Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(3), 684-697
Published Online May 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.143067

Hans汉斯

大语言模型融合知识图谱的装备问答系统研究

王美华[1]，王兴芬[2*]，张友星[1]

[1]北京信息科技大学计算机学院，北京
[2]北京信息科技大学管理科学与工程学院，北京

收稿日期：2025年2月27日；录用日期：2025年5月21日；发布日期：2025年5月30日

摘　要

在大数据时代，海量的互联网信息飞速增长，人们对信息获取的精准度与效率提出了更高的要求。随着企业信息化和装备管理现代化的不断推进，对海量企业装备信息进行有效的提炼、管理与利用，对于提升企业装备知识的应用价值以及企业资源的利用效率具有重要意义。本研究提出了一套融合大语言模型自然语言处理能力的系统，可智能理解用户查询并提供精准的装备信息。通过采用P-Tuning v2方法对大语言模型进行微调，大幅提升了其在企业装备领域对关键词的识别和提取能力。同时，借助企业装备知识图谱作为本地知识库，为模型提供行业领域知识，使其能够将相关信息作为问题的上下文进行学习。基于此，还设计了提示工程来引导模型生成更准确的回复，并对结果进行了效果评估。实验结果表明，相较于直接使用大语言模型，该基于知识图谱增强的大语言模型在企业装备领域的智能化回复准确率更高，为企业装备问答系统的建设提供了有力支持。

关键词

大语言模型，知识图谱，P-Tuning v2方法，企业装备

Research on Equipment Question-Answering System Integrating Large Language Models with Knowledge Graphs

Meihua Wang[1], Xingfen Wang[2*], Youxing Zhang[1]

[1]Computer School, Beijing Information Science and Technology University, Beijing
[2]School of Management Science and Engineering, Beijing Information Science and Technology University, Beijing

Received: Feb. 27th, 2025; accepted: May 21st, 2025; published: May 30th, 2025

[*]通讯作者。

文章引用: 王美华, 王兴芬, 张友星. 大语言模型融合知识图谱的装备问答系统研究[J]. 人工智能与机器人研究, 2025, 14(3): 684-697. DOI: 10.12677/airr.2025.143067

Abstract

In the era of big data, the volume of Internet information is growing at an astonishing rate, and people have put forward higher requirements for the accuracy and efficiency of information acquisition. With the continuous advancement of enterprise informatization and modernization of equipment management, effectively extracting, managing and utilizing massive enterprise equipment information is of great significance for enhancing the application value of enterprise equipment knowledge and improving the efficiency of enterprise resource utilization. This study proposes a system that integrates the natural language processing capabilities of large language models, which can intelligently understand user queries and provide precise equipment information. By using the P-Tuning v2 method to fine-tune the large language model, its ability to recognize and extract keywords in the field of enterprise equipment has been significantly enhanced. At the same time, with the help of the enterprise equipment knowledge graph as a local knowledge base, industry-specific knowledge is provided to the model, enabling it to learn relevant information in the context of the question. Based on this, prompt engineering is designed to guide the model to generate more accurate responses, and the results are evaluated. Experimental results show that compared with directly using large language models, the knowledge graph-enhanced large language model has a higher accuracy rate in intelligent responses in the field of enterprise equipment, providing strong support for the construction of enterprise equipment question-answering systems.

Keywords

Large Language Model, Knowledge Graph, P-Tuning v2 Method, Enterprise Equipment

1. 引言

问答系统(Question Answering, QA)是一种能够自动响应自然语言查询的技术，融合了信息检索与自然语言处理两个领域的研究成果。将知识图谱(Knowledge Graph, KG)与问答系统融合，正确理解用户语义是一大挑战[1]。虽然知识图谱问答能够通过对问题进行分析理解，最终获取答案，但面对自然语言的灵活性与模糊性，如何处理复杂问题的语义信息，如何提高复杂推理问答的高效性仍是研究难点[2]。

近年来，大语言模型(Large Language Model, LLM)取得了长足的进步，许多高效的模型相继被提出，如 ChatGPT [3]、LLaMA [4]和中文模型 BaiChuan [5]等。这些模型在各个领域都展现出出色的表现力，充分证明了它们的广泛表达能力。针对垂直领域，如果在基础模型之上结合领域专用数据集进行预训练与微调，就能为模型赋予该领域的对话能力。然而，尽管经过微调，由于缺乏大规模专业知识的支撑，大语言模型在垂直领域对专业知识的理解仍显不足，其表达能力因此受到明显限制。

在专业领域，大模型有时会生成错误的回答或推理，甚至出现所谓的“幻觉”现象。此外，即使经过微调训练，模型也可能发生“遗忘”，丧失部分原本具备的对话生成能力。在企业装备等关键领域，对模型的准确性和安全性要求远高于普通应用场景。因此，如何在保留大模型既有对话能力的前提下，提高其在专业领域回答的准确性和可靠性，已成为大模型在垂直领域应用中亟需解决的关键挑战。

随着信息技术的快速发展，许多企业装备数据库因结构松散而难以充分利用，导致管理效率低下、数据混乱等问题。基于知识图谱的企业装备问答系统为此提供了有效的解决方案：该系统通过自然语言

理解模块解析用户的口语化短句，并利用知识图谱检索相应答案，能够实时、准确地满足用户查询需求。然而，由于此类系统面对的用户输入往往简短且口语化，给意图识别和实体链接等自然语言理解环节带来巨大挑战。本研究针对企业装备管理领域，构建了专业的知识图谱系统，旨在实现领域知识的系统化整合与存储。研究采用双轨策略：首先，通过对企业装备语料库的深度处理，运用 P-Tuning v2 技术对大语言模型进行领域适应性微调；其次，将知识图谱的语义网络优势与大语言模型的推理对话能力相结合，构建了面向企业装备管理的智能问答平台。研究还设计了相应的评估体系，对知识图谱增强的大语言模型进行了系统性评估，以验证该方法的可行性与应用价值。

2. 相关研究

2.1. 装备知识图谱研究现状

知识图谱最早由 Google 在 2012 年提出[6]，随着人工智能和语义网[7]的不断发展，知识图谱作为一种重要的数据处理技术和数据表示格式[8]，近年来的研究不断进步，其在理解和应用复杂领域知识方面展现出了独特优势。知识图谱作为一种语义网络结构，采用“实体–关系–实体”的三元组形式组织知识，其中节点表示实体，边表示实体间的语义关联，从而构建起结构化的知识网络。和传统的数据存储方法相比，知识图谱能够使用户很好地分析数据之间的联系，并具有语义推断能力，基于知识图谱的知识问答[9]、知识搜索以及数据分析等已经广泛应用到医疗[10]、电子商务、金融等领域。

关于复杂装备体系的研究可追溯到 20 世纪 90 年代，核心任务在于探讨装备内部各子系统之间的相互关联与协同方式，即研究功能上互联、性能上互补、操作上互通的装备在何种结构下实现有效集成[11]。在领域知识图谱与问答系统的相关研究中，国内外学者已提出多种解决方案。比如，张克亮等人[12]设计了一个面向航空领域的问答系统，他们将用户问题分为若干类，并采用结构化的语义信息提取方法将问句转换为 SPARQL 查询语句进行检索。窦小强等人[13]则介绍了一个基于模板的智能问答系统，通过对用户问题进行分词，再将其匹配到预先构建的问题模板中，为每一类问题生成相应的知识图谱查询方法，最终给出答案。然而，该研究尚未提供可实现系统关键技术的具体细节。车金立等人[14]提出了一种领域装备知识图谱的构建方法，通过网络爬虫获取百科数据，并对获取的数据进行知识抽取、融合后存储到 Neo4j 图数据库中，并持续更新修正。在技术实现层面，研究采用词典分词与模板匹配相结合的方法，将自然语言查询转换为 Neo4j 的 Cypher 查询语句进行答案检索，但在实体消歧和口语化查询处理等方面仍有待深入探索。姜成樾[15]则提出了基于依存树的问题理解模板和语义化问句解析模型，将复杂问句拆分成多个简单三元组，并生成带有多重约束的问句元组集合，通过结构化查询给出回复。阚实松[16]在结构化数据与半结构化数据的基础上，构建了飞行器领域知识图谱，并结合神经网络方法实现了知识图谱问答系统；在问题理解阶段，他使用了朴素贝叶斯分类器，并尝试利用 doc2vec 与 TextCNN 进行问句表示。刘天雅[17]研究并设计了一种由“问题语义图”和“问题意图”组成的问句表达模型，探讨通过知识库结构及问题的先验信息来减少歧义，以辅助生成更准确的模型表达。李代祎等人[18]通过 SVM 多分类器进行问句分类和模板匹配，并利用 BiLSTM-CRF 模型完成命名实体识别，最后将识别出的实体和关系填充进问句模板生成 Cypher 查询语句进行检索。该方法需要人工构建大量问答数据来支持模型训练，随着数据规模扩大，数据标注和模型调参的工作量会显著增加，不利于系统的快速迭代。综上所述，现有研究表明，知识图谱与自然语言处理技术的结合能够在装备问答领域发挥显著作用，但在处理口语化短句、进行实体消歧以及提高系统对复杂查询的适应性等方面仍需进一步探索和完善。本文在此基础上，将企业装备知识图谱与大语言模型相融合，以期在企业装备管理与问答的场景下取得更准确、高效的查询与推理效果。

2.2. 大语言模型增强的知识图谱问答研究现状

随着大语言模型的快速发展，其在语义理解和文本生成方面的卓越能力为知识图谱问答(KGQA)带来了新的突破。目前，基于大语言模型的 KGQA 方法主要可分为两大类别：基于语义解析的方法和基于信息检索的方法。

语义解析技术的核心在于将输入问题翻译为能在图上执行的查询语句。为了解决传统方法检索效率低下的问题，Luo 等人[19]提出了基于微调大语言模型的生成检索框架 ChatKBOA。Li 等人[20]同样利用微调的大语言模型来进行逻辑查询语句的转换，提出一个结合两条平行工作流的统一框架 UniOOA。Taffa 等人[21]提出一种少样本提示逻辑表达式生成方法。该方法采用基于 BERT 的句子编码器，通过识别与测试问题最相关的 top-n 训练问题，检索其对应的 SPARQL 查询。随后，将这些相似问题的 SPARQL 查询作为示例，与测试问题共同构建提示模板，输入大语言模型以生成 SPARQL 查询语句。最终，在开放知识图谱上执行生成的查询并返回结果。Chen 等人[22]提出了基于大语言模型的 COL 生成框架。该方法采用多模块协同的工作机制：辅助任务模型预测 COL 的结构信息，专有名词匹配器提取问题中的实体和关系，示例选择器基于关键信息相似度筛选适配示例，提示构造器整合示例、问题及先验知识形成提示文本。大语言模型基于这些输入生成 COL，集成模型通过多答案投票机制提升结果准确性。Jiang 等人[23]提出的 ReasoningLM 框架，将知识图谱的子图序列化，通过将大语言模型与子图感知的自注意力机制结合，利用自注意力机制模拟图神经网络的推理能力，使得大语言模型能够感知知识图谱中的实体和关系，生成可执行的推理路径。

基于信息检索(Information Retrieval, IR)的 KGQA 方法是一种利用信息检索技术，通过自动化处理自然语言问题，并从知识图谱中定位和提取相关信息来提供答案的方法。Tan 等人[24]网提出了一个新框架 McL-KBQA，使用基于排名的 KBOA 方法枚举候选逻辑形式。Kim 等人[25]提出了一个名为 KG-GPT 的框架。该研究提出的框架采用三阶段处理机制：首先，在语句解析阶段，系统将输入的自然语言问题分解为与知识图谱中单个三元组相对应的独立子句，这种分解策略有效降低了多跳推理的复杂度。其次，在图谱匹配阶段，框架为每个子句识别并检索相应的语义关系，同时构建包含所有已识别实体的候选证据网络。最后，在逻辑推理阶段，系统利用构建的证据网络进行演绎推理，从而实现对给定命题的验证或问题的解答。Wu 等人[26]提出了 Retrieve-Rewrite-Answer 框架。该框架处理流程包含三个核心环节：子图定位、知识图谱到文本的转换以及知识增强推理。具体而言，该方法首先通过分析问题的语义特征预测所需的推理步数和关系路径，据此从知识图谱中定位相关子图结构。随后，对子图中的三元组进行选择性采样，生成与问题语义相匹配的结构化知识表示。在此过程中，系统充分利用大语言模型的语义理解能力，确保子图定位的精确性和结果的可解释性。这种多阶段处理机制不仅提高了知识检索的效率，还增强了系统对复杂问题的处理能力。Baek 等人[27]提出了 KAPING (Knowledge Augmented Language Model PromptING)框架，该框架通过有效的子图生成和语义筛选机制，确保注入的知识高度相关，从而提升推理精度并减少计算开销。Sun 等人[28]提出了 ToG (Think-on-Graph)框架，它将大语言模型与知识图谱紧密结合，以实现深度和负责任的推理。Dong 等人[29]的则提出了 EQA (Efficient Question-Answering)框架，引入大语言模型辅助知识图谱问答的推理过程。EQA 方法充分发挥大语言模型的语义理解和任务分解能力，将复杂问题分解为具有逻辑关联的子问题集，形成初始推理路径，覆盖问题的核心逻辑。

当前知识图谱增强大语言模型的研究呈现多元化技术路线，本文从知识融合机制、交互策略和领域适配性三个层面与主流方法展开对比，凸显本研究的创新价值。在知识融合层面，现有方法可分为直接注入式、检索增强式和协同推理式三大类。直接注入方法通过序列化知识子图输入模型，虽能保留图谱结构信息，但易受无关知识干扰且计算成本较高；检索增强方法依赖语义匹配动态筛选知识片段，虽提

升知识相关性，但对检索模块精度要求苛刻；协同推理方法通过图遍历与模型推理交替迭代，虽能处理复杂逻辑，但领域迁移成本显著。本文创新性地提出两阶段优化框架：第一阶段通过参数高效微调强化模型对领域语义的理解能力，第二阶段设计结构化提示模板将知识图谱检索结果动态整合生成上下文。这种“微调 + 检索”的协同机制，既避免了知识噪声干扰，又通过提示工程引导模型建立问题与知识的深度关联，在保证推理效率的同时降低了对高精度检索组件的依赖。

在交互策略方面，现有研究普遍面临知识整合粒度与计算效率的权衡难题。单阶段端到端方法直接将图谱信息拼接至输入序列，导致模型难以区分问题语义与补充知识；多阶段流水线方法通过任务分解提升可扩展性，但模块间的信息损失可能影响最终推理连贯性。本文提出的三段式提示结构(问题描述－知识上下文－回答要求)创新性地将知识图谱的结构化特征与大语言模型的生成能力相结合。通过层级化组织检索结果，模型能够自主识别关键实体关系，并在生成过程中动态调整注意力分布。实验表明，这种策略在处理装备参数查询、多跳推理等复杂任务时，较传统方法展现出更强的语义连贯性和事实一致性。

领域适应性方面，现有方法在垂直场景落地时普遍面临两大挑战：专业术语歧义消解和领域知识动态更新。通用增强方法依赖大规模预训练数据，难以捕捉装备领域细粒度特征；全参数微调方案虽能提升领域表现，但会导致模型“灾难性遗忘”通用能力。本文的领域适配方案通过参数高效微调技术，在极低参数增量下实现专业术语识别能力的显著提升，同时结合知识图谱的动态检索机制，有效解决术语歧义和新知识整合问题。相较于需要重构模型架构或全参数微调的方法，本文方案在计算成本和领域迁移效率上具有明显优势，为垂直领域知识问答提供了一种轻量化适配范式。

3. 方案设计

3.1. 装备知识图谱的构建

在知识图谱构建方法方面，主要存在自底向上和自顶向下两种范式。自底向上方法从具体数据出发，通过知识抽取逐步构建图谱，适用于通用领域；自顶向下方法则基于高层次概念，先构建框架再填充数据，更适合特定领域。本研究针对企业装备领域的特点，采用自顶向下方法：首先定义领域本体(包括实体类型、关系类型和属性等)，然后从相关数据源获取数据，通过匹配映射将实体与关系纳入预定义本体，最终形成装备知识图谱。其中，领域本体的合理设计是构建高质量知识图谱的关键。

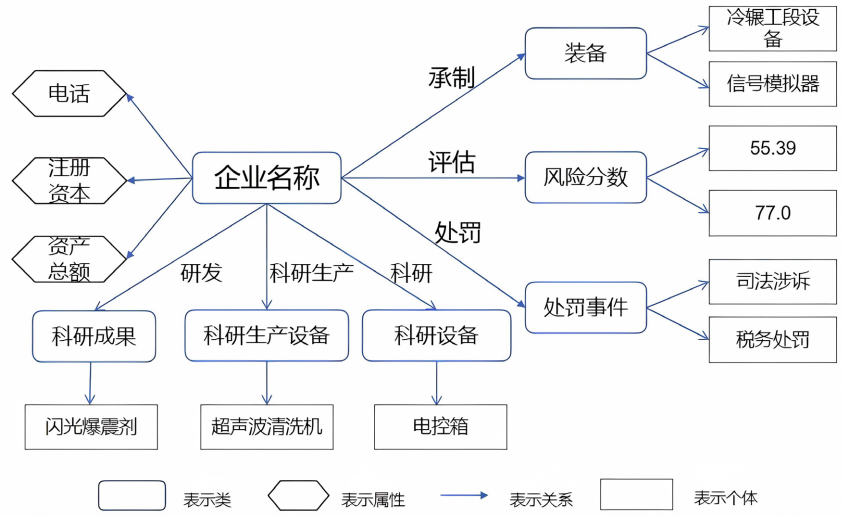

Figure 1. Equipment knowledge graph ontology model
图 1. 装备知识图谱本体模型

在复杂的企业装备领域中，设计合理的领域本体是构建知识图谱的关键一步。针对装备问答的应用场景，本文在参考现有装备本体和相关标准的基础上，将实体分为七类，实体关系分为六类，并定义了五类属性。装备知识图谱的本体模型如图 1 所示。基于所整理的装备数据集，利用 Python 调用 py2neo 库连接 Neo4j 数据库，自动化地执行构建知识图谱的脚本语句，从而在 Neo4j 数据库中形成最终的装备知识图谱结构，示例如图 2 所示。

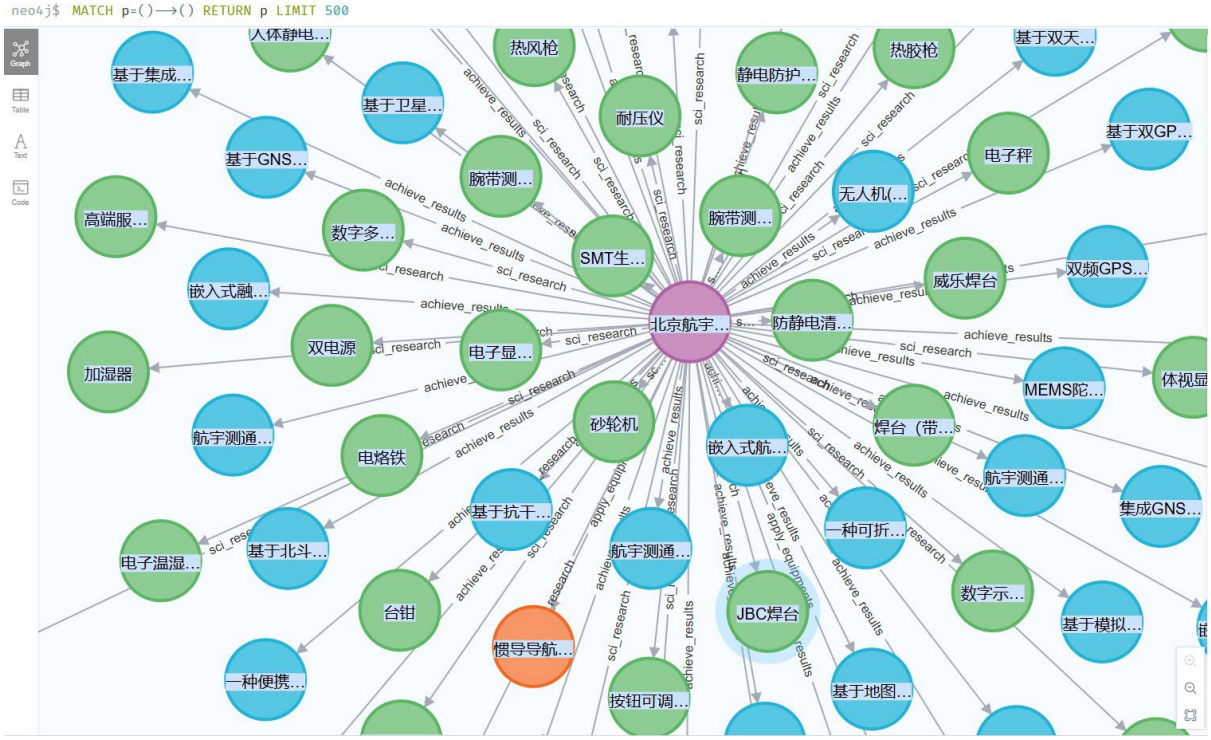

Figure 2. Equipment knowledge graph
图 2. 装备知识图谱

3.2. 大语言模型的选择

随着大规模语言模型的持续发展，全球众多研究机构正致力于改进这些模型，并发布了多个开源版本。然而，由于计算资源的限制，参数量庞大的语言模型在实际部署和调试过程中常常遇到效率问题。为此，学术界和工业界正在积极探索如何在资源有限的环境中优化这些开源的大规模语言模型，并促进其实际应用。表 1 列出了几种常见的适用于低资源环境的开源大规模语言模型，供读者参考。

Table 1. Some common low-resource open-source large language models
表 1. 部分常见低资源开源大语言模型

模型名称	发布单位	模型规模
Llama-2-7B-32k	Meta	支持的上下文长度(context length)为 32,000 个 token
ChatGLM2-6B	清华大学	1.4 TB 中英标识符约在 1.2 万亿 tokens 的语料上进行训练
MPT-7B	MosaicML	数据规模约为 1 万亿 tokens
Alpaca-7B	Stanford University	原始训练数据规模约为 1.4 万亿 tokens
Bloom-7B	BigScience	训练数据规模约为 1.6 万亿 tokens

在表 1 中，Llama-2-7B-32k 由 Meta 公司发布，是其 LLaMA 系列模型的一部分。Meta 致力于推动人工智能领域的开源发展，LLaMA 系列模型是其在大语言模型领域的重要贡献之一。相比于更大的模型(如千亿参数模型)，7B 规模的模型在计算资源需求上更为友好，适合在低资源环境中部署。Llama-2-7B-32k 作为中等规模模型，Llama-2-7B-32k 可以在需要时进一步微调或扩展，以适应特定任务的需求。但是 Llama-2-7B-32k 作为一个以英文为主要训练数据的大语言模型，在中文任务上的表现可能相对有限。

ChatGLM2-6B 是清华大学发布的第二代中文和英文双语对话预训练大语言模型。它是 ChatGLM-6B 的升级版本，继承并优化了前一代模型的架构和能力，特别在中文对话生成和模型性能方面进行了显著提升。该版本的模型进行了多项性能优化，推理速度比前一代提升了约 42%，可以更高效地处理用户请求。ChatGLM2-6B 在上下文长度上有所扩展，从 2000 tokens 扩展到了 32,000 tokens，这使得它能够更好地理解和生成长对话内容。

MPT-7B 是由 MosaicML 发布的开源大语言模型，属于 MPT (MosaicML Pretrained Transformer)系列的一部分。MPT-7B 以其高效性和灵活性在开源社区中受到了广泛关注，特别适合在低资源环境中部署。但 MPT-7B 作为一个以英文为主要训练数据的大语言模型，在中文任务上的表现可能相对有限。

Alpaca-7B 是由斯坦福大学(Stanford University)发布的开源大语言模型，基于 Meta 的 LLaMA-7B 模型进行指令微调(Instruction Tuning)。Alpaca-7B 专注于指令跟随任务，旨在提供一个高效、轻量且易于使用的模型。Alpaca-7B 专注于指令跟随任务，能够准确理解并执行用户指令，适合构建对话系统和任务型应用。但其微调数据主要基于英文指令，可能导致模型在非英文任务(如中文)上的表现不够理想。

Bloom-7B 由 BigScience 项目发布。BigScience 是一个由全球研究机构、高校和企业共同参与的大型协作项目，参与者包括 Hugging Face、法国国家科学研究中心(CNRS)等。该项目致力于推动多语言大语言模型的研究和应用。Bloom-7B 支持 46 种语言，包括中文、英语、法语、西班牙语等，适合多语言任务。Bloom-7B 作为一个多语言模型，在中文任务上的表现优于许多以英文为主的模型，但在特定领域上，模型的表现可能仍需进一步提升。

在实验设计方面，综合考虑中文问答效果和硬件配置要求，研究选用 ChatGLM2-6B 作为基础大语言模型。

3.3. P-Tuning v2 微调大语言模型

P-Tuning v2 是一种基于预训练模型的微调方法，其基本原理是在预训练模型的基础上，通过添加少量的可训练参数，对模型的输出进行微调，如图 3 所示。在模型优化层面，采用前缀调优技术：在语言模型的每一层添加 l 个可训练的注意力键和值嵌入，将原始键向量 $K \in Rl \times d$ 和值向量 $V \in Rl \times d$ 分别与可训练向量 Pk、Pv 连接，从而改进注意力机制的计算方式：

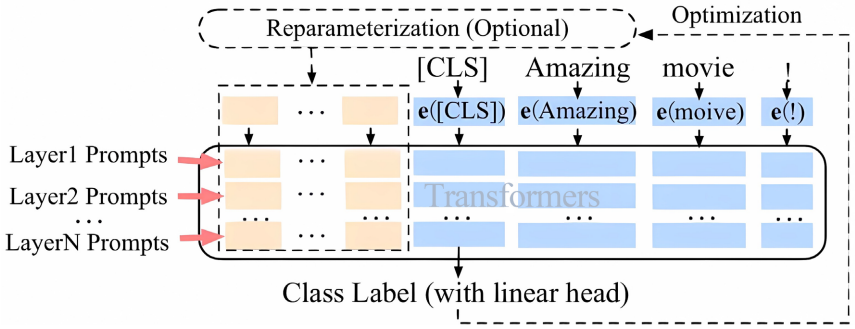

Figure 3. P-Tuning v2 basic principles
图 3. P-Tuning v2 基本原理

$$\text{head}_i(x) = \text{Attention}\left(xW(i), \left[Pk(i) : K(i)\right], \left[Pv(i) : V(i)\right]\right)$$

其中，下标 i 代表向量中与第i个注意力头对应的部分，通过这种方法来微调大型语言模型。

在大语言模型领域适配技术中，除 P-Tuning v2 外，LoRA (低秩自适应)和 Adapter Tuning (适配器调优)是两类主流参数高效微调方法。LoRA 通过低秩矩阵分解注入可训练参数，在通用领域任务中表现出良好的参数－性能平衡，但其线性特征重构机制对装备领域复杂的非线性关系建模能力有限。Adapter Tuning 通过在 Transformer 层间插入小型神经网络实现领域适配，虽能较好保持模型通用能力，但层级适配器的串行结构会引入额外推理延迟，难以满足实时问答场景的响应要求。

相较之下，P-Tuning v2 的创新性体现在三方面：其一，采用层级前缀调优策略，在每层注意力机制中注入可训练的前缀键值向量，通过跨层知识传递增强模型对领域特征的捕获能力；其二，引入动态提示编码器，利用序列模型自动学习最优提示表示，显著提升少样本场景下的微调效果；其三，通过极低参数增量实现领域知识的高效注入，在保持模型通用对话能力的同时，避免传统微调方法导致的“灾难性遗忘”问题。实验表明，该方法在装备领域问答任务中，既能达到接近全参数微调的精度，又大幅降低训练资源消耗，为垂直领域应用提供了更优的性价比选择。

这种技术优势源于其独特的参数更新机制：P-Tuning v2 仅在注意力空间调整特征分布，而非直接修改模型权重。这种“特征重组”模式既保留了预训练阶段习得的通用语言理解能力，又通过注意力聚焦强化领域特定模式的学习。相较于需要修改前向传播路径的 Adapter，或受限于低秩假设的 LoRA，P-Tuning v2 在装备领域复杂语义关系建模中展现出更强的适应性，尤其是在处理装备型号关联性分析、技术参数对比等需要深层语义推理的任务时，生成结果具有更高的事实准确性和逻辑一致性。

3.4. 基于装备知识图谱增强的大语言模型

尽管本文已对大语言模型进行了 P-Tuning v2 微调以增强其在装备领域的表现，然而大语言模型仍可能产生幻觉问题[30]，即模型会遗忘已训练的事实或知识，导致其给出错误的回答，从而降低模型的可靠性，影响其在装备领域的实际应用。

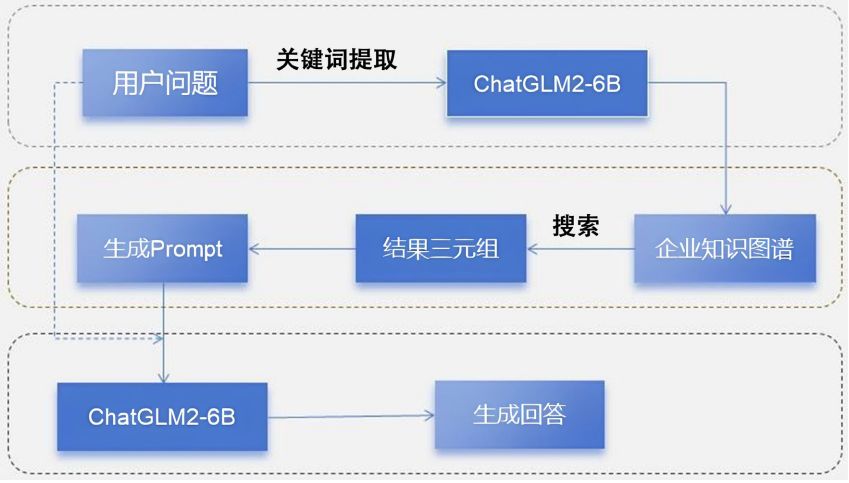

Figure 4. Overall flow chart
图 4. 整体流程图

针对上述问题，本研究设计了一种创新性的解决方案：通过构建装备领域的专用知识图谱，并将其与大语言模型深度融合，从而显著提升模型回答的准确性和可靠性。知识图谱采用三元组的形式对领域

知识进行结构化存储，不仅能够高效组织海量专业知识，还支持动态更新机制，可及时整合用户提供的新知识。具体实施过程分为两个主要阶段：首先，运用 P-Tuning v2 技术对大语言模型进行领域适应性微调，增强其对装备领域专业术语和命名实体的识别能力；其次，系统梳理装备领域的专业知识体系，构建结构化的装备知识图谱。

在实际应用过程中，当用户提交查询时，系统首先利用微调后的大语言模型进行关键词提取。随后，基于提取的关键语义信息生成 Neo4j 图数据库的查询语句，在装备知识图谱中进行精准检索。根据检索结果，系统构建特定的提示模板，该模板旨在引导大语言模型结合知识图谱的检索结果生成回答。具体而言，系统将检索结果嵌入提示模板的特定位置，形成结构化的上下文信息。这种提示工程策略将知识图谱的检索结果与用户原始问题有机结合，共同作为大语言模型的输入。最后，系统对模型生成的初步回答进行后处理，输出最终的用户答案。图 4 展示了该方案的整体实现流程。

4. 实验与分析

4.1. 数据集

本研究为模型微调构建了包含两类任务的训练数据集：命名实体识别任务和装备业务知识问答任务，共计 1500 条样本数据。这些数据严格遵循 P-Tuning v2 微调大模型的标准格式进行组织和标注，具体的数据集结构示例如表 2 所示。

Table 2. Model fine-tuning training dataset
表 2. 模型微调训练数据集

数据集类型	类型示例
命名实体识别	{"instruction": "命名实体识别任务：请帮我识别这句话中的实体，请按照实体 1-实体 2-实体 3 的格式输出。", "input": "北京航宇测通电子科技有限公司能够承制哪些装备？", "output": "北京航宇测通电子科技有限公司"}
装备知识问答	{"instruction": "您现在是装备问答助手，请回答相关问题。","input": "盐城市训保军训器材有限公司科研生产设备有哪些？", "output": "盐城市训保军训器材有限公司科研生产设备包括：剪板机","数控车床","立式注塑机","电焊机","线切割","摇臂钻床"。"}

4.2. 实验环境与实验设置

本文实验所使用的硬件配置包括了一台配备 NVIDIA A10 GPU 的高性能计算机。该 GPU 拥有 24 GB 显存，能够支持大规模神经网络的训练和推理任务，尤其是在处理大语言模型微调时提供了必要的计算资源和显存支持。针对实验环境的具体配置，我们使用了 Python 3.10 作为编程语言的版本，能够充分利用现代 Python 库和工具的性能，确保程序的高效执行。

Table 3. Fine-tuning parameter settings for large models
表 3. 大模型微调参数设置

参数	中文含义	参数值
per_device_train_batch_size	每个设备在训练时使用的数据批次大小	4
per_device_eval_batch_size	每个设备在评估时使用的数据批次大小	4
max_steps	模型训练的最大步骤数，决定训练轮数	5000
save_steps	指定每隔多少步骤保存一次模型	1000
learning_rate	控制参数更新速度的学习率	1e−3
weight_decay	权重衰减	0.01

在 CUDA 版本方面，实验采用了 CUDA 12.1，这一版本优化了多种计算任务的并行执行能力，并提高了 GPU 的计算效率，从而能够加速大模型的训练过程。该配置允许充分发挥 GPU 的性能，减少了模型训练和推理过程中的时间消耗，保证了大语言模型微调任务的顺利进行。

对于大模型的微调，我们根据实验的实际需求，设置了相应的训练参数，这些参数设置的详细信息可以参见表 3。这些微调参数包括了学习率、批次大小、优化器配置、训练轮次等关键因素，旨在确保大语言模型在特定任务和领域下能够获得最佳的性能。通过精细调整这些参数，我们能够在有限的计算资源下，达到最优的训练效果。

4.3. 实验结果与分析

在模型评估阶段，将微调获得的参数配置应用于大语言模型，并采用包含 100 个样本的测试集对微调前后的模型性能进行对比分析。评估体系选用了 BLEU-4、ROUGE-1、ROUGE-2 和 ROUGE-L 四个核心指标。其中，BLEU-4 指标虽然最初用于机器翻译系统的性能评估，但本研究将其创新性地应用于衡量大语言模型生成答案与标准答案之间的语义相似度，为模型性能评估提供了新的视角，计算方法如公式所示：

$$\text{BLEU-4} = BP \times \left(\sum_{n=1}^{4} w_n \log p_n \right)$$

式中，BP 是长度惩罚因子，用于惩罚过短的翻译结果，p_n 是 n-gram 的精确度，表示机器翻译中与参考翻译匹配的 n-gram 比例，w_n 是 n-gram 的权重，通常取 1/4。

长度惩罚因子 BP 的计算公式如下：

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \le r \end{cases}$$

式中，c 是机器翻译的长度，r 是最接近 c 的参考翻译长度。

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)是一种用于评估自动摘要生成质量的自动评估指标，由 Chin-Yew Lin 在 2004 年提出。ROUGE-1 是 ROUGE 系列中最基础的指标，专注于 1-gram (单个词)的匹配情况，通过计算生成摘要与参考摘要之间的词汇重叠度来评估质量。ROUGE-1 的核心思想是基于召回率(Recall)，即生成摘要中有多少词汇与参考摘要中的词汇匹配。它也可以结合精确率(Precision)和 F1 值来综合评估。

ROUGE-1 指标主要用于评估生成文本与参考文本在词汇层面的匹配程度，通过计算两者之间单个词语的重叠比例来反映其相似性。在本研究中，该指标被用于量化微调后的大语言模型生成答案与标准答案之间的词汇重合度。其计算公式如下：

$$\text{ROUGE-1} = \frac{\sum_{S \in \{\text{GenerationTexts}\} \cap \{\text{ReferenceTexts}\}} \sum_{\text{garm}_1} \text{Count}_{\text{match}} (\text{gram}_1)}{\sum_{S \in \{\text{GenerationTexts}\}} \sum_{\text{garm}_1} \text{Count}_{\text{match}} (\text{gram}_1)}$$

式中，分子表示生成文本与参考文本共现的 1-gram 数量，分母则为参考文本中 1-gram 的总数。

ROUGE-2 指标则进一步考察生成文本与参考文本在词组层面的匹配情况，通过分析连续二元词组的重叠程度来评估文本生成质量。本研究采用该指标来衡量微调后的大语言模型生成答案与标准答案在词组级别上的匹配精度。其计算公式如下：

$$\text{ROUGE-2} = \frac{\sum_{S \in \{\text{GenerationTexts}\} \cap \{\text{ReferenceTexts}\}} \sum_{\text{garm}_1} \text{Count}_{\text{match}} (\text{gram}_2)}{\sum_{S \in \{\text{GenerationTexts}\}} \sum_{\text{garm}_2} \text{Count}_{\text{match}} (\text{gram}_2)}$$

式中，分子表示生成文本与参考文本共现的 2-gram 数量，分母则为参考文本中 2-gram 的总数。

ROUGE-L 指标采用最长公共子序列(LCS)算法，通过计算生成文本与参考文本之间最长的连续匹配序列来评估两者在语义结构和内容组织上的相似度。该指标能够有效捕捉文本间的语义连贯性和逻辑一致性，为模型性能评估提供更深层次的洞察，即：

$$\text{ROUGE-L} = \frac{\left(1+\beta^2\right) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

其中，R_{lcs} 表示最长公共子序列的召回率，P_{lcs} 表示最长公共子序列的准确率，β 为可调节参数，R_{lcs} 和 P_{lcs} 的具体计算公式分别为：

$$R_{lcs} = \frac{\text{LCS}(X,Y)}{\text{len}(Y)}$$

$$P_{lcs} = \frac{\text{LCS}(X,Y)}{\text{len}(X)}$$

在上述公式中，X 代表参考文本，Y 代表生成文本，LCS(X, Y)表示 X 和 Y 的最长公共子序列的长度，len(X)和 len(Y)分别表示参考文本和生成文本的长度。通过这一系列计算，ROUGE-L 能够全面评估生成文本与参考文本在语义和结构上的匹配程度。

本研究采用 BLEU-4、ROUGE-1、ROUGE-2 和 ROUGE-L 四种评估指标，对微调后的大语言模型与原始模型进行了全面的性能对比分析，具体评估结果详见表 4。

Table 4. Comparison of evaluation indicators
表 4. 评价指标对比

评价指标	原始大语言模型	微调大语言模型
BLEU-4	13.7852	82.3775
ROUGE-1	36.3674	87.5583
ROUGE-2	20.8832	88.0367
ROUGE-L	27.3054	87.2775

实验结果表明，基于企业装备训练数据集对大语言模型进行微调后，其在该领域的问答性能得到了明显改善。从四项评价指标的测试结果来看，经过微调的大语言模型在所有指标上均超越了原始模型，这充分证明了采用 P-Tuning v2 方法对大语言模型进行微调的有效性。

在本研究中，经过 P-Tuning v2 微调优化的大语言模型被命名为 ChatGLM2-6B-PTuningv2。实验结果表明，该模型在企业装备领域的问答性能得到了显著提升，但在涉及具体数值的问题上(如装备的最高运行速度)，仍存在生成不准确或虚构数据的现象。为此，本文建议引入装备知识图谱作为外部知识源，为大语言模型提供精准的知识支持，从而进一步提升其在企业装备领域的回答准确性。

4.4. 问答分析

为构建面向企业装备领域的智能问答系统，本研究创新性地将领域知识图谱与大语言模型进行深度融合，实现了两者的优势互补。为此，研究设计了一系列典型问题，分别向原始大语言模型、微调后的大语言模型以及知识图谱增强的大语言模型提问，并对三者的回答进行对比分析。为进一步验证知识图谱增强大语言模型的有效性，研究团队构建了一个包含 50 组问答对的测试集，并采用 BLEU-4 指标对三

种不同方法的性能进行系统性评估。具体的问答测试结果详见表 5，这些数据充分证明了知识图谱增强策略的可行性和优越性。

Table 5. Question and answer test analysis
表 5. 问答测试分析

问题	问题答案	原始大语言模型	微调大语言模型	知识图谱增强语言模型
福建毫米电子有限公司承制装备有哪些？	承制装备有：特种元器件、通用元件	福建毫米电子有限公司主要承制通信设备、电子元器件、信息系统、测量仪器、智能电子设备等。具体产品包括手机、调制解调器、传感器等。	福建毫米电子有限公司承制装备有：信号接收、传输装备、雷达对抗装备、微波、毫米波组件	福建毫米电子有限公司承制装备有：特种元器件、通用元件
北京星光凯明智能科技有限公司能够承制哪些装备？	承制装备有：驱逐舰、无人船、训练模拟器	北京星光凯明智能科技有限公司主要承制智能家居、安防、交通、工业自动化等产品和解决方案。具体包括智能音响、智能门锁、智能交通信号灯等。	北京星光凯明智能科技有限公司能够承制装备有：机电设备、机械及零部件	北京星光凯明智能科技有限公司能够承制装备有：驱逐舰、无人船、训练模拟器
廊坊维进金属制品有限公司科研设备有哪些？	科研设备：三维激光切割机、五金插件机、剪板机、螺柱焊机	该公司主要从事金属制品的生产和销售，其产品广泛应用于建筑、交通、机械等行业。如需了解该公司的科研设备详情，建议直接联系廊坊维进金属制品有限公司获取最新信息。	廊坊维进金属制品有限公司科研设备有：领创激光切割机、不锈钢焊机、加工中心、外圆磨床	廊坊维进金属制品有限公司科研设备有：三维激光切割机、五金插件机、剪板机、螺柱焊机

从表 5 的结果可以看出，原始大语言模型的输出与实际答案存在较大差距，效果不佳。经过 P-Tuning 微调后，模型的回答更接近标准答案，但仍存在一定误差。实验结果表明，采用知识图谱增强策略的大语言模型在生成答案的准确性和一致性方面表现突出，其输出结果与实际答案的匹配度显著高于其他两种对比方法，充分验证了知识图谱增强方法的有效性和实用性。

为量化评估三种方法的性能差异，本研究基于 BLUE-4 指标进行了系统性对比分析，具体评估结果如表 6 所示。

Table 6. Indicator analysis
表 6. 指标分析

评价指标	原始大语言模型	微调大语言模型	知识图谱增强大语言模型
BLUE-4	6.73	34.68	88.32

表 6 的结果表明，基于知识图谱增强的大语言模型在 BLEU-4 得分上明显优于原始模型和微调模型。这说明知识图谱的引入为模型提供了更准确的背景知识，显著提升了其性能，进一步验证了该方法的有效性。

5. 结束语

本研究提出了一种基于企业装备知识图谱增强的大语言模型优化方案，有效提升了领域知识问答的性能。通过系统性实验验证了所提方法的可行性和实用性。本研究的创新点主要体现在以下三个方面：

首先，在模型选型方面，经过对小参数量级大语言模型的对比分析，最终选定 ChatGLM2-6B 作为基础模型，该模型在性能和部署便捷性之间取得了良好平衡。

其次，在技术实现层面，采用了两阶段优化策略：第一阶段通过领域适应性微调提升模型对企业装

备专业知识的理解能力和关键词提取准确率；第二阶段引入企业装备知识图谱作为本地知识库，为模型提供精准的领域知识支持，显著提高了回答的专业性和准确性。

最后，在性能优化方面，充分挖掘大语言模型在自然语言处理方面的潜力，通过两轮迭代优化显著提升了模型整体性能。

5.1. 模型的局限性

尽管本研究在提升企业装备领域问答系统的准确性和可靠性方面取得了显著进展，但仍存在一些局限性：

复杂问题的理解能力：当前模型在处理复杂问题时，尤其是涉及多跳推理或需要跨领域知识的场景，表现仍不够理想。例如，当用户查询涉及多个装备之间的协同工作时，模型可能无法准确理解问题的深层语义，导致生成的答案不够准确。

知识图谱的覆盖范围：虽然本研究构建了企业装备知识图谱，但其覆盖范围仍有限，特别是在新兴装备或技术领域，知识图谱的更新速度可能无法跟上行业发展的步伐。这可能导致模型在面对最新技术或装备时，无法提供准确的答案。

知识图谱的更新维护：知识图谱的实时更新机制仍需进一步完善。当前的知识图谱更新主要依赖于人工干预，缺乏自动化更新机制，这可能导致知识图谱的时效性不足，影响模型的回答准确性。

5.2. 未来研究方向

针对上述局限性，未来的研究可以从以下几个方面展开：

提升复杂问题的理解能力：未来的研究可以探索引入更复杂的推理机制，如多跳推理、图神经网络等，以增强模型对复杂问题的理解能力。此外，可以结合多模态数据(如图像、视频等)，进一步提升模型在跨领域问题上的表现。

扩展知识图谱的覆盖范围：未来的研究可以探索自动化知识抽取技术，利用大语言模型的语义理解能力，自动从互联网、学术论文等数据源中抽取最新的装备知识，并将其整合到知识图谱中。这将显著提升知识图谱的覆盖范围和时效性。

自动化知识图谱更新机制：未来的研究可以探索基于大语言模型的自动化知识图谱更新机制。通过实时监控行业动态、技术发展等数据源，自动识别并更新知识图谱中的过时信息，确保知识图谱的实时性和准确性。

模型效率优化：尽管 P-Tuning v2 微调方法在低资源环境下表现优异，但随着知识图谱规模的扩大，模型的响应时间可能会增加。未来的研究可以探索更高效的微调策略，如稀疏微调、知识蒸馏等，以在保证模型性能的同时，进一步提升其效率。

通过上述改进，未来的研究可以进一步提升企业装备领域问答系统的性能，为企业装备管理的信息化建设提供更强大的支持。

参考文献

[1] 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.

[2] 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.

[3] TB OpenAI (2022) ChatGPT: Optimizing Language Models for Dialogue.

[4] Touvron, H., Lavril, T., Izacard, G., *et al*. (2024) LLaMA: Open and Efficient Foundation Language Models. arXiv: 2302.13971. https://arxiv.org/abs/2302.13971

[5] Yang, A.Y., Xiao, B., Wang, B.N., *et al*. (2024) Baichuan 2: Open Large-Scale Language Models. arXiv: 2309.10305.

https://arxiv.org/abs/2309.10305

[6] Singhal, A. (2012) Introducing the Knowledge Graph: Things, Not Strings. Official Google Blog, **5**, 3-8.

[7] Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. *Scientific American*, **284**, 34-43. https://doi.org/10.1038/scientificamerican0501-34

[8] Ji, S., Pan, S., Cambria, E., Marttinen, P. and Yu, P.S. (2022) A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, **33**, 494-514. https://doi.org/10.1109/tnnls.2021.3070843

[9] 王党炜. 基于领域知识图谱的智能问答系统研究[D]: [硕士学位论文]. 西安: 西安工业大学, 2023.

[10] 李伟, 王竣生, 秦鹏. 基于知识图谱的医疗问答系统研究[J]. 长江信息通信, 2023, 36(6): 107-109.

[11] 李肖, 刘德生. 面向武器装备体系知识图谱的本体构建[J]. 兵工自动化, 2022, 41(3): 25-30.

[12] 张克亮, 李伟刚, 王慧兰. 基于本体的航空领域问答系统[J]. 中文信息学报, 2015, 29(4): 192-198.

[13] 窦小强, 刘天雅, 张志政. 基于军事知识图谱的问答系统[C]//第六届中国指挥控制大会论文集(上册). 2018: 537-541.

[14] 车金立, 唐力伟, 邓士杰, 等. 基于百科知识的军事装备知识图谱构建与应用[J]. 兵器装备工程学报, 2019, 40(1): 148-153.

[15] 姜成樾. 一个基于语义网技术的军事信息领域自动问答系统设计与实现[D]: [硕士学位论文]. 南京: 南京大学, 2016.

[16] 阚实松. 基于知识图谱的问答系统[D]: [硕士学位论文]. 武汉: 华中科技大学, 2019.

[17] 刘天雅. 一种面向军事问答的问题理解方法[D]: [硕士学位论文]. 南京: 东南大学, 2019.

[18] 李代祎, 盛杰, 刘运星, 等. 基于知识图谱的军事武器问答系统[J]. 指挥信息系统与技术, 2020, 11(5): 58-65.

[19] Luo, H., Tang, Z., Peng, S., *et al*. (2024) ChatKBQA: A Generate-Then Retrieve Framework for Knowledge Base Question Answering with Fine-Tuned Large Language Models. arXiv: 2310.08975. https://arxiv.org/abs/2310.08975

[20] Li, Z., Deng, L., Liu, H., *et al*. (2024) UniOQA: A Unified Framework for Knowledge Graph Question Answering with Large Language Models. arXiv: 2406.02110. https://arxiv.org/abs/2406.02110

[21] Taffa, T.A. and Usbeck, R. (2024) Leveraging LLMs in Scholarly Knowledge Graph Question Answering. arXiv: 2311.09841. https://arxiv.org/abs/2311.09841

[22] Chen, Y.L., Zhang, Y.M., Yu, J.F., *et al*. (2024) In-Context Learning for Knowledge Base Question Answering for Unmanned Systems Based on Large Language Models. arXiv: 2311.02956. https://arxiv.org/abs/2311.02956

[23] Jiang, J., Zhou, K., Zhao, X., Li, Y. and Wen, J. (2023) Reasoninglm: Enabling Structural Subgraph Reasoning in Pre-Trained Language Models for Question Answering over Knowledge Graph. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 3721-3735. https://doi.org/10.18653/v1/2023.emnlp-main.228

[24] Tan, C., Chen, Y., Shao, W., *et al*. (2024) Make a Choice! Knowledge Base Question Answering with in Context Learning. arXiv: 2305.13972. https://arxiv.org/abs/2305.13972

[25] Kim, J., Kwon, Y., Jo, Y. and Choi, E. (2023) KG-GPT: A General Framework for Reasoning on Knowledge Graphs Using Large Language Models. Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6-10 December 2023, 9410-9421. https://doi.org/10.18653/v1/2023.findings-emnlp.631

[26] Wu, Y.K., Hu, N., Bi, S., *et al*. (2024) Retrieve-Rewrite-Answer: A KG-to-Text Enhanced LLMs Framework for Knowledge Graph Question Answering. arXiv: 2309.11206. https://arxiv.org/abs/2309.11206

[27] Baek, J., Aji, A.F. and Saffari, A. (2024) Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. arXiv: 2306.04136. https://arxiv.org/abs/2306.04136

[28] Sun, J., Xu, C., Tang, L., *et al*. (2024) Think-on-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph. arXiv: 2307.07697. https://arxiv.org/abs/2307.07697

[29] Dong, Z.X., Peng, B.Y., Wang, Y.F, *et al*. (2024) EffiQA: Efficient Question-Answering with Strategic Multi-Model Collaboration on Knowledge Graphs. arXiv: 2406.01238. https://arxiv.org/abs/2406.01238

[30] 王翼虎, 白海燕, 孟旭阳. 大语言模型在图书馆参考咨询服务中的智能化实践探索[J]. 情报理论与实践, 2024, 46(8): 96-103.