

中国图书资料分类号： TP391
密 级： 公开

学校代码： 10679
学 号： 20212015120031



大理大学
DALI UNIVERSITY

专业硕士学位论文

融合大语言模型的中药领域知识图谱问答系
统的研究

**Research on the Integration of Large
Language Models with Traditional Chinese
Medicine Domain Knowledge Graph
Question Answering System**

学 院	数学与计算机学院
专 业	计算机技术
研究方向	知识图谱
研 究 生	贾宗荣
指导教师	刘文志 讲师
答辩日期	2024 年 5 月 17 日

摘 要

随着大语言模型问答系统的大量出现，这些系统在问答领域展现出了卓越的自然语言处理能力。然而，在准确性和可解释性方面，以及特定专业领域的问答中，大语言模型问答系统与知识图谱问答系统相比，仍有一定的差距。同时，现有的知识图谱问答系统研究在语义理解和答案生成方面也面临局限。鉴于这两种方法各有优缺点，并且它们之间存在互补性，本研究提出了一种通过融合大型语言模型和知识图谱来构建智能问答系统的解决方案。

本方案利用图数据库技术，基于丰富的中药开源数据，利用知识图谱的知识提取和知识融合技术，精心设计并构建了一个全面的中药领域知识图谱。该图谱涵盖了中药的性状、功效、临床应用等多个维度，囊括了各种实体、属性以及它们之间的复杂关系，为问答系统提供准确的数据支持，并通过该知识图谱所包含的结构化数据实现问答系统中的数据检索。此外，围绕该知识图谱，本方案还构建了大量问答对数据集，对大语言模型进行微调以优化大语言模型，加强在问句解析方面的能力。通过结合该模型并设计合理的提示词，利用大语言模型出色的语义解析能力实现高效的问句解析，并生成对应的查询语句以检索知识图谱。针对检索结果，再次调用大语言模型并设计适当的提示词，利用其优秀的文本生成能力，通过读取查询结果生成流畅且丰富的自然语言回答。最后，本方案利用最新的开发框架成功实现了该问答系统，并以可视化的形式展现了这一解决方案，为用户提供了一个直观、易用的问答平台。

为了验证该策略的有效性，本研究通过在中药知识问答数据集上进行对比实验，将该融合策略开发的问答系统与仅依赖大语言模型的问答系统进行比较，并设计指标对实验结果进行评估。实验结果表明，融合知识图谱与大语言模型的方法显著提高了问答系统的准确性和可解释性。该方案同时利用了大语言模型在自然语言处理方面的优势，以及知识图谱高效的数据查询能力，证明了该方案在智能问答系统中的有效性，尤其是在特定专业领域的应用中。

关键词：知识图谱；大语言模型；问答系统；中药

Abstract

As large language model question-answering systems have proliferated, they have demonstrated exceptional natural language processing capabilities in the QA domain. However, when compared to knowledge graph-based QA systems, these models still have gaps in accuracy, interpretability, and domain-specific expertise. Given the strengths and weaknesses of both approaches and their complementary nature, this study proposes a solution that integrates large language models with knowledge graphs to construct an intelligent QA system.

Utilizing graph database technology, this solution leverages rich open-source data on traditional Chinese medicine to meticulously design and build a comprehensive knowledge graph in this domain. The graph covers multiple dimensions of traditional Chinese medicine, including characteristics, effects, and clinical applications, and encompasses a variety of entities, attributes, and their complex interrelations, providing accurate data support for the QA system. Furthermore, this approach involves creating a large dataset of QA pairs to fine-tune the language model, enhancing its capability in parsing questions. By integrating this model and designing appropriate prompts, the system utilizes the language model's robust semantic parsing capabilities to efficiently analyze questions and generate corresponding query statements to retrieve information from the knowledge graph. Using the retrieval results, the language model is called upon again with carefully designed prompts to utilize its advanced text generation capabilities, producing fluent and rich natural language responses. Finally, this solution employs the latest development frameworks to successfully implement the QA system, presenting it in a visual format to offer users an intuitive and easy-to-use platform.

To validate the effectiveness of this strategy, the study conducts comparative experiments on a dataset of traditional Chinese medicine knowledge questions, comparing the hybrid system to one relying solely on a large language model. The results, evaluated with carefully designed metrics, show that integrating knowledge graphs with large language models significantly improves the accuracy and interpretability of the QA system. This approach effectively harnesses the strengths of large language models in natural language processing and the efficient data querying capabilities of knowledge graphs, confirming its effectiveness in intelligent QA systems, especially in specialized domains.

Keywords: Knowledge Graph; Large Language Model; Question-Answering System; Traditional Chinese Medicine

目录

摘 要.....	I
Abstract.....	II
第 1 章 绪论.....	1
1.1 研究背景及意义.....	1
1.2 国内外研究现状.....	3
1.2.1 知识图谱研究现状.....	3
1.2.2 大语言模型研究现状.....	5
1.3 本文研究内容.....	6
1.4 本文的组织结构.....	6
第 2 章 相关研究工作.....	8
2.1 知识图谱.....	8
2.1.1 知识图谱概述.....	8
2.1.2 知识图谱数据模型.....	8
2.1.3 知识图谱构建技术.....	9
2.1.4 知识图谱问答系统.....	10
2.2 大语言模型.....	10
2.2.1 大语言模型概述.....	10
2.2.2 大语言模型的技术原理.....	11
2.2.3 大语言模型的微调.....	12
2.2.4 Prompt 管理.....	12
2.2.5 大语言模型在智能问答中的作用.....	12
2.3 本章小结.....	13
第 3 章 中药领域知识图谱的构建.....	14
3.1 数据准备.....	14
3.1.1 数据获取.....	14
3.1.2 数据预处理.....	15
3.2 知识抽取.....	16

3.2.1 实体抽取.....	16
3.2.2 关系抽取.....	17
3.2.3 属性抽取.....	17
3.3 知识融合	18
3.3.1 实体对齐.....	18
3.3.2 本体层构建.....	18
3.4 知识图谱构建	20
3.4.1 数据模型	20
3.4.2 图数据库管理	21
3.4.3 数据导入	22
3.4.4 质量评估	24
3.5 知识图谱展示	25
3.6 本章小结	26
第 4 章 大语言模型的优化	27
4.1 模型选择.....	27
4.2 数据集构建.....	28
4.2.1 问答对构建.....	28
4.2.2 数据集构建.....	29
4.3 微调 ChatGPT-3.5 Turbo 模型.....	30
4.4 验证实验.....	32
4.4.1 实验环境.....	33
4.4.2 实验目的与评价指标.....	33
4.4.3 实验结果与分析	33
4.5 本章小结.....	34
第 5 章 融合大语言模型的知识图谱问答	35
5.1 大语言模型与知识图谱对比.....	35
5.1.1 大语言模型的优劣势.....	35
5.1.2 知识图谱的优劣势.....	35
5.1.3 融合大语言模型和知识图谱问答	36

5.2 Prompt 管理.....	38
5.2.1 Prompt 原理.....	38
5.2.2 问句解析.....	38
5.2.3 答案生成.....	39
5.3 对比实验.....	40
5.3.1 实验目的.....	40
5.3.2 实验准备.....	41
5.3.3 实验过程.....	42
5.3.4 实验结果与分析.....	42
5.4 本章小结.....	44
第 6 章 系统实现.....	45
6.1 开发工具与技术.....	45
6.2 系统设计.....	45
6.2.1 系统功能.....	45
6.2.2 系统流程.....	46
6.2.3 问句输入模块.....	47
6.2.4 问句解析模块.....	48
6.2.5 数据检索模块.....	49
6.2.6 答案生成模块.....	49
6.3 系统展示.....	49
6.4 本章小结.....	50
第 7 章 总结与展望.....	51
7.1 总结.....	51
7.2 展望.....	51
参考文献.....	53
致 谢.....	56
攻读学位期间发表的学术论文和研究成果.....	57

第 1 章 绪论

1.1 研究背景及意义

回顾新时代中医药走过的十年，习近平总书记把中医药工作摆在更加重要的位置，领导中医药传承创新发展取得历史性成就、发生历史性变革。十年来，习近平总书记对中医药工作作出一系列重要论述，深刻回答了新时代如何认识中医药、如何发展中医药、发展什么样的中医药等根本性、长远性问题。促进中医药传承创新发展成为新时代中国特色社会主义事业的重要内容，成为中华民族伟大复兴的大事，这是习近平总书记为中医药事业划定的新时代坐标。全体中医药人不忘初心，勇担使命，以习近平总书记重要论述为根本遵循，跑出了中医药振兴发展的加速度^[1]。

为了认真落实习近平总书记关于中医药工作的重要论述，促进中医药传承创新发展，中共中央、国务院提出的《关于促进中医药传承创新发展的意见》指出：以信息化支撑服务体系建设。实施“互联网+中医药健康服务”行动，鼓励依托医疗机构发展互联网中医医院，开发中医智能辅助诊疗系统，推动开展线上线下一体化服务和远程医疗服务^[2]。

随着信息技术的快速发展，知识图谱作为一种创新的数据处理方法，已经成为连接复杂知识和信息的重要桥梁。在中医药领域，知识图谱的应用尤为重要，它不仅能够系统化和结构化地组织中医药的丰富知识，包括药物、疾病、治疗方法等，还能支持复杂的查询和推理，为医生提供准确的辅助诊疗决策支持。知识图谱（Knowledge Graph）在 2012 年由 Google 正式提出，其初衷是为了优化搜索引擎返回的结果，增强用户搜索质量及体验。随后 Airbnb^[3]、亚马逊^[4]、eBay^[5]、Facebook^[6]、IBM^[7]、领英^[8]、微软^[9]、优步^[10]进一步宣布了知识图谱的开发。近年来，知识图谱作为结构化人类知识的一种形式引起了学术界和工业界的高度关注。知识图谱是事实的结构化表示，由实体、关系和语义描述组成^[11]。通过构建中药知识图谱，可以有效地提升中医药服务的质量和效率，促进个性化治疗方案的生成，进一步推动中医药的现代化和国际化进程。

基于知识图谱的问答系统（Knowledge-Based Question Answering, KBQA）能够提供准确、可靠的答案，并具有较强的解释性，但其挑战在于知识图谱的构建和维护，以及对复杂自然语言查询的理解能力有限。知识图谱技术的出现，为基于知识图谱的问答系统（KBQA）的实现提供了重要技术支撑。基于知识图谱的问答系统（KBQA）是以自然语言问题作为输入，并使用结构化知识库（如 Freebase^[12]、YAGO^[13]和 DBpedia^[14]）返回事实答案的任务^[15]。

大语言模型（Large Language Models, LLMs）近年来已成为技术革新的前沿。特别是

ChatGPT 等模型的推出和应用，标志着智能问答系统进入了一个新的发展阶段。这些模型通过在海量文本数据上进行训练，掌握了复杂的语言理解和生成能力，能够在多种任务，包括文本生成、语义理解、翻译等领域展现出卓越的性能。最近，大语言模型的应用已经扩展到智能问答系统，其中它们通过理解用户的自然语言查询并生成准确、相关的回答，极大地提升了用户体验和信息检索的效率。然而，尽管大语言模型在处理广泛的话题和生成流畅的回答方面展现出强大的能力，它们在特定领域的知识准确性、深度理解以及回答的可靠性和解释性方面仍存在挑战。

因此，本研究旨在通过大型语言模型与知识图谱的融合策略，利用两者的优势，以克服当前智能问答领域面临的挑战。首先，通过搜集的中药开源数据，本研究精心设计并构建了一个中药领域的知识图谱，并利用图数据库技术进行存储，从而为问答系统的开发提供了坚实的数据基础。此外，本研究通过大语言模型的微调技术和 Prompt 技术的应用，对大语言模型进行了优化，以适应系统中的具体需求。随后，本研究利用大语言模型卓越的语义解析和文本生成能力进行问题的解析和答案的生成，同时依托于精心设计的中药领域知识图谱进行准确的数据检索。最终，通过精心设计的实验，验证该融合策略的有效性。这种融合方法不仅充分发挥了大型语言模型在自然语言处理（NLP）领域的先进技术优势，而且有效利用了知识图谱提供的结构化专业知识，从而为智能问答系统的优化和提升提供了强有力的技术支撑。通过该融合策略，能够有效补充单一技术的不足，显著提升系统回答的质量、准确性和解释性。该融合策略的实施方案流程如图 1.1 所示：

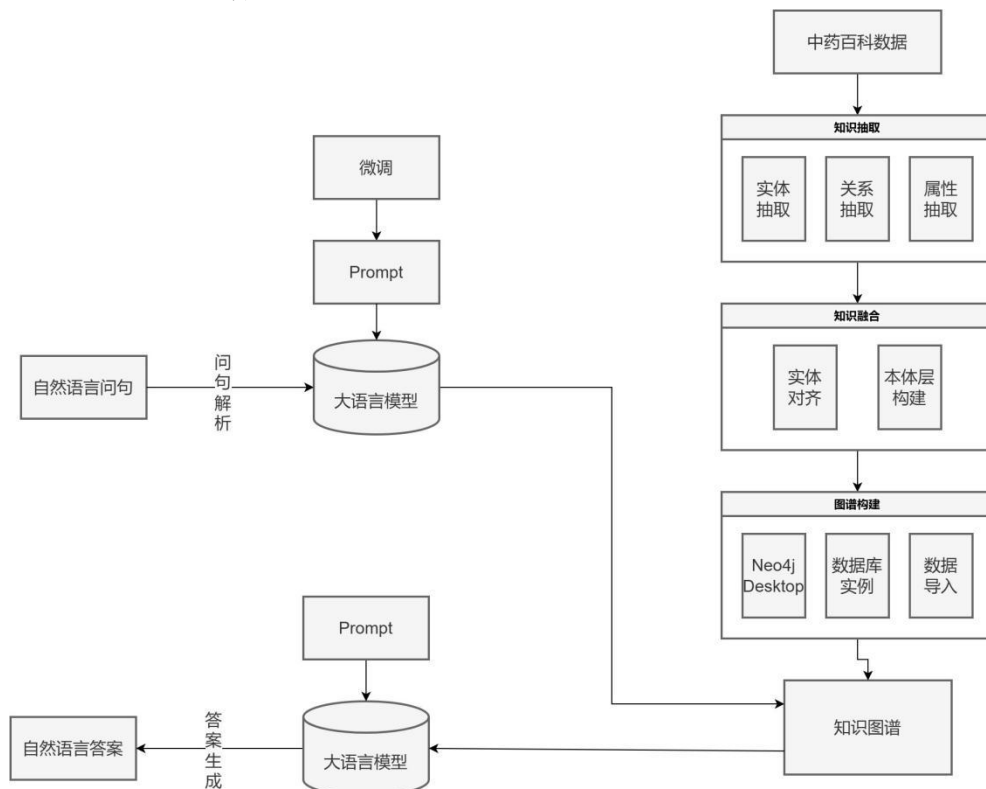


图 1.1 融合策略实施流程图

在中药等专业领域，这种融合策略显得尤为重要，使得智能问答系统能够更深入地理解和处理专业知识，为用户提供更精准和可解释的回答。通过整合最新的大语言模型技术，并以独立构建的中药领域知识图谱为支撑，开发出更先进、智能化且高效的知识图谱问答系统，有效地提升中医药服务的质量和效率，促进个性化治疗方案的生成，进一步推动中医药的现代化和国际化进程。中药领域问答系统还有助于中医药知识的传承和创新，为新时代中医药的发展注入新的活力。

1.2 国内外研究现状

1.2.1 知识图谱研究现状

中药领域知识图谱的应用能够有效地描述和挖掘实体间的关系，实现大规模知识的规范存储和高效应用，促进中医药资源的有效整合。这不仅为知识服务相关研究奠定了基础，也为中医药的传承和发展开辟了新思路^[16]。

在探索中医药基础理论知识图谱领域，众多研究已经揭示了其在各个应用方面的巨大潜力。举例来说，石燕等人^[17]通过构建中医体质知识图谱，可视化地展示了中医体质辨识在治未病和全民健康管理中的应用潜力。陈陵芳^[18]运用 Citespace 知识图谱可视化软件分析“病机十九条”相关文献，构建了该领域的知识库模型。张莹莹^[19]则开发了一个基于中医药知识图谱的舌象诊疗系统，该系统能够基于用户提供的症状描述和舌象图片，输出诊断建议和药物方案。卢克治^[20]运用深度学习技术对中医经典文献进行了精确的实体和关系提取工作，构建了一个知识图谱，并借助图数据库技术开发了一个可视化查询平台，实现了基于 Web 端的知识探索和获取功能。

在中医临床知识图谱研究方面，多项工作已经展示了知识图谱在深化临床诊疗理解和辅助决策中的重要价值。例如，牟梓君^[21]成功实现了通过知识图谱技术对小儿脑瘫相关知识进行可视化，覆盖了疾病诊断、证候、症状、治疗方法及其疗效，进而为制定和优化治疗方案提供了科学依据。石英杰^[22]则通过模拟临床诊疗路径，建立了基于病机辨证的胸痹知识图谱，为智能化辅助诊断系统的开发提供了参考模型。胡嘉元^[23]利用 Gephi 软件构建的中医病机知识图谱，进一步促进了中医临床个体化诊疗决策支持系统的建立，特别是在心血管疾病的治疗中，展现了其实际应用价值。郑子强^[24]专注于慢性肾脏病的诊疗过程，开发了中医诊疗知识图谱学习与推理系统的原型，为疾病管理提供了新的技术手段。孙明俊等人^[25]的研究则侧重于类风湿性关节炎，通过构建辅助诊疗系统，提供了诊疗指南和药方推荐，支持临床决策。张雨琪等人^[26]选择了著名医师赵炳南、朱仁康的治疗经验作为研究对象，构建了包含疾病、证候、症状、治法、方药等多维度概念及其相互关系的知识框架，并通过图数据库技术实现了知识的可视化表示。刘凡等人^[27]则基于 Neo4j 数据库，以

姚乃礼医师的脾胃病诊疗经验为基础,构建了名老中医临床经验知识图谱,实现了知识的可视化展示和高效语义搜索功能。

在中药知识图谱的方法学研究领域,一系列研究成果已经为中医药知识的组织与应用提供了新的视角和方法。例如,于彤等人^[28]采用了本体和语义网络技术,以中医药学语言系统(TCMLS)作为基础,初步构建了中医药知识图谱,并通过可视化技术展示了其结构,进而探索了如中医药维基百科系统和知识地图等应用的可能性。张德政等人^[29]利用本体知识表示法成功构建了中医核心知识图谱,并基于此探讨了中医临床经验等领域的应用。贾李蓉等人^[30]从多个维度,如资料来源、研究内容及展示形式等,构建了中医药知识图谱,并讨论了基于此的知识检索系统的应用前景。上海曙光医院的团队则构建了包含疾病、证候、方药等多个库的中医药知识图谱,旨在通过知识问答和辅助开药应用来提升临床服务的效率和质量^[31]。在NLP技术方面,屈倩倩等人^[32]基于 BertBiLSTM-CRF 模型对《伤寒论》文本中的关键实体进行识别,展现了高准确率。高佳奕等人^[33]则通过 LSTM-CRF 模型进行国医大师医案中症状实体的抽取,发现加入 Peep-hole 机制的双向 LSTM 模型在识别效果上更为优异,为中医文本自动化抽取提供了有效方法。此外,中国中医科学院中医药信息研究所已经开发并构建了包括中医药学语言系统、中医经方知识图谱、中医美容知识图谱在内的九个知识图谱,这些图谱已经被集成到中医药知识服务平台 TCMKB 中,并以可视化的形式对外展示,进一步推动了中医药知识的传播与应用。

通过这些研究工作,中药知识图谱不仅在理论构建上取得了显著进展,同时在实际应用中也展现出巨大潜力,为中医药学科的发展和知识传承提供了坚实的技术支持。

随着知识图谱技术的进步,基于知识图谱的问答系统(Knowledge Base Question Answering, KBQA)也迅速发展。知识图谱问答系统通过理解和分析自然语言问题,使用知识图谱中的数据回答问题,能够让不熟悉知识库具体数据结构的用户迅速且准确地获取所需的信息或答案^[34]。自谷歌在 2012 年首次引入知识图谱的概念以来,由于其能够以图形方式呈现实体及其相互关系的特性,知识图谱已被广泛应用于智能问答系统。基于知识图谱的问答系统通过对自然语言问题进行语义解析、知识提取和知识推理等步骤,使得用户可以更精确、更直接地获得问题的答案。

目前 KBQA 领域常用的方法有以下几种: 1) 基于规则的方法: 这些方法依赖于手工编写的规则来解析自然语言问题,并将其映射到知识图谱上的查询操作。这种方法的优势在于能够提供精确的控制和较高的解释性,但其缺点是扩展性和灵活性较差,维护成本高。2) 基于模板的方法: 模板方法通过定义一系列问题模板来识别和解析用户的查询,每个模板都对应着一种查询类型和相应的知识图谱查询语句。这种方法相比规则方法有更好的灵活性和一定的扩展性,但依然需要大量的人工工作来定义和优化模板。3) 基于语义解析的方法: 语义解析方法试图将自然语言问题转换为相应的逻辑表达式(如 Lambda-DCS^[35]等),然后将这个逻辑表达式映射到知识图谱上进行查询。这种方法通常依赖深度学习模型来进

行语义解析,能够处理更复杂的查询,但需要大量的标注数据来训练模型。(4)基于深度学习的方法:端到端的方法直接将自然语言问题转化为知识图谱上的查询,无需显式的语义解析过程。这通常通过深度神经网络来实现,例如使用注意力机制的序列到序列模型。这种方法能够自动学习问题到查询的映射规则,但对数据和模型的要求较高。

在数据集方面,目前在知识图谱问答系统中使用较多的是 SimpleQuestions 和 WebQuestions 数据集^[36]。Simple-Questions 数据集是由 Bordes 等^[37]构造,该数据集包含了 10W 多条问答对,其中包括训练集 79590 条、验证集 10845 条、测试集 21687 条,该数据集只包含简单问题,也称为 single-relation 问题。WebQuestions 数据集是由 Berant 在 2013 年提出的^[38],其中训练集包含 3782 条问答对,测试集包含 2037 条问答对。以上两种数据集中都是较为简单的问题。在 2016 年 Bao 等^[39]提供了一个多限制问题的数据集 ComplexQuestions,用来测试复杂问题的知识图谱问答系统的系统性能,其中包括了 2100 条问答对。

1.2.2 大语言模型研究现状

目前大语言模型并没有公认统一的概念定义。NCarlini 等学者指出大语言模型由具有大量参数(通常为数十亿以上)的神经网络组成^[40],并使用自监督或半监督学习在大量未标记文本上进行训练,具备通用能力,可以执行广泛的自然语言处理任务,包括文本摘要、翻译、情感分析等。Wayne Xin Zhao 等学者在综述中指出^[41],大模型指的是在海量文本语料上训练、包含至少数十亿级别参数的语言模型,例如 GPT-3、PaLM、LLaMA 等。

在大语言模型的领域中,OpenAI 开发的 GPT 系列无疑是最著名的。GPT-1 模型,首次亮相于 2018 年,是基于 Transformer 解码器架构的单向语言模型,拥有约 1.17 亿个模型参数^[42]。紧接着,GPT-2 在 2019 年推出,模型参数数量增至约 15 亿^[43]。这一时期,语言模型主要由基于编码器、基于解码器以及编解码器双向架构的模型构成。2020 年,OpenAI 的 Brown 等研究者发布了具有革命性意义的 GPT-3 模型,其参数量激增至 1750 亿^[44]。研究发现,模型的性能随着参数和数据规模的增加而提高,特别是 GPT-3 展现出了所谓的“涌现”能力^[45],即在达到一定的参数规模后,模型表现出复杂的推理能力。

近年来,一系列引领技术潮流的开源大模型相继亮相。2022 年 4 月,Google 推出了 PaLM 模型^[46],这一模型基于解码器架构,通过引入 SwiGLU 激活函数、并行层技术、多查询注意力机制和旋转向量嵌入等创新技术进行了深度优化。Google 还发布了 Flan T5^[47],这是一款基于 T5 模型的自回归语言模型,通过指令微调 and 思维链技术增强了其性能。2023 年 2 月,Meta 发布了 Llama^[48]模型,同样采用解码器架构并融合预正则化、SwiGLU 激活函数和旋转向量嵌入技术。紧接着在 7 月,Meta 进一步推出 Llama2^[49],该模型在 Llama 的基础上引入了 Ghost Attention 算法,旨在提升模型在多轮对话中的一致性。同年 5 月,Technology Innovation Institute 发布了 Falcon 模型^[50],这一模型同样采用解码器架构,并通过旋转位置

编码、多查询注意力和 Flash Attention 等技术提高了其性能。而清华大学在 6 月推出的 ChatGLM2 则采用 GLM 架构，集成了 Flash Attention 算法、多查询注意力机制、混合目标函数和人类偏好对齐训练等技术，以实现更优的模型性能。最后，百川智能在 2023 年 7 月发布了 BaiChuan 模型，采用与 Llama 相似的解码器架构，并利用 FlashAttention 算法、多查询注意力机制、RMSNorm 和混合目标函数等技术进行了细致优化。

1.3 本文研究内容

本文的研究内容主要集中在融合大语言模型的中药领域知识图谱问答系统的设计与实现，以及通过大语言模型和知识图谱的结合来提升问答系统性能的方法。研究内容具体包括以下几个方面：

- 1) 知识图谱的构建与应用：首先，本文详细介绍了中医药知识图谱的构建过程，包括数据来源的选择、数据预处理、实体识别与关系抽取等关键步骤。此外，还探讨了知识图谱在中医药辅助诊疗系统中的应用，如如何通过知识图谱支持复杂的医学查询和智能推理。
- 2) 大语言模型的优化：本文研究了如何利用大语言模型的微调技术，优化大语言模型提高其在特定领域的语义解析和文本生成能力。
- 3) 大语言模型与知识图谱的融合策略：本文探索了将大语言模型与知识图谱融合的新策略，旨在利用大语言模型的强大语义理解能力和知识图谱的结构化知识，共同提升问答系统的性能。通过实验验证了结合使用两种技术比单独使用任一技术都能显著提高问答的准确性和效率。

1.4 本文的组织结构

本文共分为七个章节。

第一章：绪论。介绍了本研究的背景及其重要性，并分析了当前国内外在相关研究领域的发展现状，旨在为读者提供研究的背景信息和研究的必要性。

第二章：相关研究工作。本章深入探讨了本研究所涉及的关键技术和理论基础，主要包括知识图谱的和大语言模型的原理技术。

第三章：中药领域知识图谱的构建。详细阐述了中药领域知识图谱知识图谱的构建流程，包括数据收集、知识抽取、知识融合及知识图谱构建过程，为后续研究奠定了坚实的数据基础。

第四章：大语言模型的优化。介绍了通过微调优化大语言模型，同时设计实验来验证微调模型的有效性和可行性。

第五章：融合大语言模型的知识图谱问答。探讨了大语言模型与知识图谱融合的可行

性和融合策略，最后通过设计对比实验验证了这种集成方法的有效性。

第六章：系统实现。详细介绍了基于知识图谱的中医药辅助诊疗系统的开发过程，展示了系统实现的细节和技术选择。

第七章：总结与展望。首先对本研究的关键发现和贡献进行总结，然后基于研究的现有不足，对未来的研究方向提出了展望。

第 2 章 相关研究工作

2.1 知识图谱

2.1.1 知识图谱概述

知识图谱 (Knowledge Graph, KG) 是一种用于组织和表达知识的数据结构, 采用图形表示方法, 其中节点代表实体 (如人、地点、物品、概念等), 边表示实体之间的各种关系。它提供了一种将现实世界中的关系网络以数学图形的形式进行建模的方法, 使得复杂的实体关系可以被直观地表示和处理。知识图谱的引入, 极大地推动了语义网的发展, 改进了搜索引擎的准确性, 增强了人工智能系统的理解能力。通过对大量数据进行结构化, 知识图谱使得信息检索变得更加高效, 为数据驱动的决策提供支持。

2.1.2 知识图谱数据模型

知识图谱的有效存储和查询对其应用至关重要。图数据库和 RDF 存储是两种常用的知识图谱数据模型, 它们为存储大规模的图结构数据提供了高效的支持。图数据库是一种专门设计来处理图形数据结构的数据模型, 它将数据存储为节点 (entities)、边 (relationships) 以及边上的属性。与传统的关系数据库不同, 图数据库专注于存储实体之间的关系和连接模式。这种数据模型使得图数据库非常适合处理复杂的查询和深层次的关系分析。本研究使用的图数据库是 Neo4j, 它以图的形式存储数据, 使用节点 (Nodes)、关系 (Relationships) 和属性 (Properties) 来表示和存储数据。节点代表实体, 关系表示节点之间的连接, 并且每个节点和关系都可以有多个属性。Neo4j 特别适合处理复杂的查询和深层链接数据, 因为它可以高效地遍历网络结构的数据。Cypher 查询语言作为 Neo4j 的查询语言, 针对的是图形数据, 能够直观地表示节点、关系和路径。如图 2.1 是本研究中 Neo4j 数据的可视化展示:

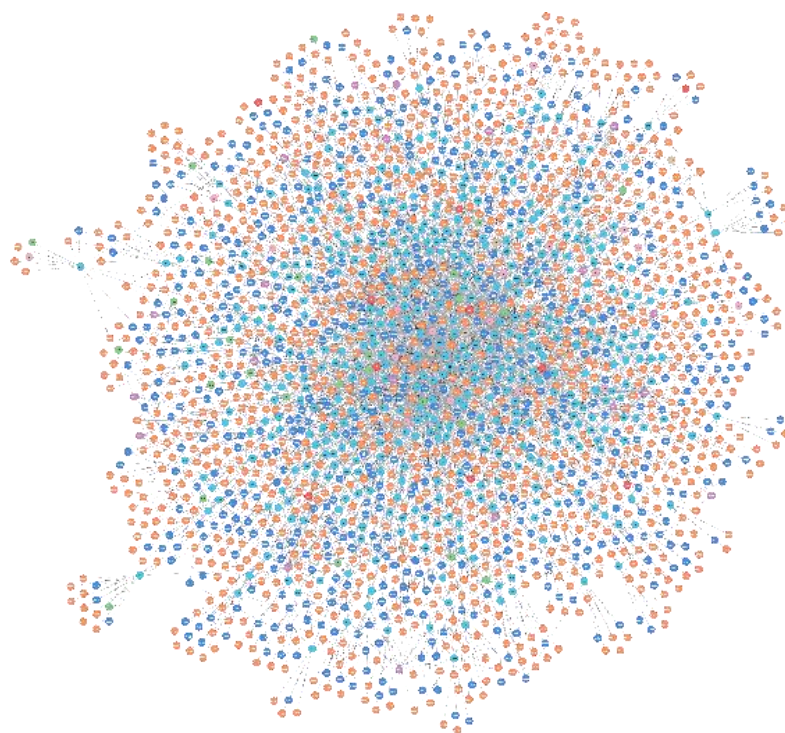


图 2.1 Neo4j 数据的可视化效果图

而 RDF (Resource Description Framework) 是一种用于描述 Web 资源的模型和语言, 其采用的查询语言 SPARQL 语言是一种类似 SQL 的查询语言。RDF 基于三元组 (triple) 的概念, 每个三元组包括主体 (subject)、谓词 (predicate) 和宾语 (object), 分别对应于资源、资源间关系的性质以及关系的目标或属性值。比如: 在知识图谱中查询麻黄的功效有哪些? 查询结果用三元组可以表示为: <麻黄, 功效, 发汗解表><麻黄, 功效, 宣肺平喘><麻黄, 功效, 利水消肿>。如图 2.2 是麻黄功效在知识图谱中三元组形式的展示:

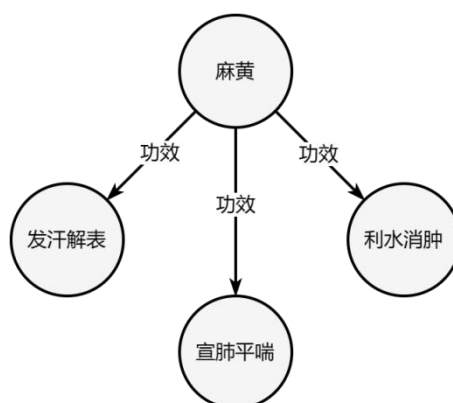


图 2.2 麻黄功效三元组效果图

2.1.3 知识图谱构建技术

知识图谱构建技术的目标是将现实世界中的知识以结构化的形式表示出来, 形成可查

询、可分析的知识库。构建知识图谱是一项涉及多个技术的复杂任务，主要包括以下几个核心步骤：数据采集：通过网络爬虫从互联网上抓取大量数据；利用现有的数据库获得结构化数据。实体识别（Named Entity Recognition, NER）：这一步骤涉及到识别文本中的具体实体，如人名、地点、组织机构名等。实体消歧：实体可能会有多重含义，实体消歧的任务是确定文本中提到的实体到底指的是哪一个，例如“苹果”可能指的是果实或是公司。关系抽取：识别实体间的关系，例如“马克·扎克伯格”和“Facebook”的关系是“创始人”。知识融合：将从不同来源抽取的知识合并在一起，解决知识来源的冗余和矛盾问题。知识存储：将提取的知识结构化存储在图数据库中，如 Neo4j，使其可以高效地进行查询和管理。知识推理：利用逻辑推理技术，根据现有的知识图谱推断出新的知识关系。知识更新与维护：随着外部数据的不断更新和变化，知识图谱也需要不断地进行更新和维护。这些技术不仅需要深厚的数据处理和自然语言处理技能，还需要利用到机器学习、图论等多个领域的方法。本研究将利用图数据库 Neo4j 的先进技术进行知识图谱的构建，构建的知识图谱以结构化的形式描述中药领域中的概念、实体及关系，将会提高搜索问答的准确性好效率。

2.1.4 知识图谱问答系统

基于知识图谱的智能问答系统通过结合知识图谱的丰富结构化信息和自然语言处理技术，提供精确的答案查询服务。该过程从查询解析开始，利用自然语言处理技术分析用户问题的语义，以识别关键实体和查询意图。接着，系统将解析后的问题映射到知识图谱中，构造并执行相应的图查询以检索相关的实体和关系。最后，通过自然语言生成（NLG）技术，根据查询结果生成准确、流畅的自然语言答案反馈给用户。

尽管如此，构建和运用这些系统仍面临一系列挑战。处理复杂或模糊的用户查询，需要更深入的语义理解，这可以通过采用先进的自然语言处理和大语言模型来部分实现。同时，保持知识图谱的覆盖性和时效性，确保系统的准确性和适用范围，是一项持续的挑战，这可以通过动态更新机制和采用开放域知识图谱来改善。通过面对这些挑战，基于知识图谱的智能问答系统能够不断进步，更好地满足用户需求。

在中药领域知识图谱的问答中，知识图谱问答系统能够提供中药功效、中药临床应用查询等服务。例如，系统能根据用户的输入的自然语言问题，通过查询中药医知识图谱使用自然语言回答用户的问题。

2.2 大语言模型

2.2.1 大语言模型概述

大型语言模型（LLMs），如 ChatGPT、BERT 等，代表了自然语言处理（NLP）领域的一项革命性进展。这些模型通过在广泛的文本数据上进行预训练，学会了理解和生成自然语言的能力。这种预训练过程使得模型能够捕捉到语言的深层语义结构和丰富的上下文信息，从而在多种 NLP 任务上展现出卓越的性能。

2.2.2 大语言模型的技术原理

大语言模型背后的核心技术是 Transformer 架构和自注意力机制。Transformer 是一种专为处理序列数据设计的深度学习模型，它由编码器（Encoder）和解码器（Decoder）两部分组成。编码器由一系列相同的层组成，每个层包括两个主要部分：多头自注意力机制（Multi-Head Attention）和前馈网络（Feed Forward Network）。每个子模块的输出都会与其输入相加（残差连接），然后进行层规范化（Normalization）。这种设计帮助 Transformer 模型缓解了深层网络训练中的梯度消失问题。解码器则包含三个主要部分：遮蔽多头自注意力机制（Masked Multi-Head Attention）、多头注意力机制（Multi-Head Attention）和前馈网络（Feed Forward Network）。

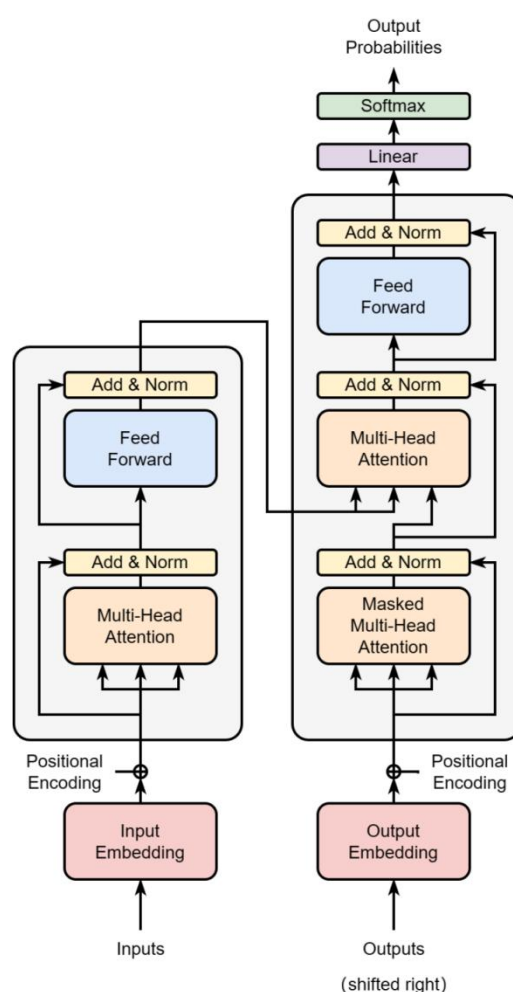


图 2.3 大语言模型工作流程图

整个模型的工作流程如图 2.3 所示, 首先, 输入序列首先通过词嵌入 (Input Embedding) 转换为固定维度的向量表达, 然后加入位置编码 (Positional Encoding) 提供序列中单词的位置信息, 编码器逐层处理这些嵌入, 构建一个上下文富含的输出, 解码器在编码器处理完所有输入后开始工作, 它接收编码器的输出和之前已生成的输出作为输入。解码器的每一层都生成当前步骤的输出, 并预测下一个单词, 直到序列结束。最终, 解码器输出通过一个线性层 (Linear) 和 Softmax 层转换成最终的概率分布, 用于预测下一个单词。整个模型是端到端训练的, 可以适应各种语言处理任务。

2.2.3 大语言模型的微调

微调 (Fine-tuning) 是一种在深度学习中常见的技术, 特别是在自然语言处理 (NLP) 任务中。它涉及到在一个已经预训练的大型模型 (如 ChatGPT-3.5 Turbo) 上进行额外训练, 以使其更好地适应特定的应用或任务。通过在特定任务的数据集上进行训练, 微调可以调整模型的权重, 从而提高其在该任务上的表现, 而无需从头开始训练一个全新的模型。这种方法既节省了时间和资源, 又能利用预训练模型在广泛数据上学习到的丰富知识。

2.2.4 Prompt 管理

Prompt 作为一种任务引导的机制, 其核心原理在于将模型面对的任务隐式地转化为一个或一系列文本生成问题。通过给定一段特定的文本 (Prompt), 模型根据这段文本生成一个续写, 这个续写被期望是对 Prompt 的回应或解答。这种方式利用了模型的语言生成能力, 让模型在没有显式任务指示的情况下完成特定的任务。通过设计合适的 Prompt, 大语言模型能够在没有或几乎没有特定任务训练样本的情况下, 完成该任务。这是因为模型在预训练过程中已经学习到了广泛的语言模式和知识, Prompt 使模型能够将这些知识应用于特定的任务上。大语言模型的另一显著特点是能够进行零样本 (Zero-shot) 或少样本 (Few-shot) 学习。通过设计合适的 Prompt, 模型能够在没有或几乎没有特定任务训练样本的情况下, 完成该任务。这是因为模型在预训练过程中已经学习到了广泛的语言模式和知识, Prompt 使模型能够将这些知识应用于特定的任务上。

2.2.5 大语言模型在智能问答中的作用

在智能问答系统中, 大语言模型展现了理解用户查询和生成准确回答的强大能力。通过预训练, 模型已经掌握了大量的世界知识和语言规则, 能够准确解析用户的问题, 并生成相关性强、信息丰富的回答。此外, 通过针对特定问答任务的微调, 模型可以进一步优化以更好地理解领域特定的术语和上下文, 从而在专业领域如中医药辅助诊疗中提供有效

支持。这种能力不仅提高了问答系统的效率，也极大地改善了用户体验，使得智能问答系统能够在医疗咨询、客户服务、在线教育等多个场景中发挥重要作用。

2.3 本章小结

本章深入探讨了本研究的两大核心技术：知识图谱（KG）和大语言模型（LLMs）。通过详细探讨知识图谱和大语言模型在本文中将会使用的技术原理，为后续研究融合大语言模型的知识图谱问答系统提供了坚实的理论基础。

第3章 中药领域知识图谱的构建

对于中药领域的知识图谱问答，知识图谱的构建具有重要意义。中药学是一门历史悠久的学科，拥有丰富而复杂的理论体系和大量的实践知识，包括药材、功效和临床应用等多个方面。传统上，这些知识以文本形式记录在各种古籍、医案、药典中，其表达形式多样，缺乏统一的结构化表示，给知识的检索、共享和应用带来了挑战。知识图谱通过构建实体之间的关系网络，将这些分散的、非结构化的中医药知识转化为结构化的、易于计算机处理的形式。这不仅有助于系统地组织和存储大量的中医药知识，还可以支持复杂的知识查询和智能推理，从而提高中医药知识的可访问性和可应用性。本研究的构建流程如图3.1所示：

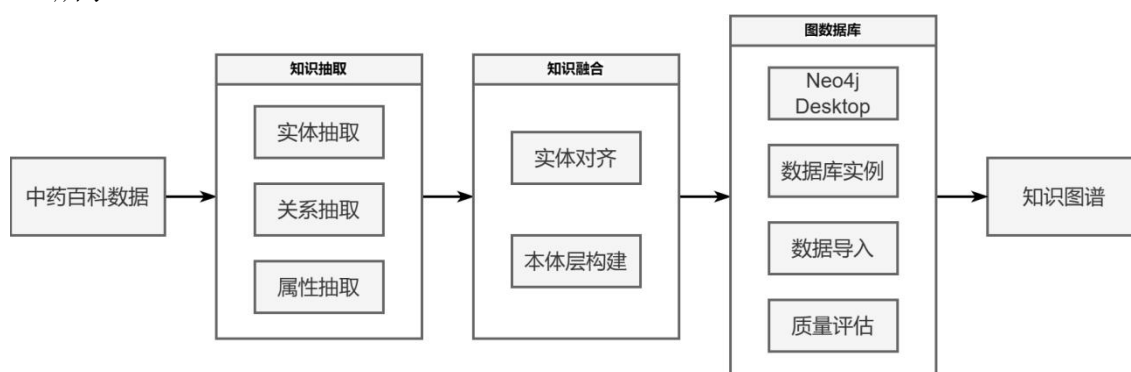


图 3.1 知识图谱构建流程图

3.1 数据准备

3.1.1 数据获取

在知识图谱中，数据通常被分为结构化数据、半结构化数据、非结构化数据。结构化数据指的是按照固定格式排列，便于存储、查询和分析的数据，如 CSV 文件中的数据，它们包含在严格定义的表格里，每个表都有固定的列和行。半结构化数据则没有结构化数据那么严格的格式，但仍包含一些组织属性，比如标签或其他标记，以区分数据的不同部分。这类数据可以包含与结构化数据相同的类型信息，但其格式更加灵活。XML 和 JSON 都是半结构化数据的示例。非结构化数据指的是没有预定义数据模型的数据，也就是没有清晰结构的数据。这类数据不易通过传统的数据库和数据模型进行管理和分析。非结构化数据的例子包括文本文件、图片、音频和视频。只有结构化数据可以直接录入知识图谱中，所以在构建知识图谱的过程中，通常需要将半结构化数据和非结构化数据进行知识抽取，即实

体抽取、关系抽取、属性抽取等，以便能够通过图谱中的节点和边关系加以表示和连接。

本研究所使用的中药数据来源于开放的中药查询网站（www.zhongyoo.com）。该平台特别丰富地涵盖了中医药领域的信息，包括中药种类、中药功效、中药性状等，为构建知识图谱提供了一个全面的数据基础。本研究通过编写爬虫代码的方式获取到了大量的中药数据，包括文本描述、相关图片和参考资料等。这种方式的数据获取既方便快捷，又确保了数据的合法性和可靠性。

通过百爬虫代码获取的中药数据，本研究经过编写代码以及人工整理将这些数据精心组织成表格形式，直接呈现了中药的基本属性，如药名、性味归经、功效和临床应用等，极大地便利了后续的数据处理和知识图谱构建工作。该数据具有如下优点：

细致的分类和归类：数据不仅包含了广泛的中草药实体，而且已经按照其药理作用和应用领域进行了细致的分类和归类。例如，解表药类别中详细列出了凡是用于发散表邪、解除表证为主要作用的药物，如麻黄、桂枝等，并对每种药物的主要作用进行了说明。

丰富的药物描述：每种中草药条目不仅提供了基本的药物信息，还深入解释了其性味、归经路径和具体的临床应用，这为深入研究中草药的治疗机理和应用提供了宝贵的信息源。此外，对药物的功效和应用场景的描述，加深了对中医药治疗原理的理解。

方便的数据格式：数据以表格形式提供，使得信息呈现清晰、易于阅读和分析。这种格式的数据直接支持了知识图谱构建过程中的实体识别、属性提取和关系构建，极大简化了从原始数据到知识图谱实体和关系映射的工作量。

3.1.2 数据预处理

虽然获取到的数据已经是半结构化数据，但是仍然存在一些混乱不能直接导入知识图谱中，需要进一步对数据进行处理。所以，在获取到的原始中医药数据集上，需要对数据进行一系列的数据处理，为知识图谱的构建打下坚实基础。

经由爬虫获取的中药表格数据包含丰富的信息，但也夹杂着一些多余或无关的内容。为此，首先进行了人工审核和清洗，手动删除了无用的列和行，如广告信息、重复条目等，同时整理表格格式，根据中药表格中的数据，首先将草药按中药种类进行分类，把不同种类的草药分成不同的表格，共分成了42个表格。其中41种草药的基本信息表将每一列分别设置为药名、性、毒性、味、归经、功效和临床应用。最后一个表格则是草药的分类信息表格。这一步骤虽费时费力，却是保证数据录入准确性的关键环节。如图3.2所示是数据预处理后的部分效果图：

	A	B	C	D	E	F	G
1	药名	性	毒性	味	归经	功效	临床应用
2	麻黄	温	无毒	辛、微苦	肺、膀胱	发汗解表、宣肺平喘、利水消肿	用于风寒表实证、用于哮喘实证、用于风水水肿
3	桂枝	温	无毒	辛、甘	肺、心、膀胱	发汗解肌、温经通脉、通阳化气	用于外感风寒表证、用于寒凝血滞的辨证、脘腹冷痛、痛经、经闭等证、用于胸痹、痰饮、水肿及心动悸、脉结代
4	紫苏	温	无毒	辛	肺、脾	发汗解表、行气宽中、解鱼蟹毒	用于外感风寒证、用于脾胃气滞证、用于食鱼蟹中毒
5	生姜	微温	无毒	辛	肺、胃、脾	发汗解表、温中止呕、温肺止咳	用于外感风寒表证、用于多种呕吐证、用于风寒咳嗽
6	香薷	微温	无毒	辛	肺、胃、脾	发汗解表、化湿和中、利水消肿	用于阴暑证、用于水肿
7	荆芥	微温	无毒	辛	肺、肝	祛风解表、透疹止痒止血	用于外感表证、用于麻疹透发不畅、风疹瘙痒、用于疮疡初起兼有表证、用于吐衄下血
8	防风	微温	无毒	辛、甘	膀胱、肝、脾	祛风解表、胜湿止痛止痉	用于外感表证、用于风寒湿痹证、用于破伤风
9	羌活	温	无毒	辛、苦	膀胱、肾	发汗解表、祛风止痛	用于外感风寒表证、用于风寒湿痹证
10	独活	温	无毒	辛	膀胱、肝	祛风散寒、胜湿止痛	用于外感风寒、颈项头痛、用于风寒湿痹
11	白芷	温	无毒	辛	肺、胃	祛风散寒、通窍止痛、消肿排脓、燥湿止带	用于风寒感冒、头痛、牙痛、用于鼻塞、鼻渊、用于疮疡肿毒、用于寒湿带下
12	细辛	温	小毒	辛	肺、肾、心	祛风解表、散寒止痛、温肺化饮、通窍	用于外感风寒及阳虚外感证、用于头痛、痹痛、牙痛等痛证、用于寒饮咳嗽
13	苍耳子	温	小毒	辛、苦	肺	祛风解表、宣通鼻窍、除湿止痛	用于风寒表证及鼻渊、用于痹证
14	葱白	温	无毒	辛	肺、胃	发汗解表、散寒通阳	用于外感风寒表证、用于阴盛格阳证
15	胡荽	温	无毒	辛	肺、胃	解表透疹、健胃消食	用于麻疹透发不畅、用于胃寒食滞
16	柝柳	平	无毒	辛	肺、胃、心	解表透疹、祛风除湿	用于麻疹透发不畅、用于风寒湿痹
17	辛夷	温	无毒	辛	肺、胃	发汗解表、宣通鼻窍	用于风寒头痛鼻塞、用于鼻渊头痛
18	鹅不食草	温	无毒	辛	肺、肝	祛风散寒、宣通鼻窍、化痰止咳	用于风寒头痛及鼻渊鼻塞、用于湿疮肿毒、用于寒痰咳嗽证
19							
20							
21							
22							
23							
24							
25							
26							
27							

图 3.2 数据预处理部分效果展示图

3.2 知识抽取

在数据预处理之后，接下来的步骤是知识抽取，这一环节对于构建中药领域知识图谱至关重要。知识抽取是从半结构化的数据中识别和提取有价值的信息，转换为知识图谱可以使用的形式。本研究的知识抽取阶段分为三个主要任务：实体抽取、关系抽取和属性抽取。这些任务的目标是构建出反映中药学领域复杂网络的节点和边，即图谱中的实体和关系。以下是知识抽取的详细步骤和方法。在知识抽取过程中使用的实验环境如表 3.1 所示：

表 3.1 数据处理实验环境

环境	具体配置
Pycharm	2021.1.1
Python	3.9.7
虚拟环境	Conda
包管理	pip
Pandas	1.2.4

3.2.1 实体抽取

实体抽取是指从数据集中识别出可以独立存在的信息单元。在本研究中，实体抽取首先需要定义中药领域中的核心实体类别，然后通过编程方法从表格中抽取这些实体。利用 Python 和 Pandas 库，对每个已经分类的表格进行处理，将每个表中每一个数据作为一个独立的实体。通过这种方式，实体抽取涉及从 41 个草药的基本信息表中识别草药实体及

其性味归经、功效和临床应用等。这些数据提供了定义实体特性的关键信息，能够快速准确地从大量数据中识别出所需的实体。此外，还有一个单独的草药种类信息表格，其中的每个类名也构成了一个单独的实体。这些类名实体通过其描述性属性提供了对各个草药分类的概览和解释。

3.2.2 关系抽取

关系抽取涉及定义实体之间的语义联系。在中药领域的知识图谱中，例如，草药实体与其性质、毒性、味道、归经路径、功效和临床应用之间的关系被细致地标识和提取。通过编程脚本分析和处理数据，这些关系被自动提取出来，并格式化为可供图数据库使用的结构。这一步不仅确保了数据的精确性，也为知识图谱的丰富性和查询效率打下了基础。

3.2.3 属性抽取

属性抽取是知识抽取的最后一步，旨在为实体和关系附加详细信息。在本研究中，每个草药实体拥有名称和分类两个属性，指向其所属的具体草药类别。这些属性为知识图谱提供了丰富的语义信息，增加了知识图谱的查询能力。通过对预处理后的数据进行进一步的分析，每个实体和关系的属性被准确抽取，并在后续步骤中转化为图数据库的属性值。

在完成以上知识抽取之后，这些数据被整理和输出成为 CSV 格式的结构化数据，便于直接录入知识图谱。这种结构化的数据格式不仅方便数据的管理和更新，也优化了数据查询和后续分析的效率。如图 3.3 所示生成的结构化文件效果图：

药名	性	毒性	味	归经	功效	临床应用	
麻黄	温	无毒	辛	微苦	肺、膀胱	发汗解表、宣肺平喘、利水消肿,用于风寒表实证、用于咳嗽实证、用于风水水肿	
桂枝	温	无毒	辛	甘	肺、心、膀胱	发汗解肌、温经通脉、通阳化气,用于外感风寒表证、用于寒凝血滞的痹证、腕腹冷痛、痛经、经闭等证、用于胸痹、痰饮、水肿及心动悸、脉结代*	
紫苏	温	无毒	辛	肺、脾	发汗解表、行气宽中、解鱼蟹毒	用于外感风寒证、用于脾胃气滞证、用于食鱼蟹中毒	
生姜	微温	无毒	辛	肺、脾	发汗解表、温中止呕、温肺止咳	用于外感风寒表证、用于多种呕吐证、用于风寒咳嗽	
香薷	微温	无毒	辛	肺、胃、脾	发汗解表、化湿和中、利水消肿	用于阴暑证、用于水肿	
荆芥	微温	无毒	辛	肺、肝	祛风解表、透疹止痒止血	用于外感表证、用于麻疹透发不畅、风疹瘙痒、用于疮疡初起非有表证、用于吐衄下血	
防风	微温	无毒	辛	甘	膀胱、肝、脾	祛风解表、胜湿止痛止痉	用于外感表证、用于风寒湿痹证、用于破伤风
羌活	温	无毒	辛	苦	膀胱、肾	发汗散风寒、胜湿止痛	用于外感风寒表证、用于风寒湿痹证
独活	温	无毒	辛	膀胱、肝	祛风散寒、胜湿止痛	用于外感风寒、用于风寒湿痹	
白芷	温	无毒	辛	肺、胃	祛风散寒、通窍止痛、消肿排脓、燥湿止带	用于风寒感冒、头痛、牙痛、用于鼻塞、鼻渊、用于疮疡肿毒、用于寒湿带下	
细辛	温	小毒	辛	肺、肾、心	祛风解表、散寒止痛、温肺化饮、通窍	用于外感风寒及阳虚外感证、用于头痛、痹痛、牙痛等痛证、用于寒饮咳嗽	
苍耳子	温	小毒	辛	苦	肺	祛风解表、宣通鼻窍、除湿止痛	用于风寒表证及鼻渊、用于痹证
葱白	温	无毒	辛	肺、胃	发汗解表、散寒通阳	用于外感风寒表证、用于阴盛格阳证	
胡荽	温	无毒	辛	肺、胃	解表透疹、健胃消食	用于麻疹透发不畅、用于胃寒食滞	
柃柳	平	无毒	辛	肺、胃、心	解表透疹、祛风除湿	用于麻疹透发不畅、用于风寒湿痹	
辛夷	温	无毒	辛	肺、胃	发汗散风寒、宣通鼻窍	用于风寒头痛鼻塞、用于鼻渊头痛	
鹅不食草	温	无毒	辛	肺、肝	祛风散寒、宣通鼻窍、化痰止咳	用于风寒头痛及鼻渊鼻塞、用于湿疮肿毒、用于寒痰咳嗽证	

行 12, 列 68

922 个字符

100%

Windows (CRLF)

UTF-8

图 3.3 结构化文件效果展示图

3.3 知识融合

知识融合,通常理解为将来自多个数据源的信息整合到一个统一的表示系统中的过程,在构建知识图谱的背景下尤其关键。它涉及同步、合并和协调不同数据集中的信息,以创建一个全面、一致和易于访问的知识库。在本研究中,知识融合是对两个结构化的中药数据集进行整合。这两种数据集来源于相同的知识体系但被存储在不同的结构化文件中,如草药的基本信息表和草药的分类信息表。这种融合的目的是确保所有相关的信息能够被系统地组织在一起,从而提供一个完整的、互联的知识视图,支持更复杂的查询和数据分析。

3.3.1 实体对齐

实体对齐是知识融合过程中的一个关键步骤,特别是当涉及到将中药的基本信息表与种类信息表整合时。在本研究中,实体对齐不仅确保了数据的一致性,也为中药领域的知识图谱构建提供了一个准确的信息基础。本节将详细介绍如何利用 Neo4j 数据库的 Cypher 查询语言中的 MERGE 语句来实现这一过程。

在中药领域,每种草药都可以有多个属性描述,如药名、功效、用途等,这些信息通常记录在基本信息表中。同时,草药还可以归属于不同的分类,如按照药效分类,这类信息则记录在种类信息表中。在不同的表中,同一名称的草药会出现多次,在录入草药实体时会出现重复,以及不同草药实体对应的性味归经、功效和临床应用会出现相同,这时创建多个实体会造成数据冗余并且影响查询效率,所以实体对齐是至关重要的。

在 Neo4j 中, Cypher 语言的 MERGE 语句可以确保一个节点或关系在图中只被创建一次。如果该节点或关系已经存在,则 MERGE 将匹配这个现有的节点或关系,否则将创建一个新的。通过 MERGE,可以确保不会因为来自两个不同表格的信息而重复创建同一个草药节点。该方法能够有效地将中药的基本信息表和种类信息表中的数据融合。这不仅提高了数据的准确性,也为后续的知识图谱查询和分析提供了坚实的基础。正确的实体对齐策略确保了知识图谱的高质量 and 可用性,是构建有效知识图谱的关键步骤。

3.3.2 本体层构建

在成功实现了中药的基本信息表和种类信息表之间的实体对齐之后,下一步是构建本体层。本体层的构建是知识图谱建设中的核心环节,它不仅为整个图谱提供了结构化的知识表示,还确保了数据与查询的语义一致性。本体描述了领域内实体之间的各种可能关系以及实体属性的种类,为描述复杂的领域知识提供了一种形式化的方法。在中药领域的知识图谱中,本体层的构建不仅帮助明确各种中药的分类、属性和用途,还规定了这些元素之间应如何相互作用和连接。

本体层构建对于准确捕捉和表述中医药知识至关重要。在本研究中,构建知识图谱的努力侧重于从经过仔细处理的数据集中提炼出的丰富的实体和关系类型。这一过程而是通过充分利用图数据的表现力,更加精确和规范地描述中药知识,以增强数据的可理解性和可用性。通过这种方法,研究确保所构建的知识图谱能够在保持数据真实性的同时,提供对中医药传统知识的现代表达,进而支持更高级别的查询和智能化应用。

为了确保本研究构建的中药领域知识图谱能够有效支持问答系统,本研究根据先前经过精心处理的结构化数据,细致构建了如表 3.2 所示的实体层。

表 3.2 实体层设计表

实体	含义	示例
Medicine (草药)	代表每种草药的节点	麻黄、桂枝、紫苏等
Property (性)	代表草药的性质	温、寒、平等
Taste (味)	代表草药的气味	辛、甘、苦等
Toxicity (毒性)	代表草药是否具有毒性	无毒、小毒等
Meridian (归经)	代表草药作用的所属和定位	肺、胃、脾等
Effect (功效)	代表草药的预期治疗效果	发汗解表、宣肺平喘等
Application (临床应用)	代表草药可以用于的证候	用于风寒表实证等
Sort (药类型名)	代表草药所属的种类	解表药、清热药等

在确立实体层之后,知识图谱的完整性依赖于实体间精确定义的关联性。这些关系不单补全了实体的信息维度,而且铺展了知识图谱的联结网,极大地增强了问答系统在检索和查询过程中的效能。因此,本研究精心构筑了如表 3.3 所示的关系层:

表 3.3 关系层设计表

关系	实体-关系-实体
HAS_PROPERTY	(Medicine)-[:HAS_PROPERTY]->(Property)
HAS_TASTE	(Medicine)-[:HAS_TASTE]->(Taste)
HAS_TOXICITY	(Medicine)-[:HAS_TOXICITY]->(Toxicity)
HAS_MERIDIAN	(Medicine)-[:HAS_MERIDIAN]->(Meridian)
HAS_EFFECT	(Medicine)-[:HAS_EFFECT]->(Effect)
HAS_APPLICATION	(Medicine)-[:HAS_APPLICATION]->(Application)
BELONGS_TO	(Medicine)-[:BELONGS_TO]->(Sort)

在设计完成实体层和关系层后,本知识图谱还对实体的属性层进行了设计,其中设计的属性层以及含义如表 3.4 所示:

表 3.4 属性层设计表

属性	含义	实体	示例
id	唯一标识符	ALL	n:Medicine {id: 43}
name	该实体的名称	ALL	n:Medicine {name: '白芷'}
introduce	介绍草药分类的依据	Sort	n:Sort {introduce: '气味芳香,性偏温燥,化湿运脾'}
type	草药类别, Sort 的更细致分类	Medicine	n:Medicine {name: '发散风寒药'}

通过精准定义这些实体、关系和属性，所构建的知识图谱得以准确地捕捉和体现中药知识的深度和广度。这种结构化的设计不仅提高了知识检索的效率，而且显著增强了问答系统在提供精确回答方面的能力。

3.4 知识图谱构建

3.4.1 数据模型

数据模型是知识图谱构建的理论基础，它定义了数据的组织方式，包括实体之间的关系以及实体的属性。知识图谱中常见的数据模型为三元组和图数据库，相对于三元组来说，图数据库是更适合本研究的数据模型。三元组与图数据库的特性对比如表 3.5 所示：

表 3.5 三元组与图数据库的特性对比表

特性	图数据库	三元组
数据模型	直观的图形模型，节点表示实体，边表示关系，便于设计和实现复杂的查询	基于主体-谓词-对象的三元组，强调资源之间的语义关系
性能优化	对连接查询和深度遍历操作进行了优化，执行复杂的图查询时更加高效	需要对大量的三元组数据进行匹配和联接操，在大规模数据集上会影响查询性能
查询语言	提供专门的图查询语言，使得表达和执行图查询更为方便和强大	SPARQL 查询语言学习曲线相对陡峭，编写和优化查询更加复杂
可扩展性	支持高效的图算法和复杂关系分析	适合于需要严格遵循 RDF 标准、与语义网技术栈集成的应用

综上，将采用图数据库作为数据模型，其原因是：

直观的数据表示：图数据库通过节点、边和属性直观地表示实体及其关系，这使得数据模型更加符合自然世界的结构。

高效的关系查询：图数据库专门为处理高度连接的数据而设计，非常适合存储复杂的网络关系，如中药之间的相互作用、药材和疾病之间的关联等。它们能够高效地执行深层次的关系查询和模式匹配，这在关系数据库中可能效率较低甚至难以实现。

强大的查询语言：图数据库通常提供专门的图查询语言，这些查询语言专为图数据设计，使得表达和执行图查询更为方便和强大。

灵活的数据模型：中药领域的数据通常是高度多样化和复杂的，图数据库允许灵活的数据模型，可以随时添加新的关系或属性，而不需要事先定义固定的模式。这种灵活性使得图数据库非常适合动态变化的知识图谱。

为了实现上述数据模型，本研究选择了 Neo4j 作为图数据库的实现工具。Neo4j 是一个高性能的图形数据库管理系统，它专为存储和处理大量复杂的关系网络而设计。比较 Neo4j 与其他图数据库，如：JanusGraph、ArangoDB、Amazon Neptune 和 Azure Cosmos DB，

它们相对于 Neo4j 的潜在缺点如表 3.6 所示:

表 3.6 其他数据库相比 Neo4j 缺点表

图数据库	缺点
JanusGraph	查询语言 Gremlin 相对复杂, 社区和生态系统虽不如 Neo4j 成熟和广泛
ArangoDB	查询语言 AQL 在图查询上不如 Cypher 直观, 图数据处理和查询优化方面相对较差
Amazon Neptune	作为托管服务, 导致更高的成本和较低的灵活性
Azure Cosmos DB	在全球分布式和多模型数据管理方面表现出色, 在图数据查询和分析能力上相对较差

综上, 以下是选择 Neo4j 的主要理由:

成熟稳定: Neo4j 是最流行和成熟的图数据库之一, 具有广泛的社区支持和丰富的文档资源。它的稳定性和成熟的功能可以为中药领域知识图谱问答系统提供坚实的技术基础。

性能优势: Neo4j 针对图数据的存储和查询进行了高度优化, 特别是在处理复杂查询和大数据量时, 能够提供出色的性能和响应时间。

查询语言: Cypher 是 Neo4j 特有的查询语言, 设计直观且易于学习。Cypher 查询语言为表达和执行复杂的图查询提供了巨大的便利, 使得从复杂关系中提取信息变得简单高效。

社区支持: 作为最流行的图数据库之一, Neo4j 拥有一个庞大且活跃的开发社区, 提供了大量的学习资源、工具和插件。

生态系统: Neo4j 的生态系统包含了数据库服务器、客户端库、管理工具和数据可视化工具等, 全面支持了从数据建模到部署的完整工作流程。

图数据模型的直观和灵活以及 Neo4j 数据库的强大功能使其成为构建中医药知识图谱的理想选择, 为中医药知识的深入分析和应用提供了坚实的技术支持。能够为中医药知识图谱构建提供一个健壮和可扩展的基础架构, 满足问答系统等应用的需求。

3.4.2 图数据库管理

本研究选取的图数据库 Neo4j 提供了多种数据管理工具, 这些工具可以进行数据库的查询、监控、管理和可视化。Neo4j 的主要工具如表 3.7 所示:

表 3.7 关系类型设计表

工具	类型	描述
Neo4j Browser	基于 Web 的图形界面	执行 Cypher 查询语句、可视化查询结果, 并查看数据库的图形表示
Neo4j Bloom	可视化工具	通过搜索、浏览和查询来发现数据之间的关系, 无需编写 Cypher 查询
Neo4j Cypher Shell	命令行界面 (CLI) 工具	执行 Cypher 查询语言命令和管理 Neo4j 数据库
Neo4j Desktop	桌面应用	本地或远程管理 Neo4j 数据库实例, 安装和管理多个 Neo4j 版本和插件
Neo4j Admin	命令行工具	执行数据库的低级管理和维护操作

对于本研究项目的需求,选择的工具为 Neo4j Desktop。Neo4j Desktop 是一个为本地开发环境量身打造的桌面应用程序,它的灵活性和易用性使其成为开发、测试和学习图数据库的理想选择。它提供了一个便捷的界面,可以在本地环境中自由实验,轻松创建和管理多个数据库实例,这为复杂查询和数据操作提供了一个安全和隔离的环境。此外,Neo4j Desktop 的集成工具和插件系统支持了与多种开发工具的无缝对接,如数 Neo4j Browser、Neo4j Bloom 等工具,大大增强了其在学术研究和应用开发中的实用性。

在选择了 Neo4j Desktop 作为数据库管理工具之后,接下来首先要做的是创建数据库实例,通过 Neo4j Desktop,新建数据库实例并进行基本设置。此环节包括分配资源、配置参数等,以确保数据库实例满足项目需求,并为后续的数据导入和操作提供良好基础。下图 3.4 是创建好的数据库示例:

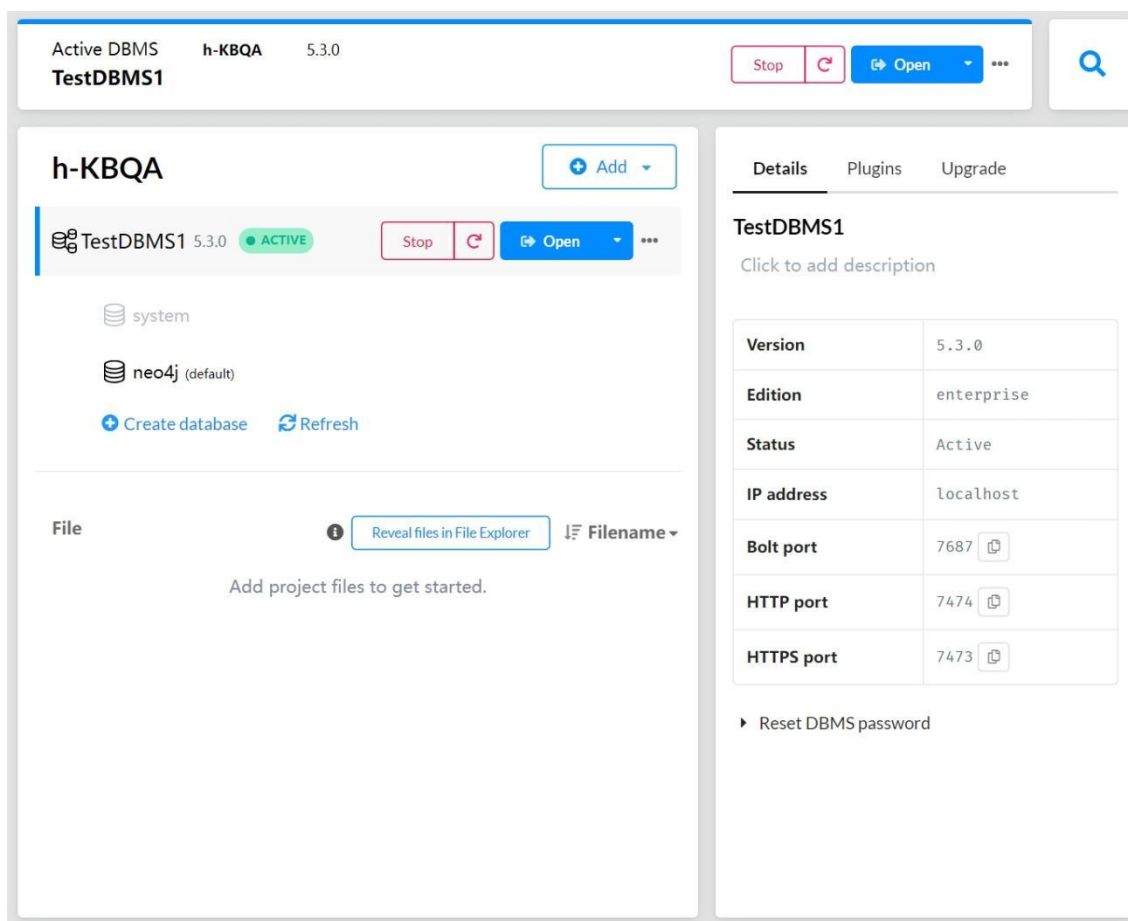


图 3.4 本项目 Neo4j 图数据库示例

3.4.3 数据导入

数据导入工作将在 Neo4j Desktop 集成的 Neo4j Browser 工具中操作。Neo4j Browser 是 Neo4j 数据库的官方 Web 界面,提供了一个用户友好的图形界面 (GUI) 用于直接与 Neo4j 图数据库进行交互。Neo4j Browser 允许用户直接在界面中输入并执行 Cypher 查询语

句，快速获取查询结果。这对于开发和测试 Cypher 查询非常有用。并且 Neo4j Browser 的查询结果可以以图形的形式展示，节点和关系会根据其类型和连接自动布局。这使得理解复杂的图结构变得直观简单。

1) 准备导入文件：对于大规模数据导入，Neo4j 提供了 LOAD CSV 命令，可以从 CSV 文件中读取数据并导入 Neo4j 中，为数据准备提供了便利。在导入前，需确保 CSV 文件已根据 Neo4j 的要求进行格式化并存放在指定目录下。此步骤旨在保证数据格式的标准化，以便顺利映射到预期的数据库结构中。图 3.5 所示是 Neo4j 指定的 CSV 文件导入文件夹：



anshen-1.csv	CSV 文件	2 KB
anshen-2.csv	CSV 文件	1 KB
badu.csv	CSV 文件	1 KB
buxu-1.csv	CSV 文件	4 KB
buxu-2.csv	CSV 文件	5 KB
buxu-3.csv	CSV 文件	2 KB
buxu-4.csv	CSV 文件	4 KB
huashi.csv	CSV 文件	2 KB
huatan-1.csv	CSV 文件	2 KB
huatan-2.csv	CSV 文件	3 KB
huatan-3.csv	CSV 文件	2 KB
huoxue.csv	CSV 文件	5 KB
jiebiao-1.csv	CSV 文件	3 KB
jiebiao-2.csv	CSV 文件	3 KB
kaiqiao.csv	CSV 文件	2 KB
liqi.csv	CSV 文件	3 KB
lishui-1.csv	CSV 文件	2 KB
lishui-2.csv	CSV 文件	2 KB
lishui-3.csv	CSV 文件	2 KB
pinggan-1.csv	CSV 文件	2 KB
pinggan-2.csv	CSV 文件	2 KB
qingre-1.csv	CSV 文件	3 KB
qingre-2.csv	CSV 文件	2 KB

图 3.5 CSV 文件展示图

2) Cypher 查询语言导入：通过 Neo4j Browser 工具，执行 Cypher 查询语言编写的数据库导入命令。Cypher 语句的编写涉及：定义节点创建命令，为每个实体及其属性指定格式。建立关系及属性，确保实体间的逻辑连接正确反映。精心构建的 Cypher 语句不仅关系到数

据的准确导入，还影响后续查询的效率和结果的准确性。图 3.6 所示是部分 CSV 文件导入成功图：

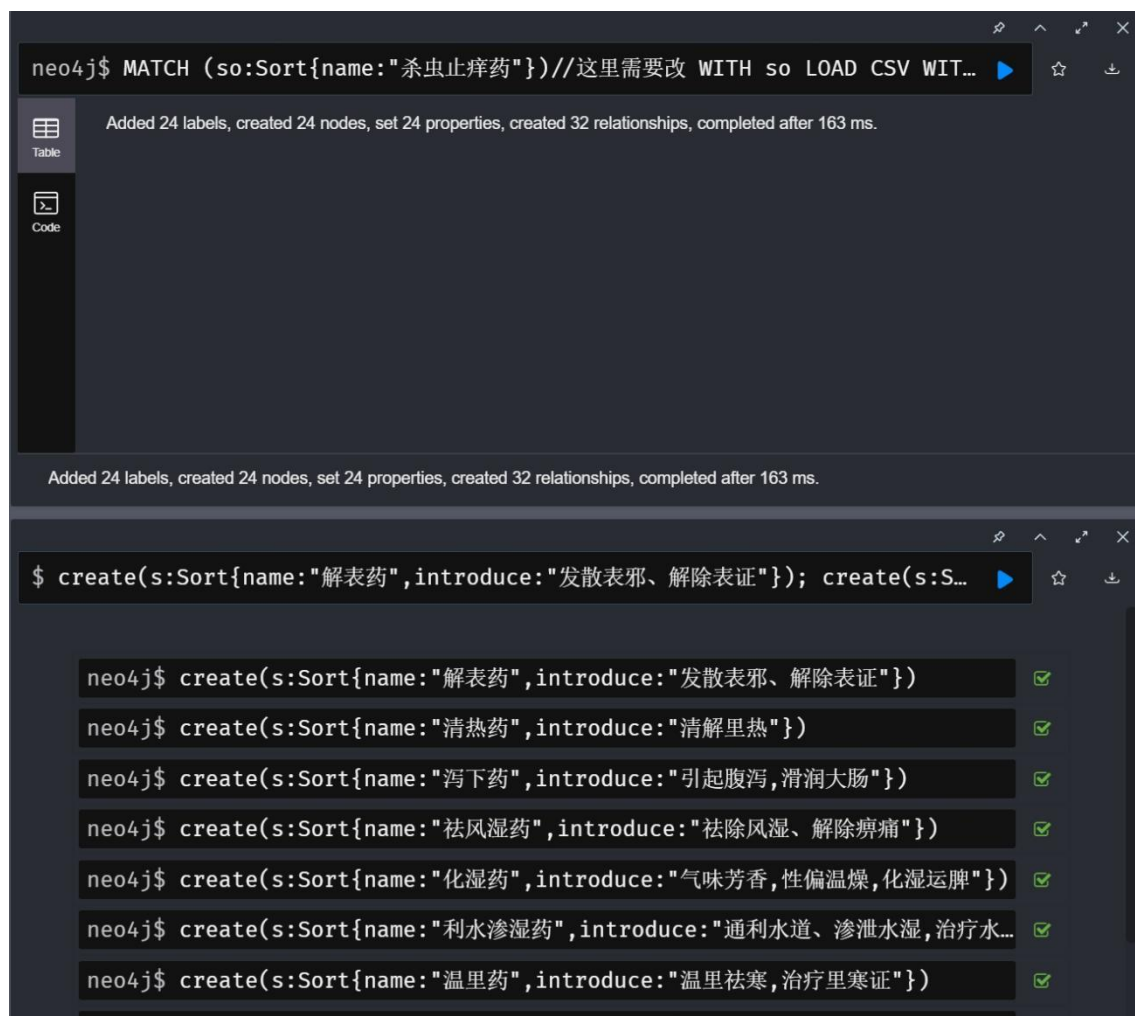
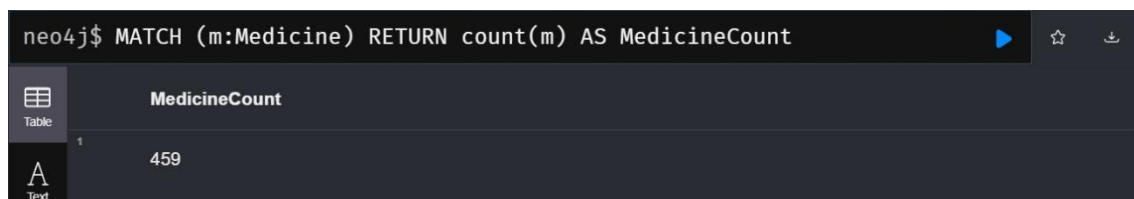


图 3.6 部分 CSV 文件导入成功图

通过选用合适的工具和细致的方法，本研究顺利完成了数据导入环节，为构建中医药知识图谱奠定了坚实基础。这一过程不仅关键在于知识图谱的构建，也为后续的分析和应用提供了价值数据，确保了问答系统等应用的高效和准确。

3.4.4 质量评估

数据成功导入后，通过执行系列 Cypher 查询进行数据完整性的校验。此步骤包括：确认节点和关系是否完整创建。校验实体属性和关系属性的准确性。测试典型查询案例，以评估数据的可用性和查询性能。通过这些验证操作，可以确保知识图谱的数据基础既稳固又可靠，为深入分析和应用提供坚实支撑。如图 3.7 所示是使用 Cypher 查询语句验证查询后得到的结果：



The screenshot shows a Neo4j query interface with the command `neo4j$ MATCH (m:Medicine) RETURN count(m) AS MedicineCount`. The result is displayed in a table with one row showing a count of 459.

MedicineCount	
1	459

图 3.7 验证查询部分结果图

3.5 知识图谱展示

在完成 Neo4j 图数据库中数据的录入之后, 本研究成功构建了一个功能强大的图数据库, 这是整个中医药辅助诊疗系统的数据基础和核心。通过精心的设计与实施, 该图数据库集成了中医药领域的丰富知识与信息, 形成了一个包含 44,112 个节点和 291,165 条关系的庞大知识网络。这些节点代表了中药领域内的各种草药的名称、性味归经和功效等, 而这些实体之间的关系则反映了它们之间的相互作用和联系, 如草药与其功效之间的关联。构建这样一个知识图谱, 不仅为系统提供了一个全面、细致的数据支撑, 也极大地增强了系统解析自然语言查询、执行精确数据检索的能力。通过这个知识图谱, 系统能够深入理解用户查询的意图, 快速而准确地检索到相关的中医药知识, 从而为用户提供科学、准确的中药知识解答。图 3.8 是图数据库中解表药所包含草药种类的展示图:

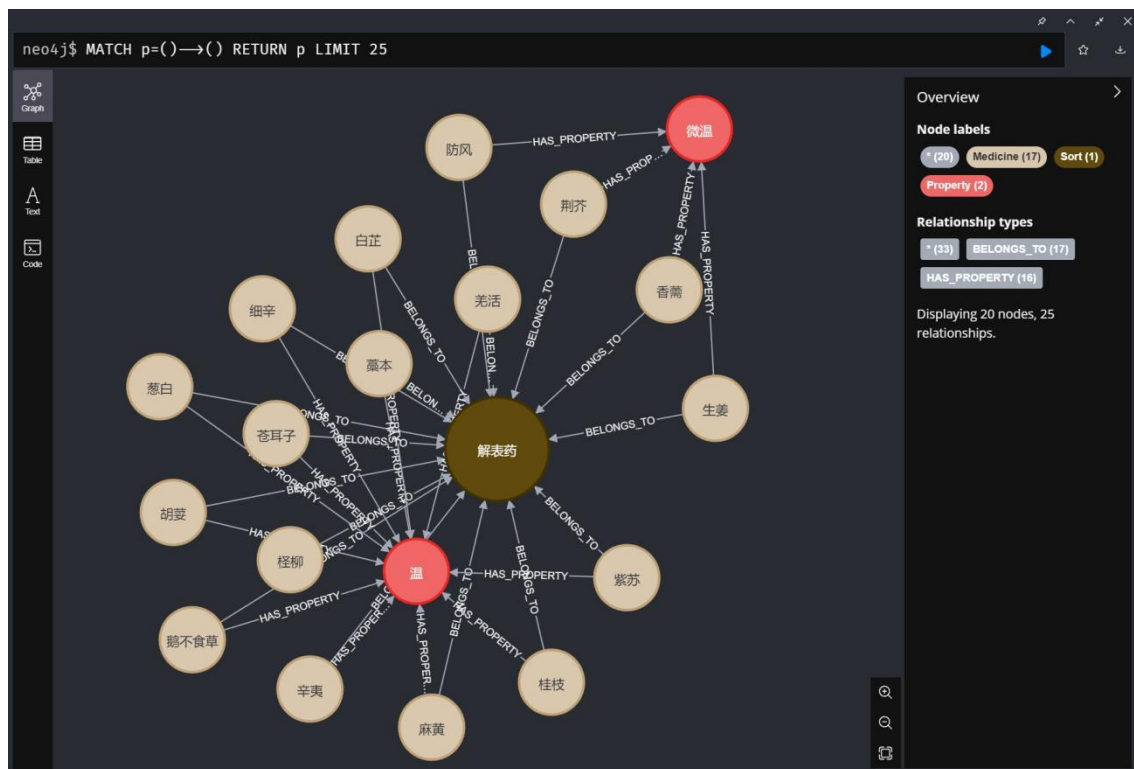


图 3.8 部分 Neo4j 数据展示图

3.6 本章小结

本章着重介绍了中医药知识图谱的构建过程，从数据源的获取与处理到知识图谱的构建，每一步都经过了精心设计和实施。利用爬虫技术获取丰富而精确的中医药数据，成功转化并导入了这些数据至 Neo4j 图数据库中，建立了一个包含 44,112 个节点和 291,165 条关系的庞大知识网络。这一过程不仅展示了中医药知识的深度整合与结构化表达，而且强调了知识图谱在促进中药知识检索、共享和应用中的重要作用。本章的内容体现了在处理大规模、复杂知识体系时，知识图谱技术的强大能力与潜在价值，为后续的系统实现和应用提供了坚实的数据支持和理论基础。

第4章 大语言模型的优化

本研究旨在提出一种创新的方法，通过借助大语言模型强大的语义理解能力来增强问句解析过程，旨在解决传统知识图谱问答系统在该环节的限制。该方法的核心在于，使用之前构建的中药领域知识图谱构建合适的问答数据集，对大语言模型进行微调，使之适应特定领域的需求。通过微调可以进一步优化模型，让它更精确地解析针对特定知识图谱库的自然语言问句，并生成高度适配的 Cypher 查询语句。通过这一过程，目标是优化模型的能力，使其在面对与该知识图谱库相关的查询时，能够准确识别出关键信息，并根据这些信息构造出正确的查询命令，从而在实际应用中提供更为准确和可靠的信息检索结果。

4.1 模型选择

在选择合适的大语言模型方面，本研究通过查询 OpenAI 官方文档发现，目前 OpenAI 提供了具有各种能力的模型，可以根据需求选择对应的模型进行微调，具体模型如表 4.1 所示：

表 4.1 OpenAI 提供的模型

模型	描述
GPT-3.5 Turbo	一组从 GPT-3.5 升级后的模型，能够生成自然语言和代码
DALL·E	能够根据自然语言提示词生成和编辑图片的模型
TTS	一组可以将文本转换成自然语音语言的模型
Whisper	可以将语音转换成文本的模型
Embeddings	一组可以将文件转换成数字形式的模型
Moderation	检测文本是否敏感或安全的审核模型
GPT base	一组没有指令的情况下也能生成自然语言和代码的模型

通过对比发现 ChatGPT-3.5 Turbo 继承了原有大语言模型的强大语言理解与生成能力，能够通过特别优化的微调功能，满足个性化需求。这种优化使得该模型能够根据特定需求进行细致的调整，精确地处理和响应特定查询。由于本研究将把微调后的模型用作问句解析，即根据自然语言问句生成相应的 Cypher 查询语句，所以选定了 ChatGPT-3.5 Turbo 模型。

此外，ChatGPT-3.5 Turbo 的选择也基于其提供的高质量 API 接入能力，极大地简化了模型在各种应用场景中的集成流程，允许其无缝嵌入到问句解析的过程中。该模型还在响应速度和运行效率方面进行了优化，这对于需要实时处理用户查询并生成准确答案的应用场景至关重要。通过 API 调用，ChatGPT-3.5 Turbo 支持快速的交互式体验，能够深入理解

用户的自然语言查询，为提升智能问答系统的整体性能和用户体验提供了坚实基础。

4.2 数据集构建

为了优化大语言模型以便其输出的 Cypher 查询语句更贴合特定的知识图谱库，本研究精心设计了一个根据本研究构建的知识图谱定制化的数据集，包括训练集和测试集。该数据集专注于包含与知识库直接相关的自然语言问句及其相应的 Cypher 查询语句，目的在于支持模型精确解析这些问句，并生成有效的查询命令。在构建数据集的过程中，重点考虑了能够显著代表知识库内容的问句，为它们匹配了准确的 Cypher 语句，从而确保查询语句能够准确反映出查询的意图。通过这种方式微调后的模型，旨在提高其针对该知识图谱库生成查询语句的准确度和效率。

4.2.1 问答对构建

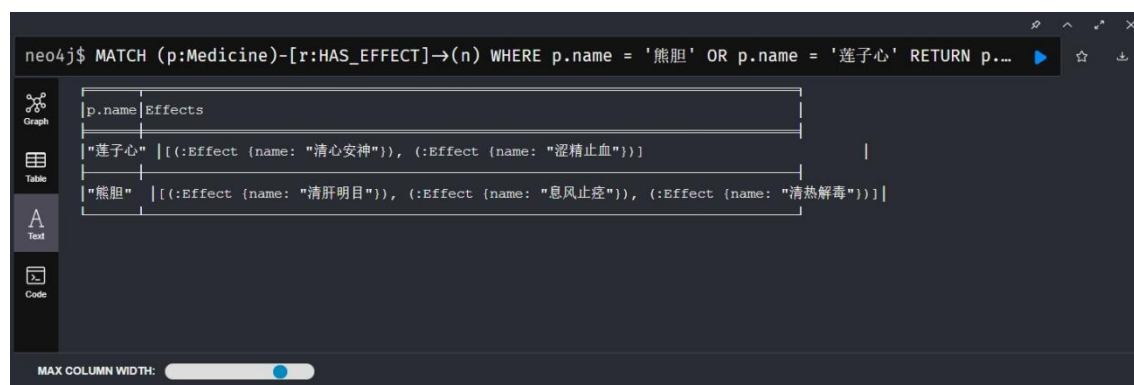
在这一阶段，通过分析知识图谱库的结构和内容，确定了与知识图谱库直接相关的查询需求。这包括诸如草药的属性、功效、临床应用等方面的查询，这些均是知识图谱库中的重要组成部分。基于这些需求，收集和编制了一系列与知识库对应的自然语言问句。这些问句旨在精确模拟出用户可能基于该知识库提出的实际查询，确保了数据集的实用性和目标的准确性。

对于每个自然语言问句，手动编写相应的 Cypher 查询语句。这一步骤紧密结合之前构建的知识图谱，确保每个查询语句都能准确地反映问句的意图，并能有效地从知识图谱中检索到相关信息。表 4.2 是编写的部分问答数据集：

表 4.2 问答数据集

问句	Cypher 查询语句
你知道解表药有哪些吗？	<pre>MATCH p()-[r:BELONGS_TO]->(n:Sort {name:'解表药'}) RETURN p</pre>
我现在有两种草药：熊胆和莲子心，想知道他们的功效是什么？	<pre>MATCH (p:Medicine)-[r:HAS_EFFECT]->(n) WHERE p.name = '熊胆' OR p.name = '莲子心' RETURN p.name, collect(n) AS Effects</pre>
最近有些上火，哪些药可以清热解毒？	<pre>MATCH (p:Medicine)-[r:HAS_EFFECT]->(n:Effect {name:'清热 解毒'}) RETURN p</pre>
心神不宁怎么办？	<pre>MATCH (m:Medicine)-[r:HAS_APPLICATION]->(e:Application) WHERE e.name CONTAINS '心神不宁' RETURN m</pre>
甘酸无毒的药有哪些？	<pre>MATCH (m:Medicine)-[r:HAS_TOXICITY]->(t:Toxicity {name:' 无毒'}),(m)-[r:HAS_TASTE]->(e1:Taste {name:'甘 酸'}),(m)-[r:HAS_TASTE]->(e2:Taste {name:'酸'}) RETURN m</pre>

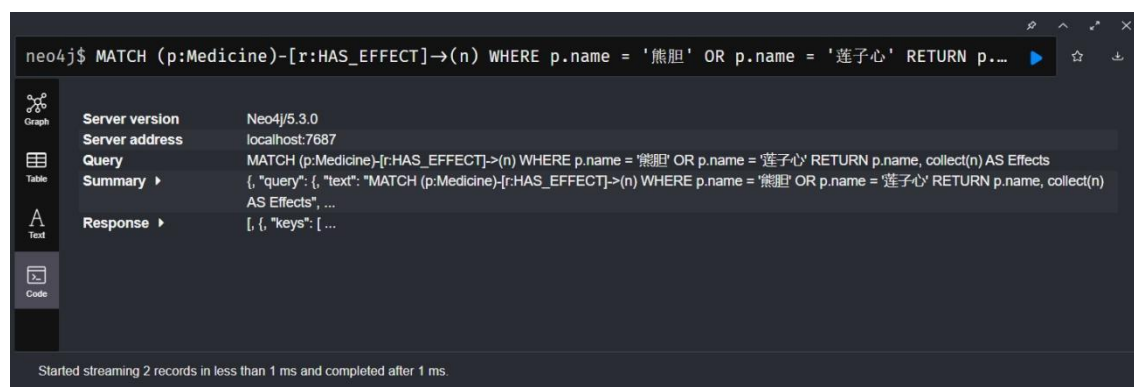
为了确保数据集的质量,每个问句和相应的 Cypher 查询语句都经过了多轮的审核和测试。审核过程中,除了语法和语义的准确性外,还重点考察了查询语句的执行效率和返回结果的相关性。图 4.1 所示是验证 Cypher 语句“MATCH (p:Medicine)-[r:HAS_EFFECT]->(n) WHERE p.name = '熊胆' OR p.name = '莲子心' RETURN p.name, collect(n) AS Effects”准确性验证结果,根据结果可以看出该 Cypher 查询语句返回了正确结果,即莲子心的功效是“清心安神”和“涩精止血”,熊胆的功效是“清肝明目”、“息风止痉”和“清热解毒”。



p.name	Effects
"莲子心"	[(:Effect {name: "清心安神"}), (:Effect {name: "涩精止血"})]
"熊胆"	[(:Effect {name: "清肝明目"}), (:Effect {name: "息风止痉"}), (:Effect {name: "清热解毒"})]

图 4.1 Cypher 查询准确性验证结果图

图 4.2 所示是验证 Cypher 语句“MATCH (p:Medicine)-[r:HAS_EFFECT]->(n) WHERE p.name = '熊胆' OR p.name = '莲子心' RETURN p.name, collect(n) AS Effects”的执行效率验证结果。根据日志结果可以看出,该查询共犯 5 条记录,在查询操作后的 1ms 内响应,并用时 2ms 完成查询,证明该 Cypher 查询语句执行效率非常高。



Server version	Neo4j/5.3.0
Server address	localhost:7687
Query	MATCH (p:Medicine)-[r:HAS_EFFECT]->(n) WHERE p.name = '熊胆' OR p.name = '莲子心' RETURN p.name, collect(n) AS Effects
Summary	{ "query": { "text": "MATCH (p:Medicine)-[r:HAS_EFFECT]->(n) WHERE p.name = '熊胆' OR p.name = '莲子心' RETURN p.name, collect(n) AS Effects", ...
Response	[{ "keys": [...

Started streaming 2 records in less than 1 ms and completed after 1 ms.

图 4.2 Cypher 查询执行效率验证结果图

4.2.2 数据集构建

根据 OpenAI 的官方文档,数据集中的每个示例都应该与 Chat Completions API 格式相同的对话。为确保数据集严格符合微调过程的格式要求,对数据进行了与处理工作,包括将收集的问句及其相应的 Cypher 查询语句以特定的 JSON 结构进行组织。数据集中的每个条目包含三个主要部分:系统指示(system)、用户查询(user)、以及模型响应(assistant)。

系统指示（system）部分明确了模型的任务，即根据用户的提问生成精确的 Cypher 查询语句。为此，所有数据点中的 system 部分的 content 被统一设定为：“作为一个 Cypher 查询语句生成器，你将根据我的问题书写出准确的 Cypher 查询语句。除 Cypher 查询语句外，不需回复任何解释或其他信息。”用户查询（user）部分包含了经过精心设计的用户提问。模型响应（assistant）部分则是对应于每个自然语言问句的 Cypher 查询语句，展示了模型应如何准确地理解和回应用户的查询。通过这样的结构化安排，确保了数据集不仅满足了模型微调的格式要求，同时也支持了模型在理解用户查询和生成正确 Cypher 查询语句方面的学习和提升。图 4.3 所示是其中一条 message 的效果展示图：

```
{
  "messages": [
    {
      "role": "system",
      "content": "作为一个Cypher查询语句生成器，你将根据我的问题书写出准确的Cypher查询语句。除Cypher查询语句外，不需回复任何解释或其他信息。"
    },
    {
      "role": "user",
      "content": "你知道解表药有哪些吗？"
    },
    {
      "role": "assistant",
      "content": "MATCH p=()-[r:BELONGS_TO]->(n:Sort {name:'解表药'}) RETURN p"
    }
  ]
}
```

图 4.3 message 效果展示图

为了保证数据格式的准确性和一致性，通过编写 Python 脚本来自动检查每个数据点。这个过程包括遍历数据集中的每个条目，检查它们是否包含三条消息（系统提示、用户问题、助理回答），并验证每条消息的 role 和 content 字段。如果所有检查都通过，脚本将输出“数据集格式正确。”；否则，它将指出数据集中存在的问题。该数据集旨在优化模型的性能，使其生成的 Cypher 查询语句更加贴合本研究创建的知识图谱库。

4.3 微调 ChatGPT-3.5 Turbo 模型

在 OpenAI 的官方博客中，开发人员通过早起测试发现，GPT-3.5 Turbo 的微调版本在某些狭窄的任务上可以匹配甚至优于基础 GPT-4 级别的功能。在 OpenAI 平台提供的官方微调指南中，可以查询到 ChatGPT-3.5 Turbo 的步骤分为：上传训练数据、训练新的微调模型、评估结果使用新的微调模型。

4.3.1 上传训练数据

通过查阅 OpenAI 的官方文档，了解到 OpenAI 提供了一个专用的 Python 库，便于用户上传微调所需的训练和验证数据。利用这一库，可以编写 Python 代码来读取准备好的训练和验证数据文件，并将它们以微调（Fine-tune）的目的上传至 OpenAI 的服务器。在上传过程中，每个文件会被赋予一个唯一的标识符（`model_id`），这一标识符对于后续启动和管理微调任务至关重要。上传操作完成后，通过服务器的响应，将能够获取到这些文件的 `model_id`，为微调过程的下一步做好准备。

4.3.2 训练微调模型

在数据上传之后，下一步是微调 ChatGPT-3.5 Turbo 模型以适应中医药领域的特定需求。为了启动微调过程，首先设定了一组初始参数，作为训练的起点。选择的学习率（Learning Rate）为 $2e-4$ ，基于经验这个值既能保证足够的学习速度，又能避免训练初期的不稳定性。考虑到数据集的规模，批大小（Batch Size）设置为 32，旨在平衡训练效率和模型性能。由于数据集较小，为确保模型能充分学习到数据中的信息，训练轮数（Epochs）设置为 10 轮，这样可以让模型多次遍历整个数据集，加深学习效果。

在训练过程中，还引入了预热步数（Warm-up Steps），设置为训练总步数的 10%，帮助模型在训练初期逐渐适应学习率，从而避免训练早期的大幅度波动。同时，采用了 0.01 的权重衰减（Weight Decay）策略，以减轻过拟合的风险。梯度裁剪（Gradient Clipping）阈值设定为 1.0，防止训练过程中可能出现的梯度爆炸问题。为了在训练过程中适当降低学习率，引入了线性衰减的学习率调度器，使学习率随着训练进度逐渐减小，有助于模型在接近最优解时更精细地调整参数。

微调模型的训练过程是通过 OpenAI 提供的 API 实现的。在确定所有必要的参数后，通过编写 Python 脚本并利用 API，将这些参数连同之前上传的数据文件 ID 一起提交至 OpenAI 平台，以启动微调任务。为了有效监控训练过程，开发了专门的监控脚本，该脚本能够实时追踪训练和验证过程中的损失值，从而提供对模型学习进度的即时反馈。这种监控机制不仅允许我们时刻观察模型的训练状态，还使得根据实时数据做出调整成为可能，以确保训练过程按照预期方向发展。通过这种方法，可以在训练过程中及时识别并应对可能出现的问题，如过拟合或未充分学习，通过调整训练参数或采用早停策略来优化模型性能。当训练完成后，通过在验证集上评估模型的表现，来验证微调后的模型是否达到了处理特定查询解析和 Cypher 查询生成任务的预定目标。如图 4.4 所示，训练损失随时间推移而减少，这表明模型正在有效地从训练数据中学习：

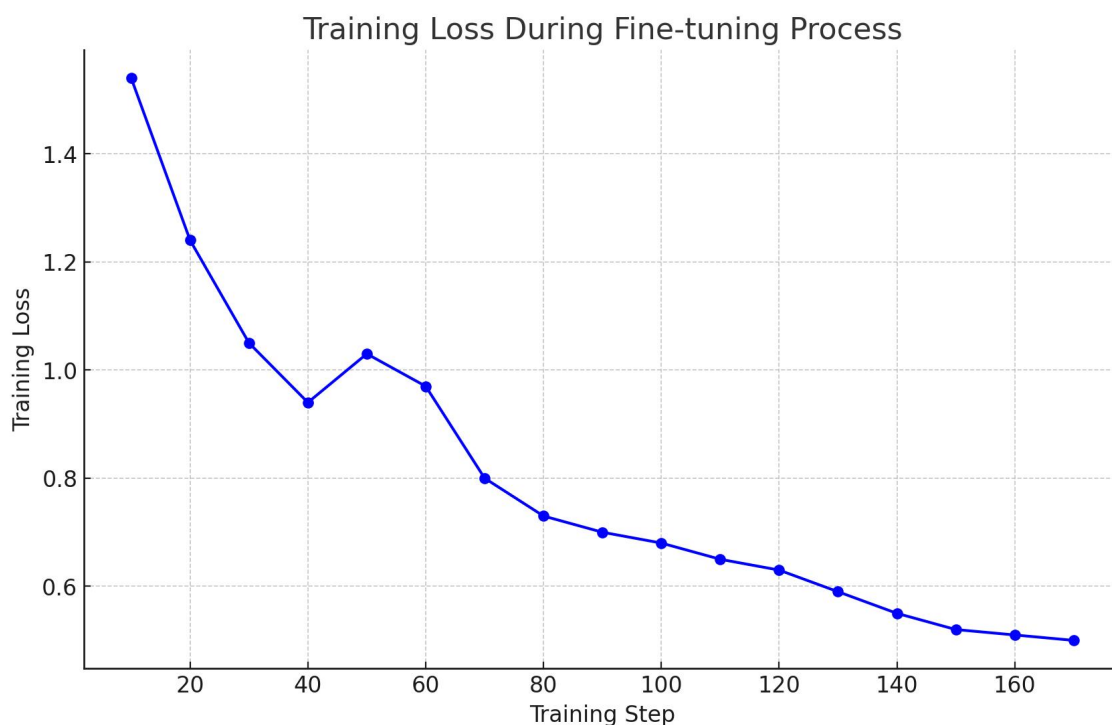


图 4.4 模型微调过程中的训练损失走势图

这一阶段的实时监控和灵活调整策略对于微调模型的成功至关重要，确保了模型能够在维持高效学习的同时，避免过拟合，从而在验证集上表现出良好的性能。通过这样的过程，微调后的 ChatGPT-3.5 Turbo 模型将具备更强的能力去准确解析自然语言问句，并生成符合预期的 Cypher 查询语句，为中药知识图谱的应用提供坚实的技术支持。

4.3.3 使用新模型

在模型微调完成后，通过编写 Python 代码即可调用微调后的模型。此过程只需按照 OpenAI 提供的代码示例进行操作，填入相应的 API 密钥和微调后模型的唯一标识符（model_id）。完成此操作后，即可实现对模型的调用。

4.4 验证实验

由于该模型的微调是针对于特定领域的专项任务，所以，为了验证微调后的大语言模型的在问句解析方面，本研究采用对比实验的方法，对比微调前后的大语言模型生成的 Cypher 查询语句的准确性和可用性对微调效果进行评估。对于微调后模型的效果，本研究采用对比实验验证而不是交叉验证的原因如下：首先，由于本研究构建了足够庞大的知识图谱，可以构建足够的数据集来划分训练集与测试集；其次，对比实验采用相同的测试数据集进行评估，可以更加直观的显示模型微调前后的性能对比；最后，交叉验证需要多次

训练模型和评估性能，而对比实验可以简化流程并减少实验的复杂性，特别是在问句解析这一环节，对比实验更为直接。

4.4.1 实验环境

本实验基于 AutoDL 提供的云服务器进行，具体实验环境如表 4.3 所示：

表 4.3 实验环境

硬件环境		软件环境	
CPU	Intel Xeon(R) Gold 6130	开发环境	Jupyter Notebook
GPU	NVIDIA Tesla V100	Python 版本	3.9.7
显存	32GB	Neo4j 版本	Neo4j Desktop 1.5.8
内存	25GB	虚拟环境管理	Conda
存储空间	50GB	依赖管理	pip

4.4.2 实验目的与评价指标

通过比较微调前后模型在生成 Cypher 查询语句任务上的性能，本研究旨在评估微调对模型效能的具体影响。为此，通过之前构建的问答对测试集，分别利用微调前的基线模型与微调后的模型进行答案生成，并详细记录两者的输出结果。

通过比较这些结果的准确度和实用性，进行性能评估。考虑到生成的 Cypher 查询语句将被直接用于 Neo4j 图数据库查询中，因此，评估的首要标准是确保生成结果的精确性，即查询语句能够正确匹配预期结果。其次，评估标准也包括生成结果的可用性，即查询语句需可直接在 Neo4j 中执行，且不包含多余的话术或语法错误。

4.4.3 实验结果与分析

为了评估微调对模型性能的具体影响，本研究设计了 126 条涉及中医药领域常见问题的问句集，用以测试模型的应答质量。通过编写 Python 代码，并利用 py2neo 库连接 Neo4j 图数据库，直接将大语言模型产生的答案输入 Neo4j 中执行。依据 Neo4j 返回的结果，本研究对比了微调前后模型在准确度和实用性两方面的表现。

图 4.5 展示了微调前后模型在准确度和实用性上的对比柱状图。通过这一比较，可以明显看出，在准确度上的提升有限，但是提高了模型输出的直接可用性：

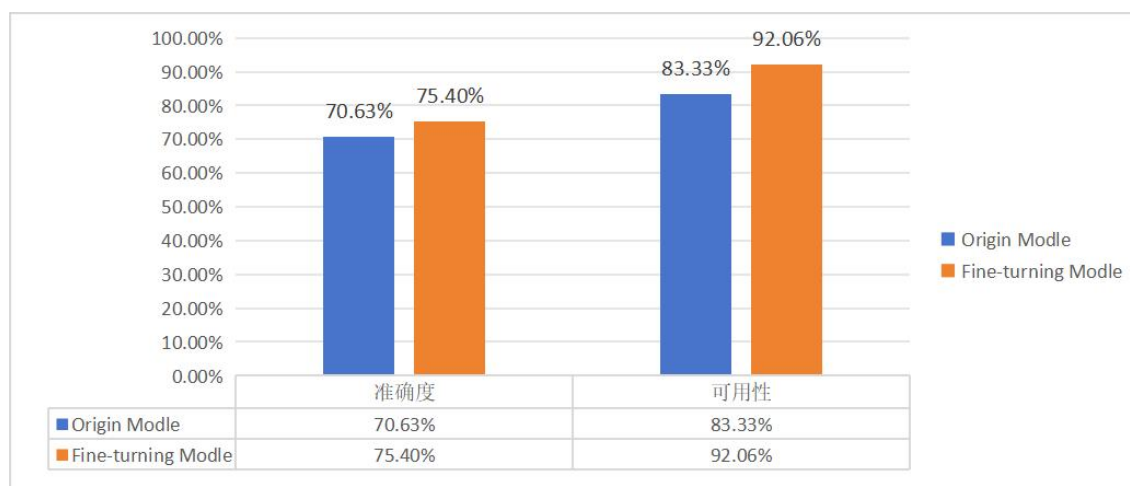


图 4.5 微调前后对比柱状图

实验结果显示，在准确度方面，微调前后的模型分别达到了 70.63%和 75.40%。这一细微的提升表明两者在准确度上的差异较小，且整体水平并不高。这主要是因为模型对知识库中存在的实体类型及其关系类型的了解有限。后续章节将通过优化 Prompt 设计来有效提高模型的准确度。

在实用性方面，微调前后模型的表现分别为 83.33%和 92.06%。微调后的模型在实用性上的表现显著优于微调前，这归因于微调过程中大量语句的引入，这些语句明确指导模型专注于生成 Cypher 查询语句。因此，经过微调的模型已经形成了明确的角色定位——作为一个 Cypher 查询语言的生成器，极少产生与任务无关的输出。相反，由于大语言模型固有的不确定性和创造性，未经微调的原始模型在输出结果时可能包含多余的语句或符号，导致无法直接使用且出现语法错误。

4.5 本章小结

本章深入探讨了通过大语言模型的微调来优化中药领域知识图谱问答系统的方法。选定 ChatGPT-3.5 Turbo 模型基于其优异的语言理解与生成能力，并针对特定领域需求进行了细致的微调，提升了模型针对中药知识图谱生成准确 Cypher 查询语句的能力。通过精心设计的数据集和实验验证，微调后的模型在准确度上和可用性方面均有一定程度的提升，验证了微调在提高智能问答系统性能方面的有效性。这一过程不仅展示了微调策略的重要性，也为下一步优化智能问答系统提供技术支撑。

第5章 融合大语言模型的知识图谱问答

近期,随着 ChatGPT 的问世,大语言模型在智能问答领域的应用变得广泛,迅速获得了公众的广泛认知。大语言模型因其在处理自然语言问答任务上的高效能而受到重视,相比之下,基于知识图谱的问答系统(KBQA)似乎逐渐被大语言模型的光芒所掩盖,有时甚至被认为是被后者替代的技术。然而,这一观点忽略了知识图谱问答系统在特定场景下的独特优势和应用价值。接下来,本节将深入分析大语言模型和知识图谱各自的优势与不足,以探讨如何结合两者的优点,以提升智能问答系统的整体性能。

5.1 大语言模型与知识图谱对比

5.1.1 大语言模型的优劣势

大语言模型凭借其卓越的语义理解能力、生成性、学习能力以及文本创作能力,在自然语言处理领域内占据了一席之地。它们能够深入挖掘自然语言文字内容的意义与关系,生成多样化形式与风格的文本。基于庞大的语料库训练得来的能力让它们在面对新的输入时,能合理响应,并产生新颖、连贯且流畅的文本输出。然而,大语言模型也有其局限性,尤其是在可解释性、可靠性、可追溯性、安全性、时效性和领域专业性方面。其决策过程的不透明性,让模型在解释其输出上显示出不足;而且,由于输出基于大规模数据而非确定的知识点,有时难以追踪到信息来源。模型的开放性增加了信息泄露的潜在风险,而知识更新的周期较长则影响了数据的时效性。虽然大语言模型在众多领域的覆盖广泛,但对某些特定专业知识的覆盖不尽人意,这在处理专业问题时可能导致误差,例如在中医药领域,不准确的答案可能引起严重后果。

5.1.2 知识图谱的优劣势

与大语言模型相比,知识图谱所具有的优势正好与之相反。知识图谱依托于明确的语义结构进行查询和分析,因此具备更好的可解释性。在知识图谱中,每个实体和关系都明确可追溯至其来源,并能通过专家校验,赋予了其较强的可追溯性和可验证性。特定领域的知识图谱因其深入的专业知识,能够提供更高的领域覆盖率,尽管这可能会牺牲一定的通用性。然而,知识图谱也有其局限性,特别是在语义理解、自然语言处理和文本生成能力方面。由于依赖于特定的知识库,知识图谱在处理复杂或混乱逻辑的自然语言时可能不

如大语言模型灵活。在传统的基于知识图谱的问答系统（KBQA）中，问句解析的通常方法为：首先，系统对问句中识别的实体和关系执行实体抽取和关系抽取。接下来，这些提取出的三元组（实体、关系、实体）在知识图谱中进行查询，以匹配到相应的结果。这种方法依赖于问句中明确表达的信息和知识图谱中预定义的结构，从而在处理明确和结构化良好的查询时效率较高。然而，当面对逻辑复杂或表达模糊的查询时，这种结构化的解析方法可能不足以准确理解用户的真实意图，导致查询效果不理想。

5.1.3 融合大语言模型和知识图谱问答

在完成对大语言模型与知识图谱各自的优缺点进行深入比较分析之后，可以发现在应用场景中，大语言模型与知识图谱的擅长领域如表 5.1 所示：

表 5.1 大语言模型与知识图谱擅长应用领域

应用领域	大语言模型	知识图谱
智能对话	√	
内容生成	√	
内容加工	√	
作品创作	√	
意图识别	√	
智能检索	√	√
智能推荐		√
辅助决策		√
知识管理		√

大语言模型与知识图谱是存在一定的互补关系。知识图谱首先能够为大语言模型提供专业领域的知识支撑，弥补大语言模型在专业领域的不足；其次，可以通过知识图谱的知识构建问答对，对大语言模型进行微调，然后对大语言模型进行评估，降低事实性错误的发生概率；最后，可以根据知识图谱的结构化数据，提高检索效率，实现知识查询。提高大语言模型在专业领域的适应能力。大语言模型对知识图谱，首先，可以利用自身的语义理解能力解析自然语言，提高问句解析的准确度；其次，可以通过语义理解能力解析知识图谱，生成更准确的查询语言；最后，辅助提升知识图谱的输出效果，生成更加合理、连贯、可读性更强的内容。图 5.1 所示是两者的互补关系：

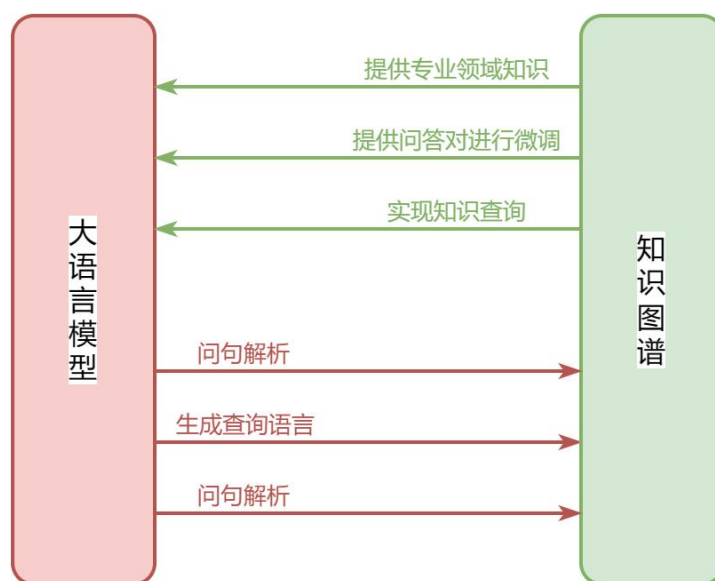


图 5.1 大语言模型与知识图谱互补关系图

本研究提出了一种创新的结合策略，旨在融合大语言模型的强大语义理解与文本生成能力，以及知识图谱的高度可解释性和精确的领域专业知识。此方法的核心在于，通过整合两种技术的优势，使得智能问答系统在处理广泛用户查询的同时，既保持了高度的灵活性和创新性，又能在需要深度领域知识时，确保所提供信息的准确性与可信度。为了能够完美实现融合，就需要使用到 **Prompt** 技术，它能够作为融合知识图谱与大语言模型的桥梁。在智能问答系统的研究领域，问句解析与答案生成构成了系统核心功能的关键环节，直接决定了系统提供答案的准确性。通过出色的 **Prompt** 管理，大语言模型能够对问句中的自然语言进行解析，并且生成相应的查询语句，知识图谱利用查询语句检索结果，将生成的结果返回给大语言模型，大语言模型则由于良好的 **Prompt** 能够根据查询结果生成问句对应的答案。具体的流程图如图 5.2 所示：

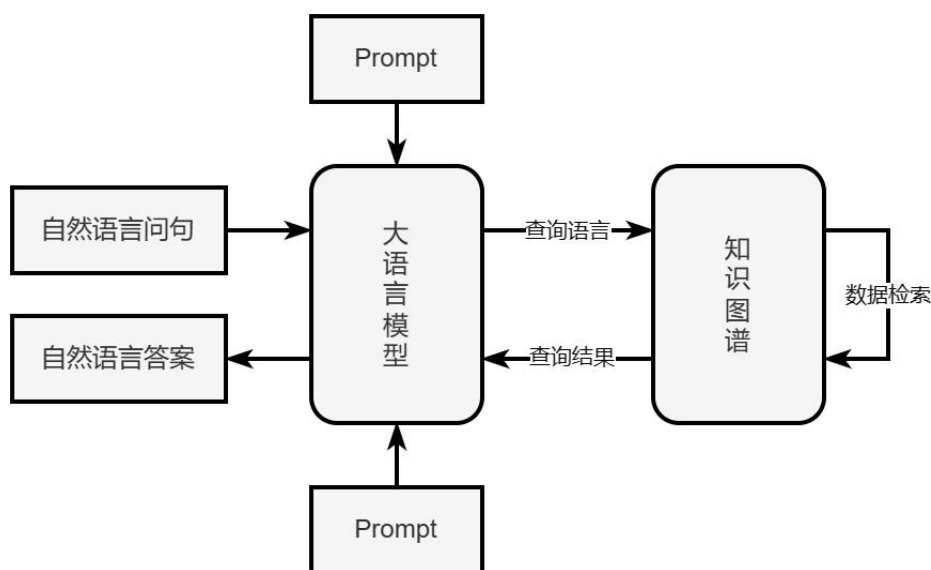


图 5.2 融合大语言模型的知识图谱问答流程图

本研究设计的智能问答系统旨在通过这种融合技术，强调在设计先进的问答系统时，知识图谱与大语言模型融合使用的重要性。

5.2 Prompt 管理

Prompt 管理在大语言模型应用中扮演着至关重要的角色，尤其在那些对任务具有明确指向性、并且需要模型理解及生成特定内容的场景下显得尤为关键。Prompt 可被视作一种指令或引导，其目的在于明确告知模型所需执行的任务类型，或期望模型以何种形式对输入数据给予响应。借助 Prompt 以有效地引导模型输出，已经被证明是在特定任务上提升模型表现的一项关键技术。

5.2.1 Prompt 原理

1) 自注意力与 Prompt: 从算法角度深入探讨，Prompt 的有效性部分归功于模型内置的自注意力机制。该机制使得模型在生成响应时，能够综合考量与 Prompt 中的每一个词元之间的关系，从而使生成的文本更加紧密地与 Prompt 相关联。这一过程极大地增强了模型捕捉用户意图的准确性。

2) 语言模型概率分布: 从数学的视角来看，大语言模型的文本生成基于条件概率的原理，即给定一段前文（Prompt），模型预测下一词出现的概率分布。Prompt 的精心设计直接影响到这一条件概率分布，进而引导模型文本生成的方向。

3) Embedding 调整: 在模型微调过程中，模型词嵌入（Embedding）的调整是根据特定任务的需求进行的。通过在特定任务上训练模型，词嵌入会相应地调整，使得模型在遇到类似 Prompt 时能够产生更加精准的响应。这一过程实质上是将特定领域的知识注入到模型的知识库中，从而在该领域任务上显著提升模型的表现能力。

5.2.2 问句解析

问句解析的主要目标是利用大语言模型对用户提出的问题进行解析，并生成对应的 Cypher 查询语句以便在知识图谱中检索信息。根据先前进行的模型微调前后对比实验，观察到微调对于提升生成 Cypher 查询语句的准确度影响有限。这一现象的原因在于模型在生成 Cypher 查询语句过程中，缺乏对知识图谱中存在的实体类型及其关系类型的明确了解，从而难以产生准确匹配的查询语句。然而，通过精心设计的 Prompt，该问题能够得到有效缓解。为了配合微调后的模型，进一步提升问句解析的效果，本研究提出了一种针对性的 Prompt 设计策略。该策略旨在通过以下几个方面的指导，优化模型生成 Cypher 查询语句的能力：

1) 模型身份定义

首先，需要明确指示模型其被赋予的新身份——一个专门针对用户问句生成 Cypher 查询语句的模型。这一步骤旨在强化模型的任务意识，确保其在处理输入时集中于查询语句的生成，避免产生与任务无关的输出。

2) 数据类型介绍

其次，为模型提供知识图谱中所有节点类型和关系类型的详细信息，包括每种类型的名称及其含义。这不仅有助于模型理解和区分不同的实体和关系，还能使模型在构造查询语句时更加精确地引用这些类型，从而提高查询语句的匹配度和准确性。

3) 任务说明

最后，明确告知模型其任务目的——基于用户的问句生成相应的 Cypher 查询语句。通过这样的 Prompt 设计，可以直接引导模型聚焦于从用户询问中捕捉关键信息，并转化为准确的查询语句。

通过 Prompt 的编写，再次使用之前构建的 126 条问答对，对问句解析模型进行测试，得到的结果是准确度达到了 90.48%，与之前的不使用 Prompt 的微调模型和原模型相比提升明显，具体折线图如图 5.3 所示，根据这个结果可以得出 Prompt 的编写对于问句解析的有效性。

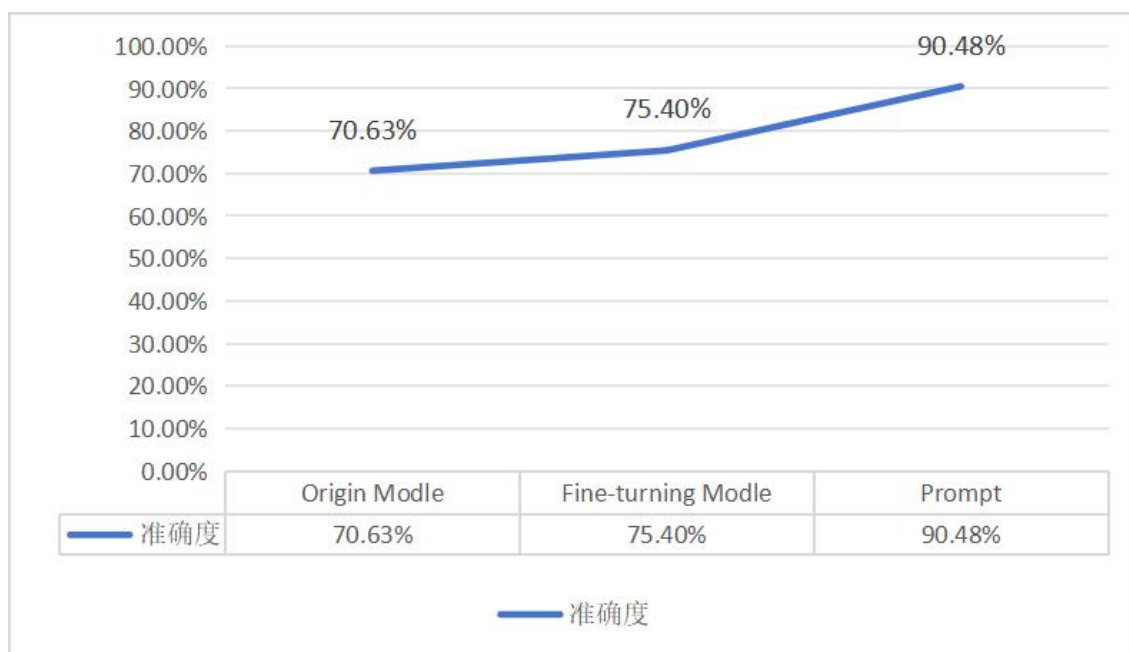


图 5.3 问句分析准确度提升折线图

5.2.3 答案生成

答案生成阶段是智能问答系统中至关重要的一环，它基于从知识图谱检索得到的数据完成。此过程中，未经微调的原始大语言模型直接解析 Neo4j 的查询结果，核心任务在于

将问句解析阶段产生的 Cypher 查询语句执行后,从 Neo4j 图数据库中检索到的结果,通过大语言模型转化为用户可理解的自然语言答案。为了实现这一转换过程的优化,精心设计的 Prompt 发挥着至关重要的作用。以下为答案生成阶段 Prompt 设计的具体策略:

1) 模型身份定义

对模型进行身份定义,将其定位为专门针对 Neo4j 查询结果生成答案的医疗咨询模型。这一定义有助于模型准确识别其执行的核心任务,从而专注于根据知识图谱的检索结果生成具体且相关的答案。这样的身份确认,不仅提升了模型针对性的任务执行能力,也优化了答案的生成质量。

2) 数据类型介绍

由于答案生成与问句解析调用的是不同的模型,所以需要再次对为模型提供知识图谱中所有节点类型和关系类型的详细信息,包括每种类型的名称及其含义。

3) 任务说明

通过明确的任务说明,指导模型根据 Neo4j 的查询结果生成响应。这一步骤的目的在于确保模型能够清楚理解其任务范畴,即从图数据库检索的信息中提取关键数据,转换为简洁明了的答案。这种任务明确性的引导对于提高答案生成的准确性和相关性至关重要。

4) 引用问句

考虑到答案生成阶段所使用的模型与问句解析阶段的模型不同,有必要将用户的原始问句告知当前模型。这样做可以确保模型在理解查询结果的同时,将答案更好地与用户的原始查询意图相对齐。这一策略有助于生成更加针对性强、用户满意度高的答案,避免因缺失上下文信息而产生的答案不相关或偏离主题的问题。

5) 错误解析

考虑到在执行 Cypher 查询时,Neo4j 可能无法检索到任何结果,因此,大语言模型依据自身的数据库生成回答。在这种情况下,模型应明确声明所生成的答案不是直接来自知识图谱的查询结果。这一策略确保了即使在知识图谱检索失败的情况下,用户仍能获得有价值的信息,同时保持了回答的透明度和可信度。

对于答案生成的准确度将通过对比实验证明其有效性。

5.3 对比实验

5.3.1 实验目的

为了验证所提方法的有效性,本研究将融合大语言模型和知识图谱问答系统与纯大语言模型驱动的问答系统(例如 ChatGPT3.5、文心一言等)进行性能比较。通过此对比实验的设计与实施,本研究旨在展示融合大语言模型和知识图谱问答系统的明显优势与实际效

用。此方法的成功验证将明确表明，虽然大语言模型在自然语言处理和问答任务中显示出强大的能力，但知识图谱的融入为系统提供了不可或缺的结构化知识和深度专业理解，从而提高了问答系统的准确性和可靠性。

5.3.2 实验准备

本实验的准备工作主要涉及数据集的筹备与评估指标的制定两大部分，旨在确保实验设计的科学性与实验结果的有效性。

1) 数据准备

为了满足本次实验的需求，特别关注于构建一个既能体现中药领域专业性，又能覆盖广泛用户自然语言查询的测试集。因此，本研究通过在调查医药卫生考试的中草药知识问答题库，选取了题库中常见的 210 条针对中草药知识的问答对。这一步骤旨在确保实验数据不仅具有高度的专业性，同时也反映了真实世界中用户可能提出的各类查询，为测试问答系统的综合性能提供了坚实的基础。如表 5.2 是其中 5 条数据：

表 5.2 中草药知识问答数据

问题	答案
藤乌头功效是什么？	镇痉，降压，发汗，利尿。
地椒功效是什么？	清热除湿，解毒散瘀。
橄榄可以用来治疗什么？	咽喉肿痛，烦渴，咳嗽吐血，酒伤昏闷
纤毛婆婆纳功效与作用分别是什么？	清热解毒，祛风利湿。
亚麻子功效是什么？	风湿关节炎，肝炎，胆囊炎，荨麻疹 平肝，顺气，通肠，解毒，止痛。

2) 评估指标

考虑到实验结果的准确性与有效性至关重要，本研究细致选择了评估指标，旨在全面反映问答系统的性能。通过对比实验结果与题库问题的标准答案，计算准确度和 Jaccard 相似度这两个指标。

准确性得分：通过对比问答系统的答案与题库答案，将准确度的评估指标分为如表 5.3 所示的 4 个指标：

表 5.3 中草药知识问答数据

完全错误	0	与题库答案完全不一致
部分正确	1	只有部分答案正确
多余回答	2	回答部分答案，但是有多余回答
完全正确	3	与题库答案完全一致

Jaccard 相似度（Jaccard Similarity）：是衡量两个集合相似度的指标。它通过比较两个集合中共同元素的数量与总元素的数量（即两个集合的并集大小）来计算相似度。Jaccard 相似度的值也在 0 到 1 之间，其中 1 表示两个集合完全相同，0 表示两个集合没有共同元

素。其公式为 (5.1)，其中，A 是问答系统中答案的集合，B 是测试集中答案的集合， $|A \cup B|$ 是两个集合的交集元素数量， $|A \cap B|$ 是两个集合的并集元素数量。

$$J(A, B) = \frac{|A \cup B|}{|A \cap B|} \quad (5.1)$$

这种方法不仅充分确保评估的专业性和可信度，也为深入理解问答系统在实际应用中的表现提供了有力的支持。

5.3.3 实验过程

1) 系统配置

融合大语言模型的中药领域知识图谱问答系统：本研究构建的系统，通过融合大语言模型与知识图谱技术，旨在提供准确且可解释的中医药领域问答服务，作为主要的实验对象与其他模型进行性能对比。

ChatGPT-3.5：作为当前领先的大语言模型之一，被选作对照组之一，用以验证本研究构建系统的性能优势。

文心一言 3.5：另一个先进的大语言模型，同样作为对照组参与实验，进一步检验本研究系统的综合表现。

2) 实验步骤

从中草药知识问答题库中精心挑选了 210 条关于中医药常识的问答，用作评估测试。每个系统被要求处理这些预选问答，它们的回答被详细记录并与题库中的标准答案对照，以确保结果的可靠性。

依据先前设定的准确性得分评估标准，每个回答被赋予相应分数，并进行了统计分析。为了深入比较答案之间的细节 Jaccard 相似度，本研究采用了 Python 环境中广泛使用的 jieba 库来进行中文文本分词处理。通过 jieba 的精确模式，答案被转换成了各自的词语集合。分词过程中，发现了一些不理想的分词结果，通过添加自定义字典，以优化分词的完整性和准确性。经过一系列的尝试和调整，分词结果得以准确反映问答内容。使用调整后的分词结果，本研究对各系统回答与标准答案之间进行了 Jaccard 相似度计算，这一步骤旨在量化答案间的相似度，为比较分析提供了另一层次的评价维度。

5.3.4 实验结果与分析

通过三组对比实验，在经过检查标准答案后，得出了如图 5.4 的结果：实验结果表明，本研究开发的融合大语言模型的中药领域知识图谱问答系统在中草药知识问答测试中获得的 Jaccard 相似度为 0.87，高于 ChatGPT-3.5 与文心一言 3.5 的 Jaccard 相似度 0.84 和 0.73，在准确性得分方面，本系统的得分 263 也要高于另外两者的 256 和 221。这一结果表明了

本研究系统在处理中医药领域问答上的优势，特别是在准确性方面超越了纯粹依赖于大语言模型的系统。

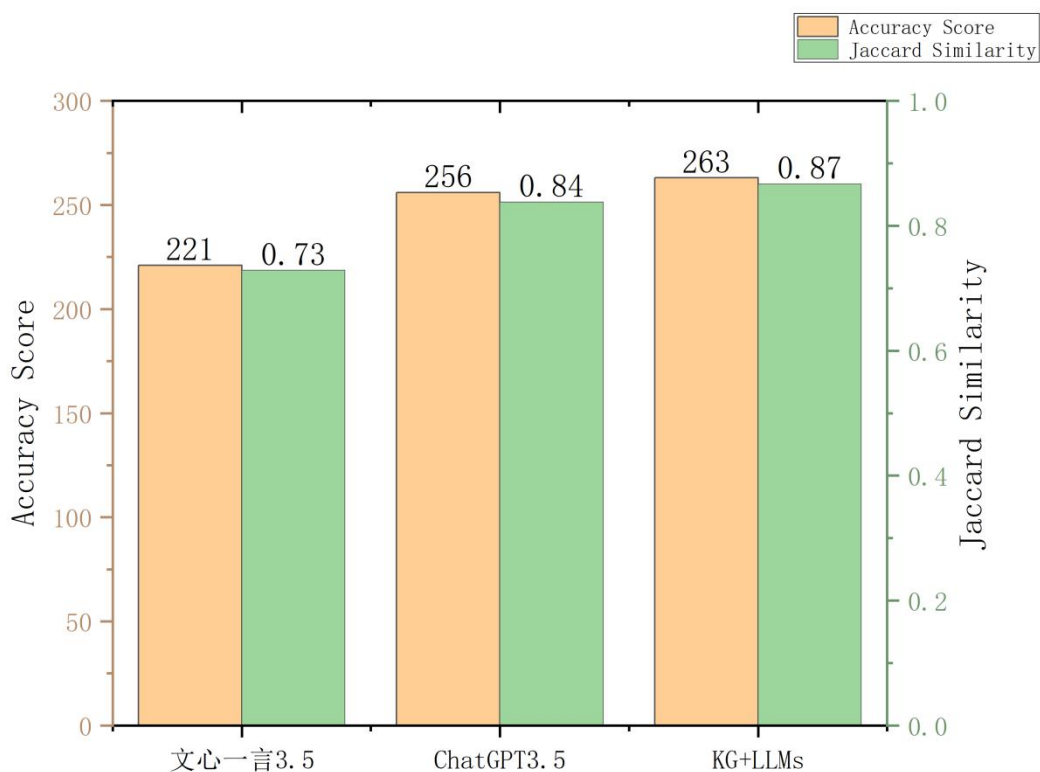


图5.4 问答系统对比结果

同时，得益于知识图谱的结构化数据特性，本问答系统的答案可解释性显著提升。在知识图谱的支撑下，问答系统能够通过引用具体的实体关系和属性，直接地向用户展示如何得出特定的结论。这种方法不仅增强了用户对系统输出的信任，而且在需要对复杂问题进行深入分析时，这种结构化的答案更具优势。相比之下，大型语言模型，因其内部复杂的神经网络结构，常常被比喻为黑盒，它们的答案虽然流畅，但在可追溯性和明确性上不如基于知识图谱的方法。本系统可以通过后台运行结果查看到每次问答的知识图谱检索数据，如图5.5所示：

```
Human(please input your question): 心神不宁怎么办?

生成的Cypher查询语句:
MATCH (m:Medicine)-[:HAS_APPLICATION]->(e:Application) WHERE e.name CONTAINS '心神不宁'

查询结果:
+-----+-----+
| m |
+-----+-----+
| (:Medicine {name: "朱砂", type: "重镇安神药"}) |
+-----+-----+
| (:Medicine {name: "磁石", type: "重镇安神药"}) |
+-----+-----+
| (:Medicine {name: "龙骨", type: "重镇安神药"}) |
+-----+-----+
| (:Medicine {name: "琥珀", type: "重镇安神药"}) |
+-----+-----+
| (:Medicine {name: "珍珠", type: "重镇安神药"}) |
+-----+-----+
| (:Medicine {name: "珍珠母", type: "平抑肝阳药"}) |
+-----+-----+

摘要:
针对心神不宁的症状,传统医药推荐了几种药物作为治疗选项。包括朱砂、磁石、龙骨、琥珀和珍珠,这些都是被分
```

图 5.5 后台运行结果图

该对比实验的结果强调了知识图谱在提供领域专业知识方面的重要性,尤其是在需要深度专业理解和精确信息提供的中药领域。融合大语言模型的语义理解与文本生成能力,与知识图谱的结构化知识库相结合,本研究系统能够提供更加准确且具有高度可解释性的答案。此外,实验结果也突显了持续探索和优化知识图谱与大语言模型结合应用的必要性,以便更好地满足特定领域内的智能问答需求。通过本次对比实验,证实了本研究所提出的融合知识图谱与大语言模型的中医药辅助诊疗系统在专业性强的领域问答中的有效性。

5.4 本章小结

本章深入探讨了大语言模型与知识图谱在智能问答系统中的结合的可能性。通过详尽的比较分析,本研究揭示了大语言模型的语义理解与文本生成能力和知识图谱的高度可解释性及精确的领域专业知识之间的互补性。实验结果表明,融合大语言模型和知识图谱的问答系统在中医药领域的应用中,相较于单独使用大语言模型,显示出更高的准确性和可解释性。这一发现不仅强调了在构建专业领域内智能问答系统时,综合利用这两种技术的重要性和有效性,也证实了在智能问答领域中,知识图谱仍然占据着重要位置,大语言模型并非问答系统的唯一解决方案。

第 6 章 系统实现

6.1 开发工具与技术

开发环境与工具如表 6.1 所示:

表 6.1 开发环境与工具

环境	具体配置
前端开发环境	VScode
前端开发框架	Vue3、Element Plus
前端开发语言	HTML、CSS、Typescript
后端开发环境	Pycharm
后端开发语言	Python3.9.7
虚拟环境与包管理工具	Conda、pip
数据库管理	Neo4j Desktop 1.5.8
数据库语言	Cypher

6.2 系统设计

中药领域知识图谱问答系统旨在为用户提供一个通过自然语言问题进行交互的问答平台。用户可以方便地查询中药的相关知识,包括但不限于草药的性味归经、种类、功效及其在临床上的应用等信息。系统背后的核心机制是通过深度理解用户的查询意图,并在庞大的知识图谱数据库中精确检索,从而快速返回准确的答案。

6.2.1 系统功能

系统的功能模块构成了本平台的核心,涵盖了四个关键功能,如图 6.1 所示。以下部分将详细介绍这四大功能:

问句输入功能: 本系统提供了一个直观的用户界面,使用户能够轻松地输入自己的查询。通过这一功能,用户可以直接向系统提交中医药领域的相关问题。

问句解析功能: 一旦接收到用户的自然语言问题,系统即启动其问句解析机制。该功能的主要任务是深度理解用户提出的问题,并将其转换成对应的 Cypher 查询语句。这一步骤是确保准确检索到相关信息的关键环节。

数据检索功能: 随后,系统利用生成的 Cypher 查询语句在 Neo4j 图数据库中进行精确查询。该数据检索功能是系统的核心,它通过高效地检索知识图谱中的数据,为生成准确

答案奠定了基础。

答案生成功能：最后，系统根据 Neo4j 的查询结果，针对用户的初始问题生成详尽的答案。该功能将检索到的数据以用户易于理解的形式呈现，从而完成了从问句输入到答案输出的整个流程。

通过这四个精心设计的功能模块，本系统为用户提供了一个全面、便捷的中医药辅助诊疗问答平台，大大简化了获取中医药知识的过程，并提高了信息检索的准确性和效率。

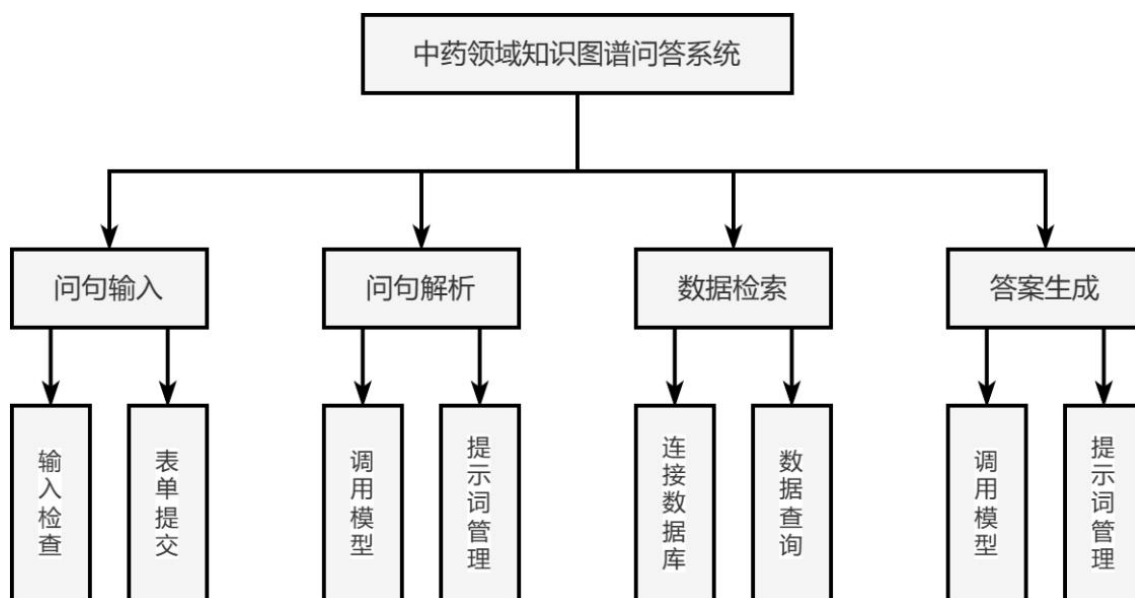


图 6.1 系统功能模块图

6.2.2 系统流程

本系统采用了一套精细化的工作流程，以确保用户能够以自然语言形式高效获取中医药相关知识。首先，当用户通过 Web 端提交自然语言问题后，系统将该问句上传至服务器。服务器端设有三个核心模块：问句解析、数据检索和答案生成，它们协同工作以处理用户的查询。问句解析模块首先通过集成的大语言模型对用户问题进行深度分析，理解其核心意图，并转换成相应的 Cypher 查询语句。此后，数据检索模块根据生成的 Cypher 语句，在 Neo4j 图数据库中进行精确查找，快速定位相关信息。最终，答案生成模块将检索到的数据处理并转换成自然语言格式的答案，通过客户端呈现给用户。这一流畅的处理流程不仅显著提升了查询的响应速度，还确保了提供给用户的答案既准确又具有实用价值，进一步优化了用户体验，使得获取中医药知识变得更加简便和高效。本系统的工作流程图如图 6.2 所示：

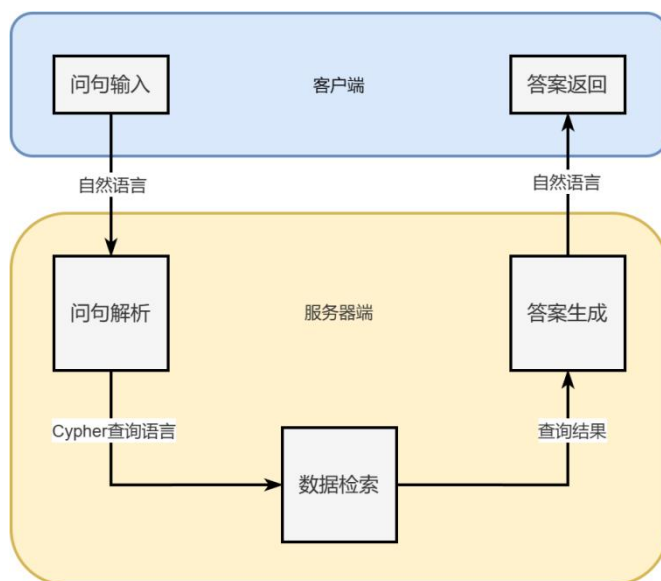


图 6.2 系统工作流程图

6.2.3 问句输入模块

由于本系统采用了 Vue3 框架，因此在问句输入模块，使用 Vue 官方提供的组件库 Element Plus 组件直接进行开发。首先使用 Form 组件进行问句输入，该组件提供了多种输入类型的选择并可以一起打包提交，包括文本输入框、单选按钮、多选按钮、下拉框、提交按钮、重置按钮。根据本系统的需要，选择 Form 组件中的表单校验组件，该组件提供了表单验证功能，可以检查用户输入内容是否为空，如果为空，则不能提交，并发出警告。如图 6.3 是该表单的效果图：

该图展示了一个带有表单校验功能的用户界面。表单包含以下字段和控件：

- * Activity name: 文本输入框，值为 "Hello"。
- * Activity zone: 下拉选择框，值为 "Activity zone"。
- * Activity count: 下拉选择框，值为 "Activity count"。
- * Activity time: 日期选择器（显示 "Pick a date"）和时间选择器（显示 "Pick a time"）。
- Instant delivery: 开关按钮，当前处于关闭状态。
- * Activity type: 包含四个复选框：☐ Online activities, ☐ Promotion activities, ☐ Offline activities, ☐ Simple brand exposure。
- * Resources: 包含两个单选按钮：☐ Sponsorship, ☐ Venue。
- * Activity form: 富文本编辑器。

底部有两个按钮：蓝色的 "Create" 按钮和灰色的 "Reset" 按钮。

图 6.3 表单校验组件效果图

为了实现问答系统，需要将用户输入与本系统回答以问答对的方式实现，因此通过使用 Scrollbar 组件来实现该功能，该组件本质上是一个滚动条模块，可以将每条文本从上到下排列，并可以通过滚动条来浏览温恩。如图 6.4 所示是 Scrollbar 组件效果图：

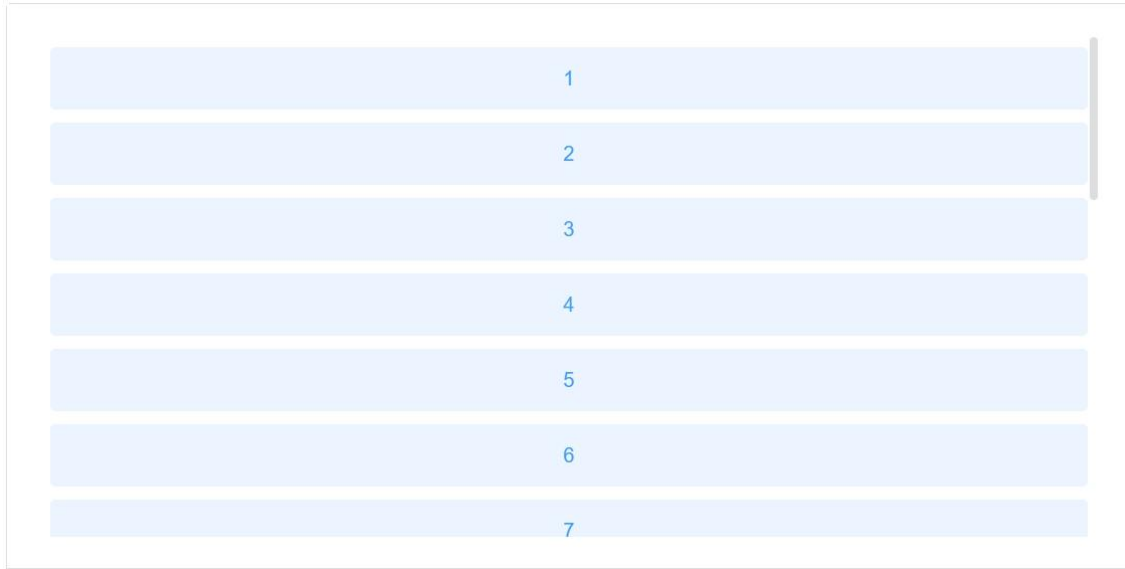


图 6.4 Scrollbar 组件效果图

6.2.4 问句解析模块

本系统的问句解析模块采用了微调后的大语言模型，精心设计其角色，使之专门用于根据用户提出的问题生成相应的 Cypher 查询语句。在微调完成后，通过使用 OpenAI 提供的 API，利用微调模型的唯一标识符来调用此模型。随后，通过精心构造的 Prompt，指导微调后的大语言模型深入理解知识图谱中定义的实体类型与关系类型。这种方法使得大语言模型在处理用户的自然语言问句时，能够更精确地生成符合查询需求的 Cypher 语句，从而大幅提高了系统检索知识图谱的准确性和效率。该模块的工作流程图如图 6.5 所示：

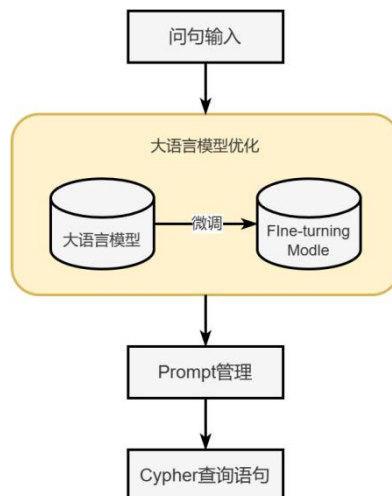


图 6.5 问句解析模块工作流程图

6.2.5 数据检索模块

本系统的数据检索模块采用了 Python 编程语言，并利用 py2neo 库实现了与 Neo4j 图数据库的连接。该模块的核心功能是接收问句解析模块生成的 Cypher 查询语句，通过精心编写的代码将这些查询语句输入到 Neo4j 数据库中进行高效的数据检索。一旦查询执行完毕，模块便会收集并整理 Neo4j 返回的结果，准备将这些信息传递给答案生成模块。此过程不仅确保了从数据库中检索信息的准确性，同时也保障了系统对用户查询的响应速度，从而显著提升了整个系统的性能和用户体验。

6.2.6 答案生成模块

本系统的答案生成模块利用大语言模型的语义解析能力和文本生成能力，将 Neo4j 返回的查询结果生成相应的答案。这一步通过 OpenAI 提供的 API 调用 ChatGPT3.5 turbo 模型，然后编写 Prompt 让模型更加准确理解意图，

本系统的答案生成模块结合了大语言模型的语义解析与文本生成能力，对 Neo4j 返回的查询结果进行加工，生成面向用户的答案。该过程通过 OpenAI 提供的 API 调用 ChatGPT-3.5 turbo 模型实现，其中精心设计的 Prompt 确保了模型能够准确理解用户的意图和查询结果的上下文，从而产生贴切、易于理解的答案。这一步骤体现了系统在智能问答技术方面的先进应用，旨在提供更加人性化和高质量的用户体验。该模块的工作流程图如图 6.6 所示：

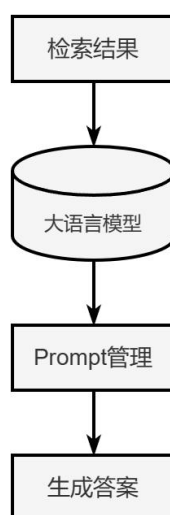


图 6.6 答案生成模块工作流程图

6.3 系统展示

为展示本系统在实际应用中的效能，本节以“麻黄的功效”查询作为示例，详细阐述

系统如何处理此类自然语言问句，并准确返回相关信息。用户通过系统界面提交查询“麻黄的功效是什么？”后，系统首先利用微调后的大语言模型对问句进行解析，转化为相应的 Cypher 查询语句。随后，该查询语句在 Neo4j 图数据库中执行，以检索“麻黄”的相关功效信息。得到查询结果后，系统再通过答案生成模块，将结果转化为易于理解的自然语言答案，展示于用户界面的就是系统生成的答案。此过程不仅展现了系统对用户自然语言问句的准确理解，也展示了从问句解析到数据检索，再到答案生成的高效流程。为直观展示该查询过程及其结果，本文嵌入了相应的展示图，如图 6.7 所示：



图 6.7 系统查询效果展示图

6.4 本章小结

本章深入展示了基于知识图谱的中医药辅助诊疗系统的构建与实现，覆盖了开发环境、工具选择、以及系统设计的关键方面。详细阐述了系统的四个主要功能模块：问句输入、问句解析、数据检索、与答案生成，每个模块均经过精心设计，确保系统能够准确理解并处理用户的自然语言查询。整个章节的叙述强调了系统在技术实现细节和用户体验设计上的考量，展现了其作为智能问答平台的强大功能和应用前景。

第7章 总结与展望

7.1 总结

在本研究中,探索了基于知识图谱的中医药辅助诊疗系统,通过整合先进的大语言模型技术,实现了一个高效、准确的智能问答平台。通过构建中医药知识图谱并与大语言模型相结合,本系统不仅能够理解和解析自然语言问句,还能在庞大的知识图谱中精确检索相关信息,为用户提供科学、精确的诊疗建议和知识解答。

知识图谱作为存储大量中医药知识的数据库,其结构化的数据性质非常适合中医药领域的复杂信息组织需求。通过明确定义的实体和关系,知识图谱能够有效捕捉和表述中医药领域内的丰富知识体系,包括药材、疾病、治疗方法等关键信息,使之成为中医药辅助诊疗的强大支持工具。

随着大语言模型的广泛应用,智能问答领域似乎被这些模型所主导。然而,尽管大语言模型在理解和生成自然语言方面表现出色,它们在特定领域内的知识精准度和可解释性方面仍存在限制。本研究发现,通过将知识图谱的结构化知识与大语言模型的强大语义理解能力相结合,可以有效克服这些限制。具体而言,通过微调和精心设计 Prompt,使得大语言模型能够在问句解析阶段根据自然语言生成精确的 Cypher 查询语句,然后利用这些语句在 Neo4j 中检索具体的知识点,最终通过大语言模型将检索结果转化为易于理解的答案,返回给用户。

为了将这一理念转化为实际的应用服务,开发了基于知识图谱的中医药辅助诊疗系统。该系统不仅体现了知识图谱和大语言模型结合的理论价值,更通过实际应用展示了其在中医药辅助诊疗领域的实用性和有效性。通过为用户提供准确、快速的问答服务,本系统在推广中医药知识、提高公众健康意识方面发挥了重要作用。

7.2 展望

展望未来,本研究在基于知识图谱的中医药辅助诊疗系统方面还有进一步的发展潜力。首先,考虑到当前知识图谱中的中医药数据无法全面覆盖该领域的广泛知识,未来工作将重点放在对知识图谱的持续补充与更新上。通过引入更多来源的专业数据,如古籍、现代研究论文等,可以显著提升系统的知识覆盖度和答案的准确性。

其次,考虑到知识图谱与大语言模型结合的问答系统还有很大的探索空间,可以进一步研究如何优化微调过程和 Prompt 的编写,以便更准确地理解和回答用户的专业性问题。

此外，探索新的模型结构和算法，以提高系统的理解和生成能力，也是未来的研究方向之一。

还有，随着人工智能技术的不断进步，引入更先进的自然语言处理技术，如情感分析、意图识别等，来提升系统与用户交互的自然度和友好度，也是值得探索的方向。通过这些技术，系统能够更好地理解用户的需求，提供更个性化和贴心的服务。

最后，本研究还计划探索多模态交互的可能性，允许用户通过上传视频、图片、音频等多种形式的数据进行辅助诊疗。这种多模态交互方式将使系统能够处理更丰富的数据类型，提供更全面的诊疗建议，同时也使用户体验更加丰富和便捷。

参考文献

- [1] 余艳红. 深入学习贯彻党的二十大精神 在新征程上奋力开创中医药传承创新发展新局面. 中国中医药网, CN11-0153, 2022-10-24.
- [2] 中共中央国务院关于促进中医药传承创新发展的意见, 2019-10-20
- [3] Spencer Chang. 2018. Scaling Knowledge Access and Retrieval at Airbnb. AirBnB Medium Blog.
- [4] Arun Krishnan. 2018. Making search easier: How Amazon's Product Graph is helping customers find products more easily. Amazon Blog.
- [5] R. J. Pittman, Amit Srivastava, Sanjika Hewavitharana, Ajinkya Kale, and Saab Mansour. 2017. Cracking the Code on Conversational Commerce. eBay Blog.
- [6] Natasha F. Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale Knowledge Graphs: Lessons and Challenges. ACM Queue 17, 2 (2019), 20.
- [7] Deepika Devarajan. 2017. Happy Birthday Watson Discovery. IBM Cloud Blog.
- [8] Qi He, Bee-Chung Chen, and Deepak Agarwal. 2016. Building The LinkedIn Knowledge Graph. LinkedIn Blog.
- [9] Saurabh Shrivastava. 2017. Bring rich knowledge of people, places, things and local businesses to your apps. Bing Blogs.
- [10] Ferras Hamad, Isaac Liu, and Xian Xing Zhang. 2018. Food Discovery with Uber Eats: Building a Query Understanding Engine. Uber Engineering Blog.
- [11] Shaoxiong Ji, Shirui Pan, Erik Cambria, Senior Member, IEEE, Pekka Marttinen, Philip S. Yu, Fellow, IEEE. A Survey on Knowledge Graphs--Representation, Acquisition and Applications[J].
- [12] Bollacker K, Evans C, Paritosh P, et al. Freebase:a collaborativelycreated graph database for structuring human know ledge [C] //Pro-ceedings of the 2008 ACM SIGM OD International Conference onM anagement of Data(SIGM OD), 2008:1247-1250.
- [13] Suchanek F M, Kasneci G, Weikum G. Yago:a core of semanticknow ledge [C] //Proceedings of the 16th International Conferenceon World Wide Web(WWW), 2007:697-706.
- [14] Lehmann J, Isele R, Jakob M, et al. DBpedia-a large-scale, multilin-gual know ledge base extracted from Wikipedia [J] . Semantic Web, 2015, 6(2):167-195.
- [15] Kangqi Luo, Fengli Lin, Xusheng Luo, Kenny Q.Zhu. 2018-EMNLP-Knowledge Base Question Answering via Encoding of Complex Query Graphs[J].
- [16] 王松,李正钧,杨涛,胡孔法.中医药知识图谱研究现状及发展趋势[J].南京中医药大学学报,2022,38(03):272-278.DOI:10.14148/j.issn.1672-0482.2022.0272.
- [17] 石燕, 何黎, 任秋静, 等. 中医体质知识图谱分析:基于 VOS-viewer 和 CiteSpace 的计量分析 [J] . 世界科学技术—中医药现代化, 2021, 23(9):3415—3423.
- [18] 陈陵芳. 《黄帝内经》病机十九条知识表示研究 [D] . 长沙:湖南中医药大学, 2018.
- [19] 张莹莹. 基于知识图谱的舌像诊疗系统研究与构建 [D] . 成都:电子科技大学, 2019.
- [20] 卢克治. 基于中医古籍的知识图谱构建与应用 [D] . 北京:北京交通大学, 2020.
- [21] 牟梓君. 小儿脑瘫中医诊疗知识图谱构建及其隐性知识显性化研究 [D] . 北京:中国中医科学院, 2021.
- [22] 石英杰. 基于病机模型的胸痹病中医智能辅助诊断方法研究 [D] . 北京:中国中医科学院, 2021.

- [23] 胡嘉元. 病机主导的中医临床个体化诊疗模式及决策支持系统构建 [D]. 北京:北京中医药大学, 2020.
- [24] 郑子强. 面向慢性肾脏病中医医案的知识图谱学习与推理研究 [D]. 成都:电子科技大学, 2020.
- [25] 孙明俊, 张丹, 郑明智, 等. 基于人工智能的类风湿性关节炎中医辅助诊疗系统 [J]. 模式识别与人工智能, 2021, 34(4):343-352.
- [26] 张雨琪, 李宗友, 王映辉, 等. 赵炳南、朱仁康皮肤科流派用方经验知识图谱构建 [J]. 中国中医药图书情报杂志, 2021, 45(2):1-5.
- [27] 刘凡, 王明强, 李凌香, 等. 名老中医临床经验知识图谱构建方法探索 [J]. 中华中医药杂志, 2021, 36(4):2281-2285.
- [28] 于彤, 刘静, 贾李蓉, 等. 大型中医药知识图谱构建研究 [J]. 中国数字医学, 2015, 10(3):80-82.
- [29] 张德政, 谢永红, 李曼, 等. 基于本体的中医知识图谱构建 [J]. 情报工程, 2017, 3(1):35-42.
- [30] 贾李蓉, 刘静, 于彤, 等. 中医药知识图谱构建 [J]. 医学信息学杂志, 2015, 36(8):51-53, 59.
- [31] 阮彤, 孙程琳, 王昊奋, 等. 中医药知识图谱构建与应用 [J]. 医学信息学杂志, 2016, 37(4):8-13.
- [32] 屈倩倩, 阚红星. 基于 Bert-BiLSTM-CRF 的中医文本命名实体识别 [J]. 电子设计工程, 2021, 29(19):40-43, 48.
- [33] 高佳奕, 杨涛, 董海艳, 等. 基于 LSTM-CRF 的中医医案症状命名实体抽取研究 [J]. 中国中医药信息杂志, 2021, 28(5):20-24.
- [34] 王守会. 知识库问答系统研究进展. 1000-1220(2021)09-1793-09.
- [35] Liang P, Jordan M I, Klein D. Learning dependency-based compositional semantics [J]. Computational Linguistics, 2013, 39(2):389-446.
- [36] Diefenbach D, Lopez V, Singh K, et al. Core techniques of question answering systems over knowledge bases: a survey [J]. Knowledge and Information Systems, 2018, 55(3):529-569.
- [37] Bordes A, Usunier N, Chopra S, et al. Large-scale simple question answering with memory networks [J]. arXiv preprint arXiv:1506.02075, 2015.
- [38] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013:1533-1544.
- [39] Bao Jun-wei, Duan Nan, Yan Zhao, et al. Constraint-based question answering with knowledge graph [C]//Proceedings of COLING2016, the 26th International Conference on Computational Linguistics: Technical Papers (COLING), 2016:2503-2514.
- [40] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models [C]//30th USENIX Security Symposium (USENIX Security 21). 2021: 2633-2650.
- [41] Zhao W X, Zhou K, Li J, et al. A survey of large language models [OL]. arXiv preprint arXiv:2303.18223, 2023.
- [42] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [J]. 2018.

- [43] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
- [44] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1 877-1901.
- [45] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[OL]. arXiv preprint arXiv:2206.07682, 2022.
- [46] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[OL]. arXiv preprint arXiv:2204.023 1 1, 2022.
- [47] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models[OL]. arXiv preprint arXiv:2210.1 1416, 2022.
- [48] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[OL]. arXiv preprint arXiv:2302.13971, 2023.
- [49] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[OL]. arXiv preprint arXiv:2307.09288, 2023.
- [50] Penedo G, Malartic Q, Hesslow D, et al. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only[OL]. arXiv preprint arXiv:2306.01 1 16, 2023.