软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(9):4056−4071 [doi: 10.13328/j.cnki.jos.007247] [CSTR: 32375.14.jos.007247]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于相关性提示的知识图谱问答*

马 杰[1,2], 孙望淳[1,2], 王平辉[1,2], 张若非[1,2], 李帅鹏[2], 苏 洲[1,2]

[1](西安交通大学 网络空间安全学院, 陕西 西安 710049)

[2](智能网络与网络安全教育部重点实验室 (西安交通大学), 陕西 西安 710049)

通信作者: 马杰, E-mail: jiema@xjtu.edu.cn

摘 要: 大语言模型 (large language model, LLM) 随着不断发展, 在开放领域取得了出色的表现. 然而, 由于缺乏专业知识, LLM 在垂直领域问答任务上效果较差. 这一问题引发了研究者的广泛关注. 现有研究通过“检索-问答”的方式, 将领域知识注入大语言模型, 以增强其性能. 然而该方式通常会检索到额外的噪声数据而导致 LLM 的性能损失. 为了解决该问题, 提出基于知识相关性的知识图谱问答方法. 具体而言, 将噪声数据与回答问题所需要的知识进行区分, 在“检索-相关性评估-问答”的框架下, 引导大语言模型选择合理的知识做出正确的回答. 此外, 提出一个机械领域知识图谱问答的数据集 Mecha-QA, 包含传统机械制造以及增材制造两个子领域, 以推进该领域大语言模型与知识图谱问答相关的研究. 为了验证所提方法的有效性, 在 Mecha-QA 和航空航天领域数据集 Aero-QA 上进行实验. 结果表明, 该方法可以显著提升大语言模型在垂直领域知识图谱问答的性能.

关键词: 大语言模型; 知识图谱; 垂直领域; 问答系统; 知识检索

中图法分类号: TP18

中文引用格式: 马杰, 孙望淳, 王平辉, 张若非, 李帅鹏, 苏洲. 基于相关性提示的知识图谱问答. 软件学报, 2025, 36(9): 4056–4071. http://www.jos.org.cn/1000-9825/7247.htm

英文引用格式: Ma J, Sun WC, Wang PH, Zhang RF, Li SP, Su Z. Knowledge Graph Question Answering Based on Relevance Prompts. Ruan Jian Xue Bao/Journal of Software, 2025, 36(9): 4056–4071 (in Chinese). http://www.jos.org.cn/1000-9825/7247.htm

Knowledge Graph Question Answering Based on Relevance Prompts

MA Jie[1,2], SUN Wang-Chun[1,2], WANG Ping-Hui[1,2], ZHANG Ruo-Fei[1,2], LI Shuai-Peng[2], SU Zhou[1,2]

[1](School of Cyber Science and Engineering, Xi’an Jiaotong University, Xi’an 710049, China)

[2](Ministry of Education Key Laboratory for Intelligent Networks and Network Security (Xi’an Jiaotong University), Xi’an 710049, China)

Abstract: As large language models (LLMs) continue to evolve, they have shown impressive performance in open-domain tasks. However, they exhibit limited effectiveness in domain-specific question-answering due to a lack of domain-specific knowledge. This limitation has attracted widespread attention from researchers in the field. Current research attempts to infuse domain knowledge into LLMs through a retrieve-answer approach to enhance their performance. However, this method often retrieves additional, irrelevant data, leading to a degradation in LLM effectiveness. Therefore, this study proposes a method for knowledge graph question answering based on the relevance of knowledge. This method focuses on distinguishing essential knowledge required for specific questions from noisy data. Under a framework of retrieval-relevance assessment-answering, this method guides LLMs to select appropriate knowledge for accurate answers. Moreover, this study introduces a dataset named Mecha-QA for question-answering using a mechanical domain knowledge graph, covering traditional machinery manufacturing and additive manufacturing, to promote research that integrates LLMs with knowledge graph question answering in this field. To validate the effectiveness of the proposed method, experiments are conducted on the Aero-QA dataset in the aerospace domain and the Mecha-QA dataset. Results demonstrate that the proposed method significantly improves the performance of

* 基金项目: 国家重点研发计划 (2021YFB1715600); 国家自然科学基金 (62306229)
马杰和孙望淳为共同第一作者.
收稿时间: 2024-01-31; 修改时间: 2024-05-04; 采用时间: 2024-06-26; jos 在线出版时间: 2024-12-31
CNKI 网络首发时间: 2025-01-02

LLMs in knowledge graph question answering in vertical domains.
Key words: large language model (LLM); knowledge graph (KG); vertical domain; question answering system; knowledge retrieval

智能问答要求机器对用户的自然语言问题进行正确作答[1,2]. 这一过程需要机器对问题语义有较强的理解能力且具备丰富的知识储备和逻辑推理能力, 以便准确把握问题的含义并生成准确的回答[3]. 近年来, 大语言模型(large language model, LLM) 呈现井喷式发展, 例如 ChatGPT (https://chat.openai.com)、ChatGLM[4,5], 并在越来越多的自然语言处理任务上取得出色表现[6]. 业界开始使用 LLM 作为智能问答的“基石”, 利用其强大的语义理解能力、逻辑推理能力以及知识储备, 构建功能强大的智能问答模型. 得益于开放域大规模语料上的预训练, 上述 LLM 对开放域知识拥有强大的理解与应用能力, 能够处理复杂性高且多样化的问题, 进而为智能问答提供更为全面与精准的支持.

然而, 在垂直领域, 由于缺乏相应的训练数据, LLM 的表现与其在开放领域的表现差距较大, 常会产生“幻觉”[7]现象. 针对这一问题, 研究者通过微调 LLM 或是引入外部知识 (如知识图谱) 的方式, 提升垂直领域 LLM 的问答性能, 减少“幻觉”现象[8]. 相对于后者, 微调 LLM 需要大量数据作为支撑, 在低资源场景下难以实现. 知识图谱作为一种组织和表示信息的结构化方式[9], 用于捕获实际场景下实体之间的关系, 由大量形如“(*subject*, *relation*, *object*)”形式的三元组构成, 它可以提供丰富的精细化知识. 因此, 越来越多的研究尝试将 LLM 与知识图谱进行结合[10−13], 利用知识图谱中的知识, 减少 LLM 的“幻觉”现象, 进而提升智能问答模型的性能.

已有研究[14−17]主要针对医疗、金融、法律等相关领域, 在保持国家经济活力、带动国民经济和科技发展至关重要的机械制造、航空航天等领域进行深入探索较少. 此外, 已有研究主要采用“检索-问答”框架进行答案预测. 为了确保关键知识条目被输入大模型, 它们通常会设置较大的知识召回量, 从而不可避免地引入噪声数据, 导致其常会产生不准确、不相关的回答[6]. 如图 1 所示, 将“KTZ700--02”的一阶子图均输入 LLM, 会使模型的回复中包含“试样直径”等无关信息, 不便于提问者理解及快速找到答案.

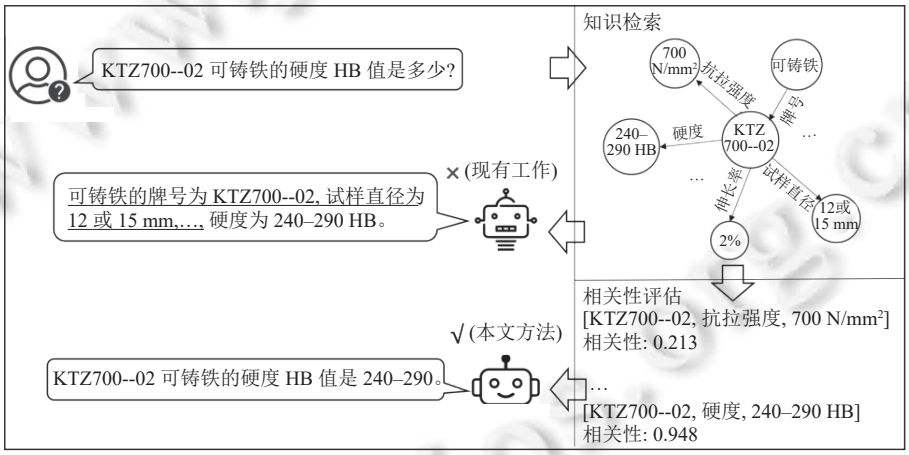


图 1 知识相关性增强大语言模型问答流程

为了解决上述难题, 本文首先以机械制造领域专业书籍[18]作为知识源, 构建领域知识图谱, 并采用“多轮采样三元组”的策略, 基于 LLM 自动生成对应的知识图谱问答数据集. 该数据集包括机械制造领域的材料特性、制造工艺特点等知识. 此外, 针对噪声数据引入的问题, 本文提出基于相关性提示的知识图谱问答方法. 如图 1 所示, 与现有工作不同, 本文未直接将检索到的三元组输入 LLM, 而是采用了“检索-相关性评估-问答”的框架. 具体而言: 在检索到三元组后, 首先, 本文方法通过额外训练相关性计算模型进行相关性评估; 其次, 将检索到的三元组按照相关性进行排序; 最后, 将该相关性作为 prompt 的一部分输入 LLM, 帮助其选择合适的知识进行问题作答. 为了验证所提方法的有效性, 本文在机械制造领域与航空航天领域的数据集 Mecha-QA 和 Aero-QA[19]上使用不同的 LLM 作为基座模型进行实验. 此外, 现有生成式问答的评价方法通过参考答案与生成答案的精确匹配[10]或是额外

训练评价模型[20]来实现, 存在精度低、标注成本高等问题. 针对上述问题, 本文设计了 LLM-Score 评价指标, 利用大语言模型从答案语义的角度实现零样本评估. 最终, 实验结果表明, 本文提出的方法在两个垂直领域的数据集上均有出色表现.

本文的主要贡献包括 3 个方面.

(1) 提出了一个机械制造领域的知识图谱与问答数据集. 据我们调查所知, 该数据集是机械领域较早的知识图谱问答相关的工作, 主要包含两个子领域: 传统机械制造和增材制造.

(2) 提出了一种知识相关性增强大模型在垂直领域知识图谱问答的方法. 该方法考虑了检索到的知识与问题之间的相关性, 采用“检索-相关性评估-问答”的框架, 提升 LLM 检索知识进行问答的能力. 此外, 该方法可兼容不同类型 (开源模型 (如 ChatGLM-6B 等) 与未开源模型 (如 ChatGPT 等)) 的 LLM.

(3) 本文在机械制造领域数据集 Mecha-QA 与航空航天领域的数据集 Aero-QA 上进行实验, 在图谱问答任务上对比其他基线模型, 均取得了较好的效果. 本文的数据集和代码均已开源至 github (https://github.com/reml-group/Mecha-QA).

1 相关工作

在自然语言处理领域, 近年来涌现了一系列强大的通用语言模型, 如 ChatGPT、百度文心一言 (https://yiyan.baidu.com/)、阿里通义千问 (https://qianwen.aliyun.com/)、ChatGLM[4,5]、LLaMA[21]等. 这些模型在通用开放领域展现出出色的语义理解和知识问答能力, 引发了学术界和工业界的广泛兴趣.

然而, 在垂直专业领域问答任务上, 它们往往由于缺乏专业知识而表现出巨大的局限性, 常常出现“幻觉”现象. 这一局限性限制了模型在处理领域特定问题时的准确性和可靠性.

为了应对这一挑战, 学术界和工业界的研究人员正致力于探索使大型语言模型更好地适应特定专业领域的方法. 主流方法可以分为两类: 基于数据微调的方法和基于知识检索增强的方法[22,23]. 基于数据微调的方法通常涉及使用特定领域的标注数据对通用语言模型进行微调. 通过在特定领域的数据集上训练, 模型可以学习到领域特定的语言模式和知识. 这种方法的关键在于收集足够多的、高质量的领域特定数据, 以便模型能够从中学习到有用的信息. 微调后的模型在处理与训练数据相似的领域问题时, 可以提供更加准确和专业的回答. 基于知识检索的垂直领域问答涉及将知识检索与语言模型相结合, 以提高模型在特定领域的知识问答能力. 知识检索的目的是从大量的非结构化或结构化知识库中检索出与问题相关的信息. 这些知识库可能包括百科全书、专业文献、知识图谱等. 当模型接收到一个问题时, 它首先使用检索系统找到最相关的知识片段, 然后利用这些知识来生成回答. 这种方法可以有效地提高模型在处理领域特定问题时的准确性和可靠性, 因为它可以直接利用已有的专业知识来辅助回答. 下面将对两种方法的相关工作进行具体介绍.

1.1 基于数据微调的垂直领域问答

在垂直领域, 通过 LoRA[24]、p-tuning[25]、p-tuning v2[26]等参数高效微调的方法构建任务相关数据集, 微调 LLM, 实现基于数据微调的垂直领域问答. HuaTuo[27]、DISC-MedLLM[15]在 CMeKG 等中医药数据集的基础上进行整合与处理, 微调本地 LLM, 得到可以用于医药领域问答的领域大模型. FinGPT[17]收集了金融领域的媒体数据以及公开数据集, 利用 LoRA 和人类反馈强化学习 (RLHF)[28]进行微调, 得到金融领域大模型, 可以实现智能投资顾问、量化交易等相关工作.

在数据微调的框架下, 可以训练得到对领域知识了解程度较高的 LLM, 从而实现垂直领域的问答. 然而, 尽管可以使用参数高效微调的方法来减少训练模型需要的资源, 但这种方法对数据规模的依赖性依旧较大, 且对于知识无法做到及时更新. 本文采取基于知识检索的垂直领域问答方法.

1.2 基于知识检索的垂直领域问答

1.2.1 基于非结构化知识增强的垂直领域问答

通过引入外部知识, 可以使模型生成更准确的回复[29,30], 有效减少“幻觉”现象. 对非结构化文本, 目前通用的

方法是采用 BM25[31]或是稠密向量检索的方式, 用问题检索相关的非结构化文本, 以此作为外部知识, 再依据预先定义的模板构建 prompt 提示词, 输入 LLM, 实现知识增强的问答. Levonian 等人[32]在数学领域, 基于“检索-问答”的框架实现了知识增强问答. 此外, 有部分研究将大模型引入了检索过程, 通过大模型增强检索时的 query, 提升检索的精确性. Chatlaw[16]通过大模型提取问题中的关键词, 将关键词与问题向量化表示后联合检索证据文档, 并构建 prompt, 实现法律领域的问答. Yu 等人[33]利用 LLM 自身生成相关文档, 通过提示词挖掘 LLM 的内部知识, 作为回答的依据. Ma 等人[34]通过大模型的反馈重写问题, 获得更精确的文档检索结果. Gao 等人[35]利用 LLM 在不输入外部知识的条件下预先生成一个伪答案, 将该答案作为 query 进行检索, 再将结果作为外部知识进行第 2 轮的答案回复. Huang 等人[36]对问题进行分类, 并检索相似问题及该相似问题的答案作为 prompt 的一部分, 以增强 LLM 在政务领域的表现.

上述方法均采用非结构化知识增强 LLM 在垂直领域的问答性能. 与知识图谱相比, 非结构化知识的组织和管理较为困难, 这使得在大量数据中快速定位和检索特定信息变得困难, 因而检索的准确性和效率较低, 问答性能不佳.

1.2.2 基于知识图谱增强的垂直领域问答

目前已有一些基于知识图谱增强 LLM 问答性能的相关研究[37−39]. Baek 等人[10]将知识图谱的三元组进行嵌入表示, 通过问题的表示向量检索最相关的 k 个三元组, 组成 prompt 输入 LLM, 实现了零样本场景下的知识图谱问答. Kim 等人[11]对 LLM 进行多轮问询, 确定与问题最相关的三元组, 并作为知识输入 LLM 提升问答效果. Wu 等人[40]在检索到问题对应的子图后, 未直接将子图输入 LLM, 而是增加了图转换为自然语言的模块, 将检索到的子图谱用自然语言的形式表述, 再进行 prompt 构建, 以增强 LLM 对检索知识的理解. Soman 等人[41]采用类似的方法, 将生物医学知识图谱与 GPT 4 相结合, 使其可以在生物医学领域进行专业问答. 此外, 还有一些工作 (如文献 [12,42–46]) 未直接利用 LLM 进行问答. 相反, 它们专注于利用 LLM 根据问题生成对应的知识图谱查询代码或数据库查询语句. 在这种方法下, LLM 被用作一个工具来生成精确的查询, 这些查询随后用于检索与问题直接相关的知识三元组.

在“检索-问答”的框架下, 无论是非结构化外部知识还是结构化外部知识都可以实现灵活的领域问答而不需要大量的外部数据对大模型进行微调. 然而, 由于该方法受限于检索召回精度的影响, 通常会引入大量噪声以确保召回率. 这些噪声知识与问题对应的知识被同等地位地组成了 prompt 的一部分输入 LLM, 从而导致 LLM 在回答时准确率低, 常常包含许多不相关的部分. 因此, 本文提出了知识相关性增强的垂直领域知识图谱问答大模型, 通过将检索到的三元组的相关性“告知”LLM, 以提升其回复的质量.

2 基于相关性提示的知识图谱问答方法

2.1 问题描述

在智能问答场景中, 对于一个给定问题 q, 使用 LLM 生成答案 a 的条件概率 P 可以表示为:

$$P(a|q)=\prod_{i=1}^{|a|}P(y_i|y_{<i},q) \tag{1}$$

其中, y_i 表示答案 a 中的第 i 个 token.

引入外部知识图谱 $\mathcal{G}$ 后, 令 $\mathcal{G}=\{(s,r,o)|s,o\in\mathcal{E},r\in\mathcal{R}\}$, 其中, (s,r,o) 表示在头实体 s 和尾实体 o 之间存在关系 r; $\mathcal{E},\mathcal{R}$ 分别表示实体集合和关系集合. 假设在问题 q 中对应的实体为 e_q, 在 $\mathcal{G}$ 中 e_q 的一阶子图为 $\mathcal{G}_s=\{(s,r,o)|s,o\in\mathcal{E},r\in\mathcal{R},s=e_q\vee o=e_q\}$, 则生成答案 a 的条件概率可以进一步表示为:

$$P(a|q)=\prod_{i=1}^{|a|}P(y_i|y_{<i},q,\mathcal{G}_s) \tag{2}$$

在检索到的一阶子图 $\mathcal{G}_s$ 中, 每一个三元组 (s,r,o) 与问题的相关程度不同, 对最终答案的贡献也应该不同. 因此, 本文训练一个相关性计算模型 ReM, 输入三元组 (s,r,o) 与问题 q 可以计算对应的相关性分数 rs:

$$rs = ReM((s,p,o),q) \tag{3}$$

根据相关性分数 rs, 可以将一阶子图 $\mathcal{G}_s$ 转换成包含对应相关性分数的四元组 T_{rs}:

$$T_{rs} = \{(s,r,o,rs) | s,o \in \mathcal{E}, r \in \mathcal{R}, s = e_q \vee o = e_q\} \tag{4}$$

其中, rs 由公式 (3) 计算得到.

最终, 对于给定的问题 q 以及外部知识图谱 $\mathcal{G}$, 生成答案 a 的条件概率可以表示为:

$$P(a|q) = \prod_{i=1}^{|a|} P(y_i|y_{<i}, q, T_{rs}) \tag{5}$$

2.2 模型架构

在垂直领域, 针对 LLM 易产生“幻觉”的问题, 常采用检索增强生成 (retrieval augmented generation, RAG) 的方式, 通过检索引入外部知识, 作为提示词 (prompt) 的一部分, 以减少“幻觉”现象. 然而, 已有 RAG 方法, 将检索到的所有知识无差别地输入给 LLM, 常常会导致一些相关性较低的知识也被 LLM 所利用, 进而生成的回复中包含这部分不相关的内容. 因此, 本文提出基于知识相关性增强垂直领域知识图谱问答的方法. 通过单独训练相关性计算模型, 计算检索到的知识图谱三元组按照其与问题的相关性, 并有差别地输入 LLM, 以帮助 LLM 更好地判断关键三元组, 增强其回答问题的准确率.

本文考虑不同三元组对模型回答问题的贡献不同, 提出了基于相关性提示的知识图谱问答方法, 模型框架如图 2 所示, 主要包括 3 个阶段: (1) 子图检索; (2) 三元组相关性计算; (3) prompt 构建与回复生成. 其中, 子图检索阶段需要在知识图谱 $\mathcal{G}$ 中检索问题中实体对应的一阶子图 $\mathcal{G}_s$. 将 $\mathcal{G}_s$ 输入下一模块, 通过训练相关性计算模型 ReM, 对 $\mathcal{G}_s$ 中包含的每一条三元组, 计算其与问题 q 的相关性 rs, 得到包含该分数的四元组集合 T_{rs}. 将 T_{rs} 输入第 3 个模块, 将四元组与问题构建合适的 prompt, 作为 LLM 的输入, 最终生成问题对应的答案 a. 通过上述流程, 将检索到的多个三元组通过添加额外信息拓展为四元组的方式, 有序、有差别地构建了提示词. 当提示词输入 LLM 之后, 与“检索-问答”的范式相比, LLM 可以根据该部分额外信息, 更简单地判断出问题对应的关键知识, 进而生成更高质量、更准确的答案.

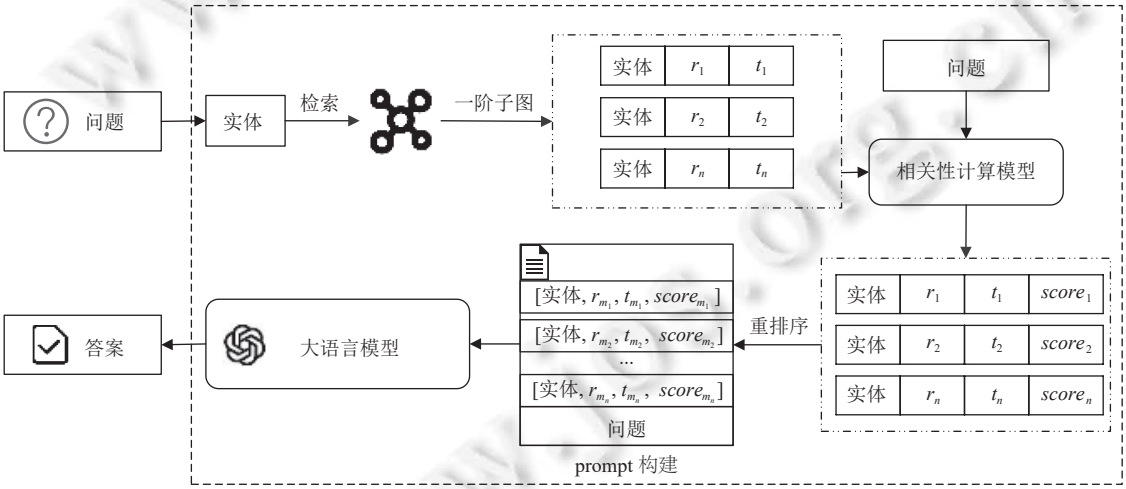


图 2 基于相关性提示的知识图谱问答框架

2.3 相关性计算模块

为了量化三元组与问题之间的相关性, 本文首先对三元组和问题分别进行语义表示, 再计算两者表示的相似度作为它们的相关性程度. 具体地, 如图 3 所示, 为了将三元组转化为包含语义信息的表示, 需要先将其序列化. 本文采用线性序列化的方式, 将一个三元组 (s,r,o) 表示为文本形式的 “[s, r, o]”. 接着, 进行关于问题的正负样本对构

建. 对于一个给定问题 q, 其中包含的实体为 e_q, q 对应的答案为 a, 则定义 (q,a) 在图谱 $\mathcal{G}$ 中对应的 n 个三元组 $[(s_1,r_1,o_1),\ldots,(s_n,r_n,o_n)]$ 中每个三元组 (s_i,r_i,o_i) 序列化后的文本 $t_{pos_i}=$“$[s_i,r_i,o_i]$”为 q 对应的正样本, 其中 i 为下标, 表示第 i 个三元组. 可以得到, q 对应的正样本集合 Ω_{pos_q} 为:

$$\Omega_{pos_q}=\left\{\left(q,t_{pos_i}\right)|1\leqslant i\leqslant n\right\} \tag{6}$$

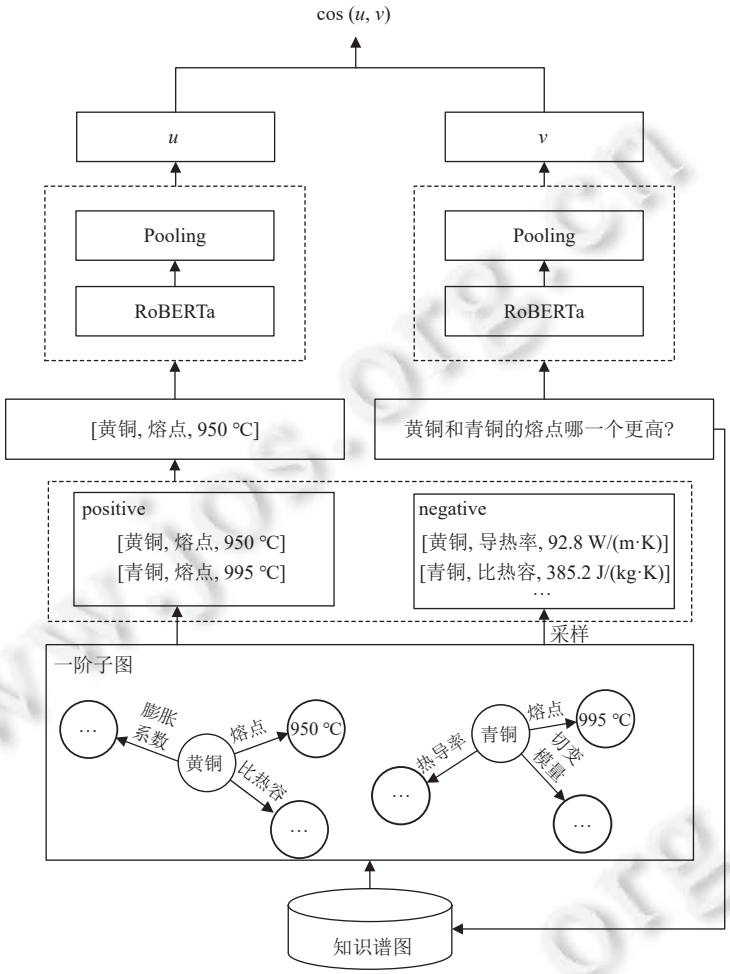


图 3 相关性计算模块

在负样本构建时, 定义问题 q 中实体 e_q 对应的一阶子图 $\mathcal{G}_s$ 中非正样本的三元组均为负样本. 考虑到在子图 $\mathcal{G}_s$ 中, 对给定问题 q, 只有极少一部分的三元组为正样本, 因此, 在实际构建时, 对所有的负样本三元组进行采样, 最后得到 q 对应的负样本集合 Ω_{neg_q} 为:

$$\Omega_{neg_q}=sample\left(\complement_{SubG_e}^{\Omega_{pos_q}}\right) \tag{7}$$

最终通过利用正负样本对, 微调 m3e 模型 (https://huggingface.co/moka-ai/m3e-base), 实现相关性计算模块. m3e 是一个基于 RoBERTa 的在中英文数据集上进行大规模训练的句子表示模型, 参考文献 [47], 本文微调过程的 loss 函数为:

$$sim_{\text{dis}}=\sum_{\substack{(i,j)\in\Omega_{pos},\\(k,l)\in\Omega_{neg}}}\exp\left(\frac{\cos(u_i,u_j)-\cos(u_k,u_l)}{T}\right) \tag{8}$$

$$L=-\log(sim_{\text{dis}}+1) \tag{9}$$

其中, Ω_{pos} 表示正样本对集合, Ω_{neg} 表示负样本对集合, $u_{\#}$ 表示 m3e 模型输出的句子表示向量, T 为超参数.

2.4 prompt 构建模块

在得到各三元组及其对应与问题的相关性分数的四元组 (s,r,o,rs) 后, 为了将该四元组输入 LLM 并使其理解相关性分数 rs 的意义, 将四元组序列化为以下形式: "[s, r, o, 与问题的相关性: rs]". 文献 [48] 中指出: 在信息辅助 LLM 进行回复生成的场景下, 当相关的信息出现在上下文的中间时, 模型的性能会出现明显的降低. 因此, 本模块依据上一模块得到的相关性分数对所有的四元组进行升序排序, 最终构建的提示词 prompt 模板如表 1 所示, 其中 $rs_1<rs_2<\ldots<rs_n$.

表 1 问答提示词模板

提示	提示词模板
P	三元组 [{s_1}, {r_1}, {o_1}, 与问题的相关性: {rs_1}] [{s_2}, {r_2}, {o_2}, 与问题的相关性: {rs_2}] … [{s_n}, {r_n}, {o_n}, 与问题的相关性: {rs_n}] 问题: {*question*} 答案:

根据提示词模板构建提示词 P 后, 将 P 输入 LLM 即可得到最终的答案.

由此, 基于相关性提示的知识图谱问答方法具体流程如算法 1 所示, 其中, *ReM* 为第 2.3 节中的相关性计算模型.

算法 1. 基于相关性提示的知识图谱问答算法.

输入: 问题 q, 知识图谱 $\mathcal{G}$, 问题中实体 e;
输出: 答案 a.

1. 初始化四元组集合: $T_{rs}=\{\}$
2. $\mathcal{G}_s \leftarrow search(\mathcal{G},e)$
3. **for** (s,p,o) in $\mathcal{G}_s$ **do**
   $rs \leftarrow ReM((s,p,o),q)$
   $T_{rs} \leftarrow T_{rs} \cup \{(s,p,o,r_s)\}$
   end
4. $T_{rs} \leftarrow$ 对 T_{rs} 按 rs 升序排序
5. $prompt \leftarrow ConstructPrompt(T_{rs},q)$
6. $a \leftarrow LLM(prompt)$)
7. **return** a

3 Mecha-QA 数据集构建

在机械制造领域, 目前缺少领域知识图谱及其对应的问答数据集. 因此, 本文构建了一个小规模的机械制造领域数据集 Mecha-QA. 为了验证模型的泛化能力以及鲁棒性, 在 Mecha-QA 中额外构建了一个增材制造子领域的测试集 Mecha-QA-3D. 本节主要介绍数据集构建的全过程, 如图 4 所示, 主要包括知识图谱构建和基于知识图谱的问答对构建.

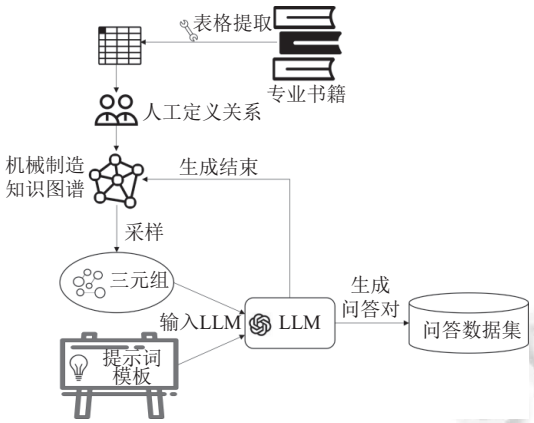


图 4 Mecha-QA 数据集构建流程

3.1 知识图谱构建

(1) 数据来源. 本文构建知识图谱的原始数据主要来自机械制造领域的专业书籍, 尤其是书籍中的结构化表格. 这些表格包含了材料的物理性质、制造工艺的特性、专业术语的定义等相关知识.

(2) 构建流程. 首先, 使用表格内容识别工具 (https://cloud.tencent.com/product/ocr) 对专业书籍中的结构化表格进行识别. 然后, 对识别后的表格, 人工定义各列之间的关系. 在这一步中, 需要根据表格的内容和专业知识, 定义表格中各列之间的逻辑关系, 例如材料类型与物理性质之间的关系, 制造工艺与工艺特性之间的关系等. 最后将表格中的数据按照定义的规则进行映射, 生成相应的头实体、关系以及尾实体, 形成三元组, 最终构成知识图谱.

3.2 基于知识图谱的问答对构建

在构建完成机械领域知识图谱之后, 需要基于知识图谱构建问答对. 本文采用基于提示词模板的方法, 利用 LLM 自动化生成问答对. 具体地, 首先定义生成问答对的提示词模板, 这些模板包含了问答对中的需要生成的关键信息, 如问题、答案、实体等. 然后, 从知识图谱中每次采样多个三元组, 利用这些提示词模板, 组成提示词输入 LLM, 生成大量的问答对. 使用的提示词如表 2 所示.

表 2 问答对构建的提示词模板

提示	提示词文本
P	你是机械领域的专家, 根据提供的三元组, 生成一些问题, 要求: 问题不要太简单, 避免形式和问题结构单一, 以JSON格式返回, 包括的key有问题、答案、实体、对应的三元组. 问题数量尽可能多且实体需要在问句中包含. 提供的三元组: {三元组 1}{三元组 2}…{三元组 n}

基于上述提示词和流程, 本文完成了机械制造领域知识图谱问答数据集的构建. 下面是一些构建的数据集中的示例.

- Q: 在铅和锡中, 哪种材料的熔点更高?
- A: 铅的熔点更高, 为 327 °C.
- 对应知识三元组: “[铅, 熔点, 327 °C]”“[锡, 熔点, 232 °C]”.
- Q: 聚三氟氯乙烯 (PCTFE、F-3) 的长期使用温度是多少?
- A: 聚三氟氯乙烯 (PCTFE、F-3) 的长期使用温度是在−195 °C–190 °C 之间.
- 对应知识三元组: “[‘聚三氟氯乙烯 (PCTFE、F-3)’, ‘特性’, ‘耐热性、电性能和化学稳定性仅次于 F-4, 在 180 °C 的酸、碱和盐的溶液中亦不溶胀或侵蚀. 机械强度、抗蠕变性能、硬度都比 F-4 好些, 长期使用温度为 −195 °C–190 °C 之间, 但要求长期保持弹性时, 则最高使用温度为 120 °C. 涂层与金属有一定的附着力, 其表面坚韧、耐磨, 有较高的强度’]”.

- Q: 15Cr 材料的热处理方式是什么, 它的齿面硬度值是多少?
- A: 热处理方式是渗碳、淬火、回头, 齿面硬度为 55–60 HRC.
- 对应知识三元组: “[15Cr, 材料的热处理方式, 渗碳、淬火、回头]”“[15Cr, 材料的齿面硬度, 55–60 HRC]”.

在构建 Mecha-QA 时, 主要采用 GPT-3.5 作为 LLM, 并以传统机械制造业为领域. 为了进一步验证模型的泛化能力和鲁棒性, 并消除可能存在的模型偏好问题 (即在后续实验中使用相同的模型对可能带来的偏倚进行问答), 本研究引入多种大型语言模型, 包括通义千问、GPT-3.5 以及 ChatGLM, 针对机械制造的子领域——增材制造, 构建了 Mecha-QA-3D 测试集作为补充.

最终, Mecha-QA 和 Mecha-QA-3D 数据集中问题对应的关键知识数量的占比如图 5 所示, 在 Mecha-QA 中, 89.2% 的问题对应唯一关键知识, 而在 Mecha-QA-3D 中则为 51.35%. 图 6 展示了 Mecha-QA 和 Mecha-QA-3D 中问题对应的噪声数据量与关键知识条目比的分布情况, 其中密度图的宽度反映了对应比值的频数. 图 6 表明, 两个数据集中该比值的主要分布范围为 2–9, 即: 在 Mecha-QA 和 Mecha-QA-3D 中, 每一个问题可以检索到的噪声数据量为关键知识条目的 2–9 倍. 结合图 5 及图 6 可知, 在 Mecha-QA 上使用传统“检索-问答”的方法进行知识图谱问答时会引入较多的噪声数据, 可能对结果产生较大干扰, 从而增加该数据集上的推理难度, 而与 Mecha-QA 相比, Mecha-QA-3D 中噪声数据与关键知识条目比更多, 推理难度进一步加大.

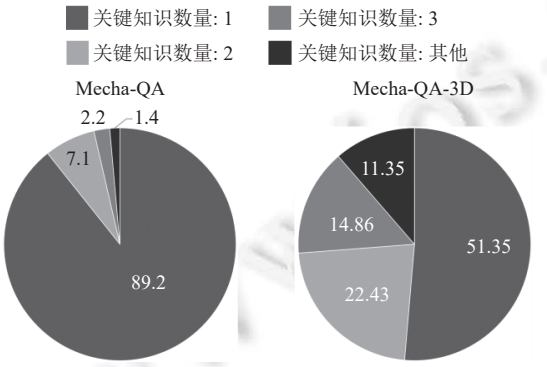


图 5 Mecha-QA 和 Mecha-QA-3D 数据集中关键知识数量占比情况 (%)

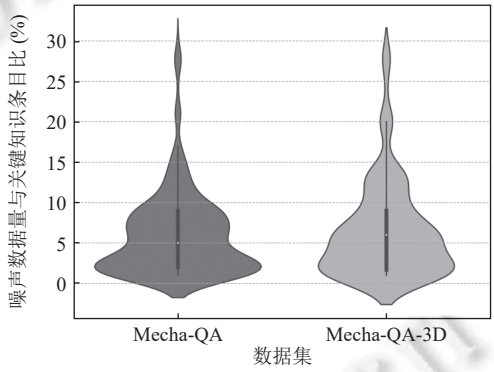


图 6 Mecha-QA 和 Mecha-QA-3D 数据集中噪声数据与关键知识条目比分布情况

4 实验结果及分析

4.1 实验数据集

本文采用的数据集是自构建机械制造领域知识图谱中文问答数据集 Mecha-QA 以及航空航天领域英文问答数据集 Aero-QA[19]. 其中, Mecha-QA 是本文基于机械制造领域三元组, 利用 LLM 生成的问答数据集. 除测试集外, 还包含了 Mecha-QA-3D 测试集, 以更全面地评估模型性能. Aero-QA 是文献 [19] 中构建的针对 AviationKG[49] 知识图谱的问答数据集. 两个数据集的相关情况如表 3 所示.

表 3 所用数据集统计情况

数据集	三元组数	训练集	验证集	测试集
Mecha-QA	1767	571	—	142
Mecha-QA-3D	838	—	—	370
Aero-QA	96686	17038	2130	2131

4.2 评价指标及基准模型

在选择评估指标时, 一方面, 与传统基于知识图谱的问答系统在知识图谱中直接检索答案不同, 在本文的实验

设定下, 答案由生成式语言模型输出, 无法直接与标准答案进行匹配, 计算准确率以进行评估, 因此, 需要从语义的角度对待评估答案与标准答案进行比较并进行量化; 另一方面, 以 ChatGPT 为代表的 LLM 在对话、问答等领域表现出色, 展示出强大的语义理解能力. 越来越多的研究使用 ChatGPT 作为生成式问答的评估者, 并取得了与人类专家一致的效果[15,50,51]. 因此, 本文采用 ChatGPT 作为评估者, 构建合适的 prompt 提示词, 对待评估的答案进行评估, 得到范围为 1–5 的 LLM-*Score* 指标. 答案评估时使用的提示词如表 4 所示.

表 4 答案评估的提示词模板

提示	提示词文本
P	你是一位专业、公正且严格的评分员. 以下是用户和AI助手在{domain}领域的问答对话. 根据下面的标准, 根据问答和参考答案, 对助手的表现在1–5的范围内进行评分. 只需要提供分数, 不需要解释. *Accuracy*: 助手提供的答案准确无误, 根据参考答案没有事实错误. 请确保你不受文本长度的影响, 努力保持客观. 你的评分应该足够严格, 不要轻易给出满分. 请按照以下格式输出评分结果: *Accuracy*: x [start of question-answering] {问题} {答案} {参考答案} [end of question-answering]

此外, 本文额外引入两个评价指标, 以更全面、可信地评估模型性能. (1) 为了提升可复现性, 本文参考文献 [10] 的做法, 计算知识图谱问答的准确率 (*Accuracy*), 即: 计算模型在测试集上生成的答案中, 包含问题对应的答案实体的比例; (2) 为了全面地评估系统性能, 本文引入用户满意度 (customer satisfaction score, CSAT)[52]. 具体计算方式为: 抽取 100 条测试问答对, 由两名人类专家进行评估, 并给出 1–5 的评分. 计算 4 分或 5 分的占比, 作为 CSAT, 如公式 (10)、(11) 所示.

$$score_i = \min(expert1_i, expert2_i) \tag{10}$$

$$CSAT = \frac{\sum_i^{100} I(score_i \geqslant 4)}{100} \tag{11}$$

其中, $expert\#_i$ 为专家 # 对第 i 个问题对应答案的分数, $1 \leqslant expert\#_i \leqslant 5$, $I(\cdot)$ 为指示函数.

参考文献 [8,9] 中的实验设置, 本文使用以下基线模型设置进行对比实验.

(1) m3e: 基于 RoBERTa 的预训练句子编码模型, 使用 m3e 模型匹配与问题最接近的三元组的头实体或尾实体作为问题的答案.

(2) GPT 3.5-turbo: 将问题直接输入给 GPT 3.5-turbo 得到对应的答案.

(3) LLM + KG-Triples: 将问题中实体对应的一阶子图输入给 LLM, 在本文中采用 ChatGLM-6B、p-tuning v2 微调后的 ChatGLM-6B、GPT 3.5-turbo 这 3 种 LLM (采用不同的 LLM 以验证本文方法在不同 LLM 为基座模型的条件下的有效性).

4.3 主要实验结果与分析

在相同的实验数据集、实验设置和评估指标下, 模型对比实验结果如表 5 所示. 从实验结果可以看出: 在机械制造领域与航空航天领域, 本文提出的知识图谱融入 LLM 的方法在 LLM-*Score* 和 *Accuracy* 的指标上, 对比相同设置的基线实验, 均取得了最优的结果.

具体地, 在 Mecha-QA 数据集上, 本文提出的方法在 LLM-*Score* 指标上相较于基线实验取得了最优结果. 当使用未微调的 ChatGLM-6B 与 p-tuning v2 微调后的 ChatGLM-6B 作为基座模型时, LLM-*Score* 相较于基线分别提升了 0.056 和 0.028, *Accuracy* 分别提升了 8.18% 和 1.4%; 而使用 GPT 3.5-turbo 作为基座模型时, LLM-*Score* 和 *Accuracy* 的提升达到了 0.190 和 5.64%. 这表明, 通过将检索到的三元组与问题的相关程度输入 LLM, 可以有效地帮助模型更好地利用证据三元组, 从而提高其回答准确性. 另一个测试集 Mecha-QA-3D 上取得了相似的效果提升.

表 5 对比实验结果

模型	Mecha-QA		Mecha-QA-3D		Aero-QA	
	LLM-*Score* (↑)	*Accuracy* (↑) (%)	LLM-*Score* (↑)	*Accuracy* (↑) (%)	LLM-*Score* (↑)	*Accuracy* (↑) (%)
m3e (RoBERTa for sentence embedding)	3.5634	28.59	3.1574	25.66	3.0512	23.71
GPT 3.5-turbo	2.7040	1.41	2.0000	7.57	—	—
ChatGLM-6B (w/o finetuning) + KG-Triples	4.0845	29.85	3.3595	49.46	3.3660	20.88
ChatGLM-6B (p-tuning v2) + KG-Triples	4.2113	40.85	—	—	3.8944	45.94
GPT 3.5-turbo + KG-Triples	4.3380	47.18	3.8405	72.43	4.6612	79.26
Ours (ChatGLM-6B w/o finetuning)	4.1408	38.03	3.7541	60.81	3.9437	28.67
Ours (ChatGLM-6B p-tuning v2)	4.2394	42.25	—	—	4.3111	49.60
Ours (GPT 3.5-turbo)	**4.5282**	**52.82**	**4.0297**	**76.22**	**4.6753**	**79.40**

注: 加粗表示同一数据集上的最好结果

在 Aero-QA 数据集上, 本文提出的方法同样取得了最优结果. 在直接使用 ChatGLM-6B 作为基座模型时, LLM-*Score* 和 *Accuracy* 提升分别达到了 0.57 和 7.79%, 而在使用 p-tuning v2 微调后的 ChatGLM-6B 作为基座模型时, 对应指标的提升同样可以达到 0.418 和 3.66%, 在使用 GPT 3.5-turbo 作为基座模型时, LLM-*Score* 和 *Accuracy* 相较于基线提升了 0.014 和 0.14%. 这进一步证明了本文所提方法的有效性.

实验结果表明, 无论是在机械制造领域还是航空航天领域, 本文提出的方法都能显著提升 LLM 的性能. 通过将检索到的三元组与问题的相关程度输入 LLM, 在多模型与多数据场景下, 使其能够更好地利用证据三元组, 从而提高其推理能力和回答问题的准确性, 证明了本文提出方法的有效性.

此外, 在 LLM-*Score* 提升的显著性方面, 观察到 Aero-QA 上的 3 种 LLM 的提升差异性较大. 这可能是因为 Aero-QA 数据集为英文数据集, 且三元组规模较大, 但问题相对简单, 因此在 ChatGLM-6B (w/o finetuning) 以及 ChatGLM-6B (p-tuning v2) 上, 本文提出的方法提升较为显著, 而对于 GPT-3.5 模型, 在基线实验中已经取得了较好的效果 (4.6612), 因此本文方法提升幅度较小.

为了更全面地衡量问答模型的性能, 本文引入了用户满意度 CAST 指标, 在 Mecha-QA 数据集上使用 GPT 3.5-turbo 模型的结果如表 6 所示. 随机从 Mecha-QA 的测试集中抽取 100 条数据, 由两位人类专家独立对模型的回复进行评分, 两位专家评分的一致性指标 k 为 0.7714. 从表中结果可以发现, 本文提出的方法可以有效提升用户对问答系统的满意度.

表 6 在 Mecha-QA 数据集上 CAST 实验结果 (%)

模型	CAST
GPT 3.5-turbo + KG-Triples	83
Ours (GPT 3.5-turbo)	87

4.4 消融实验结果分析

本文进行了一系列的消融实验, 以进一步验证 prompt 中三元组与问题的相关程度以及排序模块的有效性. 具体来说, 为了验证三元组与问题的相关性程度的作用, 在构造 prompt 时, 仅根据三元组的相关性进行排序, 而不将相关性的具体数值作为 prompt 的内容输入 LLM, 即 w/o score. 为了验证排序模块的作用, 在构造 prompt 时, 仅将三元组和对应的相关性输入 LLM 而不进行排序, 即 w/o sort. 最终, 若既不输入三元组与问题的具体相关性程度也不进行排序, 则将乱序三元组直接构建 prompt 输入 LLM, 即 w/o scores & w/o sort. 消融实验的实验结果如表 7 所示, 将 4 种方式构建的 prompt 在各基座模型和数据集下的表现进行排名, 得到的平均排名如表 8 所示. 表 7 中未使用 ChatGLM-6B (p-tuning v2) 在 Mecha-QA-3D 上进行实验, 这是因为 Mecha-QA-3D 不包含训练数据, 而该方

法需要进行训练, 因此没有进行对应的实验. 可以看到, 若仅将三元组与问题的相关性程度输入 LLM 而不进行排序 (即 w/o sort), 则该部分信息会成为噪声, 对 LLM 的回复造成干扰, 从而降低模型生成答案的准确性. 同时, 若仅根据相关性进行排序而不将对应的相关性数值输入模型 (w/o scores), 相较于既不输入具体相关性数值和排序 (即 w/o scores & w/o sort) 或仅输入具体相关性数值 (即 w/o sort), 在模型的准确性上有一定程度的提升. 而将两者进行结合, 把三元组与问题的相关性程度作为 prompt 的一部分, 并基于此进行排序 (w/ scores & w/ sort), 可以取得最优的效果.

表 7 消融实验结果

模型	提示词	Mecha-QA		Mecha-QA-3D		Aero-QA	
		LLM-*Score* (↑)	*Accuracy* (↑) (%)	LLM-*Score* (↑)	*Accuracy* (↑) (%)	LLM-*Score* (↑)	*Accuracy* (↑) (%)
ChatGLM-6B (w/o finetuning)	w/ scores & w/ sort	**4.1408**	**38.03**	**3.7541**	**60.81**	3.9437	**28.67**
	w/o sort	3.9648	30.99	3.5730	55.68	3.1103	16.38
	w/o scores	4.0634	32.39	3.6324	56.49	**4.0009**	27.36
	w/o scores & w/o sort	4.0845	29.85	3.3595	49.46	3.3660	20.88
ChatGLM-6B (p-tuning v2)	w/ scores & w/ sort	**4.2394**	**42.25**	—	—	**4.3111**	**49.60**
	w/o sort	4.1901	39.44	—	—	3.6260	42.48
	w/o scores	4.1620	37.32	—	—	4.2689	46.97
	w/o scores & w/o sort	4.2113	40.85	—	—	3.8944	45.94
GPT 3.5-turbo	w/ scores & w/ sort	**4.5282**	**52.82**	**4.0297**	**76.22**	**4.6753**	**79.40**
	w/o sort	4.4225	51.11	3.8865	75.95	4.6349	74.66
	w/o scores	4.4930	51.41	3.9541	75.14	4.6434	71.56
	w/o scores & w/o sort	4.3380	47.18	3.8405	72.43	4.6612	79.26

注: 加粗表示同组实验中的最好结果, 下划线表示同组实验中的最差结果

表 8 消融实验中平均排名

提示词	平均排名
w/ scores & w/ sort	**1.0625**
w/o sort	3.3125
w/o scores	2.5
w/o scores & w/o sort	3.125

值得注意的是, 在 Aero-QA 的实验中, 使用未微调的 ChatGLM-6B 时, 不添加相关性分数 (w/o scores) 在 LLM-*Score* 上取得了比添加分数且排序 (w/ scores & w/ sort) 更好的结果. 这可能是由于 ChatGLM-6B 参数量只有 6B, 模型能力特别是英文能力较弱, 而 Aero-QA 是一个英文数据集, 所以在该数据集上, 不微调进行实验时, 相较于仅对三元组排序 (w/o scores) 的方法, 引入了额外的分数信息, 对模型产生了一定程度的干扰, 从而导致效果略有下降.

4.5 案例分析

图 7 展示了采用本文提出的方法增强 LLM 在 Mecha-QA 数据集上进行问答与直接输入检索到的三元组到 ChatGLM-6B 中进行问答的效果对比, 其中, 加粗的检索到的三元组表示问题对应的关键知识.

从表中的结果可得, 相较于直接输入检索的三元组, 本文提出的方法可以显著减少答案中包含的不相关信息. 这是由于检索到的三元组包含大量的噪声数据, 若直接将其与关键知识一同输入 LLM, 则 LLM 不能有效判断出关键知识, 从而回复受到噪声数据的干扰, 最终导致回复中包含大量不相关的信息, 甚至回复错误.

然而, 当检索到的三元组中噪声数据较少时 (如图 7 中第 3 个问题, 检索到的三元组均为关键知识), 由于本文提出的方法额外引入了“与问题相关性”的信息, 则在该场景下有可能会对 LLM 的回复造成干扰, 导致回复不准确.

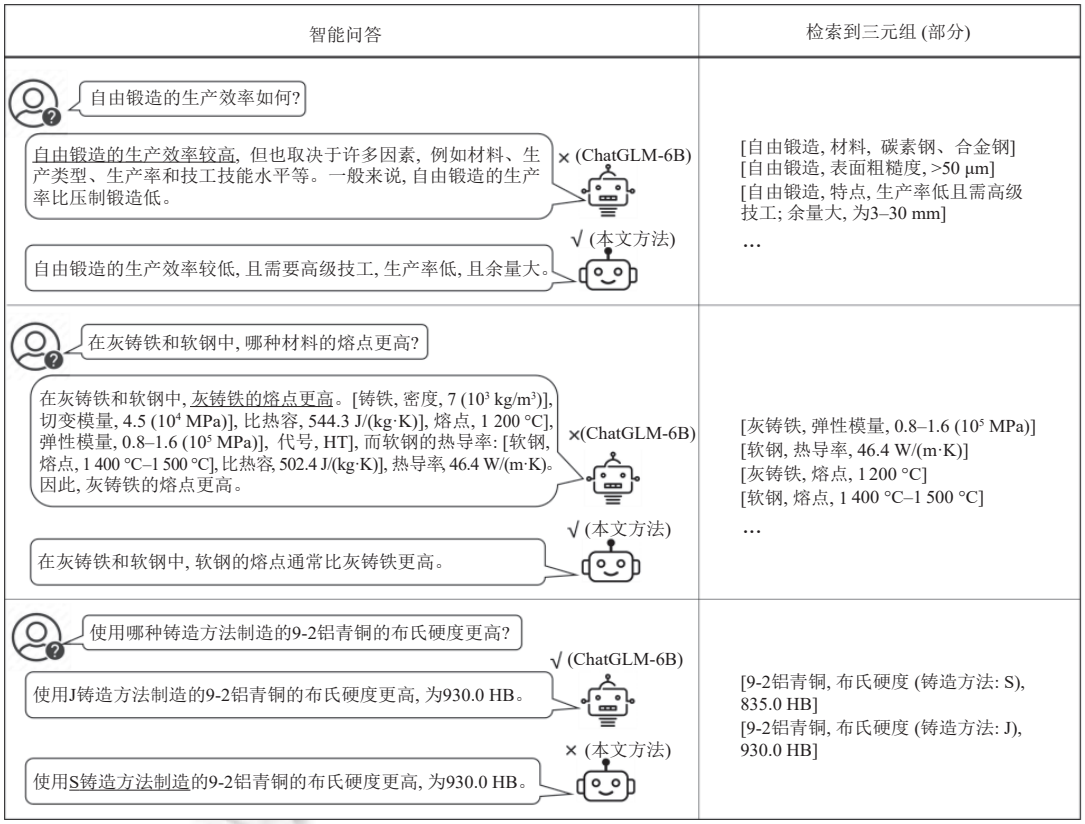


图 7 案例分析

5 总 结

本文探索了在“检索-问答”的框架下增强 LLM 垂直领域问答能力的方法. 首先, 为了促进机械制造领域知识问答的研究, 本文构建了一个机械制造领域的知识图谱与问答数据集, 即: Mecha-QA. 随后, 本文提出了一种利用知识相关性增强 LLM, 实现垂直领域知识图谱问答的方法. 该方法考虑到检索到的知识与问题的相关性不同, 对最终答案的贡献也不同. 由此提出了一个问题相关性计算模块, 将检索到的三元组与问题相关性进行量化, 并输入 LLM 生成最终的回复, 从而将利用大模型进行问答的框架变为“检索-相关性评估-问答”.

在未来的工作中, 我们将进一步对 Mecha-QA 数据集进行扩充, 增加其规模. 此外, 我们将研究如何更精确地检索知识, 从检索的角度减少知识增强大模型中的噪声, 提升 LLM 的问答性能.

References:

[1] Wang ZY, Cheng JP, Wang HX, Wen JR. Short text understanding: A survey. Journal of Computer Research and Development, 2016, 53(2): 262–269 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20150742]

[2] Cao SL, Shi JX, Hou L, Li JZ. Question answering over knowledge base: An overview. Chinese Journal of Computers, 2023, 46(3): 512–539 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2023.00512]

[3] Fan YF, Zou BW, Xu QT, Li ZF, Hong Y. Survey on commonsense question answering. Ruan Jian Xue Bao/Journal of Software, 2024, 35(1): 236–265 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6913.htm [doi: 10.13328/j.cnki.jos.006913]

[4] Zeng AH, Liu X, Du ZX, Wang ZH, Lai HY, Ding M, Yang ZY, Xu YF, Zheng WD, Xia X, Tam WL, Ma ZX, Xue YF, Zhai JD, Chen WG, Liu ZY, Zhang P, Dong YX, Tang J. GLM-130B: An open bilingual pre-trained model. In: Proc. of the 11th Int'l Conf. on Learning Representations. 2023. https://iclr.cc/virtual/2023/poster/11329

[5] Du ZX, Qian YJ, Liu X, Ding M, Qiu JZ, Yang ZL, Tang J. GLM: General language model pretraining with autoregressive blank infilling. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational

Linguistics, 2022. 320–335. [doi: 10.18653/v1/2022.acl-long.26]

[6] Shu WT, Li RX, Sun TX, Huang XJ, Qiu XP. Large language models: Principles, implementation, and progress. Journal of Computer Research and Development, 2024, 61(2): 351–361 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.202330303]

[7] Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsbug: Association for Computational Linguistics, 2020. 1906–1919. [doi: 10.18653/v1/2020.acl-main.173]

[8] Agrawal G, Kumarage T, Alghamdi Z, Liu H. Can knowledge graphs reduce hallucinations in LLMs?: A survey. arXiv:2311.07914, 2024.

[9] Zhang TC, Tian X, Sun XH, Yu MH, Sun YH, Yu G. Overview on knowledge graph embedding technology research. Ruan Jian Xue Bao/Journal of Software, 2023, 34(1): 277–311 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6429.htm [doi: 10.13328/j.cnki.jos.006429]

[10] Baek J, Aji AF, Saffari A. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In: Proc. of the 1st Workshop on Matching from Unstructured and Structured Data. Toronto: Association for Computational Linguistics, 2023. 70–98. [doi: 10.18653/v1/2023.matching-1.7]

[11] Kim J, Kwon Y, Jo Y, Choi E. KG-GPT: A general framework for reasoning on knowledge graphs using large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023. 9410–9421. [doi: 10.18653/v1/2023.findings-emnlp.631]

[12] Taffa TA, Usbeck R. Leveraging LLMs in scholarly knowledge graph question answering. In: Joint Proc. of Scholarly QALD 2023 and SemREC 2023 co-located with the 22nd Int'l Semantic Web Conf. ISWC 2023. Athens: CEUR-WS. org, 2023. 3592.

[13] Qiao SJ, Yang GP, Yu Y, Han N, Qin X, Qu LL, Ran LQ, Li H. QA-KGNet: Language model-driven knowledge graph question-answering model. Ruan Jian Xue Bao/Journal of Software, 2023, 34(10): 4584–4600 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6882.htm [doi: 10.13328/j.cnki.jos.006882]

[14] Zhang HY, Wang X, Han LF, Li Z, Chen ZR, Chen Z. Research on question answering system on joint of knowledge graph and large language models. Journal of Frontiers of Computer Science & Technology, 2023, 17(10): 2377–2388 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.2308070]

[15] Bao ZJ, Chen W, Xiao SZ, Ren K, Wu JA, Zhong C, Peng JJ, Huang XJ, Wei ZY. DISC-MedLLM: Bridging general large language models and real-world medical consultation. arXiv:2308.14346, 2023.

[16] Cui JX, Ning MN, Li ZJ, Chen BH, Yan Y, Li H, Ling B, Tian YH, Yuan L. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. arXiv:2306.16092, 2024.

[17] Wang N, Yang HY, Wang CD. FinGPT: Instruction tuning benchmark for open-source large language models in financial datasets. arXiv: 2310.04793, 2023.

[18] Wang WS, Xing M. Mechanical Manufacturing Handbook. Shenyang: Liaoning Science and Technology Publishing House, 2002 (in Chinese).

[19] Agarwal A, Gawade S, Azad AP, Bhattacharyya P. KITLM: Domain-specific knowledge integration into language models for question answering. arXiv:2308.03638, 2023.

[20] Bulian J, Buck C, Gajewski W, Börschinger B, Schuster T. Tomayto, tomahto. Beyond token-level answer equivalence for question answering evaluation. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics, 2022. 291–305. [doi: 10.18653/v1/2022.emnlp-main.20]

[21] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: Open and efficient foundation language models. arXiv:2302.13971, 2023.

[22] Gao YF, Xiong Y, Gao XY, Jia KX, Pan JL, Bi YX, Dai Y, Sun JW, Wang M, Wang HF. Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997, 2024.

[23] Che WX, Dou ZC, Feng YS, Gui T, Han XP, Hu BT, Huang ML, Huang XJ, Liu K, Liu T, Liu ZY, Qin B, Qiu XP, Wan XJ, Wang YX, Wen JR, Yan R, Zhang JJ, Zhang M, Zhang Q, Zhao J, Zhao X, Zhao YY. Towards a comprehensive understanding of the impact of large language models on natural language processing: Challenges, opportunities and future directions. SCIENTIA SINICA Informationis, 2023, 53(9): 1645–1687 (in Chinese with English abstract). [doi: 10.1360/SSI-2023-0113]

[24] Hu EJ, Shen YL, Wallis P, Allen-Zhu Z, Li YZ, Wang SA, Wang L, Chen WZ. LoRA: Low-rank adaptation of large language models. In: Proc. of the 10th Int'l Conf. on Learning Representations. 2022. [doi: 10.48550/arXiv.2106.09685]

[25] Liu X, Ji KX, Fu YC, Tam W, Du ZX, Yang ZL, Tang J. P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational Linguistics, 2022. 61–68. [doi: 10.18653/v1/2022.acl-short.8]

[26] Liu X, Ji KX, Fu YC, Tam WL, Du ZX, Yang ZL, Tang J. P-Tuning v2: Prompt tuning can be comparable to fine-tuning universally

across scales and tasks. arXiv:2110.07602, 2022.

[27] Wang HC, Liu C, Xi NW, Qiang ZW, Zhao SD, Qin B, Liu T. HuaTuo: Tuning LLaMA model with Chinese medical knowledge. arXiv:2304.06975, 2023.

[28] Griffith S, Subramanian K, Scholz J, Isbell CL, Thomaz A. Policy shaping: Integrating human feedback with reinforcement learning. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2625–2633.

[29] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proc. of the 34th Conf. on Neural Information Processing Systems. 2020. 9459–9474.

[30] Guu K, Lee K, Tung Z, Pasupat P, Chang MW. Retrieval augmented language model pre-training. In: Proc. of the 37th Int'l Conf. on Machine Learning. 2020. 3929–3938.

[31] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333–389. [doi: 10.1561/1500000019]

[32] Levonian Z, Li CL, Zhu WD, Gade A, Henkel O, Postle ME, Xing WL. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. arXiv:2310.03184, 2023.

[33] Yu WH, Iter D, Wang SH, Xu YC, Ju MX, Sanyal S, Zhu CG, Zeng M, Jiang M. Generate rather than retrieve: Large language models are strong context generators. In: Proc. of the 11th Int'l Conf. on Learning Representations. 2023. https://iclr.cc/virtual/2023/poster/12027

[34] Ma XB, Gong YY, He PC, Zhao H, Duan N. Query rewriting in retrieval-augmented large language models. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023. 5303–5315. [doi: 10.18653/v1/2023.emnlp-main.322]

[35] Gao LY, Ma XG, Lin J, Callan J. Precise zero-shot dense retrieval without relevance labels. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023. 1762–1777. [doi: 10.18653/v1/2023.acl-long.99]

[36] Huang DR, Wei ZZ, Yue AZ, Zhao X, Chen ZL, Li R, Jiang K, Chang BX, Zhang QL, Zhang SJ, Zhang Z. DSQA-LLM: Domain-specific intelligent question answering based on large language model. In: Proc. of the 1st Int'l Conf. on AI-generated Content. Shanghai: Springer, 2023. 170–180. [doi: 10.1007/978-981-99-7587-7_14]

[37] Yu DH, Zhu CG, Fang YW, Yu WH, Wang SH, Xu YC, Ren X, Yang YM, Zeng M. KG-FiD: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational Linguistics, 2022. 4961–4974. [doi: 10.18653/v1/2022.acl-long.340]

[38] Tan YM, Min DH, Li Y, Li WB, Hu N, Chen YR, Qi GL. Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family. arXiv:2303.07992, 2023.

[39] Omar R, Mangukiya O, Kalnis P, Mansour E. ChatGPT versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph Chatbots. arXiv:2302.06466, 2023.

[40] Wu YK, Hu N, Bi S, Qi GL, Ren J, Xie AH, Song W. Retrieve-rewrite-answer: A KG-to-text enhanced LLMs framework for knowledge graph question answering. arXiv:2309.11206, 2023.

[41] Soman K, Rose PW, Morris JH, Akbas RE, Smith B, Peetoom B, Villouta-Reyes C, Cerono G, Shi YM, Rizk-Jackson A, Israni S, Nelson CA, Huang S, Baranzini SE. Biomedical knowledge graph-optimized prompt generation for large language models. arXiv:2311.17330, 2024.

[42] Wang XT, Yang QW, Qiu YT, Liang JQ, He QY, Gu ZH, Xiao YH, Wang W. KnowledGPT: Enhancing large language models with retrieval and storage access on knowledge bases. arXiv:2308.11761, 2023.

[43] Hu CX, Fu J, Du CZ, Luo SM, Zhao JB, Zhao H. ChatDB: Augmenting LLMs with databases as their symbolic memory. arXiv: 2306.03901, 2023.

[44] Li ZY, Fan SQ, Gu Y, Li XX, Duan ZC, Dong BW, Liu N, Wang JY. FlexKBQA: A flexible LLM-powered framework for few-shot knowledge base question answering. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2024. 18608–18616. [doi: 10.1609/aaai.v38i17.29823]

[45] Agarwal D, Das R, Khosla S, Gangadharaiah R. Bring your own KG: Self-supervised program synthesis for zero-shot KGQA. arXiv: 2311.07850, 2024.

[46] Chen YL, Zhang YM, Yu JF, Yang L, Xia R. In-context learning for knowledge base question answering for unmanned systems based on large language models. In: Proc. of the 8th China Conf. on Knowledge Graph and Semantic Computing. Shenyang: Springer, 2023. 327–339. [doi: 10.1007/978-981-99-7224-1_26]

[47] Chu XT, Liu JP, Wang J, Wang XF, Wang YF, Wang M, Gu XX. CSDR-BERT: A pre-trained scientific dataset match model for Chinese scientific dataset retrieval. arXiv:2301.12700, 2023.

[48] Liu NF, Lin K, Hewitt J, Paranjape A, Bevilacqua M, Petroni F, Liang P. Lost in the middle: How language models use long contexts.

Trans. of the Association for Computational Linguistics, 2024, 12: 157–173. [doi: 10.1162/tacl_a_00638]
[49] Agarwal A, Gite R, Laddha S, Bhattacharyya P, Kar S, Ekbal A, Thind P, Zele R, Shankar R. Knowledge graph-deep learning: A case study in question answering in aviation safety domain. In: Proc. of the 13th Language Resources and Evaluation Conf. Marseille: European Language Resources Association, 2022. 6260–6270.
[50] Mañas O, Krojer B, Agrawal A. Improving automatic VQA evaluation using large language models. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2024. 4171–4179. [doi: 10.1609/aaai.v38i5.28212]
[51] Zheng LM, Chiang WL, Sheng Y, Zhuang SY, Wu ZH, Zhuang YH, Lin Z, Li ZH, Li DC, Xing EP, Zhang H, Gonzalez JE, Stoica I. Judging LLM-as-a-judge with MT-bench and Chatbot arena. In: Proc. of the 37th Conf. on Neural Information Processing Systems. 2023. 36.
[52] Kim Y, Levy J, Liu Y. Speech sentiment and customer satisfaction estimation in Socialbot conversations. In: Proc. of the 21st Annual Conf. of the Int'l Speech Communication Association. Shanghai: Int'l Speech Communication Association, 2020. 1833–1837.

附中文参考文献:

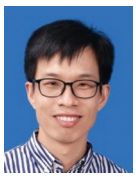
[1] 王仲远, 程健鹏, 王海勋, 文继荣. 短文本理解研究. 计算机研究与发展, 2016, 53(2): 262–269. [doi: 10.7544/issn1000-1239.2016.20150742]
[2] 曹书林, 史佳欣, 侯磊, 李涓子. 知识库问答研究进展与展望. 计算机学报, 2023, 46(3): 512–539. [doi: 10.11897/SP.J.1016.2023.00512]
[3] 范怡帆, 邹博伟, 徐庆婷, 李志峰, 洪宇. 常识问答研究综述. 软件学报, 2024, 35(1): 236–265. http://www.jos.org.cn/1000-9825/6913.htm [doi: 10.13328/j.cnki.jos.006913]
[6] 舒文韬, 李睿潇, 孙天祥, 黄萱菁, 邱锡鹏. 大型语言模型: 原理、实现与发展. 计算机研究与发展, 2024, 61(2): 351–361. [doi: 10.7544/issn1000-1239.202330303]
[9] 张天成, 田雪, 孙相会, 于明鹤, 孙艳红, 于戈. 知识图谱嵌入技术研究综述. 软件学报, 2023, 34(1): 277–311. http://www.jos.org.cn/1000-9825/6429.htm [doi: 10.13328/j.cnki.jos.006429]
[13] 乔少杰, 杨国平, 于泳, 韩楠, 覃晓, 屈露露, 冉黎琼, 李贺. QA-KGNet: 一种语言模型驱动的知识图谱问答模型. 软件学报, 2023, 34(10): 4584–4600. http://www.jos.org.cn/1000-9825/6882.htm [doi: 10.13328/j.cnki.jos.006882]
[14] 张鹤译, 王鑫, 韩立帆, 李钊, 陈子睿, 陈哲. 大语言模型融合知识图谱的问答系统研究. 计算机科学与探索, 2023, 17(10): 2377–2388. [doi: 10.3778/j.issn.1673-9418.2308070]
[18] 王宛山, 刑敏. 机械制造手册. 沈阳: 辽宁科学技术出版社, 2002.
[23] 车万翔, 窦志成, 冯岩松, 桂韬, 韩先培, 户保田, 黄民烈, 黄萱菁, 刘康, 刘挺, 刘知远, 秦兵, 邱锡鹏, 万小军, 王宇轩, 文继荣, 严睿, 张家俊, 张民, 张奇, 赵军, 赵鑫, 赵妍妍. 大模型时代的自然语言处理: 挑战、机遇与发展. 中国科学: 信息科学, 2023, 53(9): 1645–1687. [doi: 10.1360/SSI-2023-0113]

马杰(1993－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为多模态知识图谱, 多模态内容理解, 多模态内容安全.

张若非(1974－), 男, 博士, 教授, 博士生导师, 主要研究领域为机器学习, 数据挖掘, 自然语言处理, 多模态内容表示和理解.

孙望淳(2000－), 男, 硕士生, 主要研究领域为自然语言处理, 知识图谱, 大语言模型.

李帅鹏(1996－), 男, 博士生, CCF 学生会员, 主要研究领域为自然语言处理, 知识图谱, 大语言模型.



王平辉(1984－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习与数据挖掘, 自然语言处理, 移动互联网安全, 网络故障分析与诊断.

苏洲(1973－) 男, 博士, 教授, 博士生导师, 主要研究领域为通信网络安全, 物联网安全, 信息物理融合系统安全.