doi: 10.13413/j.cnki.jdxblxb.2023420

基于大语言模型的 Linux 课程问答系统

郭 东[1], 黄光强[1], 刘 颖[2]
(1. 吉林大学 计算机科学与技术学院, 长春 130012; 2. 吉林大学 公共计算机教学与研究中心, 长春 130012)

摘要: 基于国产主流大语言模型, 设计一个 Linux 课程知识问答系统. 该系统结合检索增强技术, 能根据人类反馈持续学习, 有助于解决 Linux 课程教学中如何更有效辅助学生学习的问题. 实验结果表明, 该系统提高了大语言模型回答的事实性, 能有效回答学生提问. 此外, 该系统以较低成本积累了以自然语言形式呈现的专业领域知识库, 降低了教师教学资料搜集整理的工作量.

关键词: Linux 课程; 大语言模型; 持续学习; 问答系统
中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5489(2024)06-1370-07

Linux Course Question Answering System Based on Large Language Models

GUO Dong[1], HUANG Guangqiang[1], LIU Ying[2]
(1. *College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*;
2. *Public Computer Education and Research Center, Jilin University, Changchun* 130012, *China*)

Abstract: Based on a domestic mainstream large language model, we designed a question answering system for the Linux course. This system, combined with retrieval enhancement technology, could continuously learn from human feedback, which helped to solve the problem of how to more effectively assist students' learning in the Linux course teaching. Experimental results show that the system improves the factuality of answers provided by the large language model and can effectively answer students' questions. In addition, the system accumulates a professional domain knowledge base presented in the form of natural language at a lower cost, reducing the workload of teachers in collecting and organizing teaching materials.

Keywords: Linux course; large language model; continuous learning; question answering system

近年来, 以 ChatGPT 为代表的大语言模型(large language model, LLM), 因其出色的语言生成和问答能力, 在为学生提供个性化教学和答疑辅助方面具有巨大潜力. 基于 LLM 的问答系统可以为学生提供实时、个性化的答疑服务, 上下文学习、思维链等新技术的发展可极大提升问答系统效果[1]. 目前, 在语言和基础数学等领域大语言模型的研究已取得很多成果. Gayed 等[2]基于 GPT-2 构建了 AI-KAKU 系统, 帮助学生降低英语写作难度. 面向阅读理解教学场景, Abdelghani 等[3]构建了基于 GPT-3 的人机对话系统 KidsAsk, 辅助学生提出问题, 加深对文章的理解程度. Pardos 等[4]构建了

收稿日期: 2023-10-12.

第一作者简介: 郭 东(1975—), 男, 汉族, 博士, 教授, 从事网络安全与大数据的研究, E-mail: guodong@jlu.edu.cn. **通信作者简介:** 刘 颖(1982—), 女, 满族, 硕士, 副教授, 从事智能控制与嵌入式系统的研究, E-mail: liuying329@jlu.edu.cn.

基金项目: 吉林省科技发展计划重点研发项目(批准号: 20230203037SF)和吉林省高等教育教学改革重点项目/吉林大学本科教学改革研究重点项目(批准号: 2021XZD038).

代数问题智能提示系统，通过 ChatGPT 生成代数问题的解答提示，引导学生解决问题.

尽管 LLM 在教育领域应用前景较好，但其也存在一些不足.

1) GPT 模型易产生幻觉. 在人工智能(AI)领域，幻觉是指 AI 自信地给出不符合训练数据的响应，生成错误或具有误导性的回答. 由于缺乏特定领域的专业数据，在专业课程教学领域中该问题尤为突出[5].

2) 专业领域优质数据获取困难. 如 Linux 操作系统课程教学领域，存在普通教学材料的结构和内容不适合机器学习模型训练、数据来源分散且格式不统一、人工数据标注成本较高等问题.

3) 大语言模型训练困难. 在特定领域数据上对预训练的模型进行微调时，虽然模型在特定任务的性能得到提高，但在其他任务的性能会急剧下降，即发生“灾难性遗忘”[6]. 另一方面，微调大语言模型仍需大量的计算资源和时间[7].

为增强大语言模型的事实性，减少产生幻觉，以检索外部数据增强语言模型能力的方法已广泛应用. Emily 等[8]提出了以维基百科作为外部知识库增强语言模型在开放领域问答的能力. He 等[9]提出了 Rethinking with Retrieval 方法提高语言模型的事实性. Peng 等[10]提出了基于维基百科等网络资料库和自主反馈的 LLM-Augmenter 框架，并在任务型对话和开放领域问答中验证了其有效性. Nakano 等[11]设计了一个基于文本的浏览器环境，并通过人类反馈等方式训练模型经由网络访问获取答案. 张鹤译等[12]提出了融合知识图谱的大语言模型问答系统，该系统能将自然语言转换为结构化数据，并匹配专业知识库内容. 总之，目前的研究工作一般关注于开放领域的问答系统，注重增强检索的准确性. 本文也应用了检索增强技术，但实验结果表明，对于 Linux 课程教学，有效数据的缺乏对语言模型正确率影响更大，因此本文更关注如何利用大语言模型丰富专业领域数据.

如何有针对性地修正大语言模型的错误认知，提供持续学习能力是目前该领域的研究热点之一. 部分研究[13-15]通过对模型参数进行直接修改，以增强模型的特定领域知识储备，纠正错误认知. 但该类方法的适用范围目前还相对有限. Kassner 等[16]在 BeliefBank 项目中提出了一个将检索系统先前的回答作为上下文的反馈机制. Tandon 等[17]从用户的自然语言反馈入手，探索了通过后处理方式修正模型错误认知、提升对用户指令理解的新方法. Dalvi 等[18]提出了 TeachMe 系统，该系统的记忆功能与文献[17]的工作类似，但采用了不同的方法修正模型认知. 这些研究集中于开放领域的问答系统，通过记忆用户的反馈降低模型逻辑推理的错误率. 而本文研究更注重于通过自动构建专业课程教学知识库，以增强模型的事实性.

本文研究以减少幻觉现象为核心目标，使用 LLM 自动化数据集处理，降低专业领域数据收集难度，并利用大语言模型的上下文学习(in-context learning)能力规避模型训练与微调的困难，提高跨领域泛化能力.

1　系统设计

本文系统由知识库和问答器两部分构成，系统整体架构如图 1 所示. 其中，知识库和问答器的核心基于百川智能推出的 Baichuan 系列大模型[19]，该系列模型在开源社区中受到广泛关注，并在多个基准测试中均性能优异，特别是在对话、逻辑推理和语义理解等方面. 在知识库部分，本文采用参数

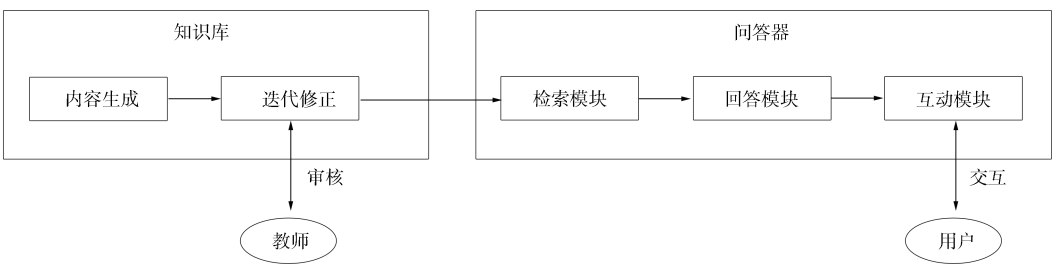


图 1　本文系统整体架构

Fig. 1　Overall architecture of proposed system

规模更大的 Baichuan 53B 模型，因其预训练更充分，幻觉现象更少. 在问答器部分，考虑到实时响应、算力和用户规模的要求，本文选择参数规模相对较小的 Baichuan 7B 模型. 问答器的主要功能是根据用户的提问，查询知识库获取相关的事实性资料，生成答案，并进一步引导学生主动思考.

1.1 知识库

提供与问题相关的事实性信息，有助于降低大语言模型的幻觉. 但传统的教学材料数据格式混乱，不利于机器检索，而完全手动构建和标注知识库成本较高. 本文利用大语言模型，将作业题库、教辅资料等专业领域知识整合并转化为问答对格式. 该格式的一个样例如图 2 所示. 采用问答对的形式有利于提高检索准确性. 通过利用大语言模型的隐含知识和能力，能实现专业领域知识库的自动化生成. 此外，为确保知识库内容的准确性，该方法以人类反馈为核心，不断进行迭代.

Q: UNIX系统的开发者是?
A: UNIX系统的主要开发者是贝尔实验室的肯·汤普逊(Ken Thompson)、丹尼斯·里奇(Dennis Ritchie)、布莱恩·柯林汉(Brian Kernighan)、道格拉斯·麦克罗伊(Douglas McIlray)和乔伊·奥桑纳 (Joe Ossanna).

图 2 问答对示例

Fig. 2 Example of a Q&A pair

1.1.1 内容生成

本文知识库内容的来源有如下两方面：

1) 模型的内隐知识. 参数量更大的 LLM 通常具有更大的预训练集和较少的幻觉现象，因此能准确回答部分题目和 Linux 相关概念. 对该特性的利用分两个阶段进行，均采用基于思维链技术的提示词以引导模型产生更详细的回复. 第一阶段，对数据集中具有明确答案的 Linux 习题，进行多次采样，只有当模型多次产生正确答案时才纳入知识库中. 第二阶段，为提高知识库的泛化性和系统在其他试题上回答的准确率，通过模型对纳入知识库中的试题提取出关键概念，并让模型进行详细解释，组成新的问答对.

2) 教辅资料. 为进一步提升知识库的泛用性和内容的多样性，从 Linux 相关教辅资料中截取片段，设计一个提示词引导模型提炼关键概念并生成问题. 对每个生成的问题，将其与教辅资料重新组合，再次使用模型获得该问题的答案. 生成内容将被用于计算 Knowledge F_1 指标，以删除与原文相关性不足的回答.

1.1.2 迭代修正

为缓解自动生成的知识库重复度较高、质量较差的问题，采用如下方法对知识库进行迭代修正，并引入人类反馈以确保生成问答对的质量. 该方法分为如下两部分：

1) 去重. 对每个新加入的问答对，在知识库中根据余弦距离检索知识点并计算相似度，若相似度高于阈值，则被标记为待合并. 对所有待合并的知识点，拼接后提交给 Baichuan 53B 模型进行合并操作. 提示词将指示 Baichuan 53B 模型不失泛化性和信息性地合并知识点.

2) 优化. 对在人工审核中发现有误的问答对，采用一种融合人工反馈与模型自我修正的策略进行优化. 审核员在标记错误的同时，提供自然语言描述，明确错误原因或改进建议. 结合原始提示词与相关反馈，模型重新生成答案. 新答案经再次审核后，如满足质量标准，则替代原始答案，否则重新进行优化. 该策略旨在保障知识库内容质量，以提高问答对的事实准确性.

1.2 问答器

问答器是系统的核心部分，负责与用户进行交互并提供答案. 为确保答案的准确性并提供与用户的有效交互，问答器由 3 个主要模块组成：检索模块、回答模块和互动模块.

1.2.1 检索模块

检索模块的主要任务是从知识库中找到与用户问题最相关的问答对. 该过程由两个子模块完成：FlagEmbedding[20] 和 Baichuan 7B. 首先，FlagEmbedding 模块将用户的问题转化为语义向量，并在知识库中检索与之相似的问答对. 检索到的问答对将与原始问题进行拼接，使用 Baichuan 7B 模块进行评估，以确定其与原始问题的相关性和有益性. 得分较高的问答对将作为参考信息提供给回答模块，否则继续判断下一个相关的问答对.

1.2.2　回答模块

回答模块的目标是生成用户问题的直接答案. 该模块主要由 Baichuan 7B 模型实现. 通过大模型思维链(CoT)技术[21]，本文设计了特定的提示词，引导 Baichuan 7B 模型利用检索模块获得的相关事实性材料对用户的问题进行回答. 在 Temperature＝0.8 的条件下，本文进行 5 次答案采样，选择出现次数最多的答案作为最终答案，并将该答案出现的次数作为置信度.

1.2.3　*互动模块*

互动模块旨在与用户进行深入交互，帮助用户更好地理解答案或相关概念，而不是直接给出解答. 该模块也由 Baichuan 7B 模型实现. 根据回答模块产生答案置信度的不同，互动模块将采用两种回答策略：1) 若置信度较低，则模块会选择解释与问题相关的概念，并拒绝对问题进行解答，避免产生误导；2) 若置信度较高，则模块采用苏格拉底式提问法，通过一系列的问题引导学生主动思考. 采用苏格拉底式提问法引导学生主动思考的问答实例如图 3 所示. 如果交互轮次超过预设的限制，则模块会直接给出回答模块得到的答案，并结束对话，以确保不会超出模型上下文限制.

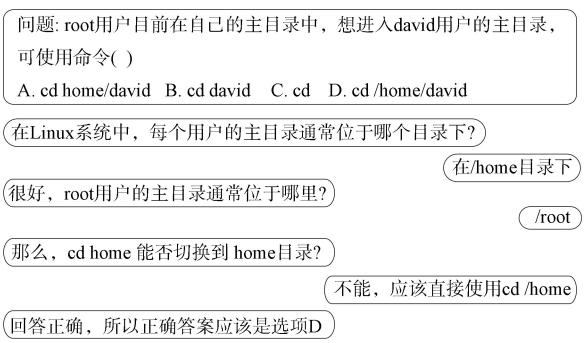


图 3　互动模块回应示例

Fig. 3　Example of response from interaction module

2　数据收集

为构建一个高效的 Linux 课程知识问答系统，首先需建立一个全面且具有代表性的题库. 题库数据不仅可以作为系统的基础知识库，还可用于后续的模型评估和验证.

2.1　数据来源

单一的数据来源可能会导致偏见或遗漏某些关键知识点. 因此，本文综合考虑两个主要的数据来源：校内超星课程的教学题库和公开的网络资料. 超星课程题库经过教师的筛选和审核，有较高的权威性和准确性. 公开网络资料涉及知识领域更广，确保了题库内容的多样性和规模.

2.2　数据集特性

本文收集的题目广泛涵盖了 Linux 的基本概念(5.3%)、命令(39.2%)、Git 与 Vim(27.6%)、系统管理与配置(19.6%)以及 Shell 及其程序设计(8.3%)等主要方面，题目组成对体现 Linux 的教学知识点具有代表性.

3　实　验

下面在收集到的 Linux 题目数据集基础上进行一系列实验，以验证本文系统的有效性，并通过消融实验检测系统中各模块的效果. 实验在配置 24 核 3.0 GHz Intel Core i9-13900K 处理器和 Nvidia GeForce RTX 4090 显卡的服务器上运行 Baichuan 7B 模型. 其他大语言模型，如 Baichuan 57B 模型等，使用官方 API 提供的接口访问. 算法部分基于深度学习框架 Pytorch 和微软大语言模型提示词工程库 Guidance 实现. 为保证稳定性，除特别标明的部分，均采用 Temperature＝0 的参数进行贪心解码. 数据集随机划分为 90%的训练集和 10%的测试集. 训练集将被用于生成知识库内容.

3.1　知识库内容生成质量

为检验知识库模块的生成质量和人工介入的作用，本文设计两个阶段的生成策略. 第一阶段完全依赖大语言模型自动化地生成知识库并进行去重处理. 第二阶段在第一阶段基础上进行人工反馈. 人工反馈包括对模型未能成功生成或生成不够详细的内容，补充事实性材料重新生成，并对模型生成的事实性错误进行修正. 实验以 GPT-4[22] 作为基线标准，从两个阶段得到的知识库中随机抽取问答对，生成对应内容. 从事实性与信息性两方面与 GPT-4 生成内容进行人工对比，实验结果如图 4 所

示. 由图4可见，知识库模块有效利用了人类反馈，在专业领域确保了内容的准确性和完整性.

3.2 问答器性能评估

为深入评估问答器的性能，本文选取几种主流的大语言模型，并在多个数据集上进行系统性的测试. 为确保测试的一致性和公正性，所有测试题目均采用选择题格式，并对每个大语言模型使用统一的思维链提示词进行提问. 在评分标准上，对于多项选择题，少选不被视为错误，但错选计入错误.

3.2.1 使用训练集进行评估

由于本文研究的应用场景在于辅助课堂Linux教学，学生用户大部分问题可能已存在训练集中. 为验证该模块是否能识别同一问题的不同表达方式，从训练集中随机选取题目，并对其进行语义上的微调，以模拟实际教学场景中学生可能采用的多样化提问方式. 图5为不同模型在训练集上的正确率. 由图5可见，尽管问题的表述发生了细微的变化，模块仍能正确识别，正确率相比原始的Baichuan 7B模型取得了显著提升.

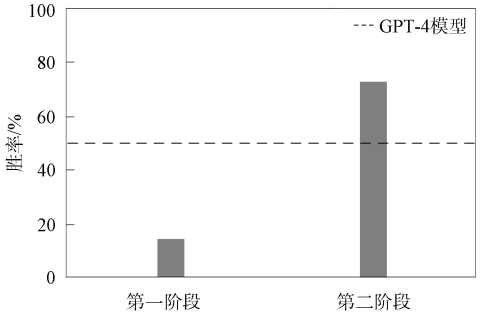


图4 知识库与GPT-4的比较胜率

Fig. 4 Comparison of winning rate between knowledge base and GPT-4

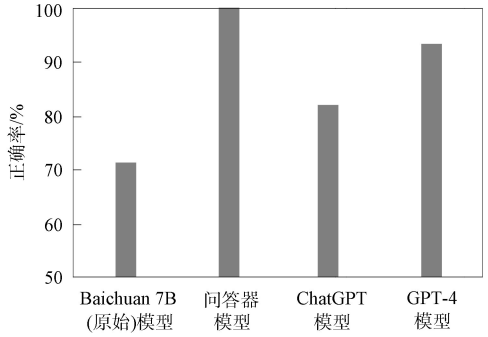


图5 不同模型在训练集上的正确率

Fig. 5 Accuracy of different models on training set

3.2.2 使用测试集进行评估

在知识库中不存在与测试集相同的题目. 图6为不同模型在测试集上的正确率. 由图6可见，使用测试集进行测试时，模块的准确率仍有上升，表明知识库内容具有一定通用性，能泛化到不同题目.

为确保系统在实际应用场景中的实用性，本文对其校准能力进行测试. 在大语言模型中，校准能力代表模型对答案的信心在多大程度上与其实际的准确率匹配，实验结果如图7所示. 由图7可见，对于该系统，高度确信的答案(置信度>4)正确率显著上升. 因此，通过规避对置信度较低的问题进行回答，能有效减少该系统在实际应用场景中的幻觉现象.

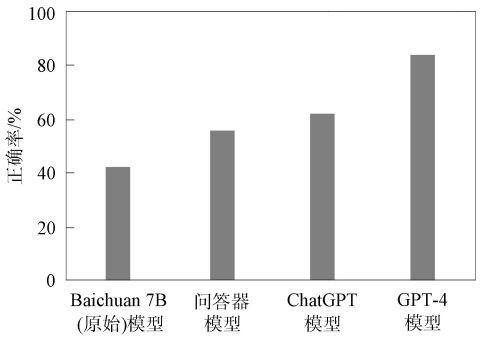


图6 不同模型在测试集上的正确率

Fig. 6 Accuracy of different models on test set

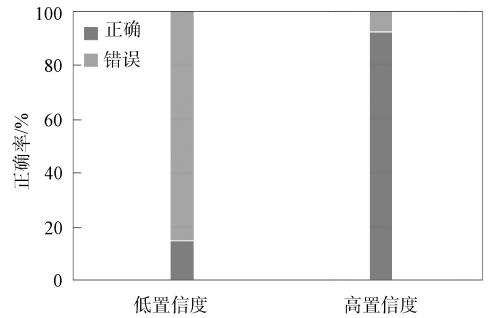


图7 本文系统在不同置信度下的正确率

Fig. 7 Accuracy of proposed system under different confidence levels

3.3 消融实验

为进一步验证本文系统中关键模块的有效性，下面进行消融实验.

3.3.1 知识库内容来源与系统泛化能力

第一组实验仅使用数据集中具有明确答案的Linux习题作为知识库，即知识库内容生成中只利用

模型内隐知识进行问答，称为 PQ. 第二组实验问答则是在习题基础上，补充基本概念的解释，并加入教辅资料，即采用本文所描述的完整方式生成知识库内容，称为 Q＋SMG. 实验在测试集上进行，结果如图 8 所示. 由图 8 可见，第二组问答实现了更全面的知识覆盖，提升了系统在未知题目中的正确率.

3.3.2　检索模块的效果

为进一步验证检索模块的性能，下面在测试集上进行实验，采用 3 种不同的检索策略.

1）直接相似度排序(DSR)：直接根据向量化后的余弦相似度进行排序，选取高于阈值的问题.

2）模型自我评估(ME)：在相似度排序的基础上，引入模型自我评估机制，即对问答对的有用性进行评分，选取高于阈值的问题.

3）人工查找(Human)：人工手动从知识库中选择最相关、最有帮助的问答对，作为基线标准.

图 9 为不同检索方式下问答器的正确率. 由图 9 可见，3 种检索方式对问答器的正确率影响较小. 但在模型自我评估策略中，有近 56%的相似问答对被检索模块拒绝采用，减少了回答模块的提示词长度，总体正确率却略有提升. 考虑到问答器在此过程中仅进行了 5 次采样，提示词长度的减少有效节省了计算资源.

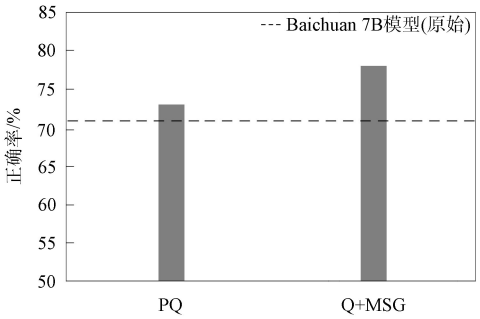


图 8　不同知识生成方式下问答器的正确率

Fig. 8　Accuracy of question answering machines under different knowledge generation methods

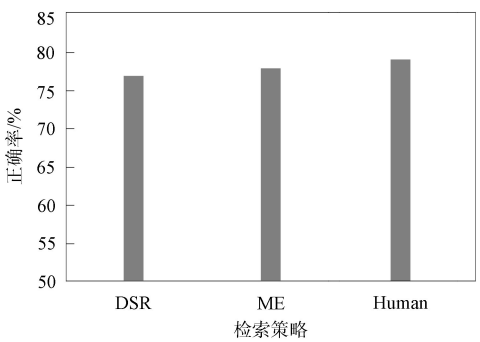


图 9　不同检索方式下问答器的正确率

Fig. 9　Accuracy of question answering machines under different retrieval methods

综上所述，基于国内主流大语言模型，本文设计了一个基于检索增强技术的 Linux 课程知识问答系统. 首先，基于国产大语言模型提出了一种自动生成专业领域知识库，并根据人类反馈进行知识迭代、持续学习的方法；其次，设计了一个针对专业课程教学领域的智能问答系统，该系统基于思维链与检索增强技术，能根据知识库中的事实进行回答，从而减少产生幻觉现象，并采用苏格拉底式提问引导学生得出答案. 实验结果表明，通过结合检索增强技术和人类反馈，可以提高系统的回答正确率，降低算力需求，使系统能有效应用于 Linux 课程的辅助教学中.

参　考　文　献

[1] 车万翔，窦志成，冯岩松，等. 大模型时代的自然语言处理：挑战、机遇与发展 [J]. 中国科学：信息科学，2023，53(9)：1645-1687. (CHE W X，DOU Z C，FENG Y S，et al. Towards a Comprehensive Understanding of the Impact of Large Language Models on Natural Language Processing：Challenges，Opportunities and Future Directions [J]. Scientia Sinica：Informationis，2023，53(9)：1645-1687.)

[2] GAYED J M，CARLON M K J，ORIOLA A M，et al. Exploring an AI-Based Writing Assistant's Impact on English Language Learners [J]. Computers and Education：Artificial Intelligence，2022，3：100055-1-100055-7.

[3] ABDELGHANI R，WANG Y H，YUAN X D，et al. GPT-3-Driven Pedagogical Agents to Train Children's Curious Question-Asking Skills [J]. International Journal of Artificial Intelligence in Education，2023，34：483-518.

[4] PARDOS Z A，BHANDARI S. Learning Gain Differences between ChatGPT and Human Tutor Generated Algebra Hints [EB/OL]. (2023-02-14)[2023-03-10]. https://arxiv.org/abs/2302.06871.

[5] JI Z W，LEE N，FRIESKE R，et al. Survey of Hallucination in Natural Language Generation [J]. ACM

Computing Surveys, 2023, 55(12): 1-38.
[6] PARISI I G, KEMKER R, PART L J, et al. Continual Lifelong Learning with Neural Networks: A Review [J]. Neural Networks, 2019, 113: 54-71.
[7] LIU H K, TAM D, MUQEETH M, et al. Few-Shot Parameter-Efficient Fine-Tuning Is Better and Cheaper than In-Context Learning [J]. Advances in Neural Information Processing Systems, 2022, 35: 1950-1965.
[8] EMILY D, STEPHEN R, KURT S, et al. Wizard of Wikipedia: Knowledge-Powered Conversational Agents [C]//Proceedings of the International Conference on Learning Representations. Piscataway, NJ: IEEE, 2019: 1-18.
[9] HE H F, ZHANG H M, ROTH D. Rethinking with Retrieval: Faithful Large Language Model Inference [EB/OL]. (2022-12-31)[2023-02-01]. https://arxiv.org/abs/2301.00303.
[10] PENG B L, GALLEY M, HE P C, et al. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback [EB/OL]. (2023-02-24)[2023-03-10]. https://arxiv.org/abs/2302.12813.
[11] NAKANO R, HILTON J, BALAJI S, et al. Webgpt: Browser-Assisted Question-Answering with Human Feedback [EB/OL]. (2022-07-01)[2023-02-10]. https://arxiv.org/abs/2112.09332.
[12] 张鹤译, 王鑫, 韩立帆, 等. 大语言模型融合知识图谱的问答系统研究 [J]. 计算机科学与探索, 2023, 17(10): 2377-2388. (ZHANG H Y, WANG X, HAN L F, et al. Research on Question Answering System on Joint of Knowledge Graph and Large Language Models [J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(10): 2377-2388.)
[13] MITCHELL E, LIN C, BOSSELUT A, et al. Fast Model Editing at Scale [EB/OL]. (2022-07-13)[2023-03-15]. https://arxiv.org/abs/2110.11309.
[14] DE CAO N, AZIZ W, TITOV I. Editing Factual Knowledge in Language Models [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACL, 2021: 6491-6506.
[15] HASE P, DIAB M, CELIKYILMAZ A, et al. Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs [EB/OL]. (2021-11-26)[2023-02-15]. https://arxiv.org/abs/2111.13654.
[16] KASSNER N, TAFJORD O, SCHÜTZE H, et al. BeliefBank: Adding Memory to a Pre-trained Language Model for a Systematic Notion of Belief [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACL, 2021: 8849-8861.
[17] TANDON N, MADAAN A, CLARK P, et al. Learning to Repair: Repairing Model Output Errors after Deployment Using a Dynamic Memory of Feedback [C]//Findings of the Association for Computational Linguistics: NAACL 2022. [S.l.]: ACL, 2022: 339-352.
[18] DALVI B, TAFJORD O, CLARK P. Towards Teachable Reasoning Systems: Using a Dynamic Memory of User Feedback for Continual System Improvement [C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACL, 2022: 9465-9480.
[19] YANG A Y, XIAO B, WANG B N, et al. Baichuan 2: Open Large-Sscale Language Models [EB/OL]. (2023-09-13)[2023-10-01]. https://arxiv.org/abs/2309.10305.
[20] FLAGOPEN. FlagEmbedding: Open-Source Embeddings [EB/OL]. (2023-08-02)[2023-09-03]. https://github.com/FlagOpen/FlagEmbedding.
[21] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.
[22] OPENAI. GPT-4 Technical Report [EB/OL]. (2023-03-15)[2023-05-10]. https://arxiv.org/abs/2303.08774.

(责任编辑: 韩 啸)