大语言模型在肿瘤诊断中的文字报告与医学影像应用研究进展

程浩然[1,2],严鸿斌[3],袁子云[4],庄泽鸿[1,2],孙学刚[5],姚学清[1,2,6]
[1]南方医科大学附属广东省人民医院(广东省医学科学院)胃肠外科普通外科,广东 广州 510080;[2]广东省人民医院赣州医院(赣州市立医院)普通外科,江西 赣州 341099;南方医科大学[3]第一临床医学院,[5]中医药学院,广东 广州 510515;[4]中山市人民医院普外科,广东 中山 528400;[6]华南理工大学医学院,广东 广州 510641

摘要:大语言模型(LLMs)作为新兴人工智能技术,凭借其优异的文字与图像处理能力,为医疗领域智能化变革提供核心支撑,显著提升临床工作效率与质量。本文系统梳理LLMs在癌症诊断领域的应用现状、技术特点及发展方向,重点聚焦两大核心场景:一是影像报告、病理报告、综合病例报告等文字报告的自动化分析与解读;二是融合文本与医学影像的多模态数据诊断。研究发现,LLMs在癌症诊断中的综合能力已可媲美普通住院医师,但在专业化诊断与精准化判断方面仍存在明显短板;同时,LLMs展现出"小参数模型适配基层场景""多语言报告分析泛用性差异"等应用层面特征。未来需进一步开发专业化、实用化的医疗专用LLMs,通过优化微调策略、构建高质量中文医疗数据集、整合视觉语言模型等方式,推动其临床落地并弥合医疗资源差距。

关键词:大语言模型;人工智能;肿瘤诊断;病理学;影像学

Research progress of large language models in tumor diagnosis: applications in textual reports and medical imaging

CHENG Haoran[1,2], YAN Hongbin[3], YUAN Ziyun[4], ZHUANG Zehong[1,2], SUN Xuegang[5], YAO Xueqing[1,2,6]
[1]Department of Gastrointestinal Surgery, Department of General Surgery, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China; [2]Department of General Surgery, Guangdong Provincial People's Hospital Ganzhou Hospital (Ganzhou Municipal Hospital), Ganzhou 341099, China; [3]First Clinical Medical School, [5]School of Chinese Medicine, Southern Medical University, Guangzhou 510515, China; [4]Department of General Surgery, Zhongshan People's Hospital, Zhongshan 528400, China; [6]School of Medicine, South China University of Technology, Guangzhou 510641, China

Abstract: Large language models (LLMs) are emerging artificial intelligence technologies with strong text and image processing capabilities, offering critical support for the intelligent transformation of healthcare and improving clinical efficiency and quality. This review summarizes the current applications, technical features, and future directions of LLMs in cancer diagnosis, focusing on two key scenarios: automated analysis of textual reports (e.g., imaging, pathology, and case summaries) and multimodal diagnosis combining text and medical images. Findings show that LLMs now perform at a level comparable to general resident physicians in cancer diagnosis but are still incapable of making specialized and precise judgments. They also exhibit application-specific traits, such as parameter-efficient models adapted for grassroots-level scenario and divergent versatility in multilingual report analysis. Future efforts should prioritize developing specialized, practical medical LLMs through optimized fine-tuning strategies, construction of high-quality Chinese medical datasets, and integration with vision-language models to promote the clinical application of these models and increase the accessibility of healthcare resources.

Keywords: large language models; artificial intelligence; cancer diagnosis; pathology; medical imaging

大语言模型(LLMs)是一种基于深度学习的人工智能技术,"大"字表示模型训练数据以及参数量远超既往模型,LLMs的核心功能是帮助计算机理解人类语言。语言模型自1950年图灵测试提出后逐步发展[1],2003年神经网络语言模型提升了语言模型处理文本的性能[2],2017年Transformer框架奠定了大规模训练基础[3],2018年BERT模型采用双向预训练过程标志着语言模型进入预训练模型阶段[4],2022年ChatGPT的问世标志着对话问答生成式的LLMs的诞生[5]。LLMs的出现不仅深刻影响了人们的生活,同时为医疗体系的现代化、智能化带来新的发展动力。

临床医学是数据密集型的领域,如何从大量的病例、文字报告、影像图像等相关资料中高效提取重要信息、做出正确判断一直是困扰临床工作者的问题,LLMs对文本以及图像的处理能力成为解决这个问题的关键钥匙。目前,Google的DeepMind-RadFusion、Meta的NLLB模型以及斯坦福大学的Drug-GPT等成果,已经初步展现出LLMs在问诊、诊断、治疗、护理等方面的巨大作用[6],提高了医疗服务的效率以及质量,但LLMs在准确性、可靠性、数据安全甚至伦理方面的隐患也逐渐凸显。

收稿日期:2025-09-02
基金项目:国家自然科学基金(82260501,82274387);广东省特支计划科技创新领军人才项目;江西省赣州市科技计划项目(202101074816)
Supported by National Natural Science Foundation of China (82260501, 82274387).
作者简介:程浩然,在读博士研究生,E-mail: chenghaoran@gdph.org.cn
通信作者:姚学清,博士生导师,主任医师,E-mail: syyaoxuqing@Scut.edu.cn

1 LLMs在文字报告方面的应用

LLMs的重要功能是将自然语言中的信息提炼成简洁高效的电子化格式。在临床工作中,文字报告承担着沟通患者与医生、临床与医技之间的桥梁作用。然而临床工作中大量的文字报告往往带来繁琐的工作,辅助科室医生需要尽快出具真实可信、有专业价值的报告,临床医生则需要在短时间内对报告进行阅读、提炼。医疗的专业性使患者对报告的理解依赖于医生,间接增加医生的工作负担。因此,推动LLMs在临床报告生成与解读中实现有效赋能,已成为医疗AI发展的重要命题。

1.1 影像学报告

影像学利用各种成像技术对人体内部结构和器官进行非侵入性检查和诊断,在各种癌症的诊断中发挥着不可或缺的作用。医学影像报告的读写是影像科医生重要的工作,临床医生需要具备专业的知识和判断力以理解影像报告的内容、做出正确的诊断。LLMs可评估影像学报告并做出诊断,根据影像检查结果高效率地生成医疗报告,包括评估肿瘤、结节的分期或者评估患者病情的变化等。

1.1.1 *应用* LLMs可以提高医生效率。LLMs辅助处理影像报告效率提升,如人工平均处理报告时间由43 s/报告缩减到16 s/报告[7],在连续腹部CT报告处理这一更复杂的任务中,从12 min/份缩短至26 s/份[8]。然而,在诊断局灶性肝病变中,LLMs可以提高初级医生的诊断能力,却无法改善中级医生的诊断能力与效率,推测原因是临床医生需要额外的时间整合LLMs的信息[9]。

LLMs在处理多中心、多语言的影像报告中也展示了一定的能力。有研究通过虚拟报告训练LLMs后,模型在外部验证集中的诊断准确率可达95%[10]。一项研究比较了LLMs和不同经验人类阅片者根据韩英混合报告进行肺癌分期的能力,LLMs对韩英报告与纯英文报告的诊断能力无差异,但医生在混合语言报告中的诊断能力随经验和语言能力波动较大[11]。另一项针对韩英报告的分析中也体现了LLMs在多语言中的泛用性[7]。而在其他语种影像报告的研究,比如中文[12]、德文[13]、意大利文/荷兰文[14]、日文[15]中,LLMs展现了在多语种分析中的泛用性。

一些研究聚焦于更加细化的任务,例如肿瘤进展评估[8]、评估癌症的TNM分期[16]、将自由文本报告转化为结构化报告[17]、甲状腺超声U评分生成[18]等任务,体现了LLMs在临床应用中的广泛性。

1.1.2 *挑战* LLMs的诊断能力不稳定。部分研究中,LLMs的诊断能力不低于甚至优于人类医师[18, 19],但另有研究仍存在显著差距[5, 14-16]。这说明当下LLMs的临床应用需要严格的人工复查,无疑增加了使用成本。

LLMs的输出不稳定。一项研究通过人工生成的肝脏MRI报告训练LLMs,然后使用真实的报告进行验证,发现32%的错误在重新运行后消失[20];一项涵盖意大利文、英文、荷兰文的研究发现,不同语言报告中的一致性存在差异,以GPT4为例,英文报告的一致性(AC1=0.58)显著高于意大利文(AC1=0.53)与荷兰文(AC1=0.45)[15];有研究发现LLMs针对前列腺MRI报告诊断的一致性远低于医师[21],另一项针对肺癌CT报告的研究也揭示了这一现象[22]。

LLMs误判肿瘤类型的问题广泛存在。前述研究发现,LLMs可能将腹水误判为与肿瘤相关,导致假阳性率升高;同时,其还存在将恶性肿瘤误判为良性肿瘤、或在难以鉴别时优先判定为良性肿瘤的倾向,进而导致假阴性率升高[8],此类假阳性与假阴性问题在其他研究中也多次出现[14, 23]。此外,LLMs高估肿瘤风险的情况也广泛存在[11, 16, 22],也存在同时高估低风险肿瘤、低估高风险肿瘤的情况,或者LLMs虚构出不存在的肿瘤分类的情况[13]。这些错误的部分原因是LLMs缺乏专业医学知识,对医学术语的理解不足,部分原因可能是LLMs会过度依靠少数证据进行诊断,还有LLMs制造不存在知识的"幻觉"现象这一问题。根源可能在于LLMs训练需要大量的语料,以GPT-3为例,其训练语料主要来源于Common Crawl这一网络爬取数据集[24],多元的网络语料包罗万象但无法保证医学相关的语料占比,这会导致LLMs的专业知识匮乏,另一方面LLMs固有的决策过程不透明、幻觉等特性也干扰着肿瘤的正确诊断,从而导致了一系列诊断错误。

1.1.3 *展望* 针对性的微调策略或提示工程可以改善LLMs的性能。有研究证明简单的提示工程可以起到作用,在研究中对比单步提示法和两步提示法(将患者信息以及影像报告分步输入),发现两步提示法即可显著提升LLMs的诊断能力,但由于文章基于单中心数据,无法证明这一方法的泛用性[9]。有学者探索了特征总结、混合检索增强生成、思维链、少样本学习4种策略如何结合才能更好提高LLMs性能,通过对比发现特征总结、混合检索增强生成、思维链三者结合在内部训练集和外部测试集中表现最佳,但该研究仅探索了肺部疾病的适用性,未验证这一提示工程组合在不同癌症种类中的泛用性[12]。有研究表明LLMs有卓越的少样本适应能力,基于CT报告训练的LLMs的系统对训练未覆盖的PET-CT/MRI报告仍保持高精准度[25]。尽管如此,有研究使用模拟报告训练的LLMs在真实报告中的表现未达预期[6]。在数据来源限制的条件下,现有研究仅仅注重某一特定癌症种类探究微调策略提示工程的效果,但是临床上更需要的是一种更加有泛用性或者自动化的微调策略和提示工程,而非面对不同条件下手动调整。

比较5篇文章中主流LLMs差异(图1),在整体上OpenAI的GPT系列LLMs表现相对较优,在不同研究之间,LLMs的诊断准确率有整体的诊断率相对较高或较低的趋势。有研究特别提示了LLMs处理不同语言中的能力差异[14],出现这样的现象主要与LLMs的训练数据集相关,另一方面则是LLMs算法本身的差异性,LLMs基于不同的Transformer框架,各个模型训练参数差异巨大,从数千亿到数万亿不等,训练方法也有无监督、有监督、端到端等不同方式。目前来看,这种差异性无法避免,不仅需要医疗领域提供足够优质、优量的公共数据集,还要求公司间进行算法之间的交流、迭代升级。一个更加符合临床实际的方案是由医生承担的职责,对各类LLMs出具的报告进行审核,择优采纳,但这能否提高医生的工作效率仍旧值得探究[4]。

在目前的技术限制下,LLMs适合作为辅助工具而不能完全代替医生,LLMs在对专业性、可靠性以及准确度要求较高的工作上表现远不如医师。在专业性方面,临床医学的高专业性对LLMs提出了更高的要求,基于现有的LLMs,结合医疗文本、临床指南对LLMs进行进一步的训练是有必要的。在可靠性以及准确度方面,尽管可以通过思维链模拟人类决策过程[26],但由于LLMs决策过程对研究人员是个"黑箱",导致无法从逻辑层面解决其"幻觉"或诊断错误问题,现有的思维链、检索增强生成等技术只能增强可解释性与可靠性,但无法解决根本问题。最新研究开发的OLMoTrace系统可以对LLMs的输出进行溯源,识别模型是否在编造事实[27]。将OLMoTrace融入医疗LLMs,可能是推动医疗LLMs临床落地的关键所在。

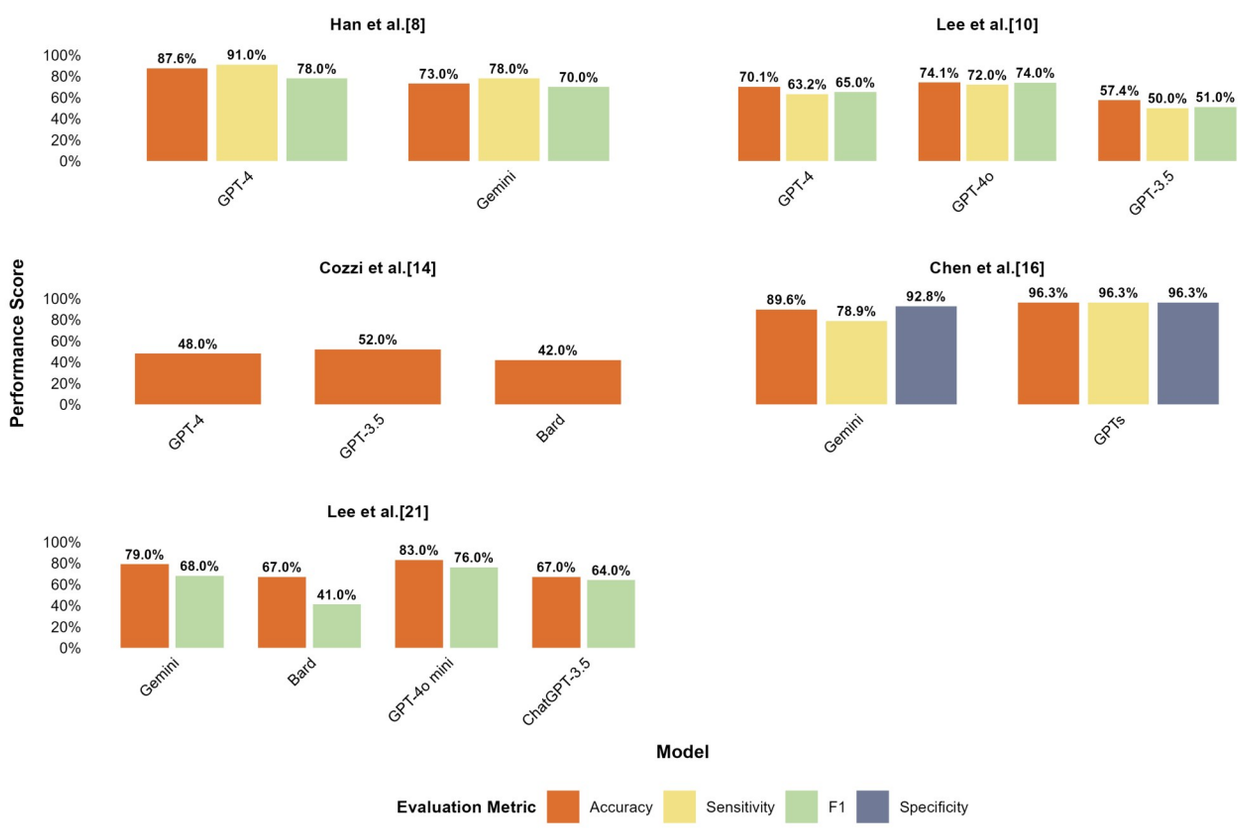

图1 主流LLMs在不同研究中的表现对比

Fig.1 Comparison of performance of the mainstream large language models (LLMs) in different studies.

1.2 病理报告

病理学是现代医学的重要组成部分,贯穿于疾病诊断、治疗、预后评估、预防以及医学研究等方面。在肿瘤的诊断中,病理是疾病诊断的金标准。在各种癌症的病例诊断中,准确的分期至关重要,通常病理分期是由病理医生人工标注的,但过程中容易出现错误,需要大量时间和精力来确保准确性并符合最新的指南。LLMs可以辅助病理医生进行病理的分期,识别患者的病理进展,高效的检查出病理报告中的错误。

1.2.1 *应用* 提示工程和微调策略可提高LLMs病理诊断性能。有研究经过思维链、低秩适应微调后的LLMs对肺癌做出了精准TN分期,与专家标注高度一致[28];另有相似研究通过提示-响应对、低秩适应的微调策略后在诊断肺癌TN分期上的准确率达到98.76%[29]。有学者探究了零样本和检索增强生成下LLMs诊断成人型弥漫性胶质瘤的性能,结果显示单独的零样本策略下的

LLMs诊断准确率较低(22.2%),不适合于临床使用,检索增强生成可以显著提升LLMs的诊断准确率(90%)[30]。

小参数LLMs(7 B)相较于大参数LLMs(70 B)病理诊断表现更优[29],相似地,添加前馈神经网络层和低秩适应微调后的小参数GPT-2模型相较于大参数LLMs有更优秀的诊断表现[31]。

1.2.2 挑战 LLMs在更加复杂的病理报告中准确率较低。有研究发现LLMs在T4分期的肺癌病理报告中预测性能较低(F1分数=0.545)[28],有研究人为在病理报告中植入错误后,发现LLMs的整体诊断能力相较于资深医师假阳性率较高,错误检出率较低[32]。

1.2.3 展望 不同数据来源与微调策略对LLMs性能存在显著影响(表1)。在Kim等[28]的研究中,LLMs判断N分期的性能优于T分期,而在Cho等[29]的研究则呈现相反趋势,且LLMs对淋巴结转移部位相显著弱于肿瘤尺寸的识别能力。出现以上情况的原因,一方面是训练集的不同,Cho等[29]研究中的淋巴结转移部位描述标注样本少、类别不平衡,而Kim等[28]研究中T4、T2b等特殊分类的样本量不足,影响了LLMs对T分期的判断性能,这一现象强调了数据样本量和类别平衡对LLMs性能的影响,Cheligeer等[31]在研究中采用SMOTE过采样阳性样本的策略是一个解决方式;另一方面,淋巴结转移部位的判断相较于肿瘤尺寸的判断更为复杂,在Cho等[29]的数据中,淋巴结位置共有48个子类别,而肿瘤尺寸仅1个,任务复杂程度影响LLMs的诊断性能,这一问题可以通过提示工程或微调策略解决,比如对复杂问题进行拆分、简化,或者通过检索增强生成提供专业性的映射,但对复杂问题处理前后的对照研究是欠缺的。Orca2相较于Mistral-7b逻辑推理能力相对较弱,临床诊断需要有强大逻辑推理能力的LLMs[28]。未微调的大参数模型相较于基础模型在性能上有所提高,但是基础模型经演绎集预训练或前馈神经网络层、低秩适应等微调策略后,在诊断方面的表现显著优于未微调的大参数模型。这一方面强调了专业性的医疗文本对LLMs在医疗领域表现的影响,另一方面强调了微调策略比参数规模更重要,经过微调的小参数LLMs有优异的性能,且方便部署、应用,适合算力相对不足的条件,这提高了LLMs在临床上的实用价值。

1.3 综合报告

在实际的临床工作中,临床医生要面对的往往不是单一检验检查结果,而是要通过综合的分析,结合患者病史、影像学评估、病理分析、分子检测以及患者的既往治疗方案,才能更好地明确患者的诊断[33]。因此对综合信息的处理才是临床工作中常用的模式。LLMs可以处理分析患者的综合信息,然后结合所有信息给出诊断。

1.3.1 应用 LLMs综合多来源文本信息诊断癌症仍具较优能力。有研究在将中文病例经专业工具翻译为英文后,评估LLMs在真实门诊、中英双语情境下诊断中枢神经系统肿瘤的能力,发现ChatGPT-4o与DeepSeek-R1有与资深医师相近的诊断能力且中、英文病例的诊断表现无差异,表明了LLMs在真实的临床场景、中英双语场景中的适用性,后续研究若进一步泛化至其他临床场景与疾病,将为LLMs临床落地提供更充分证据[33]。部分研究显示LLMs诊断能力不及资深医师[34],但结合逻辑回归分析风险模型的定制LLMs诊断能力显著提升[35]。

1.3.2 挑战 LLMs对罕见病、疑难病的诊断能力低。LLMs在中枢性神经系统肿瘤中的诊断率仅有16%[36],因此加强LLMs对罕见病、疑难病的诊断能力非常必要,可以通过增加专业的罕见病语料库解决这一问题,但这需要更多的研究和实践确定效果。

在综合更多来源的文本报告后出现了新的问题。有学者设置了不同的输入条件,如综合患者病史、生化检测、影像报告或者仅分析影像报告,发现结合更多的信息不一定能提高LLMs的诊断能力[37];有研究评估GPT-4对多份脑部MRI报告的总结质量,并生成R代码以可视化患者疾病进程,发现模型不能100%提取关键信息,但研究基于非标准化的文本信息和GPT-4,并未考虑文本标准化以及不同LLMs提取信息能力的影响[38],这可能是影响LLMs综合不同来源文本后诊断能力无法提升的原因。

1.3.3 展望 综合更多信息可以提高LLMs诊断癌症的能力,但也引入信息提取能力问题。LLMs建立于Transformer框架,基于注意力的框架在处理长文本时无法捕捉所有的细节,这可能导致了提取信息能力的不足。探索不同微调策略与提示工程的结合,或许可解决LLMs信息提取问题,如通过适当的提示工程如身份赋予、任务明确,结合思维链和检索增强生成等微调策略提升信息提取率和可解释性,或通过结构化JSON输出提升可控性与临床适配度。

2 大语言模型在图像方面的应用

医学影像作为癌症诊断的核心支柱,但传统影像诊断长期受限于解读结果的主观性差异、医生繁重的工作负荷,以及基层与偏远地区专业资源的匮乏。多模态大模型可以整合文本、图像等多类型数据,2023年3月支持图像分析的ChatGPT-4发布为重要里程碑,后续演进出了GPT-4o、Claude 3等多模态模型,拉开了LLMs分析图像辅助癌症诊断的序幕。

2.1 应用

初步研究已证实LLMs在基于多模态资料进行癌

表1 主要开源大模型及商业大模型在在不同研究中的表现对比

Tab.1 Comparison of the performance of major open-source and commercial LLMs in different studies

Literature	LLMs	Accuracy (%)	Evaluation metrics	Prompt engineering	Fine-tuning strategy
Kim et al.[28]	Orca2_13b (deductive pre-trained model)	93.4	F1-score: T classification: 0.889, N classification: 0.997	Three-step prompting (clarify task, provide glossary, structured output)	Chain-of-thought, Low-rank adaptation
	Orca2_7b (deductive pre-trained model)	91.4	F1-score: T classification: 0.886, N classification: 0.992		
	Llama2_7b (basic open-source model)	86.4	F1-score: T classification: 0.750, N classification: 0.967		
	Mistral_7b (basic open-source model)	57.2	F1-score: T classification: 0.515, N classification: 0.888		
Cho et al.[29]	Deductive Mistral-7B (deductive pre-trained model)	92.24	F1-score: Tumor max size: 0.9939, Nodule location: 0.2939, T classification: 0.9860, N classification: 0.7845	Three-step prompting (role assignment, clarify task, structured output)	Low-rank adaptation
	Orca-2 (deductive pre-trained model)	91.15	F1-score: Tumor Max size: 0.9939, Nodule location: 0.4097, T classification: 0.9905, N classification: 0.7811		
	Llama-2-70B (commercial large-parameter model)	75.78	F1-score: Tumor max size: 0.9788, Nodule location: 0.2027, T classification: 0.8776, N classification: 0.7271		
	Llama-2-7B (basic open-source model)	62.89	F1-score: Tumor max size: 0.9438, Nodule location: 0.4725, T classification: 0.8567, N classification: 0.6331		Low-rank adaptation, Deductive datasets pre-training
	Mistral-7B (basic open-source model)	19.57	F1-score: Tumor max size: 0.9543, Nodule location: 0.0568, T classification: 0.8000, N classification: 0.5188		
Cheligeer et al.[31]	GPT-2 (fine-tuned version)	95.80	Sensitivity: 95.3%, Specificity: 96%	None	Feed-forward neural network layer, Low-rank adaptation
	BART	94.40	Sensitivity: 100%, Specificity: 92%		Text chunking, Token embedding
	BioClinicalBERT	90.10	Sensitivity: 95.2%, Specificity: 86%		
	GPT-2 (base version)	88.70	Sensitivity: 85.7%, Specificity: 90%		
	GPT-2 (large version)	88.70	Sensitivity: 100%, Specificity: 92%		

症诊断方面的潜力。GPT-4系列模型对低风险结节表现出极高的特异性(99.5%),有效辅助排除良性病例,优化分诊流程[39]。虽然多项研究表示,通用的LLMs在结合多模态资料后,相对医师的诊断率仍较低[40-43],但也有学者探索了基于标准化报告提升LLMs甲状腺结节超声诊断性能的方法,并对比了3种策略:一是医生先解读图像并记录关键信息,LLMs基于记录给出诊断;二是由多任务学习分类模型解读图像转换为文本,LLMs基于文本给出诊断;三是基于卷积神经网络作为对照组,结果显示尽管卷积神经网络组诊断性能最优,但其属于"黑箱模型",而LLMs可基于TI-RADS征象生成诊断逻辑,临床医生可追溯每一步决策依据,显著提升了诊断透明度;此外,图文转换策略与高级医生-LLMs人机交互策略性能相当,足以满足临床需求[44]。LLMs对膀胱镜图像中膀胱肿瘤的识别率达92.2%,彰显了其作为筛查助手的价值[45]。有研究将每位患者的CT图像按时间顺序转换为20帧/s的视频后,发现了GPT-4o在纵向CT扫描动态评估肺结节恶性概率、大小及特征变化的可靠能力,且随纵向CT图像数量增加,GPT-4o诊断效能显著提升,表明多模态LLMs在随访中的实用性,但由于平均随访CT次数仅2.8次,无法评估LLMS在长期随访中的能力[46]。有学者采用检索增强生成的方式并精准引用可靠外部知识,大幅度提高了LLMs的诊断准确率,并能精确定位知识来源,提升了医生对模型输出的验证效率与信任度[47]。

在图像处理上,有研究通过SegResNet模型分割语义、Julich脑图谱映射的策略,构建JSON文件限制LLMs输出,将分割错误传播率从53.1%降至1.8%,再驱动LLMs输出符合临床用语且可溯源的临床报告,在脑部肿瘤中创建了分割、图谱映射、JSON结构化、LLMs报告的工作流,但由于实验基于标准化数据集,未验证其在真实临床数据集上的实用性[48]。有学者创建了4步图像处理框架,实现了CT图像从重建、预处理、分割到自动描述的整合,优于主流深度学习图像处理方法,再经LLMs提取提取视觉特征,生成连贯的描述,提高了可读性,辅助临床医生的工作[49]。有研究将LLMs与

深度学习自动分割技术深度整合，成功校正了深度学习自动分割图像分割中的所有共22例假阳性错误，减轻了放射学家的手动轮廓负担[50]。

2.2 挑战

LLMs在多模态资料分析中有肿瘤精确定位、微钙化识别等细节判断能力不足、可靠性较低等问题。在乳腺癌X线摄影分析中，LLMs虽有一定检出率(>60%)，但在判定乳房密度、肿瘤精确定位及微钙化识别等关键细节上准确率不足[43]。LLMs在口腔恶性肿瘤的诊断准确率远低于人类专家，且相较于单纯基于文本，结合图像(图像可提供额外的颜色、大小等信息)后，仅新增5例诊断病例，诊断准确率仍未实现显著提升[51]。

LLMs在医学图像深度理解方面存在显著局限。有研究评估了ChatGPT-4o识别脑部MRI的能力，即便LLMs对基础的图像特征识别能力尚可，其病灶定位、最终诊断的能力仍无法达到放射科医生水平，在脑膜瘤病例中的诊断率仅10%[40]。在骨科肿瘤病理诊断中，ChatGPT的准确率远低于住院医师团队，甚至在软组织肉瘤的诊断中完全失败(0%)[41]。LLMs通过甲状腺结节超声图像进行良恶性判别，结果显示LLMs的诊断性能显著低于初级放射科医生，不必要活检率(43%)远高于医生(12%)[42]。有研究采用GPT-4分析多器官组织病理图像，其识别肿瘤及组织起源的总体准确率和一致性较低[52]。有研究评估ChatGPT-4o分析口腔病变能力，虽然诊断准确率达85%，但在10例正常图像中，有7例被误判为病变[53]。也有学者探究了思维链、少样本提示等策略对LLMs图像信息处理能力的影响，而不同的策略均未稳定提升LLMs性能，甚至经少样本提示后，Gemini 2.5 Pro的诊断准确率从63%降至5%[54]。

2.3 展望

LLMs训练多基于文本数据，缺乏对图像像素级特征(如灰度、边缘、纹理)的理解能力，难以捕捉诊断所需的关键视觉信息，也缺乏对病灶、病理特征的专业识别能力，为其基于多模态临床资料诊断带来挑战。LLMs在图像处理方面能力不足，结合专用的视觉编码器、图像分割等对图像进行处理后，由LLMs辅助分析可以结合两者优点从而提高诊断性能，这一方向的核心发展路径为视觉语言大模型(VLM)，VLM作为新兴概念发展迅速，但仍旧存在视觉幻觉、空间关系理解、动态图像理解不足等问题，需要等待技术更加成熟才能在临床应用上落地。此外，视觉处理等算法的处理过程存在“黑箱”问题，且缺乏量化的可解释性与置信度指标，影响多模态LLMs的实际应用[49]。前述研究利用LLMs强大的文本理解能力处理影像报告信息，指导或校正视觉提取的结果，可以构建优势互补的协同工作流，证明了这一方法的可行性[50]。

LLMs展现出快速迭代升级的潜力。有研究追踪了GPT-4在半年内的表现，观察到其对管状腺瘤识别的准确率从67%显著提升至75%[52]。

3 讨论

LLMs可应用于多种临床场景，在前列腺癌、神经系统肿瘤的影像学报告及甲状腺的超声影像中，有不逊于资深医师的诊断性能，但现有研究以及LLMs仍存不足。首先，研究多基于单中心数据，未验证LLMs在多中心、多语言、多人种的泛用性，少数多中心研究显示其在训练集、测试集诊断能力有局限[18]，且对不同语言报告处理性能差异显著[14]，如一项双语研究中，ChatGPT-4o基于英文报告的诊断准确率略高，豆包基于中文报告的诊断准确率更优[33]，这说明需要构建以中文为主要训练语言的医疗数据集，开发更适配中文临床场景的LLMs。其次，LLMs在医学术语及模糊描述处理中表现不佳，诊断准确率较低。可采用专业文本预训练的LLMs处理专业任务[55]，或使用微调策略、提示工程，但不同LLMs对提示工程、微调策略反应不同[10,20]，部分微调后诊断准确性反而下降[56]，目前缺乏不同策略对各LLMs诊断提升的对比研究。另一策略是结合不同特点、侧重的LLMs完成复杂、高要求任务[57]。现有研究较少关注LLMs罕见病诊断能力，可结合罕见病公共资源库，提升其少样本学习下的罕见病诊断能力。

在实际应用方面，LLMs多模态研究贴合实际不足，临床常需结合多模态信息，但LLMs综合图像、报告等多模态资料能否提升诊断性能的评估较少，综合多模态资料是LLMs未来临床应用的关键。LLMs可从辅助筛查、高效分诊及报告自动化切入，通过提升临床医生效率与诊断能力，成为医生信赖的智能辅助伙伴，从而弥合医疗资源差距、改善患者预后。

在伦理层面，大型语言模型作为新兴技术，其应用边界与责任归属尚缺乏清晰的界定。OLMoTrace可追溯LLMs知识来源，但研究表明，仅用0.001%医疗虚假信息替换训练数据标记，LLMs传播医疗错误风险显著升高[58]。LLMs相关医疗事故的责任归属尚不明确。LLMs存在训练集导致的偏见问题，甚至可能在自杀倾向的对话中给出不当回应[59]，需引入多方面评价体系。在数据安全方面，医疗数据隐私性强，使用云端商业化LLMs存在隐私泄露、依赖第三方平台风险[60]。使用可本地部署的开源的小参数[23]或中参数LLMs[54](建议参数范围为70亿～120亿[61])，可避免数据外流，符合HIPAA个人信息保护标准，同时在成本、灵活性方面有显著优势[10]。

尽管LLMs的全面临床应用仍有距离，但利用AI医生问诊、宣教可提升医疗机构服务质量。LLMs的发

展需更多研究机构、医务人员参与，加强伦理与隐私监管，建立完善审查制度。未来，期望LLMs辅助医务人员，为患者提供智能化、精准化的医疗服务。

Declaration of interests: The authors declare no competing interests.

参考文献：

[1] Alan MT. Computing machinery and intelligence[J/OL]. Mind, 1950, 59(236): 433-60.

[2] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. J Mach Learn Res, 2003(3): 1137-55.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J/OL]. arXiv. https://arxiv.org/abs/1706.03762v7.

[4] Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J/OL]. arXiv. https://arxiv.org/abs/1810.04805.

[5] Liu Y, Han T, Ma S, et al. Summary of ChatGPT-related research and perspective towards the future of large language models[J]. Meta-Radiology, 2023, 1(2): 100017.

[6] 孙 磊, 汪安安, 宋一敏, 等. 大语言模型在临床医学领域的应用、挑战和展望[J/OL]. 解放军医学院学报, 2025, 46(1): 50-60.

[7] Di Palma L, Darvizeh F, Alì M, et al. Structured transformation of unstructured prostate MRI reports using large language models[J]. Tomography, 2025, 11(6): 69.

[8] Han NY, Shin K, Kim MJ, et al. Enhancing oncological surveillance through large language model-assisted analysis: a comparative study of GPT-4 and gemini in evaluating oncological issues from serial abdominal CT scan reports[J]. Acad Radiol, 2025, 32(5): 2385-91.

[9] Sheng LJ, Chen YD, Wei H, et al. Large language models for diagnosing focal liver lesions from CT/MRI reports: a comparative study with radiologists[J]. Liver Int, 2025, 45(6): e70115.

[10] Lee JH, Min JH, Gu K, et al. Automated resectability classification of pancreatic cancer CT reports with privacy-preserving open-weight large language models: a multicenter study[J]. J Med Syst, 2025, 49(1): 118.

[11] Lee JE, Park KS, Kim YH, et al. Lung cancer staging using chest CT and FDG PET/CT free-text reports: comparison among three ChatGPT large language models and six human readers of varying experience[J]. AJR Am J Roentgenol, 2024, 223(6): e2431696.

[12] Li RH, Mao S, Zhu CM, et al. Enhancing pulmonary disease prediction using large language models with feature summarization and hybrid retrieval-augmented generation: multicenter methodological study based on radiology report[J]. J Med Internet Res, 2025, 27: e72638.

[13] Fervers P, Hahnfeldt R, Kottlors J, et al. ChatGPT yields low accuracy in determining LI-RADS scores based on free-text and structured radiology reports in German language[J]. Front Radiol, 2024, 4: 1390774.

[14] Cozzi A, Pinker K, Hidber A, et al. BI-RADS category assignments by GPT-3.5, GPT-4, and google bard: a multilanguage study[J]. Radiology, 2024, 311(1): e232133.

[15] Suzuki K, Yamada H, Yamazaki H, et al. Preliminary assessment of TNM classification performance for pancreatic cancer in Japanese radiology reports using GPT-4[J]. Jpn J Radiol, 2025, 43(1): 51-5.

[16] Chen K, Xu WG, Li XF. The potential of gemini and GPTs for structured report generation based on free-text 18F-FDG PET/CT breast cancer reports[J]. Acad Radiol, 2025, 32(2): 624-33.

[17] Jiang H, Xia SJ, Yang YX, et al. Transforming free-text radiology reports into structured reports using ChatGPT: a study on thyroid ultrasonography[J]. Eur J Radiol, 2024, 175: 111458.

[18] Watts E, Pournik O, Allington R, et al. Enhancing diagnostic precision: utilising a large language model to extract U scores from thyroid sonography reports[J]. Stud Health Technol Inform, 2025, 328: 56-60.

[19] Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors[J]. Eur Radiol, 2025, 35(4): 1938-47.

[20] Gu K, Lee JH, Shin J, et al. Using GPT-4 for LI-RADS feature extraction and categorization with multilingual free-text reports[J]. Liver Int, 2024, 44(7): 1578-87.

[21] Lee KL, Kessler DA, Caglic I, et al. Assessing the performance of ChatGPT and Bard/Gemini against radiologists for Prostate Imaging-Reporting and Data System classification based on prostate multiparametric MRI text reports[J]. Br J Radiol, 2025, 98(1167): 368-74.

[22] Singh R, Hamouda M, Chamberlin JH, et al. ChatGPT vs. Gemini: Comparative accuracy and efficiency in Lung-RADS score assignment from radiology reports[J]. Clin Imaging, 2025, 121: 110455.

[23] Hussain S, Naseem U, Ali M, et al. TECRR: a benchmark dataset of radiological reports for BI-RADS classification with machine learning, deep learning, and large language model baselines[J]. BMC Med Inform Decis Mak, 2024, 24(1): 310.

[24] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners[J/OL]. arXiv. http://arxiv.org/abs/2005.14165.

[25] Tay SB, Low GH, Wong GJE, et al. Use of natural language processing to infer sites of metastatic disease from radiology reports at scale[J]. JCO Clin Cancer Inform, 2024, 8: e2300122.

[26] Wang ZX, Zhang Z, Traverso A, et al. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach[J]. Quant Imaging Med Surg, 2024, 14(2): 1602-15.

[27] Liu J, Blanton T, Elazar Y, et al. OLMoTrace: Tracing language model outputs back to trillions of training tokens[J/OL]. arXiv. http://arxiv.org/abs/2504.07096.

[28] Kim S, Jang S, Kim B, et al. Automated pathologic TN classification prediction and rationale generation from lung cancer surgical pathology reports using a large language model fine-tuned with chain-of-thought: algorithm development and validation study[J]. JMIR Med Inform, 2024, 12: e67056.

[29] Cho H, Yoo S, Kim B, et al. Extracting lung cancer staging descriptors from pathology reports: a generative language model approach[J]. J Biomed Inform, 2024, 157: 104720.

[30] Hewitt KJ, Wiest IC, Carrero ZI, et al. Large language models as a diagnostic support tool in neuropathology[J]. J Pathol Clin Res, 2024, 10(6): e70009.

[31] Cheligeer K, Wu GS, Laws A, et al. Validation of large language models for detecting pathologic complete response in breast cancer

using population-based pathology reports[J]. BMC Med Inform Decis Mak, 2024, 24(1): 283.

[32] Yang XW, Zhang Y, Jiang JY, et al. Harnessing GPT-4 for automated error detection in pathology reports: Implications for oncology diagnostics[J]. Digit Health, 2025, 11: 20552076251346703.

[33] Pan YF, Tian S, Guo J, et al. Clinical feasibility of AI doctors: evaluating the replacement potential of large language models in outpatient settings for central nervous system tumors[J]. Int J Med Inform, 2025, 203: 106013.

[34] Liu CX, Wei MY, Qin Y, et al. Harnessing large language models for structured reporting in breast ultrasound: a comparative study of open AI (GPT-4.0) and microsoft Bing (GPT-4)[J]. Ultrasound Med Biol, 2024, 50(11): 1697-703.

[35] Xian MF, Lan WT, Zhang Z, et al. Enhancing hepatocellular carcinoma diagnosis in non-high-risk patients: a customized ChatGPT model integrating contrast-enhanced ultrasound[J]. Radiol Med, 2025, 130(7): 1013-23.

[36] Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases[J]. Neuroradiology, 2024, 66(1): 73-9.

[37] Chen LC, Zack T, Demirci A, et al. Assessing large language models for oncology data inference from radiology reports[J]. JCO Clin Cancer Inform, 2024, 8: e2400126.

[38] Laukamp KR, Terzis RA, Werner JM, et al. Monitoring patients with glioblastoma by using a large language model: accurate summarization of radiology reports with GPT-4[J]. Radiology, 2024, 312 (1): e232640.

[39] Cabezas E, Toro-Tobon D, Johnson T, et al. ChatGPT-4's accuracy in estimating thyroid nodule features and cancer risk from ultrasound images[J]. Endocr Pract, 2025, 31(6): 716-23.

[40] Ozenbas C, Engin D, Altinok T, et al. ChatGPT-4o's performance in brain tumor diagnosis and MRI findings: a comparative analysis with radiologists[J]. Acad Radiol, 2025, 32(6): 3608-17.

[41] Baker HP, Aggarwal S, Kalidoss S, et al. Diagnostic accuracy of ChatGPT-4 in orthopedic oncology: a comparative study with residents[J]. Knee, 2025, 55: 153-60.

[42] Chen ZM, Chambara N, Wu CQ, et al. Assessing the feasibility of ChatGPT-4o and Claude 3-Opus in thyroid nodule classification based on ultrasound images[J]. Endocrine, 2025, 87(3): 1041-9.

[43] Tekcan Sanli DE, Sanli AN, Yildirim D, et al. Can ChatGPT detect breast cancer on mammography [J]. J Med Screen, 2025, 32(3): 172-5.

[44] Wu SH, Tong WJ, Li MD, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models[J]. Radiology, 2024, 310(3): e232255.

[45] Guo LF, Zuo YT, Yisha Z, et al. Diagnostic performance of advanced large language models in cystoscopy: evidence from a retrospective study and clinical cases[J]. BMC Urol, 2025, 25(1): 64.

[46] Mao YQ, Xu N, Wu YN, et al. Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images[J]. Cell Rep Med, 2025, 6(3): 101988.

[47] Tozuka R, Johno H, Amakawa A, et al. Application of NotebookLM, a large language model with retrieval-augmented generation, for lung cancer staging[J]. Jpn J Radiol, 2025, 43(4): 706-12.

[48] Valerio AG, Trufanova K, de Benedictis S, et al. From segmentation to explanation: Generating textual reports from MRI with LLMs[J]. Comput Methods Programs Biomed, 2025, 270: 108922.

[49] Ahmad Abbasi A, Farooqi AH. Integrating CT image reconstruction, segmentation, and large language models for enhanced diagnostic insight[J]. Med Biol Eng Comput, 2025. doi: 10.1007/s11517-025-03446-3.

[50] Zhu LB, Rwigema JM, Feng X, et al. Improving the precision of deep-learning-based head and neck target auto-segmentation by leveraging radiology reports using a large language model[J]. Cancers (Basel), 2025, 17(12): 1935.

[51] Pradhan P. Accuracy of ChatGPT 3.5, 4.0, 4o and Gemini in diagnosing oral potentially malignant lesions based on clinical case reports and image recognition[J]. Med Oral Patol Oral Cir Bucal, 2025, 30(2): e224-31.

[52] Ding L, Fan L, Shen M, et al. Evaluating ChatGPT's diagnostic potential for pathology images[J]. Front Med: Lausanne, 2024, 11: 1507203.

[53] Vaira LA, Lechien JR, Maniaci A, et al. Diagnostic performance of ChatGPT-4o in analyzing oral mucosal lesions: a comparative study with experts[J]. Medicina, 2025, 61(8): 1379.

[54] Laohawetwanit T, Apornvirat S, Asaturova A, et al. Evaluation of general-purpose large language models as diagnostic support tools in cervical cytology[J]. Pathol Res Pract, 2025, 274: 156159.

[55] Hang H, Yang LK, Wang ZJ, et al. Comparative analysis of accuracy and completeness in standardized database generation for complex multilingual lung cancer pathological reports: large language model-based assisted diagnosis system vs. DeepSeek, GPT-3.5, and healthcare professionals with varied professional titles, with task load variation assessment among medical staff[J]. Front Med (Lausanne), 2025, 12: 1618858.

[56] Sng GGR, Xiang Y, Lim DYZ, et al. A multimodal large language model as an end-to-end classifier of thyroid nodule malignancy risk: usability study[J]. JMIR Form Res, 2025, 9: e70863.

[57] Pan C, Lu W, Chen BL, et al. Automated literature screening for hepatocellular carcinoma treatment through integration of 3 large language models: methodological study[J]. JMIR Med Inform, 2025, 13: e76252.

[58] Alber DA, Yang ZH, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks[J]. Nat Med, 2025, 31(2): 618-26.

[59] Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery[J]. NPJ Digit Med, 2021, 4(1): 93.

[60] Shayegani E, Mamun MAA, Fu Y, et al. Survey of vulnerabilities in large language models revealed by adversarial attacks[J/OL]. arXiv. https://arxiv.org/abs/2310.10844v1.

[61] Lenz S, Ustjanzew A, Jeray M, et al. Can open source large language models be used for tumor documentation in Germany-An evaluation on urological doctors' notes[J]. BioData Min, 2025, 18(1): 48.

（编辑：郎 朗）