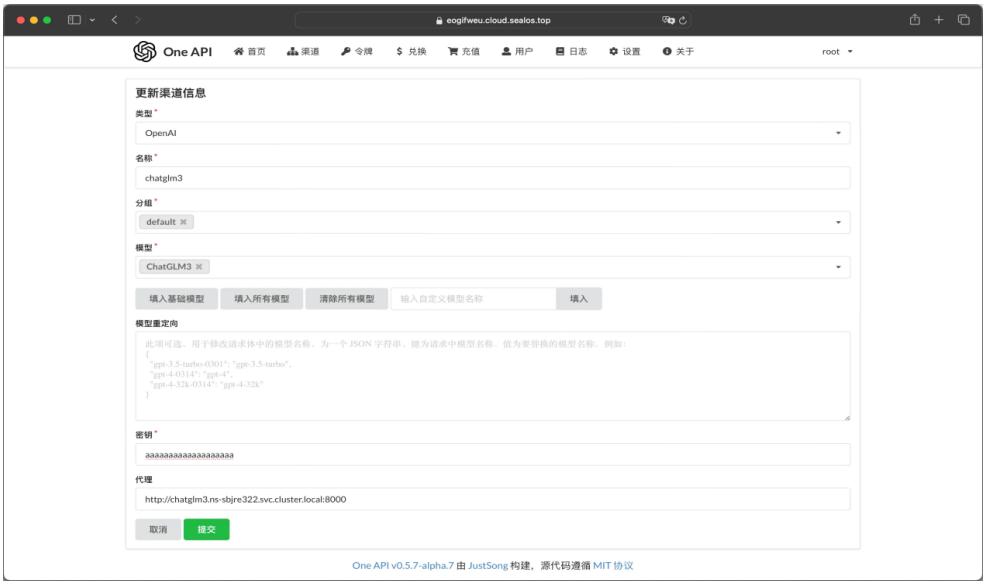

图 4-6 渠道更新配置

Fig. 4-6 Update the configuration of the channel

最后修改 FastGPT 的配置，将 ChatGLM3-130B 接入 FastGPT。首先在 FastGPT 的应用详情中点击变更：然后点击配置文件中的/app/data/config.json：修改完成后，点击确认，然后点击右上角的变更，文件配置如图 4-7 所示。

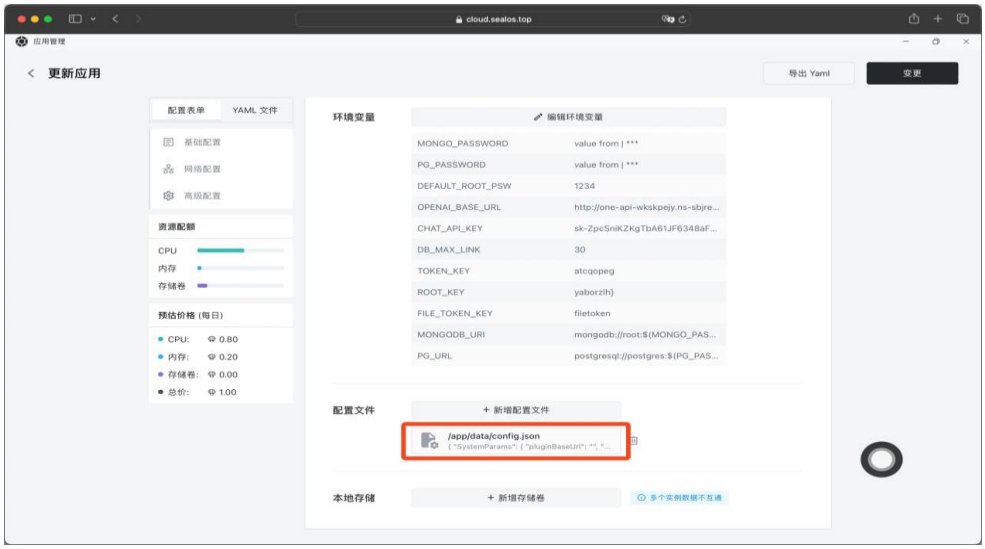

图 4-7 文件配置界面

Fig. 4-7 File configuration page

等待 FastGPT 重启完成后，再次访问 FastGPT，点击立即开始进入登录界面，接下来可以对 AI 应用进行命名以及模型选择 AI 模型选择前文接入的 ChatGLM3，然后点击保存并预览。最后，点击左上角对话打开一个聊天会话窗口，这样就接入完成，完成界面如图 4-8 所示。

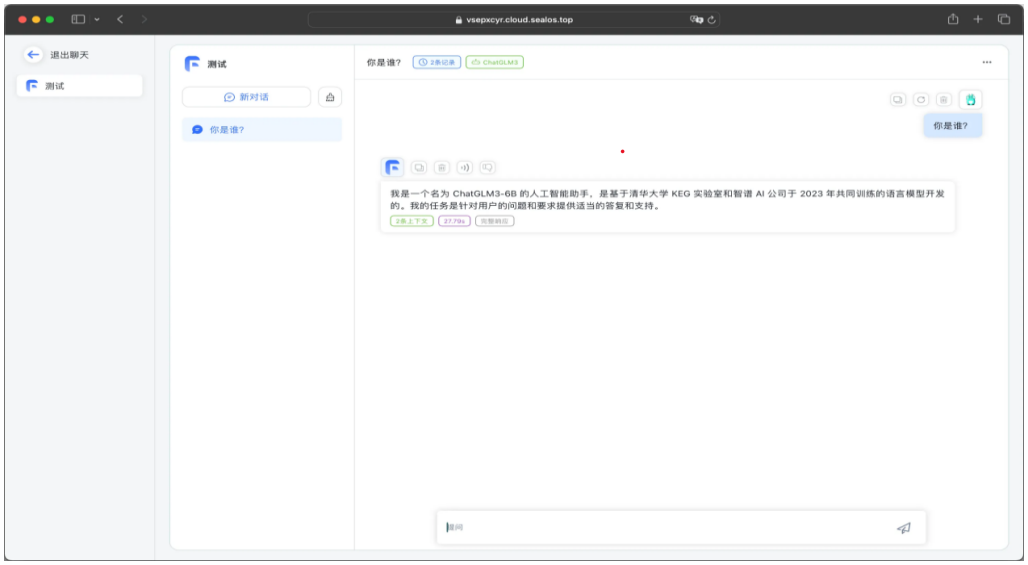

图 4-8 模型接入成功界面

Fig. 4-8 The model is successfully connected

第五章 农业知识问答系统的构建

5.1 系统概述

本文旨在探讨一种基于 ChatGLM 与 FastGPT 结合的农业知识问答系统范式，通过深度整合大型语言模型与专业知识库，以实现针对农业领域的专业问答效果和提供用户友好的交互服务。为实现这一目标，本文设计并实现了具备信息过滤、专业问答和抽取转化功能的系统。该系统通过 FastGPT 框架，成功地将农业知识与大型语言模型相融合，从而开创了大型语言模型与知识库深度结合的新模式。这一创新性的设计使得系统能够准确、高效地为用户提供专业的农业知识问答服务，为农业领域的发展提供了有力的技术支撑。系统通过接入提示词工程可以使信息过滤模块旨在减少大型语言模型生成虚假信息的可能性，以提高回答的准确性。专业问答模块通过将专业知识库与大型语言模型结合，提供专业性的回答。这种方法避免了重新训练大型语言模型所需的高硬件要求和可能导致的灾难性遗忘后果。抽取转化通过从自然语言文本数据库抽取出结构化数据，以及将结构化数据转化为自然语言文本实现进一步探索问答系统新范式而设计。一方面基于大型语言模型提取出专业知识，将结构化数据转化为自然语言文本，易于用户理解；另一方面利用知识抽取出三元组和 json 数据对比验证，可以增强大型语言模型回答专业性，除此之外，本系统还实现了用户友好的交互服务。系统实现流程图如图 5-1 所示。

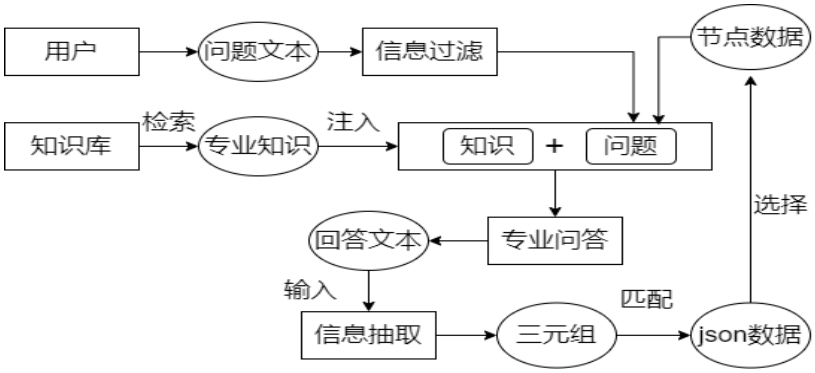

图 5-1 系统交互服务

Fig. 5-1 System Interaction Services

系统交互流程通过（1）用户向系统提出问题，问题通过信息过滤后，与知识

库中的相关专业知识组成提示，输入到专业问答模块中得到答案；（2）信息抽取模块从回答中提取出三元组，与 json 数据进行匹配，获取相关节点数据；（3）这些节点数据可以经用户选择后，同样以提示的形式输入专业问答模块得到知识库增强的回答。这种双向交互实现了大型语言模型和知识库的深度结合。

除了图像识别和知识抽取功能外，还特别关注用户交互服务的优化。系统设计了简洁明了的交互界面和流程，使用户能够轻松地提出问题、上传图像并获取相应的回答。同时，系统还提供了丰富的交互选项，允许用户根据需要对回答进行筛选、修改和补充，以满足不同用户的需求和偏好。

总的来说，本研究成功构建了一种基于大型语言模型的农业知识问答系统，该系统在图像识别、自然语言描述生成以及用户交互服务等方面表现出色。这一系统将为农业生产者、农业科技人员以及广大农业爱好者提供便捷、高效的知识获取方式，推动农业领域的信息化和智能化发展。

5.2 系统的构建方法

本节从数据收集与处理、信息过滤、专业问答、抽取转化四个方面，以农业领域的应用为例，介绍如何构建系统。该系统在构建过程中，注重数据收集、预处理以及训练流程的精细化设计，以确保其易于部署并具备高效性能。该系统具备三大核心功能：信息过滤、专业问答以及抽取转化。

在问答流程方面，首先，系统对输入的农业相关问题文本进行信息过滤，即实施文本分类操作，以判断该文本是否与农业领域相关。接着，利用 FastGPT 模型在农业知识库中检索与文本内容相关的专业知识，并以提示的方式将检索结果和问题一同输入到大型语言模型 ChatGLM-6B 中。随后，ChatGLM-6B 通过推理机制生成具备专业知识的回答。

为进一步提升答案的专业性，系统对生成的回答进行知识抽取操作，从中提取出三元组结构的信息。这些抽取出的三元组与 json 数据进行匹配验证，以确保回答的专业性和准确性。此外，系统还实现了大模型和知识库的双向转换功能。具体而言，系统能够将知识图谱中的节点以问题的形式输入到 ChatGLM-6B 模型中，获取易于理解的自然语言解释，从而实现了知识库信息的有效转化和利用。

综上所述，本文设计的农业领域问答系统通过综合运用信息过滤、专业问答

和抽取转化等核心技术，实现了对农业领域问题的精准回答和知识的有效利用，为相关领域的研究和应用提供了有力的支持。

5.2.1 数据的收集与处理

本系统的实现需要收集整理专业数据集，以支持系统的实现。本文基于多种数据构造系统所需的数据集、知识库，并对这些数据进行数据预处理。

（1） 基于已有的专业领域数据集。本文直接搜集专业领域已有的相关数据集，参考其构成，从中整理筛选出所需的数据。对于农业领域，参考以下数据集整理并构建相关专业数据。

（2） 相关介绍如图 5-2 所示

图 5-2 相关专业数据集

Fig. 5-2 Related professional datasets

农业相关的权威数据从专业书籍或权威网站收集。这部分数据来自于农业领域的专业书籍和权威网站，用于构建知识库，为大模型的回答提供专业知识支撑。对于农业领域，主要基于农业学、农业技术等专业书籍构建了农业专业知识库，同时从农业部、国家统计局、农业科研机构等专业权威网站收集农业领域的相关数据知识。

（3） 问题数据。问题数据用来训练信息过滤模型。因为某些专业领域存在问题数据缺失的情况，本文设计了一种基于提示的方法，使用大模型生成问题数据，首先从相关数据中选择一条数据用来生成提示，将提示输入大模型生成一条数据，重复以上述步骤，直到相关数据被选完。

算法：ChatGPT 生成数据
输入：相关文本 D，ChatGPT 的 API_KEY
输出：ChatGPT 生成的问题数据 R
1:ChatGPT_connection←create(API_KEY)//通过
2:API_KEY 创建与 ChatGPT 接口的链接。
3:for i:=1 to N do
4:d←select(D)//选择一条相关数据。
5:prompt←P(d)//根据选择的数据生成提示。
6:r←ChatGPT(chatgpt_connection,prompt)通过建立的链接访问 ChatGPT，输入提示生成答复。
7:R←abstract(R)//从 ChatGPT 生成结果中提取出问题数据，并汇总到 R 中
8:end for

图 5-3 ChatGPT 生成问题数据

Fig. 5-3 ChatGPT generates question data

在上图 5-3 中 D 表示所有的相关数据，d 表示一条相关数据，R 表示所有生成的问题数据，r 表示一条生成的数据。create 根据用户提供的 API_KEY 创建与 ChatGPT 的链接，select 表示选择一条数据，P 表示根据数据生成合适的提示，R 表示获取 ChatGPT 生成的回复，abstract 表示从生成回复中提取出问题数据并进行汇总。

在农业领域，如图 5-4 中所示，将提示输入 ChatGPT，生成相关问题。农业领域问答系统的问题数据，80%来自于现有的问答数据集，本文从中整理相关问题，并将其按照是否为农业专业领域添加标签。20%的农业相关问题使用大模型生成的方式构建。

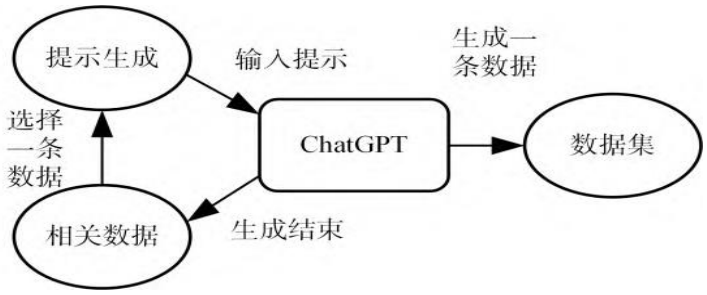

图 5-4 基于 ChatGPT 生成数据集方法

Fig. 5-4 The method for generating datasets based on ChatGPT

5.2.2 数据集构建

United States Department of Agriculture (USDA) 数据集：美国农业部提供了各种关于农业、农产品生产、土地利用等方面的数据集，包括农业普查数据、气象数据、作物种植数据等。Food and Agriculture Organization (FAO) 数据集：联合国粮食及农业组织提供了全球各种农业相关数据集，包括农产品生产、耕地利用、渔业数据等。Kaggle 上数据集：Kaggle 上有许多与农业相关的数据集，涵盖了各种主题，如作物产量预测、土壤特性、气候变化对农业的影响等。国家统计机构数据：各国的国家统计机构通常提供有关农业生产、出口、进口、就业等方面的数据。

此外，本文还收集了一些农业相关的问答数据集，爬取了 http://cloud.sinoverse.cn/index_bch.html，通过父网页爬取子网页 url，遍历子网页 url，进行爬取。通过数据处理和 re 正则，提取关键信息，整理成 csv 文件，这里的作物名，疾病名、各种属性、关系都是自定义的，想扩充知识图谱的话，需要自己添加属性、实体、关系等信息，信息越多知识图谱越丰富。大小: 15000 个问答对，样本分布: 作物种植: 5000 个问答对，病虫害防治: 4000 个问答对，农业技术: 3000 个问答对，农业政策: 2000 个问答对，其他: 1000 个问答对。

结果如图 5-5 所示：

	A	B	C	D	E	F	G	H	I	J
1	did	plant	disname	style	factors	way	type	ename		
2	445	蓖麻	蓖麻黑斑病	叶片发病初生小不规则灰绿色小斑	湿度是该病发生、蔓延的重要条件	(1)农业防治：适期播种，不宜	病害	NULL		
3	446	蓖麻	蓖麻枯萎病	(1)苗期：幼苗发病，茎部出现暗褐	品种间抗病性有差异。高温高湿有	(1)农业防治：选用抗病品种。	病害	NULL		
4	447	蓖麻	蓖麻细菌性叶斑	主要危害叶片，最初叶上产生水浸	整个生育期均可发生，生长季节高	(1)农业防治：及时摘除病叶，病害		NULL		
5	448	蓖麻	蓖麻炭疽病	幼苗发病，子叶上出现暗褐色病斑	天气潮湿及阴湿地发病重。	(1)农业防治：合理密植，注意	病害	NULL		
6	449	蓖麻	蓖麻斑点病	主要为害叶片。初生微小的黑褐色	NULL	(1)农业防治：实行3年以上轮换	病害	NULL		
7	450	蓖麻	蓖麻夜蛾	(蓖麻)幼虫食叶成缺刻或孔洞，啃	(1) 发生世代。广东、广西年生4	1.农业防治。2.物理防治。3.生	虫害	NULL		
8	451	蓖麻	蓖麻疫病	主要危害茎秆和叶片。茎秆发病初	蓖麻品种抗病性有差异。多年连种	(1)农业防治：合理灌溉，雨后病害		NULL		
9	452	蓖麻	棉铃虫	幼虫孵出后，先取食卵壳，随后蛀	(1)发生世代 在我国由北向南年发生	(1)农业防治:1.深翻冬灌，减少	虫害	cotton bollworm		
10	453	蓖麻	蓖麻白绢病	初在茎基部出现暗褐色稍凹陷的病	高温多雨易发病。沙土和酸性大的	(1)农业防治：与非寄主植物进	病害	NULL		
11	454	蓖麻	蓖麻锈病	主要为害叶片。叶片受害，叶背产	高温、高湿时发病重。	(1)农业防治：合理轮作，适当	病害	NULL		
12	455	蓖麻	蓖麻灰斑病	叶片发病初生黑褐色或茶褐色小斑	NULL	(1)农业防治：实行3年以上轮换	病害	NULL		
13	456	蓖麻	菜粉蝶	幼虫取食叶片。2龄前只啃食叶肉	(1)世代。内蒙古、辽宁、河北年发	(1)农业防治。及时清洁田园，虫害		Cabbage butterfly		
14	457	蓖麻	黄刺蛾	幼虫食叶。低龄啃食叶肉，稍大将	(1)发生世代：在东北和华北地区一	(1)农业防治：结合果树冬剪，虫害		oriental moth		
15	458	蓖麻	蓖麻四星尺蛾	幼虫食叶成缺刻或孔洞，严重者将	发生代数不详，河北8月中旬为害盛	因地制宜选育和种植抗虫品种	虫害	NULL		
16	397	茶	茶黄蜗	成蜗和幼蜗集中在寄主的幼嫩部位	(1)发生世代：一年发生多代。世代	(1)农业防治：摘好冬季防治工	虫害	Yellow tea mite		
17	398	茶	茶黄蜗	成蜗和幼蜗集中在寄主的幼嫩部位	(1)发生世代：一年发生多代。世代	防治](1)农业防治：摘好冬季防	虫害	Yellow tea mite		
18	399	茶	茶蓑蛾	幼虫食叶，低龄啮食叶肉残留叶脉	(1)世代。贵州每年发生1代，安徽	(1)农业。人工摘除蓑囊，果树	虫害	Tea bagworm		
19	400	茶	褐圆蚧	以雌成虫、若虫固着于叶片、果实	(1)发生世代：陕西汉中地区一年发	(1)农业。培养树势，提高树体	虫害	Florida red scale		
20	401	茶	茶翅蝽	成虫和若虫均吸食嫩叶、嫩茎和果	(1)发生世代。每年发生1代。(2)越	(1)农业防治：①套袋是减少蝽	虫害	Yellow-brown stink bug		
21	402	茶	茶翅蝽	成虫和若虫均吸食嫩叶、嫩茎和果	(1)世代。每年发生1～2代。(2)越冬	1.农业防治\u3000越冬期捕杀成	虫害	Yellow-brown stink bug		
22	403	茶	茶黄蜗	成蜗和幼蜗集中在寄主的幼嫩部位	(1)发生世代：一年发生多代。世代	(1)农业防治：摘好冬季防治工	虫害	Yellow tea mite		
23	404	茶	茶翅蝽	成虫和若虫均吸食嫩叶、嫩茎和果	(1)发生世代。每年发生1代。(2)越	(1)农业防治。利用成虫喜欢在	虫害	Yellow-brown stink bug		
24	405	茶	茶卷叶蛾	幼虫在芽梢上卷缀嫩叶藏在其中。	(1)发生世代:安徽、浙江1年发生4	(1)农业防治:低龄幼虫期结合日	虫害	NULL		
25	406	茶	苹果小卷蛾	越冬代幼虫出蛰后先爬到嫩芽、幼	(1)发生世代：黄河故道地区1年4代	(1)农业防治：果树休眠期至萌	虫害	Smaller apple leaf roller		
26	407	茶	柑橘粉虱	以幼虫群集于叶背刺吸汁液，粉虱	(1)发生世代：浙江一年发生3代。(	(1)农业。适量剪除虫害枝和衰	虫害	citrus whitefly		
27	408	茶	橘二叉蚜	以成蚜、若蚜在寄主植物嫩叶后丛	(1)发生世代：安徽一年发生25代以	(1)农业防治：个别发生数量多	虫害	Black citrus aphid		
28	57	大豆	豇豆荚螟	幼虫为害豆叶、花及豆荚，常卷叶	(1)发生世代。在华北地区年发生3	(1)农业防治。及时清除田间落	虫害	Bean pod borer		
29	58	大豆	胃毒蛾	以幼虫取食叶片，吃成缺刻、孔洞	(1)发生世代:长江流域一年发生3代	(1)农业防治:秋冬季节，清除田	虫害	Pear tussock moth		
30	59	大豆	豆荚斑螟	以幼虫吃食花、荚和豆粒为主。产	(1)发生世代。江苏、安徽每年发生	(1)农业防治。及时清除田间落	虫害	Limabean pod borer		

图 5-5 农业问答数据集

Fig. 5-5 Agricultural Q&A dataset

5.2.3 信息过滤

针对特定领域的问答需求，大型语言模型需精确聚焦于专业领域问题，以规避对非相关领域问题的响应。为此，本系统特别引入了基于 BERT 的文本过滤器，用以筛选和过滤问题，从而限定大模型处理问题的范畴。其他模型在遭遇专业领域的边缘问题或交叉问题时，往往产生误导性的幻觉事实，导致错误文本的生成。尽管微调方法可赋予大模型一定的问题识别能力，但它在面对与微调数据集相似的其他问题时，仍可能受到干扰，甚至对原本能正确回答的问题也产生错误回应。

因此，本文独立设计了文本过滤器以实施信息过滤。假设可输入大模型的所有问题构成集合 Q，大模型在特定专业领域能够回答的问题构成集合 R，而能够生成专业回答的问题构成集合 D。显然，Q 的规模大于 R，R 的规模又大于 D。若采用微调方式限制，可能使 R 趋近于 D，削弱模型的回答能力。而使用过滤器，则可使 Q 趋近于 R，尽可能确保所提问题落在 R 的范围内。尽管仍可能有部分 R 之外的数据进入大模型，但鉴于本系统设计的专业增强问答系统仍保留一定的通用能力，故对于 R 之外的问题也能进行无专业验证的回应。

信息过滤机制确保本系统能最大限度地回答其能力范围内的问题，从而降低产生幻觉事实的风险。在构建过滤器时，将训练数据输入 BERT，随后将 BERT 的输出传递给全连接层（FCL），以得出文本分类结果[CLS]。根据数据集中的标签，训练过程中仅需更新全连接层的参数[23]。

通常，利用 BERT 进行文本分类时，会采用 BERT 输出的分类词向量 H，并结合 softmax 构建简单分类器，以预测类别标签 L 的概率。在此，通过稍作修改，引入全连接层计算每个标签的概率。训练时，输入全连接层的向量维度为 768，包含两个隐藏层，其维度分别为 384 和 768，输出维度则与类别个数相对应。在本任务中，由于是一个二分类问题，因此输出维度为 2。最终，选择概率更高的标签作为分类结果[CLS]。在农业学领域的应用中，[CLS]用于判断问题是否与农业相关，通过过滤问题，减少生成幻觉事实的可能性，并与检索结果共同决定能否进行专业回答。

5.2.4 专业回答

为了使得大模型农业问答系统的回答更具备专业性，本文通过提示的方式

注入知识库中的专业知识，增强回答的专业性。通过检索知识库，大模型可以回答其本身能力之外的专业问题，这使得大型语言模型支持的问题边界扩大。这种方式和引入专业数据的微调方法对比，无需重新训练就可以部署一个专业领域大型语言模型。在农业领域，本文使用 FastGPT+ChatGLM-6B，生成更具备专业知识的回答。本系统基于 FastGPT 在知识库中检索与问题相关的专业知识，然后专业知识和问题文本一起构成输入大模型，最终得到答案文本，这里选择使用 ChatGLM-6B 作为大模型。

假设知识库中的第 i 个文件为 vF（F=1，2，3，…，n），基于 FastGPT 进行检索会将各个文件中的文本进行分块，vFj（F=1，2，3，…，n；j= 1，2，3，…，m.）表示对第 i 个文件的第 j 个文本块。然后对每一块文本建立为向量索引$\boldsymbol{Vi}$（i= 1，2，3，…，n×m），在检索时将问题文本向量化，得到问题文本向量 Q，最后通过向量相似度计算出和 Q 最相似的 k 个向量索引，并返回其对应的文本块。将匹配到的专业知识文本 D 和问题文本以的形式拼接，最终输入 ChatGLM- 6B 中得到大模型生成的专业回答。该过程如图 5-6 所示。

算法：专业问答过程
输入：问题文本 q，知识库文件 f
输出：大模型的回答文本 result
1:for i:=1 to N do
2:di←split(fi)//对第 i 个文件划分文本块。
3:end for
4:for i:=1 to N*M do
5:Vi←trans(di)//生成每个文本块的向量索引。
6:end for
7:Q←trans(q)//将问题文本转化为问题向量。
8:Vk←score(Q,Vn*m)//计算问题向量和索引向量的相似度，得到 k 个最相似的索引向量。
9:dk←de_trans(Vk)//根据匹配到的 k 个问题向量转化为相应的知识文本。
10:result← model(P4(q,dk))//将问题文本和专业知识文本以 P4 形式输入大模型以获取回答。

图 5-6 专业问答过程

Fig. 5-6 Professional Q&A process

在图 5-6 中，q 表示问题文本，f 表示知识库文件，d 表示知识文本块，Q 表示问题文本向量，V 表示文本块的向量索引，split 表示划分文本块的过程，trans 表示从文本转化为向量，de_trans 表示从向量转化为文本，score 将返回 k 个最相似的向量索引，model(P4(q,dk))表示将问题文本和专业知识文本以 P4 形式输入大模型 ChatGLM-6B。

5.2.5 抽取转化

利用 ChatGLM 的自然语言理解能力，从用户输入中抽取关键信息，如涉及的农业概念、技术、问题等。模型通过分析上下文、识别实体和关系来抽取相关信息[错误!未找到引用源。]。基于抽取的信息，ChatGLM 进行知识推理和生成。它可能结合自身的内部表示、已学习到的知识以及语言模型的能力，生成与输入问题相关的自然语言回答。这个过程可能涉及对农业领域知识的逻辑推理、类比推理等。对生成的回答进行必要的后处理，如调整语句结构、添加适当的解释或例子，以提高回答的可读性和准确性。此外，可以通过用户反馈和模型优化技术，不断迭代和改进系统的性能。

5.3 系统的功能设计

5.3.1 交互界面展示

FastGPT 不仅实现了对大容量预训练模型如 GPT-2 和 GPT-3 的快速推理加速，也能对国产的如 ChatGLM 和通义千问等的推理的流式输出，更提供了一种直观易用的可视化操作界面。

FastGPT 界面设计简洁明了，主要由输入框和输出区域构成。用户只需在左侧的输入框内写下您的问题或指令，点击“提交”按钮，FastGPT 就能迅速利用后台高效运行的大规模语言模型为您生成高质量的回答或完成指定任务。

该界面的一大亮点在于它对模型响应速度的优化，即使面对复杂的问题，也能在短时间内给出智能且富有洞察力的反馈。同时，FastGPT 充分考虑了用户体验，无论是在桌面端还是移动端都能流畅使用。

另外，FastGPT 还具备良好的可定制性，可以根据不同的业务场景加载特定领域的知识库，让模型能够更加精准地服务于各类专业应用需求。

以玉米作物为例，本文进行与农业知识问答系统的交流。首先，先针对玉米作物种植时间的进行询问，系统会根据适宜的季节、当地温度区间、月份进行回答。如图 5-7 所示。针对玉米的叶斑病提出了怎样选择药物治疗，系统从常见的药物选择，使用方法，怎样减少抗药性三部分进行回答。具体内容如图 5-8 所示。对玉米农作物种植的方法进行提问，系统根据土壤的选择、肥料的使用、病虫害防治、采收时间等方面进行方法推荐，根据系统提供的方法推荐，选择了病虫害防治方面进行更进一步的提问，系统列举出了常见的病虫害类别以及发病原理、时期、表现。具体内容如图 5-9、图 5-10 所示。

该系统在作物种植之前，可以根据用户的要求，提供专业的建议。包括温度的选择、土壤的选择、肥料的使用，对农户农作物种植的科学性、准确性提供了极大的帮助。可以对发生病虫害问题进行预防和治理推荐，对具体病症进行药物推荐，帮助农户做到对症下药，减少农作物的损失，维护农民的经济收益，可以促进农业的发展。

图 5-7 作物种植时间

Fig. 5-7 Planting time of crops

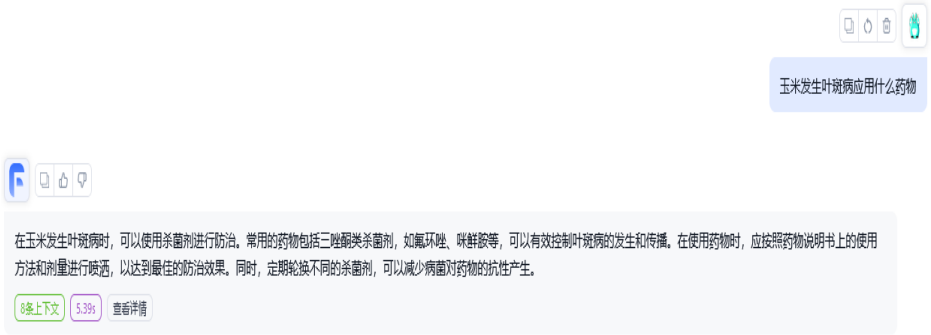

图 5-8 药物治疗

Fig. 5-8 drug therapy

图 5-9 作物种植方式

Fig. 5-9 crop planting patterns

图 5-10 病虫害防治

Fig. 5-10 Pest and disease control

5.3.2 图像识别功能

多模态模型是一种具有读图能力的模型，通常采用统一的架构来处理不同类型的输入数据，包括文本、图像、音频等。对于图生文任务，模型会首先通过专门的图像编码器对输入的图像进行特征提取和理解，将其转化为连续的向量表示。这些图像特征随后与文本相关的上下文信息结合在一起，共同输入到模型的核心 Transformer 结构中进行联合推理。

在这样的模型中，多模态能力意味着模型能够根据接收到的图像信息生成

相应的自然语言描述。具体来说，当模型接收到一张图片时，它会分析图片内容，然后根据所学到的视觉和语言之间的关联，生成与图片内容相符的文字描述。

FastGPT 同样实现了类似的多模态功能，实现过程应该也会遵循相似的设计原则和技术路径，即整合图像识别与自然语言生成技术，在一个统一的架构下进行跨模态的学习与推理。

如图 5-11、5-12 所示，农户可以对农作物以及农作物存在的病虫害问题进行拍照并通过相应的文字说明来向系统进行提问。很多情况下农户并不能对所发生的问题进行准确的文字描述，针对以上问题，系统增加了图像输入和语音输入两种输入方式，可以通过图片和文字提问的方式和语音转文字的方式对农作物的类型和作物病虫害进行提问，解决了农民不会描述、描述不清等问题，体现了系统友好的人机交互功能，方便用户操作。

图 5-11 农作物类型识别

Fig. 5-11 Crop type identification

图 5-12 病虫害识别

Fig. 5-12 Identification of pests and diseases

5.3.3 历史对话信息预览框

系统提供了历史对话信息预览框，其具体实现方式可能包括但不限于在一个滚动列表中逐条列出对话记录，或者以时间线形式展现，每条记录包含用户输入和 AI 回复等内容。

历史对话信息预览框通常扮演着重要角色，它可以：

（1） 记录对话历史

保存用户与 FastGPT 的所有交互内容，便于用户回顾之前讨论过的话题或检索过去的答案。

（2） 上下文感知

AI 模型可以根据预览框中的历史对话内容理解上下文情境，从而更好地回答后续问题，保持对话连贯性。

（3） 个性化体验

长期的历史对话记录有助于 AI 模型学习用户的偏好和习惯，提供更个性化的服务。

（4） 用户便利性

用户不必记忆之前的对话内容，只需查看预览框就能快速回忆起之前的问题和答案。

（5） 调试和优化

对于开发者而言，历史对话预览框也是重要的调试工具，可以帮助他们了解模型的表现，并据此调整优化模型参数或设计更好的对话流程。

如图 5-13 所示可以对所提过的问题进行查阅，用户不用担心对话内容的丢失，方便保持对话的连续性。

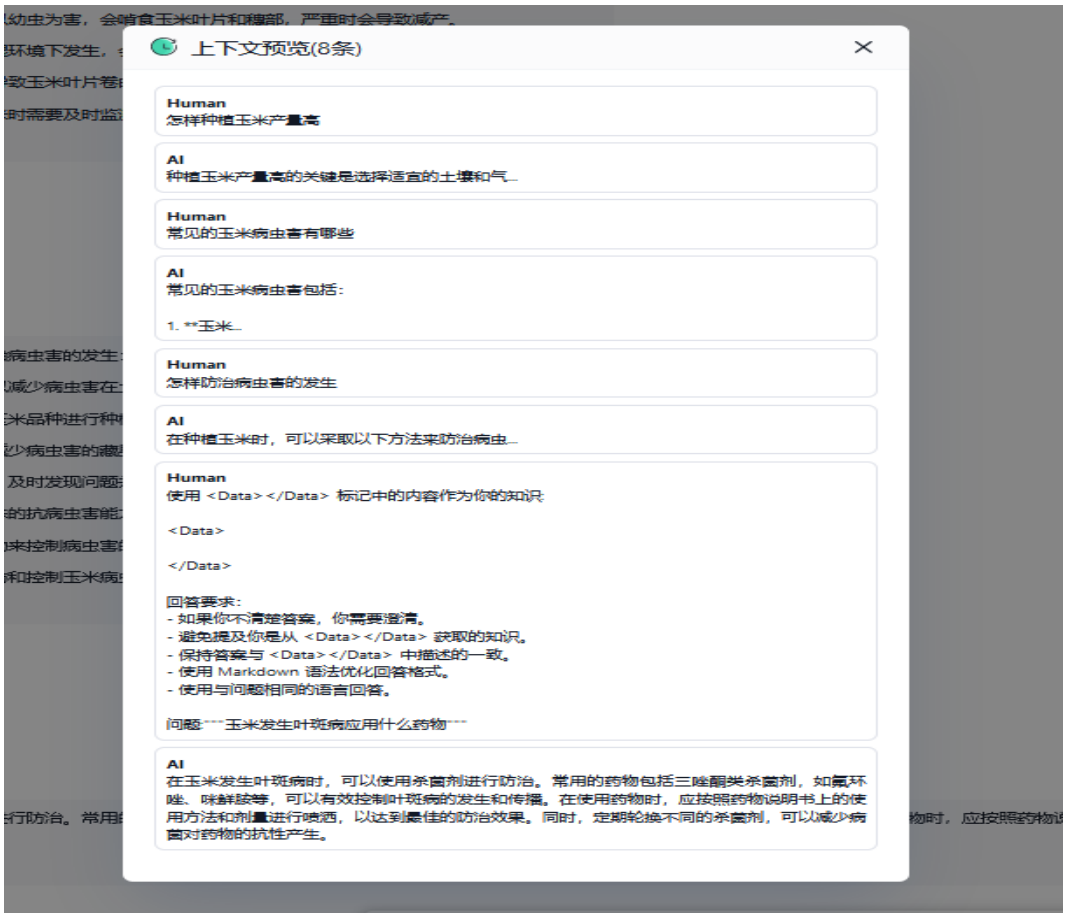

图 5-13 上下文预览展示

Fig. 5-13 shows the context

5.3.4 提示词及对话配置界面

FastGPT 是一个提供大语言模型服务的平台，它允许用户创建、配置和发布基于 LLM（Large Language Model）的应用。

在这个界面上，可以看到以下内容：

左侧导航栏：提供了应用的不同功能模块，包括“聊天”、“发布”、“设置”等。

中间部分：这是应用的简介区域，描述了该应用的基本信息和用途。

右侧部分：这是一个对话框，用于测试应用的功能。可以看到当前应用被设定为农业问答系统助手，并且能够处理农业知识及图像识别问题。

应用配置区：在中间部分下方，有一个 AI 配置区域，这里选择了 FastAI 图片识别作为模型，并提供了提示词来引导模型聊天方向。

提示词(Prompt):提示词是在应用中使用的文本，用于引导模型聊天的方向或行为。当用户与应用交互时，提示词可以帮助模型理解用户的意图，并给出相应的回复。

对话(Conversation):对话是指用户与应用之间的交流过程。在 FastGPT 平台上，您可以使用对话日志来查看历史对话记录，以便了解应用的表现和改进点。此外，您还可以立即开始新的对话，以测试应用的功能是否正常工作。配置界面如图 5-14 所示。

图 5-14 系统配置界面

Fig. 5-14 The system configuration page

提示词(Prompt):提示词是在应用中使用的文本，用于引导模型聊天的方向或行为。当用户与应用交互时，提示词可以帮助模型理解用户的意图，并给出相应的回复。本系统的提示词界面如图 5-15 所示。

设置农业知识问答系统的提示词时，需要考虑以下几个方面：

（1） 领域特定词汇

包括农业领域常用的术语、概念和关键词，例如作物种类、疾病名称、农业技术、农产品生产等。这些词汇可以帮助系统更好地理解用户提出的问题，并提供准确的答案。

（2） 常见问题关键词

梳理常见的农业问题，并提取关键词作为提示词，例如“农作物种植技术”、“病虫害防治”、“土壤肥力管理”等。这些关键词可以引导用户提出问题，并帮助系统准确识别用户的需求。

（3） 问题分类关键词

根据农业知识的不同分类，设置相应的关键词，例如“种植技术”、“农药防治”、“肥料施用”、“农业政策法规”等。用户可以通过点击或输入这些关键词来选择特定的问题类别，从而获取相关的信息和解答。

（4） 地域特定词汇

系统针对特定地区或国家的农业问题，可以设置地域特定的词汇作为提示词，例如该地区常见的作物品种、气候特点、土壤类型等。这样可以使系统更贴近用户的实际情况，提供更加个性化的服务。

（5） 提示词设计目的

简要介绍提示词的设计目的，即为了提高用户与系统的交互效率和准确性，帮助用户更快地找到所需的农业知识和解答。

（6） 提示词列表

列举设置的一些关键词和短语作为提示词，并说明其所代表的含义和作用。可以按照不同的类别进行分类展示，便于读者理解。

（7） 提示词与用户交互

描述提示词在系统中的应用场景和用户与系统的交互方式，例如用户如何使用提示词进行问题提问或分类选择，并举例说明。

（8） 优化和改进

分析当前提示词设置的优点和不足之处，提出可能的优化和改进方案，以进一步提升系统的用户体验和功能性。

（9） 提示词对系统性能的影响

讨论提示词设置对系统性能的影响，包括用户提问的准确性、系统响应时间、问题分类的精准度等方面的影响，并进行评估和分析。

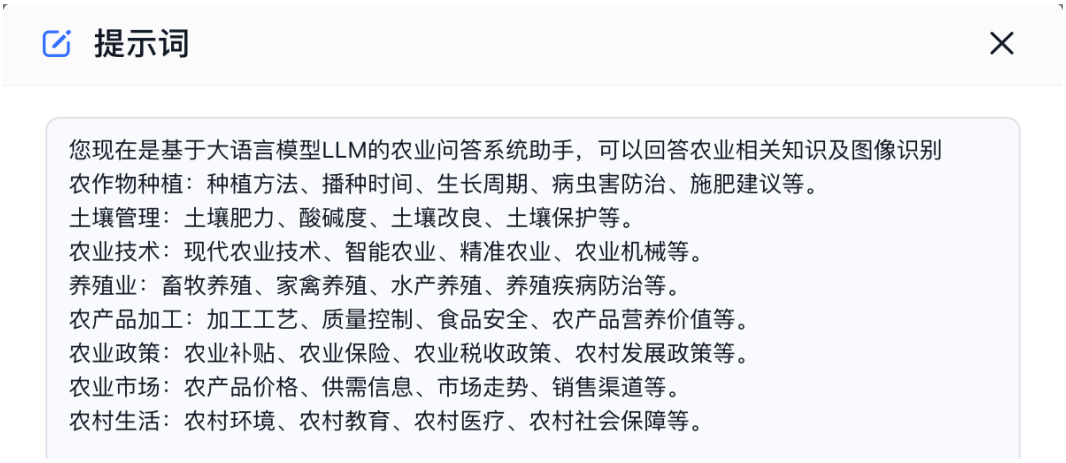

图 5-15 提示词界面

Fig. 5-15 prompt word page

5.3.5 数据集挂载

有些数据较为独特，可能需要单独的进行预处理分割后再导入 FastGPT，此时可以选择 csv 导入，可批量的将处理好的数据导入，个别数据也可以进行手动输入。文件导入界面如图 5-16 所示

图 5-16 文件导入界面

Fig. 5-16 File import page

大模型挂载数据库的作用在于为模型提供额外的数据支持，以提高模型的性能和效果。通过挂载数据库，可以实现以下几个方面的功能和优势：

数据增强：数据库中的数据可以用于增强模型的训练数据集。大规模的数据库可以提供更多样化、更丰富的数据，有助于训练模型更好地理解和泛化各种情况。

实时数据更新：数据库中的数据可以随时更新，确保模型获取到最新的信息。这对于需要实时反馈的任务非常重要，如实时推荐、实时预测等。

数据检索：模型可以通过数据库检索相关的数据，以响应用户的查询或请求。数据库提供了高效的数据存储和检索机制，可以帮助模型快速找到所需的信息。

个性化服务：基于数据库中的用户数据，模型可以提供个性化的服务或建议。例如，根据用户的偏好和历史行为推荐个性化的内容或产品。

数据管理和存储：数据库作为数据的集中存储和管理平台，可以帮助模型组织和管理海量的数据，提高数据的可用性和可访问性。

5.4 系统功能测试

本节将阐述对农业知识问答系统的各个功能模块进行测试的情况，目的是检验系统各部分的性能是否满足原定的设计标准。这一系列的测试流程，不仅对于确认系统功能的完备性具有重大意义，更是保障系统在实际使用中能够稳定运行、可信赖的不可或缺的一环。

5.4.1 登陆界面测试

该测试部分的目标是对登录模块的安全性、稳定性和用户体验进行全面评估。进行了一系列严格的安全协议测试、用户界面评价和稳定性测试，以保障登录模块在维护用户信息安全的基础上，能够为用户提供顺畅且直观的操作体验。具体测试内容如表 5-1 所示

5.4.2 数据采集和处理模块的测试

此模块对系统进行数据抓取以及预处理的功能性验证。测试的核心目标在于确保系统在数据采集与预处理流程中的精准度与高效性，从而保障数据在进入后续分析阶段前能够满足预定的质量标准和要求。通过此次测试，期望能够全面评估系统在这一关键环节的性能表现。详细的测试内容如表 5-2 所示。

5.4.3 算法接口模块测试

本部分测试的重点在于评估农业知识问答接口的精确性和响应速度。测试

旨在确保接口在处理多样化问题文本时，能够展现出卓越的稳定性与准确性。具体测试参数及预期成果已详细列于表 5-3 中。

表 5-1 登陆界面的测试

Table 5-1 Testing of the login screen

编号	测试功能	测试用例	测试数据	预期结果	是否通过
1	用户身份验证测试	验证正确和错误的凭证	正确和错误的用户名及密码	正确的用户名和密码登录成功，错误的用户名和密码拒绝访问并且提供错误提示	是
2	密码安全性测试	测试密码强度要求和加密方式	包括不同强度的密码(如简单、一般、复杂密码)以及密码加密机制	密码符合安全要求,且存储和传输中得到加密	是
3	用户界面友好性测试	测试登录页面的用户体验和响应性	包括对登录页面设计的评估如布局合理性、错误提示的清晰度和响应速度	登陆页面直观易用,有效反馈用户操作结果	是

表 5-2 数据采集和处理模块的测试

Table 5-2 Tests of the data acquisition and processing module

编号	测试功能	测试用例	测试数据	预期结果	是否通过
1	数据源抓取测试	抓取到的数据	各种平台知识库数据	成功获取并识别平台数据	是
2	数据预处理测试	执行数据清洗、格式转换和去重	针对抓取到的原始数据	数据清洗后无重复内容，格式统一	是

表 5-3 算法接口模块测试

Table 5-3 Algorithm Interface Module Testing

编号	测试功能	测试用例	测试数据	预期结果	是否通过
1	接口稳定性测试	检验 API 的连通性和响应速度	不同负载下 RESTful API 请求，包括请求频率和数据量	API 响应迅速，无中断	是
2	回答准确性测试	提交多种问题文本样本	包括多种表达形式、多种方面的文本样本，确保测试覆盖多种问题表达	分析结果与样本结果一致	是

5.5 本章小结

测试结果表明，关于登陆界面的测试，系统在用户身份、密码安全性、用户界面友好性三方面表现良好，满足登录界面的要求。关于数据采集处理模块，可以获取各平台数据，清洗后的数据没有重复，格式统一，通过测试要求。关于算法接口模块测试，API 响应迅速，没有中断，接口稳定性通过。通过提交多种问题文本样本，对系统回答的稳定性进行测试，分析结果与样本结果一致。

第六章 问答系统核心问题与解决方案

本节对模型搭建中存在的问题以及解决方案进行了归纳与总结，具体包括数据问题、行为不匹配问题、知识修改问题、LLM 脆弱评估性问题几方面。

6.1 数据问题

数据方面存在的问题包括：

（1） 难以理解的数据集。主要是因为预训练数据集太大，所以不可能去人工评估整个数据集，这可能导致隐私问题、重复数据和低质量数据影响训练效果、平衡数据集、数据闭源等一系列问题。

（2） 隐私问题：大规模数据集通常包含敏感信息，例如个人身份、位置信息等。对整个数据集进行人工评估可能涉及潜在的隐私泄露风险，这是一个严峻的问题。

（3） 重复数据和低质量数据：数据集的庞大可能导致其中包含大量的重复数据或低质量数据，这会对模型的训练效果产生负面影响。而人工评估所有数据并清理异常数据是非常耗时且困难的。

（4） 平衡数据集：在大数据集中，不同类别的样本数量可能存在巨大的不平衡，这会导致模型在训练时对于较小类别的学习不足。

（5） 数据闭源：对于一些商业公司或组织，他们的数据集可能是封闭的，不容易被外部人员理解和审查。

针对数据采取的解决方案：

（1） 隐私问题的处理：引入差分隐私技术，该技术能够在维护数据集隐私的同时，允许一定程度的统计分析，从而解决了对整个数据集进行人工评估时的难题。

（2） 重复数据和低质量数据的清理：利用自动化的数据清理技术，例如基于规则的清理、异常检测等，来自动识别和清理数据集中的问题数据。

（3） 平衡数据集的方法：使用过采样或欠采样等技术来平衡数据集中各个类别的样本数量，确保模型在训练时对每个类别都能够得到充分的学习。

（4）数据闭源的透明性：提供模型解释性工具，例如 LIME（Local Interpretable Model-agnostic Explanations），以帮助理解模型在封闭数据集上的决策过程。

面对庞大而复杂的数据集，需要综合运用技术手段和伦理原则来解决难以理解的问题。通过差分隐私、自动化清理、平衡数据和提高数据透明性等措施，能够更好地应对数据集带来的挑战，确保模型的可靠性和可解释性。这些方法的综合使用将有助于构建更加健壮和可信赖的预训练模型。

6.2 行为不匹配问题

行为不匹配问题指的是模型的实际输出或行为与用户的预期或指令之间存在显著的不一致或差异。这种不匹配可能表现为模型未能准确理解用户的指令，或者模型的输出未能满足用户的实际需求。

行为不匹配问题主要讨论对齐问题，目前主要分为两种：检测误对齐行为和对齐模型行为。

行为不匹配问题可能源于多个方面：（1）训练数据的偏差：如果训练数据不能充分覆盖模型在实际应用中可能遇到的各种情境，模型可能无法正确对应用户的期望行为。（2）标签噪声：训练数据中存在的标签错误或噪声可能导致模型学到错误的行为。（3）模型结构不足：模型结构的不足或者过于简单可能无法捕捉复杂的语境和用户意图，导致输出行为的不匹配。

对齐问题主要分为两种类型：（1）检测误对齐行为：模型错误地将某些输入对应到不正确的输出行为上。（2）对齐模型行为：模型在输出中不能很好地对应到用户期望的行为，即模型生成的行为与用户的意图不匹配。

针对行为不匹配问题采取的解决方案：

（1）数据质量与多样性

利用数据增强技术，扩充训练数据，使得模型能够更好地适应多样的输入情境，减轻训练数据偏差，对于训练数据中的标签错误或噪声，可以采用标签纠错的方法，通过专门的算法或者人工审核来修正标签。

（2）模型设计与对抗性训练

设计更为复杂的模型结构，使其能够更好地捕捉语境和用户意图，提高模型

对用户输入的理解。引入对抗性训练，通过故意引入对抗样本，让模型学会更加鲁棒和准确的映射关系，降低对检测误对齐行为的敏感性。

（3） 模型解释与可解释性

使用模型解释技术，使得模型生成的结果更具可解释性，从而更容易理解模型为何做出特定的决策。在模型训练和评估的过程中，引入用户的参与，通过用户反馈来调整模型的行为，提高模型与用户期望的对齐度。

行为不匹配问题是深度学习应用中一个复杂而关键的挑战。通过综合利用高质量、多样性的训练数据，设计复杂且鲁棒的模型结构，以及加强模型的解释性和用户参与度，可以有效地缓解行为不匹配问题，提高人工智能系统在实际应用中的可靠性和用户满意度。在未来的研究和应用中，这些建议有望为更好地解决行为不匹配问题提供有益的启示。

6.3 知识修改问题

知识修改问题主要涉及对模型内部存储的知识进行更新或修正的过程。这通常是由于模型在训练过程中可能吸收了一些过时、错误或不完整的信息，或者随着时间和领域知识的进步，需要更新模型中的知识以适应新的环境和需求。

知识修改问题具体可能来源于多个方面：（1）数据偏差：如果训练数据不能充分涵盖真实世界的各种情况和知识，模型学到的知识可能受到数据偏差的影响。（2）标签错误：训练数据中存在标签错误或噪声，这可能导致模型学到错误的知识。（3）模型结构限制：模型结构设计的不足或者过于简单，可能无法捕捉复杂的知识表示。

关于知识修改问题的解决方案：

（1） 数据多样性与纠偏

多样性数据：引入更多样性的训练数据，覆盖更广泛的场景和知识，以减轻数据偏差的影响。

纠偏技术：对训练数据进行纠偏，通过专门的算法或人工审核来修复标签错误，提高数据质量。

（2） 模型复杂性与表示能力

复杂模型设计：设计更为复杂、深层次的模型结构，以提高模型的表示能力，

使其能够更好地捕捉各种复杂的知识表示。

迁移学习：利用迁移学习的思想，将在其他领域学到的知识迁移到当前任务，提高模型的泛化能力。

（3） 模型解释与可解释性

模型解释工具：使用模型解释工具，对模型的预测结果进行解释，帮助理解模型到底学到了哪些知识。

（4） 用户参与：引入用户在模型训练和评估过程中的参与，通过用户反馈调整模型的知识表示，增加模型与用户期望知识的一致性。

（5） 持续监控与更新

监控系统：建立监控系统，对模型在实际应用中的表现进行实时监控，及时发现知识修改问题。

在线学习：引入在线学习的机制，根据新的数据和反馈不断更新模型，使其能够适应不断变化的知识需求。

知识修改问题在深度学习应用中具有一定的挑战性，但通过综合利用多样性的训练数据，设计复杂且鲁棒的模型结构，以及加强模型的解释性和用户参与度，可以有效地缓解知识修改问题，提高人工智能系统在实际应用中的鲁棒性和可靠性。未来的研究和应用中，需要更深入地挖掘和理解知识修改问题的本质，以更好地应对不断变化的知识需求和复杂的应用场景。

6.4 LLM 脆弱评估性问题

LLM 脆弱评估性问题指的是对 LLM 在特定情境下可能展现出的安全缺陷、错误输出或不稳定行为进行系统的评估和分析。这种评估通常关注模型在处理不同任务、面对不同输入时可能暴露出的脆弱性，例如模型对恶意输入、误导性信息或复杂语境的应对能力。

因为大型语言模型通常具有庞大的参数和复杂的内部机制，这使得它们在某些情况下可能产生不准确、不安全或具有误导性的输出。通过进行脆弱性评估，研究人员和开发者可以更好地了解模型的局限性，并采取相应的措施来降低潜在风险。对于大规模语言模型（LLM），如 GPT 系列等，评估其脆弱性成为当前研究中的一个关键问题，LLM 的广泛应用需要保证其在各种情境下的鲁棒性

和可信度。本文将深入分析 LLM 评估脆弱性的原因，并提出相应的解决方案，以期更好地理解和解决这一问题。

LLM 脆弱性的原因:

（1） 数据偏见：LLM 在训练过程中可能受到大规模数据集的影响，使其学到特定社会偏见、文化差异或负面刻板印象，导致输出结果具有偏见性。

（2） 对抗攻击：LLM 可能受到对抗攻击的影响，即通过有意制作的输入，使模型产生误导性输出。这可能影响 LLM 在实际应用中的可靠性。

（3） 不确定性处理：LLM 在面对不确定性时，可能输出自信度高但错误的结果，缺乏对不确定性的有效处理，降低了模型的鲁棒性。

解决 LLM 脆弱性的方案：

（1） 数据多样性与平衡

多样性数据：引入更多样性的训练数据，涵盖不同社会、文化和语境，减轻模型因数据偏见而导致的脆弱性。

平衡标签：对于可能导致偏见的标签，采取平衡策略，确保训练数据中的各种标签得到合理的表示。

（2） 对抗攻击防御

对抗训练：引入对抗训练，通过将对抗样本加入训练数据，使模型更好地抵抗对抗攻击，提高鲁棒性。

检测与过滤：建立对抗样本检测与过滤机制，剔除具有潜在攻击性的输入，减少对抗攻击的影响。

（3） 不确定性建模与解释

不确定性模型：引入更先进的不确定性建模技术，使 LLM 在输出结果中能够有效表达不确定性，提高在不确定性情境下的鲁棒性。

结果解释工具：提供结果解释工具，帮助用户理解 LLM 的输出，特别是在不确定性较高或模型输出可能存在误导性时。

（4） 持续监控与反馈

监控系统：建立 LLM 的持续监控系统，及时发现模型输出中的潜在问题，包括脆弱性，以便进行及时调整和改进。

（5） 用户反馈：引入用户反馈机制，鼓励用户提供模型输出的反馈信息，以不断优化和改进 LLM 的性能。

需要注意的是，LLM 的脆弱性评估是一个持续的过程，随着模型的不断发

展和更新，新的脆弱性可能会不断出现。因此，对 LLM 进行定期的脆弱性评估是非常重要的，以确保模型的安全性和可靠性。

6.5 本章小结

基于大模型的知识问答系统在其发展和应用过程中，面临着一系列关键问题。这些问题涵盖了技术、伦理、社会等多个层面，对于构建可靠、高效且社会负责任的系统来说，都是需要认真思考和解决的难题。

首先，大型模型所需的计算资源是构建知识问答系统时不可忽视的一个方面。这些模型通常包含数以亿计的参数，需要大量的训练数据和强大的计算能力。对于小规模团队或资源受限的研究者来说，获取和使用这些计算资源可能是一项严峻的挑战。除了硬件成本外，云计算服务的使用费用也可能成为一项不小的经济负担。

其次，数据隐私与安全问题是大型知识问答系统面临的另一个关键问题。为了训练大型模型，通常需要大规模的数据，而这可能涉及到用户个人信息的处理。因此，在设计和使用这样的系统时，保护用户数据的隐私和安全就显得尤为重要。采取有效的加密、匿名化等手段，以确保用户数据不受到滥用和泄露，成为保障系统可持续发展的一项基础工作。

此外，大型模型在处理特定领域的知识问答时可能表现出一定的局限性。尽管这些模型具有强大的泛化能力，但在某些专业领域或需要特定领域知识的问题上，它们可能不如专门设计的小型模型表现出色。这种领域特定性可能会限制系统在某些行业或学科领域的应用。

解释性问题也是大型知识问答系统的一大挑战。大型模型通常包含数以亿计的参数，其内部结构极为复杂，导致解释模型决策的过程变得更加困难。在一些应用场景中，用户对于模型是如何得出特定答案的有强烈的需求。因此，提高模型解释性，使用户能够理解模型的决策过程，是提升系统可接受度和用户信任度的关键步骤。

在样本效率方面，大型模型通常需要大规模标注数据才能充分发挥其性能。然而，获取大规模标注数据并不总是容易的，尤其在一些特定领域，数据可能相对匮乏。这可能导致模型在样本较少的情况下表现不佳，因此提高模型在小样本

数据上的效果成为一个亟待解决的问题。

综上所述，基于大模型的知识问答系统在发展过程中面临着多重挑战，涉及技术、伦理、社会等多个层面。解决这些问题需要多方面的努力，包括技术手段的创新、制定规范和政策的制定以及用户教育等方向的探索。只有综合考虑这些问题，系统才能更好地为用户提供高效、安全、可靠的知识问答服务。

第七章 结论与展望

在当前大模型时代，充分发挥大模型的“涌现”能力，并将其有效适配到具体领域场景，已经成为垂直行业建立竞争力的重要关键。在这一背景下，高质量的领域数据与专业知识的融合成为实现这一目标的不可或缺的要素。在此背景下，本文研究并构建了一种基于大语言模型的农业知识问答系统，该系统实现了ChatGLM 嵌入 FastGPT 的方式，实现了对农业知识的深度理解和高效问答。

在研究中，本文利用 ChatGLM 的强大语言处理能力，对农业领域的文本数据进行深度学习和理解，提取出丰富的农业知识。同时，本文嵌入 FastGPT 问答系统，使得系统能够更快速地响应用户的问题，并生成准确、具体的回答。本文首先使用 js 结合 Python 技术部署出了 FastGPT 平台，接下里私有化部署 ChatGLM 大模型，这个可以通过 modelscope 下载模型，使用脚本设置路径，运行出大模型的接口，然后接入 OneAPI，OneAPI 起到的作用是将 FastGPT 接入 ChatGLM 大模型，从而实现对大模型的问答，以及提示词的设置和 agent 等功能，大模型的提示词工程是比较重要的步骤，在设置提示词的时候，限定了模型只能回答农业相关问题，不允许回答除农业外的其他问题，这样子可以避免模型回答的时候出现乱回答或回答错的情况，接下来就是设置工作流，让模型能有逻辑的按照要求执行，具体流程就是先接受开场白，确定模型是一个知识问答系统助理的身份，然后读取向量数据库处理挂载的数据集，让模型学习这个数据，之后接入输入输出的节点，节点的字段格式需要一致，然后设置提示词给模型，最后完成工作流，实现模型输入输出的知识问答。

综上所述，本研究成功构建了一种基于 ChatGLM 和嵌入 FastGPT 的农业知识问答系统，实现了对农业知识的智能化处理和应用，为农业发展提供了有力支持。

参考文献

[1] 宋仕月，陈政羽，郑一凡，徐梓航，潘铖深度学习在农业病虫害智能识别方面的研究进展[J].智慧农业导刊,2023,3(04):1-4.DOI:10.20028/j.zhnydk.2023.04.001.

[2] 王其聪. 农业专家系统的发展趋势[J]. 安徽农业科学,2006,34(16):4169,4185. DOI:10.3969/j.issn.0517-6611.2006.16.152.

[3] 陈子睿, 王鑫, 王林, 等. 开放领域知识图谱问答研究综述[J]. 计算机科学与探索, 2021, 15(10): 1843-1869. CHEN ZIRUI, WANG XIN, WANG LIN, et al. Survey of Open-Domain Knowledge Graph Question Answering[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(10): 1843-1869.

[4] 萨日娜, 李艳玲, 林民. 知识图谱推理问答研究综述[J].计算机科学与探索, 2022, 16(8): 1727-1741. SA RINA, LI YANLING, LIN MIN. Survey of Question Answering Based on Knowledge Graph Reasoning[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(8): 1727-1741.

[5] GUU K, LEE K, TUNG Z, et al. Retrieval augmented language model pre-training[C]//International conference on machine learning, Messe Wien Exhibition & Congress Center, Vienna, Austria, Jul 12-18, 2020. New York: PMLR, 2020: 3929-3938.

[6] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.

[7] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J]. arXiv preprint arXiv:2206.07682, 2022.

[8] WANG Y, KORDI Y, MISHRA S, et al. Self-Instruct: Aligning Language Model with Self Generated Instructions[J]. arXiv preprint arXiv:2212.10560, 2022..

[9] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[10] OPENAI. GPT-4 Technical Report[R]. arXiv e-prints: arXiv:2303.08774, 2023.

[11] 王鑫,韩立帆,等. 大语言模型融合知识图谱的问答系统研究 [J].张鹤译.计算机科学与探索, 2023, 17 (10): 2377-2388.

[12] MAYNEZ J, NARAYAN S, BOHNET B, et al. On faithfulness and factuality in abstractive summarization[J]. arXiv preprint arXiv:2005.00661, 2020.

[13] TONEVA M, SORDONI A, COMBES R T, et al. An empirical study of example forgetting during deep neural network learning[J]. arXiv preprint arXiv:1812.05159, 2018.

[14] [1]曹亚菲. AI 新征途——万物皆可模型化 [J]. 软件和集成电路, 2023, (06): 16-20.

DOI:10.19609/j.cnki.cn10-1339/tn.2023.06.016.

[15] DU Z, QIAN Y, LIU X, et al. GLM: General language model pretraining with autoregressive blank infilling[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 22-27, 2022. Stroudsburg: ACL, 2022: 320-335.

[16] LIU X, ZHENG Y, DU Z, et al. GPT understands, too[J]. arXiv preprint arXiv:2103.10385, 2021..

[17] LIU X, JI K, FU Y, et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arXiv preprint arXiv:2110.07602, 2021.

[18] WU C, ZHANG X, ZHANG Y, et al. PMC-LLaMA: Further Finetuning LLaMA on Medical Papers[J]. arXiv preprint arXiv:2304.14454, 2023.

[19] SINGHAL K, AZIZI S, TU T, et al. Large Language Models Encode Clinical Knowledge[J]. arXiv preprint arXiv:2212.13138, 2022.

[20] YUNXIANG L, ZIHAN L, KAI Z, et al. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge[J]. arXiv preprint arXiv:2303. 14070, 2023.

[21] 郭紫月. 基于生成式方法的蒙汉机器翻译研究[D]. 呼和浩特：内蒙古大学, 2021. DOI:10.27224/d.cnki.gnmdu.2021.000423.

[22] 车万翔,窦志成,冯岩松,等. 大模型时代的自然语言处理:挑战、机遇与发展 [J]. 中国科学:信息科学, 2023, 53 (09): 1645-1687.

[23] 张文龙,胡天亮,王艳洁,等. 云/边缘协同的轴承故障诊断方法 [J]. 计算机集成制造系统, 2020, 26 (03): 589-599. DOI:10.13196/j.cims.2020.03.002.

[24] 任乐,张仰森,刘帅康.基于深度学习的实体关系抽取研究综述[J].北京信息科技大学学报(自然科学版),2023,38(06):70-79+87.