大语言模型在临床诊断中的应用前景与挑战综述

程睿

四川大学华西临床医学院，四川成都，610041；

摘要：近年来，随着相关技术迅速发展，大语言模型在医疗中的应用不断拓展，而其在临床诊断中的应用与挑战亟待分析总结。本文系统梳理了大语言模型在问诊与病历生成、临床决策支持、影像与报告生成等临床诊断领域的应用探索。随后详细阐述了大语言模型幻觉现象、可解释性、隐私与伦理问题、专科适配性方面存在的挑战与局限，并提出未来解决思路。最后得出结论，大语言模型在临床诊断中展现出显著潜力，其广泛应用仍需在各方面不断完善，最终服务于临床实践。

关键词：大语言模型；人工智能；临床诊断

DOI：10.69979/3029-2808.25.04.057

引言

近年来，人工智能在医学领域的快速发展为临床诊断模式带来深刻变革。尤其是大语言模型（Large Language Models，LLMs），凭借在自然语言理解和生成方面的突破性能力，逐渐展现出在临床信息处理、辅助诊断与决策中的应用潜力[1]。其强大的跨学科知识整合能力，使其能够快速获取、解析并生成符合临床逻辑的诊断提示，从而为医生提供辅助支持。然而，大语言模型在临床诊断应用中仍存在模型幻觉、隐私与伦理问题等挑战，需要更多研究加以验证和完善[2,3]。

1 大语言模型的现状与能力

1.1 大语言模型的概念和核心特征

大语言模型是基于深度学习和大规模语料训练的生成式人工智能模型，其核心特征包括参数规模庞大、预训练与微调结合、上下文理解与生成能力强等。其主要原理是通过Transformer架构实现对语言序列的建模，使模型能够捕捉语义关系并生成逻辑连贯的自然语言输出。与传统基于规则或小规模数据的医学自然语言处理方法相比，大语言模型在语义泛化、跨领域知识迁移方面具有显著优势。在医学领域，大语言模型能够实现电子病历摘要生成、患者问诊对话等多种任务。其优势在于，一是能够快速提取非结构化医学文本中的关键信息；二是支持多模态数据的融合应用趋势；三是具备通过少量提示执行复杂任务的能力。然而，大语言模型的性能高度依赖于训练数据的质量与多样性，且在面对医学问题时可能存在事实性错误与幻觉风险。因此，如何在确保医学安全性的前提下平衡模型的生成能力与可靠性，是大语言模型在临床诊断中推广应用的关键。

1.2 代表性医学大语言模型

近年来，国际上多款医学专用大语言模型相继提出，并在临床诊断和医学问答等任务中展现出突出表现。其中，BioGPT 基于大量的生物医学文献进行训练，在处理问答和分类任务、生物医学文本生成任务时展现出强大能力；Med-PaLM 是由谷歌公司研发的医疗大语言模型，通过指令微调优化，其对于 HealthSearchQA、LiveQA 和 MedicationQA 问题的输出结果已接近医生水平；而 BioBERT 则通过 PubMed 中的医学文献进行深度预训练，在命名实体识别、关系提取和问答中表现十分出色[4]。与此同时，GPT-4、Claude 等通用大模型在医学语料微调后同样能够展现跨任务的迁移潜力，显示了从通用模型向医学专用化演进的可行路径。这些国际代表性医学大语言模型在医学知识表达、临床推理和诊断支持中具有重要应用前景，但其临床落地仍受限于训练数据适配性、幻觉与临床验证不足等问题。

2 大语言模型在临床诊断中的应用探索

2.1 临床问诊与病历生成

大语言模型在临床问诊与病历生成中的应用已进入实质性探索阶段。其突出优势在于能够将复杂的医患交流信息快速转化为结构化文本，并生成具备临床价值的病历初稿。Chua 等人开发的 Russell-GPT 模型在出院

小结生成中表现出色，不仅显著提升了主要与次要诊断列表的准确性，并节省了大量时间，有助于提高临床医生工作效率[5]。不仅如此，Huang 等通过客观结构化临床考试模拟发现，ChatGPT-4 在临床问诊和病历书写方面的表现优异，能够较为完整地记录病史要点和体格检查结果，且在推理、病历完整性等方面达到甚至超过临床医生[6]。这些探索表明，大语言模型不仅具备生成规范化临床文书的能力，还能够在医生工作减负中发挥积极作用。虽然在细节上仍需人工检查与修订，但随着知识增强与临床语料的进一步融合，大语言模型有望成为病历生成领域的重要辅助工具，并推动医疗信息化水平迈向新阶段。

2.2 临床决策支持

大语言模型在临床决策支持中的探索日益深入，既可用于生成诊断提示，也能直接提供诊疗方案，同时在病理与实验室结果解读方面展现出重要价值，为医生提供了新的智能化工具。

在临床诊断提示方面，大语言模型展现出为医生提供辅助思路的重要潜力。与传统决策支持系统不同，大语言模型能够在接收到非结构化病史信息后，迅速生成涵盖多个可能诊断的鉴别清单，帮助医生拓展思维。Hirosawa 等的研究表明，在复杂内科病例分析中，ChatGPT-4 在前十项鉴别诊断列表中正确诊断率达到 83%，其表现与临床医生相当[7]。这一结果说明，大语言模型并非简单提供标准答案，而是具备生成合理诊断候选的能力，在临床早期决策和罕见病识别中尤具价值。更为重要的是，医生在面对信息复杂度较高或临床经验有限的场景时，可以借助大语言模型提供的提示作为补充参考，从而减少漏诊风险。尽管目前仍存在部分诊断优先级排序不理想、对患者具体特征把握不足等问题，但这些不足并不影响其作为诊断提示工具的潜力。未来，若能结合本地化病历数据库和知识增强机制，大语言模型有望成为提升诊断全面性和准确性的得力助手。

除诊断提示外，大语言模型也逐渐展现出直接生成诊断结论和治疗方案的能力。Liu 等的研究指出，ChatGPT-4 在回答眩晕相关疾病的诊断、治疗等问题上表现出优异的准确性、全面性、可信性，其诊断能力接近于一年经验的临床医生，表明其有潜力成为临床疾病诊断、管理中的实用工具[8]。这说明，大语言模型不仅能生成可能的诊断线索，还能为临床提供具体的诊断、管理路径，从而在医生工作中提供第二意见。然而，需要指出的是，模型在复杂病例诊断和跨学科治疗方案的生成中仍可能出现遗漏或逻辑错误，因而不能完全替代专业医生的判断。整体来看，大语言模型在直接诊疗方案生成上具有一定应用潜力，但仍需依赖严格的临床监督与逐步验证。

在病理和实验室诊断领域，大语言模型的应用探索也展现出积极前景。Hewitt 等提出了一种结合检索增强生成的大语言模型，用于解析自由文本病理报告并对脑肿瘤进行亚型分类，结果显示模型准确率高达 90%[9]。这一成果突出了大语言模型在实验室与病理场景中的独特优势，不仅能够处理非结构化报告信息，还能通过知识增强实现接近专家水平的分类推断。这意味着医生在面对复杂或冗长的病理报告时，可以借助模型快速获取关键诊断要点，提升诊断效率与准确性。与此同时，大语言模型的应用还有望推动跨机构病理报告的标准化，促进数据在科研与临床之间的高效利用。尽管目前仍存在少量误判与患者数据保密的问题，但大语言模型在病理、实验室诊断支持中的潜在应用价值不能忽略。

2.3 医学影像与报告生成

在医学影像领域，大语言模型的应用探索主要集中在报告生成与辅助影像诊断两方面。传统影像诊断高度依赖放射科医生的专业知识与丰富经验，然而随着检查量的不断增加，报告撰写带来了巨大的工作量。大语言模型能够基于影像相关数据与描述，生成高质量的初步报告，从而显著缓解医生在此方面的压力。现有研究显示，经过大规模临床语料微调的大语言模型已能够生成影像学报告的“Impression”板块的内容，其临床准确性、语法准确性和文体质量水平较高，表明大语言模型能够起草初步报告内容而提高放射科医生的工作效率，并提升报告质量[10]。除了报告生成外，大语言模型与影像人工智能算法的结合也展现出广阔前景。例如,大语言模型可以在影像人工智能模型生成的结构化结果基础上，进一步转化为更易理解、逻辑更清晰的解释，从而帮助医生快速把握诊断要点，也能让医患沟通更加顺畅。与此同时，大语言模型的多模态发展趋势值得关注，

其不仅能处理文字信息，还可能结合医学影像，自动生成病例总结并进行跨模态分析。这意味着未来医生在查看影像时，模型能够同步提供简明而有条理的解读。不过，目前单一依赖文本的大语言模型在复杂影像理解上仍有局限，尤其在需要精确空间判断或定量分析的情况下表现不足。总而言之，随着技术的成熟和临床研究的深入，大语言模型有望进行在影像报告撰写方面投入应用，甚至能够协助医生进行智能化影像解读。

3 大语言模型在临床诊断中的挑战与局限

3.1 幻觉与错误诊断

虽然大语言模型在自然语言理解与生成上展现出前所未有的潜力，但其内在的幻觉现象却导致其临床诊断应用存在重大风险。幻觉指模型在缺乏真实依据时生成虚构或错误信息，在医疗诊断情境下，可能导致严重的误导甚至危及患者安全。模型可能基于不完整或错误的病史生成一个看似合理但事实错误的诊断推断。即便是经过医学语料专门训练的模型，其输出中也难以完全避免事实性错误。此外，幻觉现象在临床决策流程中存在多重放大效应。若医生将其作为初筛工具，错误信息可能误导病史采集方向，降低后续诊断准确率；若直接应用于患者问诊与病历生成，虚构的病史信息或遗漏的关键症状将导致医疗文档质量下降，影响临床决策。与传统临床决策支持系统相比，大语言模型的高流畅度文本生成更易造成过度信任，这使得幻觉问题被进一步凸显。缓解此问题的途径主要包括以下几点。一是引入知识库约束和检索增强机制，通过实时调用权威医学指南和数据库来减少幻觉；二是结合临床逻辑，对模型输出进行二次验证；三是要求诊断的最终把关必须由医生完成。

3.2 可解释性不足

可解释性不足是大语言模型应用于临床的重要障碍之一。医生在诊断中需要明确的推理路径，以确保决策符合循证医学和临床逻辑。然而，大语言模型的生成过程缺乏透明的因果链条，使得医生难以判断模型输出的可靠性。大语言模型像是一种“黑箱”，其诊断推理依据往往难以追溯。这种不透明性直接影响医生对模型结果的信任度，尤其是在涉及高风险诊疗的场景中。如果模型仅给出结论，而无法提供可验证的推理证据，将难以满足临床使用的基本要求。因此，提高可解释性将成为其未来发展的核心目标之一。实现这一目标需要在模型设计中引入可追踪的推理机制，在输出结果中明确呈现诊断逻辑，并建立标准化的解释框架。只有当模型能够清晰展示其诊断结论的形成过程时，医生才可能采纳其提供的诊断信息。

3.3 隐私问题与伦理问题

大语言模型在临床诊断应用中必须直面隐私保护与伦理方面的挑战。病历、影像报告和实验室检查结果等信息均涉及患者个人健康数据，若在数据收集、存储与训练过程中缺乏严格的匿名化与安全控制机制，极易导致隐私泄露和数据滥用。同时，大语言模型在生成过程中可能通过参数记忆或信息反演重现训练语料中的隐私信息，即便在用户未输入相关内容的情况下，也存在敏感信息被重建或间接推断的风险，从而进一步放大数据保护的难度。

伦理层面的挑战主要体现为知情同意、责任归属等方面。患者与大语言模型交互的过程中，若信息来源和决策机制不透明，将削弱知情权的实现。若模型生成的诊断信息导致临床错误，则开发方、使用方及医疗机构之间的责任将难以划分。此外，由于医学大语言模型的训练语料主要集中于特定语言和人群，其在不同族群、性别或疾病类型中的诊断可能表现出系统性偏差。

应对上述问题需要多层次的协同措施。在技术层面，应降低敏感数据暴露风险，确保数据处理环节的安全性与可控性。在制度层面，应建立专门的治理框架，对大语言模型在临床诊断中的应用范围、责任划分进行系统规定。伦理审查机制应贯穿研发、部署与使用的全过程，以在推动人工智能创新的同时最大限度保障患者权益。

3.4 专科适配性问题

临床诊断的高度专科化特征对大语言模型的适配性提出了严苛要求。不同医学专科在病理机制、诊断标准和辅助检查等的解读上均具有独特性，而通用型大语言模型通常依赖于大规模通用语料训练，难以满足各专科临床诊断的深层次需求。高度专业化的诊断任务往往超出通用模型的知识覆盖与推理能力。此外，临床诊断

不仅依赖静态知识，还涉及动态指南更新与特定专科的实践经验。若模型未能及时获取最新循证医学成果，其结论可能与现行诊疗规范脱节；同时，专科诊断常需在大量相似症状中辨识细微差异，这种高精度任务对模型的知识深度与推理稳定性提出更高要求。因此，大语言模型在临床诊断的有效应用，需要基于专科知识库进行定向训练与优化。通过引入权威数据库、专病种临床指南和真实世界诊疗记录等，可以增强模型在特定专科疾病诊断中的适配性。

5 结语

大语言模型为临床诊断提供了新的智能化工具，其在病历生成、辅助诊断、以及病理与影像报告处理等方面均展现出重要潜力。它能够提升医疗信息处理效率，拓展医生的诊断思路，提高诊断的全面性与准确性。然而，幻觉与错误诊断、可解释性、隐私与伦理困境、专科适配性问题，仍然是限制其大规模临床应用的关键因素。未来的发展应在诊断推理、个性化适应及监管体系建设等方面持续推进，并以临床需求为导向，加强循证研究与专科优化。在保障安全与合规的前提下，大语言模型有望成为临床诊断的重要补充力量，为医学实践带来深层次变革。

参考文献

[1]施呈昊,屠馨怡,史佳伟,等.大语言模型临床实践应用范围综述[J].医学信息学杂志,2024,45(9):19-26.

[2]肖仰华,徐一丹.大规模生成式语言模型在医疗领域的应用:机遇与挑战[J].医学信息学杂志,2023,44(9):1-11.

[3]何剑虎,王德健,赵志锐,等.大语言模型在医疗领域的前沿研究与创新应用[J].医学信息学杂志,2024,45(9):10-18.

[4]肖建力,许东舟,王浩,等.医疗领域的大型语言模型综述[J].智能系统学报,2025,20(03):530-547.

[5]CHUA C E, LEE YING CLARA N, FURQAN M S, et al. Integration of customised LLM for discharge summary generation in real-world clinical settings: a pilot study on RUSSELL GPT[J/OL]. The Lancet Regional Health: Western Pacific, 2024, 51: 101211[2025-08-28]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11776078/. DOI:10.1016/j.lanwpc.2024.101211.

[6]HUANG T Y, HSIEH P H, CHANG Y C. Performance Comparison of Junior Residents and ChatGPT in the Objective Structured Clinical Examination (OSCE) for Medical History Taking and Documentation of Medical Records: Development and Usability Study[J/OL]. JMIR medical education, 2024, 10: e59902. DOI:10.2196/59902.

[7]HIROSAWA T, KAWAMURA R, HARADA Y, et al. ChatGPT-Generated Differential Diagnosis Lists for Complex Case-Derived Clinical Vignettes: Diagnostic Accuracy Evaluation[J/OL]. JMIR medical informatics, 2023, 11: e48808. DOI:10.2196/48808.

[8]LIU X, SHI S, ZHANG X, et al. The role of ChatGPT-4o in differential diagnosis and management of vertigo-related disorders[J/OL]. Scientific Reports, 2025, 15(1): 18688. DOI:10.1038/s41598-025-96309-8.

[9]HEWITT K J, WIEST I C, CARRERO Z I, et al. Large language models as a diagnostic support tool in neuropathology[J/OL]. The Journal of Pathology. Clinical Research, 2024, 10(6): e70009. DOI:10.1002/2056-4538.70009.

[10]SERAPIO A, CHAUDHARI G, SAVAGE C, et al. An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study[J/OL]. BMC medical imaging, 2024, 24(1): 254. DOI:10.1186/s12880-024-01435-w.