大语言模型在医疗领域的前沿研究与创新应用*

何剑虎[1]　王德健[2]　赵志锐[3]　汪　辉[1,4]　游　茂[5]

（[1] 浙江大学医学院附属妇产科医院　杭州 310006　[2] 浙江大学软件学院　杭州 310015
[3]人民卫生出版社有限公司　北京 100021　[4] 浙江省妇科重大疾病精准诊治研究重点实验室　杭州 310006
[5]国家卫生健康委卫生发展研究中心　北京 100032）

〔摘要〕　**目的/意义** 系统梳理大语言模型在医疗领域的应用与研究进展，分析关键挑战与机遇，为相关研究提供参考。**方法/过程** 采用系统性文献回顾方法，全面梳理近年来发表的相关文献，聚焦医疗大语言模型最新进展；分析大语言模型在医疗自然语言处理任务中的应用成效、研究现状以及面临的挑战。**结果/结论** 大语言模型在医疗领域应用前景广阔，未来研究重点应集中在技术进步与伦理规范完善等方面。一方面加速技术创新步伐，另一方面确保严格遵守伦理标准，共同促进医疗领域大语言模型技术可持续发展。

〔关键词〕　大语言模型；医疗；自然语言处理；人工智能

〔中图分类号〕R－058　　〔文献标识码〕A　　〔**DOI**〕10. 3969/j. issn. 1673－6036. 2024. 09. 002

The Frontier Research and Innovative Applications of Large Language Models in the Medical Field

HE Jianhu[1], WANG Dejian[2], ZHAO Zhirui[3], WANG Hui[1,4], YOU Mao[5]

[1] Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310006, China; [2] School of Software Technology, Zhejiang University, Hangzhou 310015, China; [3] People's Medical Publishing House, Beijing 100021, China; [4] Zhejiang Provincial Key Laboratory for Precision Diagnosis and Treatment of Major Gynecological Diseases, Hangzhou 310006, China; [5] China National Health Development Research Center, Beijing 100032, China

〔**Abstract**〕　**Purpose/Significance** The paper systematically reviews the application and research progress of large language models (LLMs) in the medical field, analyzes the key challenges and opportunities, and provides references for related research. **Method/Process** The systematic literature review method is adopted to comprehensively review the relevant literature published in recent years, focusing on the latest progress of medical LLMs. The application results, research status and challenges of LLMs in medical natural language processing (NLP) tasks are analyzed. **Result/Conclusion** The application of LLMs in the medical field has a broad prospect, and the future research should focus on the technical progress and the improvement of ethical norms. On the one hand, the pace of technological innovation should be accelerated, and on the other hand, strict compliance with ethical standards should be ensured to jointly promote the sustainable development of LLMs technology in the medical field.

〔**Keywords**〕　large language models (LLMs); medical treatment; natural language processing (NLP); artificial intelligence (AI)

〔修回日期〕　2024－09－12

〔作者简介〕　何剑虎，高级工程师，发表论文30 余篇；通信作者：游茂，汪辉。

〔基金项目〕　浙江省“尖兵”“领雁”研发攻关计划项目（项目编号：2023C03101）。

1 引言

近年来得益于深度学习技术进步以及基础算力提升，基于 Transformer[1] 架构的大规模语言模型（large language models，LLMs）在自然语言处理（natural language processing，NLP）领域取得革命性进展。如生成式预训练（generative pre – trained transformer，GPT）[2]、双向编码器表征（bidirectional encoder representations from transformers，BERT）[3]、T5[4] 等标志性模型，通过在海量文本数据基础上进行无监督预训练，具备了语义理解和逻辑推理能力，并在文本生成、翻译、问答等任务中表现卓越。

大语言模型的发展推动研究人员和从业者探索其在医疗领域的应用潜力[5-6]。医疗领域大量文献、电子健康记录、患者咨询文本数据为大语言模型应用提供了契机。将大语言模型应用于医疗领域能够为辅助诊断[7]、医学研究[8]、文档管理[9]、医学教育[10] 等带来创新和变革。同时医疗行业的法律伦理敏感性、知识密集性、隐私保护及准确率要求高等属性对大语言模型提出更高要求，其应用面临挑战。例如，模型产生“幻觉”可能导致错误的医疗建议[11]，知识更新不及时会限制其准确性[12]，人工智能在医疗决策中所承担的角色和责任无法绕开伦理问题等[13]。

本文系统梳理和分析大语言模型在医疗领域的最新研究进展与应用情况，以期为未来大语言模型在医疗领域的研究和应用提供参考。

2 大语言模型增强医疗 NLP 能力

2.1 信息提取

在医疗领域，大语言模型在信息提取方面发挥重要作用。信息提取过程涉及从海量医学文献、病历记录、临床指南及研究论文中自动识别并提取关键医疗信息，如疾病症状、治疗方法、药物剂量、患者预后等。有研究[14-18] 表明大语言模型能够从海量文本数据中提取结构化信息。Chen J 等[18] 运用 ChatGPT 生成标注数据，有效改善事件抽取任务的表现。针对临床实践具体需求，Wang L 等[19] 应用大语言模型从住院记录中提取语义信息。对模型信息提取能力的系统评估也成为研究热点，例如 Jethani N 等[20] 研究 ChatGPT 提炼临床记录信息的效能，强调其在促进医疗决策支持方面的作用。此外，大语言模型在揭示健康的社会决定因素（social determinants of health，SDOH）及人口统计学特征方面展现出独特价值。Ramachandran G K 等[21] 通过部署 GPT 4，在非结构化电子健康记录中自动捕获 SDOH 信息，Bhate N J 等[22] 探索如何以最少的引导信息最大化发挥 GPT 模型潜能，精准提取包括人口统计学信息、SDOH 以及家族病史在内的多维度数据。一系列研究表明大语言模型在医疗信息提取领域取得实质性进展，不仅提高了信息提取效率和准确性，还促进了医疗决策支持。基于大语言模型的信息提取在深化医疗领域数据洞察、促进健康管理个性化发展等方面有较好应用前景。

2.2 命名实体识别

医学命名实体识别是医学数据处理的关键环节，但尚需解决其依赖于大量标注数据的问题。随着大语言模型技术兴起，命名实体识别领域正在经历深刻变革。与传统方法不同，大语言模型凭借其强大的泛化能力，能够在近乎零样本或极少量示例的情况下，在通用领域实现有效命名实体识别，减弱其对数据标注的依赖。这一技术革新在医学领域展现出较强应用潜力。例如，Li M 等[23] 深入探讨大语言模型在面对少样本场景时执行医学命名实体识别任务的性能。Li M 等[24] 利用 ChatGPT 模型对宋代医学文献中记载的病例进行深度剖析与总结，了解古代中医临床实践经验，为现代中医临床应用和研究提供参考。此外针对特定场景的深入应用同样值得关注。例如，Iscoe M 等[25] 针对急诊室非结构化记录中的尿路感染症状识别，设计并实施两个任务导向大语言模型。大语言模型在医学命名实体识别领域应用取得显著进展，特别是减少了对大量标注数据的依赖。这一进步不仅简化了数据准备流程，还极大加速了医学知识发现与应用进程。

2.3 情感分析

医疗领域中，情感分析覆盖患者情绪精细管理、公众健康情绪宏观把握、医患互动质量深入理解等层面。大语言模型技术的引入显著提升此类分析能力，特别是在处理非结构化数据、实现低样本乃至零样本学习方面。Lossio – Ventura J A 等[26]验证大语言模型在健康调查数据情感分析中的高效性，特别是在少量样本或无样本学习场景下，表明大语言模型能够准确捕捉并分类公众对于健康议题的情感反应，为政策制定和公众健康干预提供科学依据。De S 等[27]开发多标签分类器，可识别社交媒体中关于疫苗接种的担忧情绪。在临床实践微观层面，Clarke P 等[28]和 Gómez – Zaragozá L 等[29]分别探索大语言模型深度应用。前者构建一个自动化情感分析模型，以评估心理治疗过程中医患对话的情感动态，加深对治疗过程中心理变化的理解；后者则聚焦于虚拟人交互场景，通过分析与抑郁症状相关的语言模式，探索技术辅助心理健康诊断的新途径。Mittal S 等[30]采取基于大语言模型的表示框架，对海量社交媒体和新闻文章中的心理健康讨论进行全面的情感和道德基础分析，揭示不同媒介心理健康话语如何影响社会对心理健康的感知和耻辱感，为减少心理健康污名化和促进公众教育提供新视角。大语言模型在医疗情感分析领域的应用获得显著成效，不仅能够监测宏观公众情绪反应，还能深入分析具体诊疗沟通情况，以及促进对心理健康文化的传播与深入理解。

3 医疗大语言模型应用

3.1 辅助诊断

在医学辅助诊断领域，大语言模型在整合多模态信息和提高诊断精度方面具有显著优势。大语言模型作为诊断辅助手段时主要处理文本信息，这些信息源自患者自述或医患交流记录，因而可能存在一定主观性偏差。为应对这一问题，Wang S 等[31]提出一项创新性解决方案，将大语言模型融入医学影像的计算机辅助诊断（computer aided diagnosis, CAD）系统中，实现技术跨界协同。通过大语言模型对自然语言文本的高效汇总与重构，有效增强 CAD 系统输出的全面性和深度。该方法联结大语言模型在医学知识、逻辑推理优势与 CAD 系统的图像解析专长，共同构建多维度、高精准度的辅助诊断框架。在心理学诊断领域，鉴于心理状态评估高度依赖患者主观报告和情感表达，大语言模型应用显得尤为关键。例如，Danner M 等[7]探索基于 BERT、GPT 3.5 和 ChatGPT 4 等大语言模型应用，从丰富的语言模式和文本信息中识别抑郁症的特征。Mazumdar H 等[32]提出一种基于 GPT 的心理检测方法 GPTFX，在精神健康障碍分类和生成解释方面表现出卓越性能。

3.2 问答系统

大语言模型凭借其语义理解能力和广泛的知识，在医疗问答系统中发挥重要作用。其能够解析患者症状描述、回答医疗问题、提供初步诊断建议并提供心理健康支持。例如，Jo E 等[33]展示一款基于大语言模型的聊天机器人，其不仅能够有效缓解用户孤独感，还能智能评估对话者精神健康状态。Akilesh S 等[34]应用 GPT 3.5 技术构建医疗聊天机器人，重新定义个人与医疗保健系统的交互方式，允许用户便捷地输入症状并即时获取精确反馈、个性化诊断及治疗推荐。Chen S 等[35]对大语言模型聊天机器人遵循美国国家综合癌症网络（national comprehensive cancer network, NCCN）指南提供癌症治疗建议的能力进行评估，强调其在复杂疾病管理方面的潜在应用前景。同时，Chen S 等[36]探讨大语言模型在协助临床医生回复患者问题方面的实用性，指出其可有效提升工作效率。Laker B 等[37]研究基于大语言模型的智能助手在回应患者咨询质量与同理心方面的表现。结果显示，智能助手不仅能有效减轻医护人员回复患者电子消息的负担，而且在维持回复质量与同情理解方面表现良好。大语言模型的应用不仅增强了医疗系统响应能力，还促进了医患沟通。随着技术的发展，大语言模型将在医疗问答系统中发挥更加重要的作用。

3.3 文档管理

3.3.1 报告生成 在医疗实践中，报告撰写工作繁重且易出错。大语言模型通过学习海量医疗文献与病例记录，能够精准提炼信息，产出格式统一、内容精练的报告，有效减轻临床工作负担。大量研究将大语言模型应用于临床报告生成的不同方向，如出院摘要[38]、患者指导[39]、临床信函[40]。有研究探索大语言模型应用于健康相关事项咨询的可行性，如回答术后问题[41]、生成健康信息[42]、提供美容手术建议[43]等。有研究[9]探索8种大语言模型在放射学报告、患者咨询、病情进展记录、医患交流4个核心临床文档总结任务中的应用，证明大语言模型应用于临床流程中的潜力，指出其可缓解医生文档工作压力，帮助其聚焦于直接的患者护理。尽管如此，大语言模型在医疗报告生成中的应用尚存在缺陷。有研究[42]指出，尽管某些模型在逻辑和语义表达上能接近人类专家水平，但在信息准确性和完整性方面尚存在差距。另有研究揭示应用大语言模型生成的医疗报告中可能出现虚构信息及表述不够精练等问题[9]。因此，未来发展方向应当集中在提升报告准确度、确保信息真实性以及增强内容紧凑性，以全面提高大语言模型在医疗报告生成领域的实用性和可靠性。

3.3.2 ICD编码整理 国际疾病分类（international classification of diseases，ICD）[44]作为一种标准化工具，用于编码临床诊断和医疗程序信息，确保全球医疗记录的一致性和可比性。在每次诊疗互动后，相应的ICD代码被录入患者电子健康记录系统中[45]。然而，人工执行ICD编码过程极为耗时，促使研究者[46-48]探索利用大语言模型实现自动化。Huang C W等[46]介绍一种创新框架，运用预训练大语言模型自动完成ICD编码任务。Ong J等[47]则聚焦于视网膜病变领域，通过ChatGPT展示一个无需额外反馈即可自动生成ICD编码序列的实例，指出该方法在减轻医生编码负担方面具有潜力。另一方面，Dong H等[48]肯定了自动化临床编码概念，深入分析该领域面临的挑战。在实际应用中，基于深度学习的临床编码方法在可解释性、一致性方面尚待提升，应列为未来研究重点。

3.4 医学教育

大语言模型在医学教育领域的应用促进教育内容个性化与实用性。有研究[49]指出，大语言模型能模拟医疗案例，促进医学生临床决策能力的培养。大语言模型还能够针对学生对实际案例的反应自动生成反馈，实时指出学生的认知盲点，促进其自我反思和改进[50]。大语言模型可以智能调节其输出文本中的医学术语密度，以适配不同背景读者。面对非专业人士，将复杂的医疗概念转化为通俗易懂的语言，确保信息的可接近性和理解度；面对医务工作者，精准运用专业术语，提供更详细的信息[49]。大语言模型在医学教育应用中应注意道德、算法偏见等问题[10,51-52]。有研究[53-54]关注大语言模型在医疗考试中的表现，例如Kung T H等[53]评估ChatGPT在美国执业医师资格考试中的表现，展示其辅助人类学习及作为临床决策辅助工具的潜力。综上所述，大语言模型在提升医学教育互动性和效率方面展现较强潜力，并预示着医学教育与临床实践深度融合的新趋势。然而在这一进程中也伴随着特定的挑战与机遇。

4 医疗大语言模型应用挑战

4.1 “幻觉”问题

“幻觉”是指大型语言模型在处理特定复杂任务时可能产生的看似合理实则错误的输出[55-56]。医疗领域作为知识密集型行业，信息量庞大、专业术语繁多且情境复杂多变。在尝试模拟临床决策、患者诊断或治疗建议时，大语言模型可能因无法完全把握医学知识的深度、广度及最新进展，而产生看似符合逻辑却偏离事实的“幻觉性”结论，这种现象尤其体现在对罕见病、特殊病例的判断上。为应对该挑战，Shuster K等[57]提出一种连接外部数据的方法，为模型提供现实世界数据参考，使其在生成内容时能够依赖于真实数据，显著减少“幻觉”，确保输出准确性和真实性，有助于提高临床决策可靠性。Roit P等[58]设计一种基于强化学习与

文本蕴涵反馈的方法，通过强化学习机制，从反馈中不断优化生成的摘要质量，减少不真实信息的产生，进一步提高模型在医疗环境的适用性。Chern I C 等[59]则引入一个结合对比学习和奖励学习的框架，使用事实性指标解决大语言模型摘要生成中的“幻觉”问题，这对于确保医疗信息准确传递至关重要。此外，Dhuliawala S 等[60]开发验证链方法为不同任务中的“幻觉”问题提供有效解决方案，不仅对基于维基数据的问题有效，而且在封闭式书籍 MultiSpanQA 和长文本生成任务中也能显著减少“幻觉”，对于提升医疗诊断安全性和准确性具有重要意义。

上述方法虽然在一定程度上缓解了“幻觉”问题，但相关研究主要聚焦于通用领域。未来应继续探索如何降低“幻觉”发生率，并开发适用于更多医疗场景的技术。此外，随着医学知识不断更新和发展，持续性的模型训练和更新也将成为确保其可靠性的关键。

4.2 缺乏基准数据集和指标

大语言模型在医疗领域展现出巨大潜力，但尚缺乏统一且全面的评估基准数据集和指标体系，如可信度、有效性和可解释性等对医学领域至关重要的指标[61]。对此有研究提出解决方案。Jin D 等[62]提出用于解决医疗问题的自由形式多项选择数据集 MedQA，从专业医学委员会考试中收集，涵盖英语、简体中文和繁体中文 3 种语言。Pal A 等[63]提出大规模多项选择题回答数据集，收集超过 19.4 万条高质量全印医学科学研究所（all India institute of medical sciences，AIIMS）和全印研究生医学入学考试（national eligibility cum entrance test for post-graduate，NEET PG）多选题，涵盖 2 400 个医疗保健主题和 21 个医学科目。为了进一步标准化评估方法，Singhal K 等[5]提出基准数据集 MultiMedQA，结合涵盖专业医学、研究和消费者查询的 6 个医学问答数据集，以及一个新的在线搜索医学问题数据集 HealthSearchQA；同时还提出一个评估框架，用于从事实性、理解性、推理能力、可能损害和偏见等维度评估模型。尽管通用大语言模型领域已有评价指标[64-65]，但尚不能直接用于医学领域。因此迫切需要更多针对医学领域的特定基准数据集和评估方法，以促进大语言模型在医疗领域的评估标准化，为模型的可信度和可靠性提供保障。

4.3 知识更新慢

医疗大语言模型的精确度与实用性高度依赖其对海量医疗数据的学习。医疗知识更新速度快，新发现层出不穷，医疗大语言模型要能够及时吸纳最新医学知识，反映最新医疗知识状态。然而大语言模型训练需要庞大的计算资源和时间成本。针对医疗大语言模型知识更新慢的问题，检索增强生成（retrieval - augmented generation，RAG）是一种有潜力的解决方案[66]。RAG 结合传统机器学习模型的生成能力和现代信息检索技术，通过检索相关新数据或知识辅助模型生成更准确、时效性更强的输出。

4.4 伦理与法律问题

医疗领域大语言模型的应用，因涉及个人健康数据处理与敏感医疗决策辅助，存在伦理与法律问题。一方面，确保患者隐私、消除算法偏见、实现医疗决策的可解释性是维护个体权益与社会公正的需求[13,67]；另一方面，随着医疗大语言模型在诊断、治疗中的角色日益重要，如何在法律框架内界定人机责任，同时激励创新与技术进步，成为亟待解决的法律挑战[68-69]。Valiña L G 等[67]主张在技术应用初期即全面考量其长远影响，辅助作出更加周全的决策。Stokel - Walker C[70]认为出于道德方面的考虑，不应在撰写生物医学研究论文时使用大语言模型。医疗大语言模型的伦理与法律问题是多维度、跨学科的挑战，需要政策制定者、医学专家、数据科学家、法律学者以及社会各界的共同努力，通过建立全面的监管体系、强化伦理指导原则、提升技术透明度和可解释性，确保技术进步在惠及人类健康的同时，妥善解决伴随而来的复杂问题。

5 结语

本文综述了大语言模型在医疗领域的应用与研究

现状，分析其在医疗自然语言处理任务中的最新研究进展，以及在辅助医疗诊断、教育创新和文档智能化管理等方面的应用成果。这些应用显著提高了医疗工作效率，提升医疗服务的质量和可及性。然而，大语言模型在医疗领域的应用面临一系列挑战，包括模型“幻觉”、评估体系不完善、知识更新滞后以及要遵守伦理、法律等问题，需要通过多学科合作来解决，以确保技术安全性、可靠性和合规性。

未来研究应当着重于技术创新与优化，持续改进模型结构和训练方法，减少“幻觉”现象，提高模型准确性和可靠性；同时，开发一套全面评估体系并逐渐完善，用于衡量模型在不同医疗场景下的表现，确保模型能够满足实际需求。此外，还需要制定明确的伦理指导原则和法律框架，保护患者隐私权和其他合法权益；构建动态知识更新机制，使模型能够及时反映最新医学知识和技术进展。通过上述努力推动医疗大语言模型技术向更加成熟和负责任的方向发展，进而为患者提供更高质量、更个性化的医疗服务，并最终实现医疗服务的整体优化。

作者贡献：何剑虎负责文献调研、论文撰写；王德健负责文献收集与整理、参与论文撰写；赵志锐参与论文撰写、论文修订；汪辉负责提供指导、论文修订；游茂负责研究设计、论文修订。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]. Long Beach: Advances in Neural Information Processing Systems 30, 2017.

2 RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre – training [EB/OL]. [2024 – 05 – 05]. https://openai.com/index/language – unsupervised.

3 DEVLIN J, CHANG M W, LEE K, et al. Bert: pre – training of deep bidirectional transformers for language understanding [EB/OL]. [2024 – 05 – 05]. https://arxiv.org/pdf/1810.04805.

4 RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text – to – text transformer [J]. Journal of machine learning research, 2020, 21 (140): 1 – 67.

5 SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620 (7972): 172 – 180.

6 SINGHAL K, TU T, GOTTWEIS J, et al. Towards expert – level medical question answering with large language models [EB/OL]. [2024 – 05 – 06]. https://arxiv.org/pdf/2305.09617.

7 DANNER M, HADZIC B, GERHARDT S, et al. Advancing mental health diagnostics: GPT – based method for depression detection [C]. Tsu: 2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE), IEEE, 2023.

8 LIU H, AZAM M, BIN NAEEM S, et al. An overview of the capabilities of ChatGPT for medical writing and its implications for academic integrity [J]. Health information & libraries journal, 2023, 40 (4): 440 – 446.

9 VAN VEEN D, VAN UDEN C, BLANKEMEIER L, et al. Clinical text summarization: adapting large language models can outperform human experts [J]. Research square, 2023, 30 (4): 1134 – 1142.

10 ABD – ALRAZAQ A, ALSAAD R, ALHUWAIL D, et al. Large language models in medical education: opportunities, challenges, and future directions [J]. JMIR medical education, 2023, 9 (1): e48291.

11 UMAPATHI L K, PAL A, SANKARASUBBU M. Med – halt: medical domain hallucination test for large language models [EB/OL]. [2024 – 05 – 07]. https://arxiv.org/pdf/2307.15343.

12 LIU F, ZHU T, WU X, et al. A medical multimodal large language model for future pandemics [J]. NPJ digital medicine, 2023, 6 (1): 226.

13 PERNI S, LEHMANN L S, BITTERMAN D S. Patients should be informed when AI systems are used in clinical trials [J]. Nature medicine, 2023, 29 (8): 1890 – 1891.

14 AGRAWAL M, HEGSELMANN S, LANG H, et al. Large language models are few – shot clinical information extractors [EB/OL]. [2024 – 05 – 10]. https://arxiv.org/pdf/2205.12689.

15 GERO Z, SINGH C, CHENG H, et al. Self – verification improves few – shot clinical information extraction [EB/OL]. [2024 – 05 – 10]. https://arxiv.org/pdf/2306.00024.

16 GOEL A, GUETA A, GILON O, et al. Llms accelerate annotation for medical information extraction [C]. New Orleans: Machine Learning for Health (ML4H), PMLR, 2023.

17 HU D, LIU B, ZHU X, et al. Zero – shot information extraction from radiological reports using ChatGPT [EB/OL]. [2024 – 05 – 09]. https://doi.org/10.1016/j.ijmedinf.2023.105321.

18 CHEN J, CHEN P, WU X. Generating Chinese event extraction method based on ChatGPT and prompt learning [J]. Applied sciences, 2023, 13 (17): 9500.

19 WANG L, MA Y, BI W, et al. An entity extraction pipeline for medical text records using large language models: analytical study [EB/OL]. [2024 – 05 – 10]. https://www.jmir.org/2024/1/e54580.

20 JETHANI N, JONES S, GENES N, et al. Evaluating ChatGPT in information extraction: a case study of extracting cognitive exam dates and scores [EB/OL]. [2024 – 05 – 10]. http://dx.doi.org/10.2139/ssrn.4520434.

21 RAMACHANDRAN G K, FU Y, HAN B, et al. Prompt – based extraction of social determinants of health using few – shot learning [EB/OL]. [2024 – 05 – 11]. https://arxiv.org/pdf/2306.07170.

22 BHATE N J, MITTAL A, HE Z, et al. Zero – shot learning with minimum instruction to extract social determinants and family history from clinical notes using GPT Model [C]. Sorrento: 2023 IEEE International Conference on Big Data (BigData), 2023.

23 LI M, ZHANG R. How far is language model from 100% few – shot named entity recognition in medical domain [EB/OL]. [2024 – 05 – 11]. https://arxiv.org/pdf/2307.00186.

24 LI M, ZHENG X. Identification of ancient Chinese medical prescriptions and case data analysis under artificial intelligence GPT algorithm: a case study of song dynasty medical literature [EB/OL]. [2024 – 05 – 11]. 10.1109/ACCESS.2023.3330212.

25 ISCOE M, SOCRATES V, GILSON A, et al. Identifying signs and symptoms of urinary tract infection from emergency department clinical notes using large language models [J]. Academic emergency medicine, 2023, 31 (6): 599 – 610.

26 LOSSIO – VENTURA J A, WEGER R, LEE A Y, et al. A comparison of ChatGPT and fine – tuned open pre – trained transformers (opt) against widely used sentiment analysis tools: sentiment analysis of COVID – 19 survey data [EB/OL]. [2024 – 05 – 12]. https://mental.jmir.org/2024/1/e50150.

27 DE S, VATS S. Decoding concerns: multi – label classification of vaccine sentiments in social media [EB/OL]. [2024 – 05 – 12]. https://arxiv.org/pdf/2312.10626.

28 CLARKE P, LEININGER C, PRINCIPATO C, et al. From a large language model to three – dimensional sentiment [EB/OL]. [2024 – 05 – 12]. https://doi.org/10.31234/osf.io/kaeqy.

29 GÓMEZ – ZARAGOZÁ L, MINISSI M E, LLANES – JURADO J, et al. Linguistic indicators of depressive symptoms in conversations with virtual humans [C]. Valencia: Working Conference on Virtual Enterprises, 2023.

30 MITTAL S, DE CHOUDHURY M. Moral framing of mental health discourse and its relationship to stigma: a comparison of social media and news [C]. Hamburg: The 2023 CHI Conference on Human Factors in Computing Systems, 2023.

31 WANG S, ZHAO Z, OUYANG X, et al. Chatcad: interactive computer – aided diagnosis on medical image using large language models [EB/OL]. [2024 – 05 – 15]. https://arxiv.org/pdf/2302.07257.

32 MAZUMDAR H, CHAKRABORTY C, SATHVIK M, et al. GPTFX: a novel GPT – 3 based framework for mental health detection and explanations [EB/OL]. [2024 – 05 – 15]. 10.1109/JBHI.2023.3328350.

33 JO E, EPSTEIN D A, JUNG H, et al. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention [C]. Hamburg: The 2023 CHI Conference on Human Factors in Computing Systems, 2023.

34 AKILESH S, ABINAYA R, DHANUSHKODI S, et al. A novel AI – based chatbot application for personalized medical diagnosis and review using large language models [C]. Chennai: 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) IEEE, 2023.

35 CHEN S，KANN B H，FOOTE M B，et al. Use of artificial intelligence chatbots for cancer treatment information［J］. JAMA oncology，2023，9（10）：1459－1462.

36 CHEN S，GUEVARA M，MONINGI S，et al. The impact of responding to patient messages with large language model assistance［EB/OL］.［2024－05－17］. https：//arxiv. org/abs/2310. 17703.

37 LAKER B，CURRELL E. ChatGPT：a novel AI assistant for healthcare messaging—a commentary on its potential in addressing patient queries and reducing clinician burnout［EB/OL］.［2024－05－18］. https：//doi. org/10. 1136/leader－2023－000844.

38 PATEL S B，LAM K. ChatGPT：the future of discharge summaries［J］. The lancet digital health，2023，5（3）：e107－e108.

39 LIU F，YANG B，YOU C，et al. Retrieve，reason，and refine：generating accurate and faithful patient instructions［C］. New Orleans：Advances in Neural Information Processing Systems 35，2022.

40 ALI S R，DOBBS T D，HUTCHINGS H A，et al. Using ChatGPT to write patient clinic letters［J］. The lancet digital health，2023，5（4）：e179－e181.

41 WAISBERG E，ONG J，MASALKHI M，et al. Chat generative pretrained transformer to optimize accessibility for cataract surgery postoperative management［J］. The Pan－American journal of ophthalmology，2023，5（1）：46.

42 LIM S，SCHMÄLZLE R. Artificial intelligence for health message generation：an empirical study using a large language model（LLM）and prompt engineering［EB/OL］.［2024－05－20］. https：//doi. org/10. 3389/fcomm. 2023. 1129082.

43 XIE Y，SETH I，HUNTER－SMITH D J，et al. Aesthetic surgery advice and counseling from artificial intelligence：a rhinoplasty consultation with ChatGPT［J］. Aesthetic plastic surgery，2023，47（5）：1985－1993.

44 World Health Organization. International classification of diseases［EB/OL］.［2024－05－21］. https：//icd. who. int/en.

45 THIRUNAVUKARASU A J，TING D S J，ELANGOVAN K，et al. Large language models in medicine［J］. Nature medicine，2023，29（8）：1930－1940.

46 HUANG C W，TSAI S C，CHEN Y N. PLM－ICD：automatic ICD coding with pretrained language models［EB/OL］.［2024－05－21］. https：//arxiv. org/abs/2207. 05289.

47 ONG J，KEDIA N，HARIHAR S，et al. Applying large language model artificial intelligence for retina international classification of diseases（ICD）coding［EB/OL］.［2024－05－21］. https：//jmai. amegroups. org/article/view/8198/html.

48 DONG H，FALIS M，WHITELEY W，et al. Automated clinical coding：what，why，and where we are［J］. NPJ digital medicine，2022，5（1）：159.

49 KARABACAK M，OZKARA B B，MARGETIS K，et al. The advent of generative language models in medical education［EB/OL］.［2024－05－28］. https：//mededu. jmir. org/2023/1/e48163.

50 AHN S. The impending impacts of large language models on medical education［J］. Korean journal of medical education，2023，35（1）：103.

51 LEE H. The rise of ChatGPT：exploring its potential in medical education［J］. Anatomical sciences education，2023，17（5）：926－931.

52 SALLAM M. ChatGPT utility in healthcare education，research，and practice：systematic review on the promising perspectives and valid concerns［J］. Healthcare，2023，11（6）：887.

53 KUNG T H，CHEATHAM M，MEDENILLA A，et al. Performance of ChatGPT on USMLE：potential for AI－assisted medical education using large language models［J］. Plos digital health，2023，2（2）：e0000198.

54 SCHUBERT M C，WICK W，VENKATARAMANI V. Performance of large language models on a neurology board－style examination［J］. JAMA network open，2023，6（12）：e2346721－e2346721.

55 RAWTE V，SHETH A，DAS A. A survey of hallucination in large foundation models［EB/OL］.［2024－06－01］. https：//arxiv. org/abs/2309. 05922.

56 MANAKUL P，LIUSIE A，GALES M J F. SelfcheckGPT：zero－resource black－box hallucination detection for generative large language models［EB/OL］.［2024－06－05］. https：//arxiv. org/abs/2303. 08896.

57 SHUSTER K，POFF S，CHEN M，et al. Retrieval augmentation reduces hallucination in conversation［EB/OL］.［2024－06－05］. https：//arxiv. org/abs/2104. 07567.

58 ROIT P，FERRET J，SHANI L，et al. Factually consistent summarization via reinforcement learning with textual entail-

ment feedback [EB/OL].[2024 – 06 – 05]. https://arxiv.org/abs/2306.00186.

59 CHERN I C, WANG Z, DAS S, et al. Improving factuality of abstractive summarization via contrastive reward learning [EB/OL].[2024 – 06 – 07]. https://arxiv.org/abs/2307.04507.

60 DHULIAWALA S, KOMEILI M, XU J, et al. Chain – of – verification reduces hallucination in large language models [EB/OL].[2024 – 06 – 10]. https://arxiv.org/abs/2309.11495.

61 XIE Q, SCHENCK E J, YANG H S, et al. Faithful AI in medicine: a systematic review with large language models and beyond [EB/OL].[2024 – 06 – 10]. https://doi.org/10.1101/2023.04.18.23288752.

62 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.

63 PAL A, UMAPATHI L K, SANKARASUBBU M. Medmcqa: a large – scale multi – subject multi – choice dataset for medical domain question answering [C]. online: Conference on Health, Inference, and Learning, PMLR, 2022.

64 LIN S, HILTON J, EVANS O. Truthfulqa: measuring how models mimic human falsehoods [EB/OL]. [2024 – 06 – 10]. https://arxiv.org/abs/2109.07958.

65 LI J, CHENG X, ZHAO W X, et al. Halueval: a large – scale hallucination evaluation benchmark for large language models [C]. Singapore: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

66 LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval – augmented generation for knowledge – intensive NLP tasks [C]. online: Advances in Neural Information Processing Systems 33, 2020.

67 VALIÑA L G, MASTROLEO I. The ethical and scientific challenges of ChatGPT in health: utopianism, technophobia and pragmatism [EB/OL].[2024 – 06 – 11]. https://doi.org/10.31219/osf.io/phtwz.

68 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 [EB/OL].[2024 – 06 – 11]. https://arxiv.org/abs/2310.05694.

69 TIAN S, JIN Q, YEGANOVA L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health [J]. Briefings in bioinformatics, 2024, 25 (1): bbad493.

70 STOKEL – WALKER C. ChatGPT listed as author on research papers: many scientists disapprove [J]. Nature, 2023, 613 (7945): 620 – 621.

(上接第 9 页)

68 KIVIKANGAS J M, FERNÁNDEZ – CASTILLA B, JÄRVELÄ S, et al. Moral foundations and political orientation: systematic review and meta – analysis [J]. Psychological bulletin, 2021, 147 (1): 55 – 94.

69 KORDZADEH N, GHASEMAGHAEI M. Algorithmic bias: review, synthesis, and future research directions [J]. European journal of information systems, 2022, 31 (3): 388 – 409.

70 RHUE L. The anchoring effect, algorithmic fairness, and the limits of information transparency for emotion artificial intelligence [J]. Information systems research, 2024 (1): 28.

71 LIU Y, DENG G, XU Z, et al. Jailbreaking ChatGPT via prompt engineering: an empirical study [EB/OL].[2024 – 02 – 28]. https://arxiv.org/pdf/2305.13860.

72 WEI A, HAGHTALAB N, STEINHARDT J. Jailbroken: how does LLM safety training fail [EB/OL].[2023 – 07 – 05]. https://arxiv.org/pdf/2307.02483.

73 DWORK C, KOHLI N, MULLIGAN D. Differential privacy in practice: expose your epsilons [J]. Journal of privacy and confidentiality, 2019, 9 (2): 1 – 22.

74 PEREZ E, RINGER S, LUKOŠIŪTĖ K, et al. Discovering language model behaviors with model – written evaluations [EB/OL].[2023 – 12 – 19]. https://arxiv.org/pdf/2212.09251.

75 ACAR A, AKSU H, ULUAGAC A S, et al. A survey on homomorphic encryption schemes: theory and implementation [J]. ACM computing surveys (CSUR), 2018, 51 (4): 1 – 35.