

文章编号: 1003-0077(2024)11-0001-12

中文医疗大模型综述：进展、评估与挑战

籍欣萌, 咎红英, 崔婷婷, 张坤丽

(郑州大学 计算机与人工智能学院, 河南 郑州 450001)

摘要: 大型语言模型 (Large Language Models, LLMs) 已在多个领域受到广泛关注, 并取得了令人瞩目的成绩。将 LLMs 应用于医学领域, 如辅助医疗诊断、影像报告生成等, 在人工智能和临床医学中都是很有前景的研究方向。由于中外医生对患者进行医疗诊断的过程存在差异, 包括治疗方式、用药习惯和用药剂量等, 特别是在传统中医领域, 因此, 为了更好地满足中文医疗领域的需求, 构建大规模真实的中文医学数据集, 开发中文医疗大模型是十分重要的。该文从医疗问诊、医学影像、心理健康三个方面对当前中文医疗大模型进行了概述, 并介绍了当前中文医疗大模型已有的评测基准, 阐述了当前中文医疗大模型面临的挑战, 如幻觉、价值对齐等。未来的研究将致力于解决这些问题, 并扩展医疗大模型的应用场景。

关键词: 大语言模型; 医疗大模型; 人工智能

中图分类号: TP391

文献标识码: A

A Survey of Chinese Large Language Models in Medicine: Progress, Evaluation and Challenges

Ji Xinmeng, ZAN Hongying, CUI Tingting, ZHANG Kunli

(School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, Henan 450001, China)

Abstract: Large language models have received widespread attention for their remarkable success in various fields. Applying large language models to the medical domain, such as assisting doctors in diagnosis and generating imaging reports, holds great promise for both artificial intelligence and clinical medicine. Due to differences between Chinese and foreign doctors in medical diagnosis, treatment methods, medication habits and drug dosages, it is important to build large-scale real Chinese medical datasets and develop Chinese large language models in medicine for the Chinese medical field. This paper overviews current Chinese large language models in medicine in three aspects: medical consultation, medical imaging, and mental health. Additionally, it introduces the evaluation benchmarks currently available for Chinese large language models in medicine and discusses the challenges like hallucinations and value alignment. Future researches of Chinese large language models in medicine are also discussed.

Keywords: large language models; medical large language models; artificial intelligence

0 引言

预训练语言模型 (Pre-training Language Models, PLMs)^[1] 是自然语言处理 (Natural Language Processing, NLP) 领域中的重要研究方向之一。它首先使用大量的无监督语料对语言模型进行预训练 (Pre-training), 再使用少量的标注语料对预训练后的语言模型进行微调 (Fine-tuning), 从而完成具体

的 NLP 任务^[2]。预训练语言模型在语音识别^[3]、机器翻译^[4-5]、情感分析^[6-7]、信息抽取^[8-9] 等多个任务上都展现出强大的性能, 如 Google AI 团队提出的预训练语言模型 BERT^[10], 在 11 项自然语言理解任务上刷新了最好指标。除了常见的 NLP 任务外, 还有专门为医疗领域量身定制的预训练语言模型。如在医疗领域, 研究人员引入了一种叫生物医学实体掩模 (Biomedical Entity Masking) 的技术, 以将更多的医学知识纳入模型中, 从而提高其在医疗相

收稿日期: 2024-01-24 定稿日期: 2024-07-03

基金项目: 国家自然科学基金 (U23A20316); 河南省科技攻关项目 (232102211041); 郑州市协同创新重大专项 (20XTZX11020)

关应用中的性能^[11]。对于医疗 PLMs,大多数模型都采用基于 BERT 架构的判别式方法,如 BioBERT^[12]、选择这种架构的原因是,生物医学领域中典型的任务,如命名实体识别、关系抽取、疾病预测等任务,归根结底都是分类任务。判别式方法在处理分类任务时具有较好的效果和适用性。

随着预训练语言模型的不断发展,模型参数规模也不断增加,当参数规模超过一定水平时,这个参数规模更大的语言模型实现了显著的性能提升,并出现了小模型中不存在的能力,如上下文学习。为了区别于 PLMs,这类模型被称为大型语言模型(Large Language Models, LLMs)。LLMs 经过预训练后,可以学习到更为丰富的通用语言表示,从而应用于各种下游任务。为了获得强大的下游性能,LLMs 可以在更小的、特定任务的数据集上进一步微调,这使得模型能够将其通用的语言表示适应于目标任务的特定需求^[13]。此外,大规模的预训练可以让 LLMs 学习广泛的语言知识,并对语言模式和概念有广泛的理解,这使得 LLMs 在“零样本”和“少样本”的学习场景中表现良好^[14]。在这些场景中,它们可以准确地执行下游任务,而不再需要人为标记的数据进行训练。在过去的几年中,各种各样的通用大型语言模型,如 PaLM^[15]、LLaMA^[16-17]、GPT 系列^[18]和 ChatGLM^[19-20]等,已经出现并推进了各种自然语言处理任务的发展,包括文本生成、文本摘要和问答等。

在通用 LLMs 取得巨大成功的启发下,医疗 LLMs 的开发和应用得到了越来越多的关注。同时,由于在低资源语言场景和医疗场景下通常无法获得用于训练的大量标记数据,LLMs 在“零样本”和“少样本”的场景中的表现显得尤为重要。Med-PaLM 2^[21]是谷歌专为医疗领域量身打造的 LLM。它是第一个可以在美国医疗执照考试风格试题的 MedQA 数据集^[22]上达到“医学专家”级别性能的 LLM,准确率超过 85%。此外,医疗 LLMs,如本草(BenTsao)^[23]、华佗 GPT^[24]、Visual Med-Alpaca^[25]等的发布,将 LLMs 微调以适应医疗保健领域的独特特征,改善了 LLMs 在医疗领域的应用。

此外,考虑到中国医学实践的独特重点和方法,中外医生在对患者进行医疗诊断的过程中存在很大差异,例如治疗方式、用药习惯和用药剂量等,特别是在传统中医领域。因此,中文医学信息处理急需

大规模真实的中文医学数据集来支撑,使用中国医学数据集来构建中文医学大语言模型也至关重要。现有的医疗 LLMs 的研究综述^[26-27]大多关注英文 LLMs 的医疗应用及其影响,为了对中文医疗大模型领域提供有价值的参考,本文从算法开发、性能评估以及 LLMs 面临的挑战等方面对中文医疗大模型相关研究进行了综述。

本文组织结构如下:首先为引言部分,第 1 节介绍目前已有的中文医疗大模型及其微调方法。第 2 节介绍和讨论现有关于中文医疗大模型的评测基准。第 3 节重点讨论目前医疗大模型面临的挑战,如数据安全、价值对齐等。最后,第四节给出本文的总结和展望。

1 中文医疗大模型

从 2023 年起,结合中文环境下的医疗问诊、中药特色以及日趋更为关键的心理医学领域的医疗大模型不断涌现,无论是大模型基座的选择、大模型微调技术、微调训练推理经济成本考量,还是医疗领域数据集构建方法、训练数据集、中文医疗大模型评价标准等各个方面,都呈现了各自的特色。

本节将分别从医疗问诊、医学影像和心理健康三个方向分别对中文医疗领域中现有的 LLMs 进行概述,强调它们各自的研究重点。部分开源中文医疗大模型相关信息如表 1 所示。

此外,国内研发机构、医疗机构也发布了较多医疗大模型,主要面向患者和医院两方面。针对患者的大模型产品,例如,阿里于 2023 年 10 月提出的通义仁心大模型和商汤发布的医疗健康大语言模型升级版,旨在为患者提供健康服务;而面向医院的大模型产品,如于 2023 年 12 月推出的 YiduCore 大模型,主要为医生在辅助诊断和病历整合等方面提供帮助。目前企业已发布的部分医疗大模型如图 1 所示。

1.1 医疗问诊大模型

医疗问诊大模型基于自然语言处理技术,能够根据患者的症状、病史和体检结果等信息,对患者进行诊断并提供治疗建议。该模型使用大量的医学文献和临床数据进行训练,能够快速准确地回答患者的问题。中文医疗问诊大模型分析如表 2 所示。

表 1 部分开源中文医疗大模型相关信息

模型	基座模型	训练方式	指令数据	应用	发布时间
DoctorGLM ^[28]	ChatGLM	SFT, IFT	问答对话数据	医疗问诊	2023.4
BenTsao ^[23]	LLaMA	SFT, IFT	问答对话数据	医疗问诊	2023.4
BianQue ^[29]	Clue-AI ^[43] / ChatYuan-large-v2	SFT, IFT	问答对话数据	医疗问诊	2023.4
HuatuoGPT ^[24]	BLOOMZ ^[44]	SFT, IFT, RLAI	问答对话数据	医疗问诊	2023.5
ChatGLM-Med ^[30]	ChatGLM	SFT, IFT	问答对话数据	医疗问诊	2023.5
MedicalGPT ^[31]	ChatGLM/LLaMA/ Baichuan/BLOOM ^[45]	SFT, IFT, RLHF	问答对话数据	医疗问诊	2023.6
ClinicalGPT ^[32]	BLOOM	SFT, IFT, RLHF	问答对话数据,电子病历	医疗问诊	2023.6
ChatMed ^[33]	LLaMA	SFT, IFT	问答对话数据	医疗问诊	2023.7
MINGMOE ^[34]	BLOOMZ	SFT, IFT	问答对话和任务数据	医疗问诊	2023.7
IvyGPT ^[35]	Baichuan/LLaMA/ Qwen/InternLM	SFT, IFT	问答对话数据	医疗问诊	2023.8
Zhongjing ^[36]	LLaMA	SFT, IFT, RLHF	问答对话数据	医疗问诊	2023.8
DISC-MedLLM ^[37]	Baichuan	SFT, IFT	问答对话数据	医疗问诊	2023.9
ShenNong-TCM ^[38]	LLaMA	SFT, IFT	中医药数据、知识图谱	中医药问诊	2023.8
XrayGLM ^[39]	VisualGLM-6B	SFT, IFT	胸部 X 光片数据集	医学影像	2023.4
Qilin-Med-VL ^[40]	LLaVA	SFT, IFT	X 射线、CT 扫描等多种 影像数据集	医学影像	2023.10
SoulChat ^[41]	ChatGLM	SFT, IFT	长文本、问答对话数据	心理健康	2023.6
MindChat ^[42]	Baichuan/InternLM/ Qwen	SFT, IFT	—	心理健康	2023.7

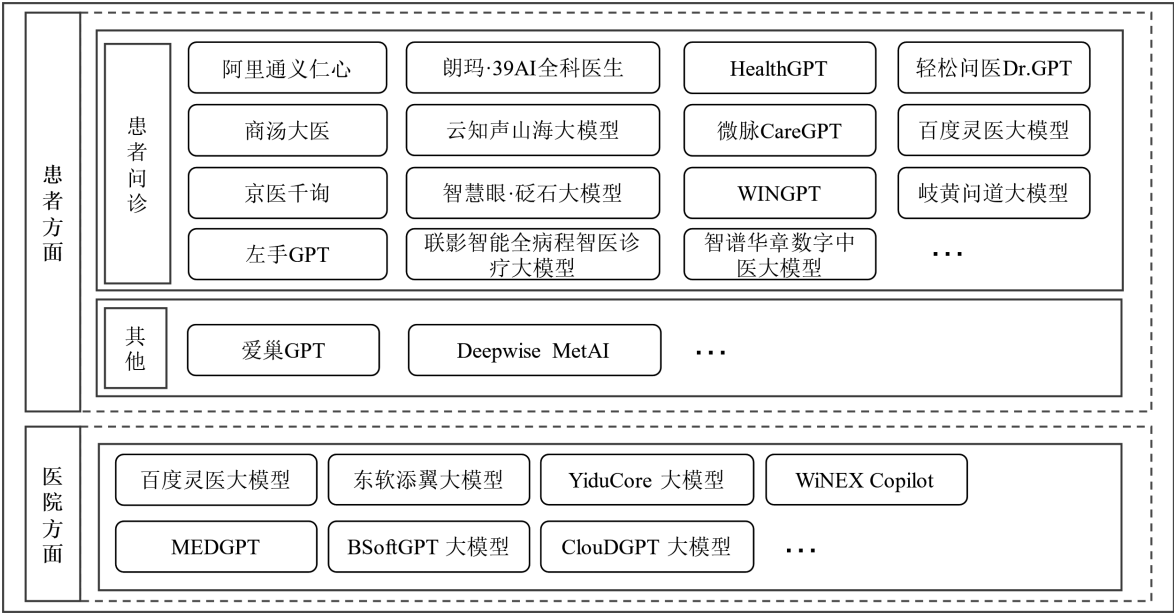


图 1 企业已发布的部分医疗大模型

表 2 现有中文医疗问诊大模型分析

模型	应用	特色	缺点
DoctorGLM ^[28]	医疗问诊	成本较低	存在能力缺失问题;量化困难;响应时间较长
BenTsao ^[23]	医疗问诊	融入了医学知识图谱 CMeKG 以及医学文献知识	仅针对“肝癌”疾病;单轮对话
BianQue ^[29]	多轮问询、健康建议	模型问询能力强,并在 2.0 版本扩充了多种医学数据	1.0 版本侧重于多轮询问,2.0 版本侧重于健康建议,需结合使用
HuatuoGPT ^[24]	多轮医疗问诊	使用的指令数据和问答数据来源丰富,并使用了 RLAIIF 来改善生成质量	数据均为问答数据,缺少多样性
ChatGLM-Med ^[30]	医疗问诊	融入医学知识图谱 CMeKG,并构建中文医学指令数据集	单轮对话;不支持合并病因和症状
MedicalGPT ^[31]	多轮医疗问诊	使用 RLHF 和直接偏好优化来提升模型效果,并提供基于知识库文件的 LLM 问答功能	部分数据来源于在线医疗百科,缺乏可靠性
ClinicalGPT ^[32]	医疗问诊、医学考试	使用的医学数据种类多样,来源丰富,并使用 RLHF 和近端策略优化以提高模型性能	数据集涵盖的科室不全
ChatMed ^[33]	医疗问诊	使用在线爬取的问诊数据,可以反映真实世界用户的问诊需求	单轮对话;训练数据的准确性和真实性没有保证
MINGMOE ^[34]	医疗问答、智能问诊、体检报告分析和影像报告分析	使用多轮情景诊断、案例分析数据和安全性数据	局限于单模态,无法直接从影像获得分析结果
IvyGPT ^[35]	多轮医疗问诊	进行了 LLM 的全流程训练,还支持网络检索功能	仅在医疗问答数据上进行微调,缺少多样性
Zhongjing ^[36]	医疗问诊	构建了真实医患对话的多轮对话数据集,首个实现了 LLM 的完整训练流程	数据多为医疗问答数据,缺乏多样性
DISC-MedLLM ^[37]	医疗问诊、治疗方案咨询	使用人工筛选构建符合人类偏好的高质量小规模行为微调样本	模型回答的准确性和帮助性没有保障
ShenNong-TCM ^[38]	中医药医疗问诊	使用中医药指令数据集和中医药知识图谱,回答中医药相关的问题的能力较强	单轮对话;中医药数据集相对较少

目前中文医疗问诊大模型主要源于通用大模型针对中文医学知识和语境进行微调,使其在医疗领域的问答效果得到显著提升。目前常用的微调方法主要包括有监督微调(Supervised Fine-Tuning, SFT)、指令微调(Instruction Fine-Tuning, IFT)等。

SFT 主要是使用人工标注的“输入,输出”文本对预训练的模型进行微调,使其更适应特定的下游任务。如华佗 GPT(HuatuoGPT)采用自构建的医用数据集,通过 SFT 方式对 Baichuan-7B^[46]和 Ziya-LaMA-13B-Pretrain-v1^[47]的模型进行微调,以提高模型在医学领域的问答能力。而 IFT 旨在提高模型的指令跟踪能力,并调整其输出以匹配给定的输出。例如,DoctorGLM^[28]通过 LoRA (Low-Rank Adaptation)^[48]和 P-Tuning v2^[49]对 ChatGLM-6B 进行微调,本草利用中文医学知识图谱和 GPT 3.5 构建了中文医学指令数据集,然后对 LLaMA-7B 进行了指令微调。

此外,在医疗领域,用户通常不能在一轮交互中

就清晰地表达自己的问题,常需要多轮询问,而前面提到的医疗问答模型,如 DoctorGLM 和本草,都侧重于解决单轮用户描述的问题,而忽略了“用户的描述可能存在不足”的情况。为了解决这个问题,扁鹊(BianQue)^[29]使用多轮问询对话对 ClueAI/ChatYuan-large-v2^[50]进行微调,增强模型的问询能力,并有效记录上下文信息,从而使模型在最后根据用户提供的信息综合给出诊断结果和建议。

在模型的准确性和有用性有一定的保障后,模型输出的结果是否符合人类偏好也是十分重要的。例如,模型的说话语气不应过于强硬和绝对、回答应简明扼要等。进行人类反馈强化学习(Reinforcement Learning from Human Feedback, RLHF)^[51]是使模型更符合人类偏好的方法之一。在人类反馈强化学习中,首先根据已有的输入文本通过大语言模型得到输出,并为生成的输出文本进行人工评分,作为奖励,根据得到的“输入文本,输出文本,奖励”三元组的数据集训练一个奖励模型,使用奖励模型提供

的奖励来训练主模型,通过奖励模型与主模型的交互来提升模型能力。例如华佗 GPT(HuatuoGPT)通过人类反馈强化学习策略设计,使得问答能够更好地适应简要问答和多轮对话的场景。Zhongjing^[36]实现了预训练、有监督微调、强化学习与人类反馈强化学习的完整训练流程,展现出了很好的泛化能力,在某些对话场景中甚至接近专业医生的专业水平。与 SFT 和 IFT 相比,RLHF 技术的应用相对较少,其主要原因可能是缺乏足够的稳定性。

这些工作不仅推动了中文医疗大模型领域的发展,也推动了 LLMs 在中医药领域的发展,提升了 LLMs 在中医药方面的知识和回答医学咨询的能力。如神农中医药大模型(ShenNong-TCM)^[38]通过引入以实体为中心的自指令(Entity-centric Self-instruct)^[52]方法,使用中医药指令数据集 ShenNong-TCM_Dataset 对中文 LLaMA-7B 模型进行 LoRA 微调,显著提高该模型在回答中医药相关问题方面的表现。

1.2 医学影像大模型

在医学领域,除了进行问诊外,影像也是重要的诊断依据。在患者进行问诊时,医生会通过询问症状、病史以及进行体格检查等方式获取信息。然而,有时候这些信息并不足以做出准确的诊断,这时就需要借助影像学来做进一步的检查和确认。影像学技术,如 X 射线、CT 扫描、MRI 等,能够提供详细的内部结构和组织的图像,为医生提供更全面的诊断信息。医生通常会综合患者的问诊信息与影像学检查结果,以制定最佳的治疗方案。因此,使模型拥有从影像中生成影像报告,并对影像报告进行分析的能力,能极大地减轻医生的负担。明医(MING)大模型^[34]使用多轮情景诊断和案例数据进行微调,可以对已有的影像报告内容进行分析,帮助患者了解诊断结果。但是其使用的数据为纯文本的输入输出,局限于单模态,无法直接根据影像进行分析。

医学影像模型是一种整合多种医学影像数据和临床数据的多模态模型。通用领域的多模态大模型(Multimodal Large Language Models, MLLMs)的研究,如 MiniGPT-4^[53]、mPLUG-Owl^[54]、Multimodal-GPT^[55]和 LLaVA^[56],是在 LLMs 的基础上进一步扩展了多模态能力,它可以接受多种类型的输入数据,如文本、图像、视频等,将它们融合到一起进行联合建模,通过学习不同模态之间的关联性和语义信息,MLLMs 能够更好地理解和处理多模态

数据^[57]。医学影像模型,是在通用领域的多模态大模型的基础上,对模型进行微调,使其可以处理多种类型的医学影像数据,如 CT、MRI、X 光等,并与其他临床数据进行结合,为医生提供更全面、更准确的诊断结果。

如剑桥团队提出的 Visual Med-Alpaca 模型,它在进行临床问诊的基础上,还能够根据输入的胸部 X 光片,直接给出影像报告结果。但是 Visual Med-Alpaca 模型使用的数据为英文诊断报告,其输出的诊断结果也是英文,不方便中国医生使用。XrayGLM^[39]是首个会看胸部 X 光片的中文医学影像大模型,它借助 ChatGPT 以及公开的数据集,构造了一个胸部 X 光影像-诊断报告对的中文医学多模态数据集,并将构造的数据集在 VisualGLM-6B^[58]上进行微调训练,在胸部 X 光片影像报告生成和诊断上表现良好。而 Qilin-Med-VL^[40]使用 X 射线、CT 扫描、回波描记术和超声检查等多种影像数据集对模型进行微调,丰富了模型对多样化的医疗场景的需求。但是 XrayGLM 和 Qilin-Med-VL 所使用的部分数据是使用 ChatGPT 由英文数据集翻译而来的,准确性有待商榷。

1.3 心理健康大模型

近年来,心理学问题也日益受到社会关注,焦虑症、抑郁症、创伤后应激障碍等心理健康难题给许多人带来困扰。心理健康大模型能够从大量的心理学文献和实际数据中进行学习和训练,应用于多种心理健康问题,减轻心理医生的负担,并为普及心理健康知识、提高心理健康服务水平和改善心理健康问题处理能力提供了新的机遇和可能性。

目前,国内也推出了一些心理健康大模型。国内首个心理大模型 Emohaa^[59]基于 GLM-4 大模型基座,包含 90 亿、320 亿、1 300 亿参数版本,上下文的记忆长度可达 8k+,能够有效覆盖各类复杂应用场景。Emohaa 基于心理学中的 Hill 助人理论,以用户为中心,与用户建立情感连接,从而满足用户的情感需求。Emohaa 还具有危机检测系统,能够对对话过程中用户的情绪状态、自伤自杀风险进行精准评估,从而提供相应的心理援助服务。灵心则(SoulChat)^[41]构建了超过 15 万规模的单轮长文本心理咨询指令与答案,在 ChatGLM-6B 上进行了全量参数的指令微调,旨在提升模型的共情能力,引导用户倾诉能力,以及提供合理建议的能力。漫谈(MindChat)^[42]是基于 Baichuan-13B、InternLM-7B^[60]

和通义千问 Qwen-7B^[61] 的微调模型,期望从心理咨询、心理评估、心理诊断、心理治疗四个维度帮助人们纾解心理压力与解决心理困惑,提高心理健康水平。

此外,在快节奏的社会生活中,不少中小學生也面临着学习压力和社交困扰等心理健康问题,为了更好地助力青少年的心理健康成长,并为他们提供个性化心理支持,目前也有专门为中小學生服务的心理健康大模型,如基于清华智谱 AI 的积极心理专用大模型“清小深”。“清小深”构建了涵盖学业、情感、社交和自我认同等多元主题的高质量语料库,对 ChatGLM3-6B 模型进行 LoRA 微调,并利用检索增强生成(Retrieval-Augmented Generation, RAG)^[62] 技术将传统的生成模型与外部积极心理知识检索组件相结合进行信息增强检索。“清小深”不仅能够为面临心理健康问题的学生提供积极的心理辅导,还能够帮助老师快速检索所需的心理辅导策略和解决方法。

2 中文医疗大模型评估进展

尽管医疗大模型有广阔的应用前景,但评估模

型仍面临着许多挑战。在医院中部署大模型可能引发重大的伦理问题,并且从实际场景中获取反馈意见相对困难。现有的模型评估通常使用自动评估方法,例如,使用 GPT-4 评估,但考虑到医疗领域对准确性和专业性的要求,这种没有参考答案的评估方法缺乏说服力。而专业人工评估则需要昂贵的人力成本,并且难以扩大规模应用。现有的评测基准如 BioLAMA^[63] 拥有参考答案,但它主要用于评估掩码语言模型而不是自回归模型;MultiMedBench^[64] 包含问题解答、报告总结、视觉问题回答、报告生成和医疗图像分类等任务,但目前仅有英文版本,直接翻译该基准一方面需要对中英文医学和文化有深入了解,另一方面也难以覆盖中医医学知识。因此,我们需要建立标准化的中文医疗评估基准,以量化医疗模型的改进程度。

2.1 中文医疗评测基准

随着中文医疗大模型的发展,已经有一些中文医疗评测基准出现,表 3 列举了目前已有的中文医疗大模型评测基准,并说明了其评估任务和优缺点。模型评估过程如图 2 所示。

表 3 中文医疗大模型评测基准分析

评估基准	评估任务	优点	缺点
CBLUE ^[65]	医学检索和术语标准化、医学文本分类、医学句子语义关系判定、医学文本理解和生成、医学 OCR 要素识别	任务类型覆盖广,数据真实且有噪音,对模型的鲁棒性要求较高	缺少多模态评估任务,除了 OCR 要素识别外均为文本相关评估
PromptCBLUE ^[66]	医学文本信息抽取、医学文本分类、医学自然语言推理、医疗对话的症状状态理解、医学文本生成	在 CBLUE 的基础上,将原有任务全部转化为基于提示的语言生成任务,任务形式统一	缺少多模态评估任务
CMB ^[67]	不同临床职业和不同职业阶段考试中的多项选择题、基于真实病例的复杂临床诊断	既有专业知识考核,也存在真实病例诊断	任务类型单一,缺少多模态评估任务
ShenNong-TCM-Evaluation-BenchMark ^[38]	基于中医执业医师资格考试题库的多类选择题	涵盖了中医药领域的多个方面,覆盖性广	均为选择题,任务类型单一;数据来源单一,缺少中医药领域相关的真实病例

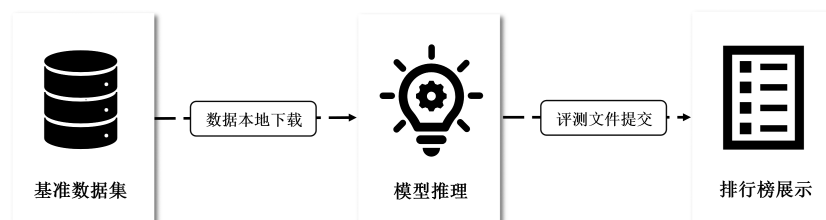


图 2 模型评测流程图

2.1.1 CBLUE

中文医疗信息处理评测基准 CBLUE(Chinese

Biomedical Language Understanding Evaluation)包括了医学文本信息抽取“实体识别、关系抽取”、医学

术语归一化、医学文本分类、医学句子语义相关性判定 4 大类常见的医疗信息处理任务,共包含 8 个子任务,CBLUE 还建立了一个用于模型评估、比较和分析的相关在线平台^[65]。CBLUE 1.0 版本的语料来源于医学教科书、医学期刊、临床试验、互联网搜索和医学论坛。CBLUE 的数据集及任务如表 4 所示。

表 4 CBLUE 的数据集及任务

数据集名称	数据集缩写	任务类型	评价指标
中文医学命名实体识别	CMeEE	命名实体识别	Micro F_1
中文医学文本实体关系抽取	CMeIE	关系抽取	Micro F_1
临床术语标准化	CHIP-CDN	术语归一化	Micro F_1
疾病问答迁移学习	CHIP-STs	语义相似度判断	Macro F_1
临床试验筛选标准短文本分类	CHIP-CTC	文本分类	Macro F_1
医疗搜索检索词意图分类	KUAKE-QIC	意图分类	Accuracy
医疗搜索查询词-页面标题相关性	KUAKE-QTR	语义相似度判断	Accuracy
医疗搜索查询词-查询词相关性	KUAKE-QQR	语义相似度判断	Accuracy

CBLUE 2.0 版本新增了电子病历(医学专家编制而成)和互联网医患对话语料集。CBLUE 2.0 在 CBLUE 1.0 的基础上新增了医学对话理解和生成任务,由 1.0 版的 4 大类扩充为 5 大类,子任务也由 8 个增加到 15 个。CBLUE 3.0 由医学文本信息抽取、医学检索和术语标准化、医学文本分类、医学句子语义关系判定、医学文本理解和生成共 5 大类任务 18 个子任务组成。此外,还引入了医学 OCR 光学要素识别任务作为多模态方向的初步尝试。

2.1.2 PromptCBLUE

PromptCBLUE 评估基准是对 CBLUE 基准进

行二次开发,将 16 种不同的医疗场景 NLP 任务全部转化为基于提示的语言生成任务,形成首个中文医疗场景的 LLMs 评测基准^[66]。PromptCBLUE 中的任务主要有医学文本信息抽取、医学文本分类、医学自然语言推理、医疗对话的症状状态理解、医学文本生成 5 大类任务。

PromptCBLUE 旨在采用一个生成式大模型完成医疗 NLP 相关的各种不同任务,如病历结构化、问诊、病例文书撰写等。PromptCBLUE 提供了更全面、细粒度的医疗场景评测内容,有助于评估大语言模型在医疗领域各个方面的性能。PromptCBLUE 的数据集及任务如表 5 所示。

表 5 PromptCBLUE 的数据集及任务

数据集	任务类型	评价指标
CMeEE-V2	医学关系抽取	Micro F_1 , Precision, Recall
CMeIE	医学关系抽取	Micro F_1 , Precision, Recall
CHIP-CDEE	医学关系抽取	Micro F_1 , Precision, Recall
CHIP-CDN	医学关系抽取	Micro F_1 , Precision, Recall
IMCS-V2-NER	医学关系抽取	Micro F_1 , Precision, Recall
CHIP-CTC	医学文本分类	Macro F_1 , Precision, Recall
KUAKE-QIC	医学文本分类	Macro F_1 , Precision, Recall
IMCS-V2-DAC	医学文本分类	Macro F_1 , Precision, Recall
CHIP-STs	医学自然语言推理	Micro F_1 , Precision, Recall
KUAKE-QQR	医学自然语言推理	Micro F_1 , Precision, Recall
KUAKE-IR	医学自然语言推理	Micro F_1 , Precision, Recall
KUAKE-QTR	医学自然语言推理	Micro F_1 , Precision, Recall
IMCS-V2-SR	医学对话理解	Micro F_1 , Precision, Recall
CHIP-MDCFNPC	医学对话理解	Micro F_1 , Precision, Recall
IMCS-V2-MRG	医学文本生成	Rouge-1, Rouge-2, Rouge-L
MedDG	医学文本生成	Rouge-1, Rouge-2, Rouge-L

2.1.3 CMB

中文综合医学基准(Comprehensive Medical Benchmark in Chinese, CMB)包括了不同临床职业、不同职业阶段考试中的多项选择题(CMB-Exam)和基于真实病例的复杂临床诊断问题(CMB-Clin),从医学考核和临床应用出发,结合选择题和复杂病历问诊来全面检验模型在医学知识与诊断能力上的表现^[67]。

CMB-Exam 的构建理念是反映真实医疗领域的考核体系,它分为医师、护理、医技、药师岗位考试,以及医学的大学学科知识考试和研究生入学考试六个类别。每个类别根据职业阶段进行子目录的划分,而子目录则针对不同科室提供了更细粒度的标签。

CMB-Clin 基于 74 个复杂真实案例,考察模型在真实诊断和治疗情境中的知识应用水平,评测该模型是否可以利用知识真正帮助解决临床复杂问题。其中,每个案例包含病历详情和 1~3 个问题,整个数据集共 208 个问题。模型需要理解病人信息,包括主诉、病史概要、体格检查等,并根据这些信息来解答与诊断、治疗相关的问题,并提供相应的解决方案。

2.1.4 ShenNong-TCM-Evaluation-BenchMark

ShenNong-TCM-Evaluation-BenchMark 建立了一个标准化、综合性的中医评测框架,该评测框架充分考虑中医药领域的复杂性和专业性,涵盖多个方面,以确保大语言模型在真实场景下的实用性和适用性。其使用的数据集为首个中医评测数据集 ShenNong-TCM-EB,该数据集基于中医执业医师资格考试题库,涵盖了中医基础理论、中医诊断学、中药学、方剂学、中医经典、中医内科学、中医外科学、中医妇科学、中医儿科学、针灸学、诊断学基础、内科学、传染病学、医学伦理学、卫生法规等 16 个知识点,共 3 个选择题类型,每类题型都包含 16 个知识点的至多 100 道题。

2.2 医疗健康行业大模型标准体系

随着医疗健康行业中大模型的快速发展,除了需要建立标准化的评估基准以度量模型的能力外,还应确保模型符合市场需求和相应的价值观念。

2023 年 9 月 25 日,医疗健康行业大模型应用技术要求《医院侧医疗服务》《医疗健康行业大模型应用技术要求》和《患者侧医疗服务》等技术要求正式发布,成为业界首个围绕医疗健康领域提出的大

模型应用技术标准规范。医疗健康行业大模型标准,从标准制定、评估测试、生态建设等方面着手,共同创建了一套符合医疗健康行业需求的评价体系,为医疗健康行业大模型行稳致远地建立坚实的评价规范与理论基础,支持大模型在医疗健康领域规模化规范应用,推动医疗健康行业大模型产业高质量发展。

3 挑战

随着一些中文医疗大模型的不断推出,面向医疗健康的中文医疗大模型的研究更加强调多样化、精准化、专业化、医疗健康数据的收集,以及数据安全和隐私保护。中文医疗大模型的发展也面临以下挑战。

3.1 模型应用挑战

3.1.1 幻觉

大模型的幻觉是指生成的输出包含不准确或非事实信息的现象^[68]。幻觉可以分为内在幻觉和外在幻觉。内在幻觉是指生成逻辑上与事实信息相矛盾的输出,例如,大模型生成错误的数学计算公式。当生成的输出无法验证时,就会发生外在幻觉,典型的例子为大模型伪造不存在的引文或回避问题。

用大模型进行诊疗尤其是中医诊疗时,常常存在不确定性和模糊性,医生能够通过丰富的经验和临床直觉来处理这些不确定的问题,做出最合适的选择。而大模型由于其情景理解能力、沟通交流能力和临床经验存在一定的欠缺,导致它们在处理模糊的医疗场景时,可能会出现回避问题或者输出不准确的现象,造成误诊。

3.1.2 数据安全

医疗大语言模型需要大量的医疗数据进行训练,如果对数据筛查和审核不到位,可能会涉及患者的个人隐私信息,如果隐私泄露,可能会引发安全和法律风险。此外,大模型数据安全和隐私隐患还体现在它对于对话内容的存储,当用户在和大模型进行互动时,他们的信息会以某些形式被记录和存储下来。

中国在数据安全和隐私保护方面有着严格的监管要求,中文医疗大模型需要符合中国相关法律法规的规定,如《个人信息保护法》等,确保用户数据的合规处理和安全存储。

3.1.3 价值对齐

价值对齐指的是让大模型的能力和行与人类的价值、真实意图和伦理原则相一致,确保人类与人工智能协作过程中的安全与信任^[69]。一个没有价值对齐的医疗大语言模型,可能输出含有违背道德伦理和法律法规的内容,从而对患者造成错误引导。

中文医疗大语言模型需要理解中文专业术语、医学概念以及文化特点,避免输出与文化、价值观不符的内容。

3.2 领域数据挑战

大型语言模型研究强调数据收集,数据质量和多样性尤为重要,而医学领域数据目前存在以下问题:

(1) 数据量大,数据类型复杂。在医院,每个患者的相关数据不仅有病历资料中包含的个人信息,还会有生化检查、多种影像或病理切片检查等生物学信息。数据类型也十分复杂,除了文本数据之外,还包括仪器检查或化验产生的数字、图片、影像、视频以及声、光、电信号数据等,这会给数据的存储、分析、处理、传输等带来很大挑战。

(2) 数据缺乏标准化、交换性。每家医院的信息化工作各不相同,对数据采集、存储等标准都自成体系。各个系统之间难以做到无缝衔接,信息共享遭遇阻碍。数据缺乏统一的标准和共享机制,使得我国的海量医疗数据有相当大的一部分无法转化为可用的数据。

(3) 临床数据整合和分析能力欠缺。在临床工作中有大量的电子病历、随访记录、临床试验等数据,想要利用计算机快速挖掘分析出这些数据潜在的价值,就要求分析人员同时具备医学知识和计算机专业知识,但熟悉医学和大数据的复合型人才十分欠缺。

此外,相较于英文而言,中文医疗数据的规模较小,且质量参差不齐。这对训练中文医疗大模型提出了更高的要求,需要通过更精细的数据清洗和标注来提升数据质量。

3.3 评测基准挑战

现有的自动评估方法缺乏参考答案,无法满足医疗领域对准确性和专业性的要求,缺乏说服力。同时,人工评估成本高昂,且难以扩大应用规模,限制了评估的广度和深度。此外,现有的中文医疗领域评估基准多局限于医疗问诊,缺乏影像相关的多

模态模型的评估基准。因此,在中文医疗领域建立更加完善的评估基准、引入更多医学专家参与评估过程是非常必要的。

3.4 应用场景挑战

目前,中文医疗大语言模型主要应用于医疗文本和图像等信息的处理和分析。然而,医疗领域的需求千变万化,如血氧浓度监测、药物研发、辅助手术等,涵盖了从基础医疗保健到前沿科研的各个方面。若这些模型不能在不同场景下提供高效解决方案,可能会影响其未来在市场上的应用。以血氧浓度监测为例,中文医疗大语言模型可以结合传感器技术,实时分析患者的生理数据,并提供个性化的健康评估和预警。在药物研发领域,这些模型可以通过挖掘海量医学文献和临床数据,加速新药物的发现和研发过程。而在辅助手术方面,它们可以通过图像识别和语义理解技术,为医生提供实时的指导和辅助,提高手术的安全性和效率。因此,中文医疗大语言模型需要不断优化和拓展,以满足医疗领域不同场景的需求,并更好地服务于医疗健康产业的发展。

4 总结与展望

4.1 总结

本文系统地介绍了大模型在中文医疗领域的发展和应用,并从多个角度进行了探讨和分析。首先,本文对医疗领域中现有的中文大语言模型进行了梳理,包括医疗问诊大模型、医学影像大模型和心理健康大模型等方面的研究进展和特点。此外,文中还针对中文医疗大模型的评估基准和医疗健康行业大模型标准体系进行了深入探讨,强调了对模型性能和应用场景的全面考量和评估的重要性。最后,本文还对中文医疗大模型面临的挑战进行了详细分析,包括模型应用中存在的幻觉、数据安全、价值对齐问题,以及在领域数据整合、评测基准和应用场景方面的挑战。

4.2 展望

随着技术的不断进步,中文医疗大模型的研究将更加注重算法和模型结构的优化,以提高模型的准确性和可靠性。例如,通过引入更多的领域专家参与模型训练,加强对医学领域知识的理解和应用,

以减少幻觉的发生。这些模型还会更加注重医疗大模型的价值对齐,通过引入伦理专家和法律专家参与模型设计和应用过程,以确保模型输出符合伦理和法律要求,减少对患者的错误引导。

未来工作中,有望出现更多针对医学领域的多模态模型,能够同时处理文本、图像、视频等多种医疗信息,并添加相应的时序信息,以满足医疗领域不同场景的需求。

参考文献

- [1] SUN K, LUO X, LUO M Y. A survey of pretrained language models[C]//Proceedings of the International Conference on Knowledge Science, Engineering and Management. Cham: Springer International Publishing, 2022: 442-456.
- [2] GE M, MAO R, CAMBRIA E. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application[J]. Artificial Intelligence Review, 2023: 1-67.
- [3] LI J. Recent advances in end-to-end automatic speech recognition[J].arXiv.preprint.arXiv.2111.01690,2021.
- [4] 冯洋,邵晨泽.神经机器翻译前沿综述[J].中文信息学报,2020,34(7):1-18.
- [5] RANATHUNGA S, LEE E S A, PRIFTI SKENDU-LI M, et al. Neural machine translation for low-resource languages: A survey[J]. ACM Computing Surveys, 2023, 55(11): 1-37.
- [6] BAO H, HE K, YIN X, et al. Bert-based meta-learning approach with looking back for sentiment analysis of literary book reviews[C]//Proceedings of the Natural Language Processing and Chinese Computing: 10th CCF International Conference, Qingdao, China, 2021: 235-247.
- [7] MAO R, LI X. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13534-13542.
- [8] HE K, HUANG Y, MAO R, et al. Virtual prompt pre-training for prototype-based few-shot relation extraction[J]. Expert Systems with Applications, 2023, 213: 118927.
- [9] 庄传志,靳小龙,朱伟建,等.基于深度学习的关系抽取研究综述[J].中文信息学报,2019,33(12):1-18.
- [10] KENTON J D M W C, TOUTANOVA L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of naacL-HLT, 2019.
- [11] PERGOLA G, KOCHKINA E, GUI L, et al. Boosting low-resource biomedical QA via entity-aware masking strategies[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 1977-1985.
- [12] LEE J, YOON W, KIM S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
- [13] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. arXiv preprint arXiv: 2303.18223, 2023.
- [14] WANG Y, YAO Q, KWOK J T, et al. Generalizing from a few examples: A survey on few-shot learning[J]. ACM Computing Surveys, 2020, 53(3): 1-34.
- [15] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.
- [16] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language model[J]. arXiv. preprint.arXiv: 2302.13971,2023.
- [17] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.
- [18] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv: 2303.08774, 2023.
- [19] DU Z, QIAN Y, LIU X, et al. GLM: General language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 320-335.
- [20] ZENG A, LIU X, DU Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv: 2210.02414, 2022.
- [21] SINGHAL K, TU T, GOTTWEIS J, et al. Towards expert-level medical question answering with large language models[J]. arXiv preprint arXiv: 2305.09617, 2023.
- [22] JIN D, PAN E, OUFATTOLE N, et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams[J]. Applied Sciences, 2021, 11(14): 6421.
- [23] WANG H, LIU C, XI N, et al. Huatuo: Tuning llama model with Chinese medical knowledge[J]. arXiv preprint arXiv:2304.06975, 2023.
- [24] ZHANG H, CHEN J, JIANG F, et al. HuatuoGPT, towards taming language model to be a doctor[C]//Proceedings of the Association for Computational Linguistics; EMNLP, 2023: 10859-10885.

- [25] SHU C, CHEN B, LIU F, et al. Visual med-alpaca: A parameter-efficient biomedicalllm with visual capabilities[EB/OL]. <https://github.com/cambridgeltt/visual-medalpaca>.
- [26] ZHOU H, GU B, ZOU X, et al. A survey of large language models in medicine: Progress, application, and challenge[J]. arXiv preprint arXiv:2311.05112, 2023.
- [27] HE K, MAO R, LIN Q, et al. A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics[J]. arXiv preprint arXiv:2310.05694, 2023.
- [28] XIONG H, WANG S, ZHU Y, et al. DoctorGLM: Fine-tuning your Chinese doctor is not a herculean task[J]. arXiv preprint arXiv:2304.01097, 2023.
- [29] CHEN Y, WANG Z, XING X, et al. BianQue: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt[J]. arXiv preprint arXiv:2310.15896, 2023.
- [30] WANG H, LIU C, ZHAO S, et al. ChatGLM-Med: 基于中文医学知识的 ChatGLM 模型微调[EB/OL]. <https://github.com/SCIR-HI/Med-ChatGLM>. [2024-06-14].
- [31] XU M. MedicalGPT: Training medical GPT model[EB/OL]. <https://github.com/shibing624/MedicalGPT>. [2024-06-14].
- [32] WANG G, YANG G, DU Z, et al. ClinicalGPT: Large language models finetuned with diverse medical data and comprehensive evaluation[J]. arXiv preprint arXiv:2306.09968, 2023.
- [33] ZHU W, WANG X L. ChatMed: A Chinese medical large language model[EB/OL]. <https://github.com/michael-wzhu/ChatMed>. [2024-06-14].
- [34] LIAO Y S, MENG Y T, LIU H C, et al. MING-MOE: Enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts[J]. arxiv preprint arxiv:2404.09027, 2024.
- [35] WANG R, DUAN Y, LAM C T, et al. IvyGPT: Interactive Chinese path way language model in medical domain[C]//Proceedings of the CAAI International Conference on Artificial Intelligence. Singapore: Springer Nature Singapore, 2023: 378-382.
- [36] YANG S H, ZHAO H J, ZHU S B, et al. Zhongjing: Enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 19368-19376.
- [37] BAO Z, CHEN W, XIAO S, et al. DISC-MedLLM: Bridging general large language models and real-world medical consultation[J]. arXiv preprint arXiv:2308.14346, 2023.
- [38] ZHU W, YUE W J, WANG X L. ShenNong-TCM: A Traditional Chinese medicine large language model[EB/OL]. <https://github.com/michael-wzhu/ShenNong-TCM-LLM>. [2024-06-14].
- [39] WANG R S, DUAN Y F, LI J, et al. XrayGLM: The first chinese medical multimodal model that chest radiographs summarization[EB/OL]. <https://github.com/WangRongsheng/XrayGLM>. [2024-06-14].
- [40] LIU J, WANG Z, YE Q, et al. Qilin-Med-VL: Towards Chinese large vision-language model for general healthcare[J]. arXiv preprint arXiv: 2310.17956, 2023.
- [41] CHEN Y, XING X, LIN J, et al. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations[C]//Proceedings of the Association for Computational Linguistics, 2023: 1170-1183.
- [42] YAN X, XUE D. MindChat: Psychological large language model[EB/OL]. <https://github.com/X-D-Lab/MindChat>. [2024-06-14].
- [43] ALTAN D, SARIEL S. Clue-AI: A convolutional three-stream anomaly identification framework for robot manipulation[J]. IEEE Access, 2023.
- [44] MUENNIGHOFF N, WANG T, SUTAWIKA L, et al. Crosslingual generalization through multitask fine-tuning[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023.
- [45] LE S T, FAN A, AKIKI C, et al. BLOOM: A 176 b-parameter open-access multilingual language model[J]. arXiv preprint arXiv: 2211.05100, 2022.
- [46] YANG A, XIAO B, WANG B, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.
- [47] YANG P, WANG J, GAN R, et al. Zero-shot learners for natural language understanding via a unified multiple choice perspective[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022: 7042-7055.
- [48] HU E J, WALLIS P, ALLEN ZHU Z, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.
- [49] LIU X, JI K, FU Y, et al. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 61-68.
- [50] ZHANG L, ZHAO K. ChatYuan: A large language model for dialogue in Chinese and English[EB/OL].

- <https://github.com/clue-ai/ChatYuan>. [2024-06-14].
- [51] BAI Y, JONES A, NDOUSSE K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback [J]. arXiv preprint arXiv:2204.05862, 2022.
- [52] WANG Y, KORDI Y, MISHRA S, et al. Self-instruct: Aligning language models with self-generated instructions [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023.
- [53] ZHU D Y, CHEN J, SHEN X, et al. MiniGPT-4: Enhancing vision-language understanding with advanced large language models [C]//Proceedings of the 12th International Conference on Learning Representations, 2023.
- [54] YE Q, XU H, XU G, et al. mPLUG-Owl: Modularization empowers large language models with multimodality [J]. arXiv preprint arXiv:2304.14178, 2023.
- [55] GONG T, LYU C, ZHANG S, et al. Multimodal-GPT: A vision and language model for dialogue with humans [J]. arXiv preprint arXiv:2305.04790, 2023.
- [56] LIU H, LI C, WU Q, et al. Visual instruction tuning [C]//Proceedings of the 37th Conference on Neural Information Processing Systems, 2023.
- [57] YIN S K, FU C Y, ZHAO S R, et al. A survey on multimodal large language models [J]. arXiv preprint arXiv:2306.13549, 2023.
- [58] DU Z, QIAN Y, LIU X, et al. GLM: General language model pretraining with autoregressive blank infilling [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 320-335.
- [59] SABOUR S, ZHANG W, XIAO X, et al. A chatbot for mental health support: Exploring the impact of Emohaa on reducing mental distress in China [J]. Frontiers in Digital Health, 2023, 5: 1133987.
- [60] TEAM InternLM. InternLM: A multilingual language model with progressively enhanced capabilities [EB/OL]. <https://github.com/InternLM/InternLM>, 2023. [2024-06-14].
- [61] BAI J, BAI S, YANG S, et al. Qwen-VL: A frontier large vision-language model with versatile abilities [J]. arXiv preprint arXiv:2308.12966, 2023.
- [62] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive tasks [J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.
- [63] SUNG M, LEE J, YI S, et al. Can language models be biomedical knowledge bases? [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 4723-4734.
- [64] TU T, AZIZI S, DRIESS D, et al. Towards generalist biomedical AI [J]. NEJM AI, 2024, 1 (3): A10a2300138.
- [65] ZHANG N, CHEN M, BI Z, et al. CBLUE: A Chinese biomedical language understanding evaluation benchmark [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 7888-7915.
- [66] ZHU W, WANG X, ZHENG H, et al. Prompt-CBLUE: A Chinese prompt tuning benchmark for the medical domain [J]. arXiv preprint arXiv: 2310.14151, 2023.
- [67] WANG X, CHEN G H, SONG D, et al. CMB: A comprehensive medical benchmark in Chinese [J]. arXiv preprint arXiv:2308.08833, 2023.
- [68] RAWTE V, SHETH A, DAS A. A survey of hallucination in large foundation models [J]. arXiv preprint arXiv:2309.05922, 2023.
- [69] SHEN T, JIN R, HUANG Y, et al. Large language model alignment: A survey [J]. arXiv preprint arXiv: 2309.15025, 2023.



籍欣萌(2000—), 硕士研究生, 主要研究领域为自然语言处理。

E-mail: xmji1014@163.com



崔婷婷(2001—), 硕士研究生, 主要研究领域为自然语言处理。

E-mail: 653099805@qq.com



咎红英(1966—), 通信作者, 博士, 教授, 主要研究领域为自然语言处理。

E-mail: iehyzan@zzu.edu.cn