国内大语言模型在烧伤辅助诊疗多跳推理任务中的性能评估与比较

张文一[1]，郭九宫[2]，郑金光[3]，王庆梅[4]，李　林[1]
[1]解放军总医院医学创新研究部，北京　100853；[2]国防大学联合勤务学院卫勤教研室，北京　100091；[3]解放军总医院第四医学中心烧伤整形医学部，北京　100048；[4]解放军总医院第一医学中心军队伤病员管理科，北京　100853

摘要：背景　烧伤救治要求迅速整合多维临床信息以支持准确决策，多跳推理技术在这一过程中扮演关键角色。**目的**　评估DeepSeek R1、DeepSeek V3、豆包和KiMi四种国内大语言模型在烧伤辅助诊疗多跳推理任务中的性能差异，为临床和急救环境中大模型的选型与优化提供理论依据。**方法**　从解放军总医院2023年1月—2025年2月出院的烧伤病例中随机抽取30例。统一输入患者基本信息、主诉、现病史、既往史、个人史及体格与辅助检查数据，通过四类模型确定疾病诊断。3名专家采用Likert 5分盲评对诊断结果的准确性进行评估。总体比较采用配伍组方差分析，亚组(问题字数、烧伤部位、面积、严重程度)采用Mann-Whitney U检验，并利用混合效应模型评估各大语言模型与亚组因素的交互作用。**结果**　专家一致性评分Cronbach’s Alpha达0.809。总体上，DeepSeek R1得分为(4.2±0.62)，显著高于DeepSeek V3(2.4±1.06)、豆包(3.2±1.31)和KiMi(1.6±0.86)($P<0.001$)。亚组分析显示，2000字以下和2000字及以上组、单部位烧伤和多部位烧伤组、10%以下烧伤和10%以上烧伤面积组、深II度以下和深II度以上烧伤程度组中DeepSeek R1均表现优异；混合效应模型表明，问题字数、烧伤部位数量与烧伤面积对模型得分存在显著交互效应(分别P=0.006、0.007、0.001)。**结论**　国内大模型在烧伤辅助诊疗多跳推理任务中存在显著性能差异，其中DeepSeek R1表现最佳，凸显了多跳推理技术在复杂临床信息整合与快速决策中的应用前景，为临床急救中大模型的优化提供了重要参考。

关键词：大语言模型；多跳推理；烧伤诊疗；人工智能；临床决策支持
中图分类号：R197；TP183　　**文献标志码：**A　　**文章编号：**2095-5227(2025)10-0988-06
DOI：10.12435/j.issn.2095-5227.25042102
引用本文：张文一，郭九宫，郑金光，等.国内大语言模型在烧伤辅助诊疗多跳推理任务中的性能评估与比较［J］.解放军医学院学报，2025，46（10）：988-993.

Performance evaluation and comparison of domestic large language models in multi-hop reasoning tasks for burn injury diagnosis and treatment assistance

ZHANG Wenyi[1], GUO Jiugong[2], ZHENG Jinguang[3], WANG Qingmei[4], LI Lin[1]
[1]Medical Innovation Research Department of PLA General Hospital, Beijing 100853, China; [2]Health Service Teaching and Research Section, College of Joint Service, National Defence University, Beijing 100091, China; [3]Burns and Plastic Surgery Medical Department, the Fourth Medical Centre of PLA General Hospital, Beijing 100048, China; [4]Military Patient Management Department, the First Medical Centre of PLA General Hospital, Beijing 100853, China
Corresponding author: LI Lin. Email: lilin@301hospital.com.cn

Abstract: Background　Burn care demands rapid integration of multidimensional clinical information to support accurate decision-making, and multi-hop reasoning plays a key role in this process. **Objective**　To evaluate the performance differences of four domestic large language models——DeepSeek R1, DeepSeek V3, DouBao, and KiMi——in multi-hop reasoning tasks for burn-assisted diagnosis and treatment, and provide theoretical reference for model selection and optimization in clinical and field emergency environments. **Methods**　A total of 30 burn cases were randomly selected from those discharged from Chinese PLA General Hospital from January 2023 and February 2025. Three burn-care experts performed a blind evaluation using a 5-point Likert scale to assess the accuracy of the diagnostic results. Overall comparisons were analyzed using randomized block ANOVA, subgroup analyses (question word count, burn site, area, severity) employed the Mann-Whitney U test, and mixed-effects models were used to assess the interaction between major language models and subgroup factors. **Results**　The experts' consensus score Cronbach's Alpha reached 0.809. DeepSeek R1 achieved a mean score of (4.2±0.62), significantly outperforming DeepSeek V3 (2.4±1.06), Doubao (3.2±1.31) and KiMi (1.6±0.86) ($P<0.001$). Subgroup analysis revealed DeepSeek-R1 consistently demonstrated superior performance metrics across all defined subpopulations: cases with word counts ≤2 000 versus ≥2 000, single-site versus multi-site

收稿日期：2025-04-21
基金项目：省部级课题
第一作者：张文一，硕士，研究实习员。Email: 895374163@qq.com
通信作者：李林，博士，主任医师。Email: lilin@301hospital.com.cn

burn injuries, total body surface area (TBSA) involvement ＜10% versus ≥10%, and burn severity below deep partial-thickness versus deep partial-thickness or greater. Mixed-effects modeling revealed significant interactions between model score and prompt length (*P*=0.006), number of burn sites (*P*=0.007), and burn area (*P*=0.001). **Conclusion** Significant performance differences exist among domestic large language models on multi-hop reasoning tasks for burn-care diagnostic support, with DeepSeek R1 demonstrating superior capability. These findings underscore the promise of multi-hop reasoning techniques for integrating complex clinical data and facilitating rapid decision-making, and they offer important guidance for optimizing large models in emergency burn-care settings.

Keywords: large language models; multi-hop reasoning; burn care diagnostics; artificial intelligence; clinical decision support

Cited as: Zhang WY, Guo JG, Zheng JG, et al. Performance evaluation and comparison of domestic large language models in multi-hop reasoning tasks for burn injury diagnosis and treatment assistance[J]. Acad J Chin PLA Med Sch, 2025, 46(10): 988-993.

烧伤是一种常见且复杂的创伤性疾病，其救治过程充满挑战，尤其在紧急救治环境中，医疗资源和时间均十分有限[1-2]。在这种条件下，医师需迅速整合多源信息(如病史、体格检查、辅助检查等)并进行多层次推理，从而做出准确的临床决策。多跳推理(multi-hop reasoning)是指大语言模型(large language models，LLMs)在不同上下文之间建立逻辑连接以得出问题或决策答案的能力，结合了分布在文档、知识库和其他资源中的多个支持事实、推论和上下文关系，需要跨越多个信息点(每一次的“跨越”称作一次“跳跃”)，做出需要综合分布式信息和超出直接陈述范围的级联推理的决策[3]。另外，LLMs因其在自动疾病分类和疾病名称预测方面的强大性能而备受关注[4-6]，如Zhou等[7]通过LLMs模型考察基于各种疾病类型、临床数据集和评估技术推断疾病诊断，Gu等[8]研究提示工程如何提高临床诊断的准确性和灵活性。近年来，LLMs在自然语言理解和推理方面展现出巨大潜力[9-10]，并已在医疗问答和影像诊断中得到应用[11-15]。然而，目前针对烧伤辅助诊疗中多跳推理任务的研究相对较少，并且基于LLMs的诊断辅助在实际临床工作中还需要考虑多样化的患者群体特征、临床使用场景以及病史、体格检查、专科检查、辅助检查等复杂数据资料进行验证。为此，本研究构建了基于烧伤病历的多跳推理任务评估数据集，系统比较了DeepSeek R1、DeepSeek V3、豆包和KiMi四种国内大模型在整合复杂临床信息，推理生成初步诊断报告中的性能差异，并探讨了问题字数、烧伤部位、面积和严重程度等变量对模型输出的影响，旨在为紧急救治中的智能决策支持提供理论依据和实践参考。

1 对象和方法

1.1 数据来源

随机选取解放军总医院2023年1月至2025年2月出院的30例烧伤患者的住院病历。纳入标准：主要诊断为ICD-10：T20-T32，烧伤后首次就诊即在该医院住院病历资料完整者。排除标准：年龄小于18岁、住院天数小于2 d。进行去隐私化、去诊断信息处理。本研究经解放军总医院医学伦理委员会审查批准(审查批号S2025-481-01)。

1.2 样本量计算

选取10个烧伤病历开展试评估，结果为，四个模型得分均值为1.9～4.5，标准差为0.5～1.5，组内相关系数为0.2。采用样本量计算软件PASS 2022，选择单因素重复测量设计的样本量计算模块，计算得到本研究所需最小样本量为6。考虑到亚组分析，根据问题字数、烧伤部位数量与烧伤面积各划分为两个亚组，为保证亚组样本量≥6，将实际总样本例数定为30。

1.3 任务设计

基于抽取的病历，本研究构建了多跳推理测试任务，要求大语言模型综合整合患者基本信息、主诉、现病史、既往史、个人史、体格检查、专科检查及辅助检查数据后，生成初步诊断，包括主要诊断和次要诊断。测试过程中采用预先设计并验证的统一提示词模板，提示词如下：您是一个专门从事疾病诊断的医疗助手，我将给您一份患者的入院记录，内容包括基本信息、主诉、现病史、既往史、个人史、婚育史、家族史、体格检查、辅助检查，请您生成初步诊断。于2025年3月1日至3月12日将病历分别输入DeepSeek R1、DeepSeek V3、豆包和KiMi 4种模型，以模拟实际急救和烧伤条件下的临床决策支持需求。其中DeepSeek R1、DeepSeek V3通过Cherry Studio应用软件在电脑上输入；豆包模型在网页版上输入，当时尚未增加深度思考功能；KiMi模型在网页版上输入，当时尚未增加长思考(1.5k)功能。

1.4 评价指标及评分办法

由于无法按照客观任务的评价方式对模型结

果进行正确与否的评判，邀请3名在大型医院工作10年以上的具有丰富烧伤救治经验的专家对各模型回答的初步诊断结果进行盲评。专家在统一培训后，使用Likert 5分量表对每个病历的回答进行评分，评分结果以均值(mean)和标准差(standard deviation，SD)表示。其中，初步诊断完全错误1分、主要诊断错误但次要诊断漏项2分、主要诊断错误但次要诊断正确3分、主要诊断正确但次要诊断漏项4分、初步诊断完全正确5分。为考察提问问题的字数、烧伤部位数量、烧伤面积、烧伤程度等因素对得分的影响，对所有问题按以上4个维度分别分类为亚组，其中问题字数根据提问问题的中位数，划分为2 000字以下和2 000字及以上两组；烧伤部位按单部位、多部位分为两组；烧伤面积按照病历数量分布，划分为10%以下、10%及以上两组；烧伤程度按照病历数量分布，划分为深II度以下、深II度及以上两组。

1.5　统计学分析

统计学分析采用SAS 9.4软件。模型总体得分差异采用配伍组方差分析进行评估。采用Mann-Whitney U检验对亚组的得分进行比较。同时，采用混合效应模型评估各亚组变量与模型得分间的交互效应。$P<0.05$为差异有显著性意义。

2　结果

2.1　总体比较

3位专家的一致性指数Cronbach’s Alpha[16]为0.809，表明评分一致性良好。4种大语言模型执行多跳推理任务生成烧伤病历初步诊断的平均得分，DeepSeek R1、DeepSeek V3、豆包和KiMi分别为4.2±0.62、2.4±1.06、3.2±1.31和1.6±0.86，得分的总体比较存在显著差异($P<0.001$)，DeepSeek得分显著高于其他模型。4个模型得分均值为2.9±0.63分，见表1。

2.2　亚组分析

按提问字数(通过函数公式统计)，在2 000字以下组中，DeepSeek R1、DeepSeek V3、豆包和KiMi的平均得分分别为4.2±0.72、2.3±0.99、2.9±1.18和2.0±1.0，总体比较差异显著($P<0.001$)。在2 000字及以上组中，各模型的得分分别为4.3±0.52、2.4±1.16、3.6±1.37和1.2±0.39，整体差异亦显著($P<0.001$)。各模型的组间比较中，DeepSeek R1与DeepSeek V3、豆包的差异未达到统计学显著水平，但KiMi在2 000字及以上组中的得分明显低于2 000字以上组，差异有统计学意义($P=0.003$)，见表1。

按烧伤部位数量，单部位组中，4个模型(顺序同上)平均得分分别为4.2±0.63、2.4±1.07、2.7±1.21和2.2±1.11，总体比较差异有统计学意义($P=0.003$)。多部位组中，4个模型的得分分别为4.3±0.63、2.3±1.08、3.5±1.32和1.3±0.58，总体比较差异有统计学意义($P<0.001$)。各模型的组间比较中，DeepSeek R1与DeepSeek V3、豆包之间无显著差异，但KiMi在单部位组的得分明显高于多部位组，差异有统计学意义($P=0.008$)，见表1。

按烧伤面积，10%以下组中，4个模型(顺序同上)平均得分分别为4.3±0.63、2.6±0.98、2.8±1.30和2.1±1.01，整体比较有统计学差异($P<0.001$)。10%及以上组中，各模型的得分分别为4.2±0.62、2.2±1.12、3.6±1.25及1.2±0.40，差异比较有统计学差异($P<0.001$)。各模型的组间比较中，DeepSeek R1、DeepSeek V3和豆包的差异不显著，但KiMi在10%及以上组的得分显著低于10%以下组，具有统计学差异($P=0.002$)，见表1。

按烧伤程度，在深II度以下组中，4个模型(顺序同上)平均得分分别为4.3±0.62、2.4±1.03、3.2±1.34和1.5±0.72，总体差异显著($P<0.001$)。在深II度及以上组中，各模型的得分为4.2±0.65、2.4±1.17、3.4±1.32和1.9±1.07，整体比较亦显示显著性($P=0.000\ 7$)。在组内比较中，各模型之间的差异均未达到统计学显著水平($P>0.05$)，见表1。

2.3　混合效应模型分析

通过混合效应模型进行亚组分析显示，烧伤严重程度指标呈现显著异质性效应，其中，烧伤面积10%及以上较10%以下、烧伤部位数量的多部位组较单部位组具有显著交互效应(P值分别为0.000 8、0.006 9)，提示烧伤面积≥10%且累及多部位的患者需作为重点干预对象。问题字数2 000字以上较2 000字以下具有显著交互效应($P=0.005\ 9$)，提示文本复杂度可能通过信息处理负荷间接影响评估结果。烧伤深度分层深II度及以上较深II度以下未显示显著交互效应($P=0.845\ 7$)，考虑因其与面积/部位存在共线性导致效应稀释。

3　讨论

经检索，越来越多的证据支持LLMs在诊断任务中的有效性，如极少的人工输入提高效率[17]、提高诊断准确性解决医疗资源不均问题[18]、患有

表 1 多跳推理任务各模型得分比较($\bar{x}$±s)

Tab. 1 Comparison of scores across models for multi-hop reasoning tasks ($\bar{x}$±s)

指标	DeepSeek R1	DeepSeek V3	豆包	KiMi	平均值	*P*值
总体(n=30)	4.2±0.62	2.4±1.06	3.2±1.31	1.6±0.86	2.9±0.63	<0.001[b]
字数						
少(n=15)	4.2±0.72	2.3±0.99	2.9±1.18	2.0±1.0	2.9±1.18	<0.001[b]
多(n=15)	4.3±0.52	2.4±1.16	3.6±1.37	1.2±0.39	2.9±1.37	<0.001[b]
各模型结果比较 *P*值	0.933	0.899	0.055	0.003[a]	1.000	
烧伤部位数量						
单部位(n=9)	4.2±0.63	2.4±1.0)	2.7±1.21	2.2±1.11	2.9±1.15	0.003[b]
多部位(n=21)	4.3±0.63	2.3±1.08	3.5±1.32	1.3±0.58	2.9±1.33	<0.001[b]
各模型结果比较 *P*值	0.679	0.627	0.152	0.008[a]	0.928	
烧伤面积						
10%以下(n=14)	4.3±0.63	2.6±0.98	2.8±1.30	2.1±1.01	3.0±1.19	<0.001[b]
10%及以上(n=16)	4.2±0.62	2.2±1.12	3.6±1.25	1.2±0.40	2.8±1.35	<0.001[b]
各模型结果比较 *P*值	0.472	0.203	0.122	0.002[a]	0.532	
烧伤程度						
深II度以下(n=20)	4.3±0.62	2.4±1.03	3.2±1.34	1.5±0.72	2.8±1.28	<0.001[b]
深II度及以上(n=10)	4.2±0.65	2.4±1.17	3.4±1.32	1.9±1.07	3.0±1.26	0.000 4[b]
各模型结果比较*P*值	0.911	0.857	0.825	0.348	0.708	

总体差异采用配伍组方差分析，亚组比较采用Mann-Whitney U检验。[a]$P<0.05$，*vs*对照组；[b]$P<0.05$，*vs*模型组。

多种合并症的老年人群的临床决策[19]等，但通常提供的是对不同临床应用的广泛概述[20]，而不是专门针对疾病诊断。目前尚未发现利用烧伤临床诊疗病历对LLMs的性能开展测评的文献报道。

本研究系统评估了国内4种LLMs在烧伤辅助诊疗多跳推理任务中的表现，结果显示模型间存在显著差异。总体来看，DeepSeek R1获得的评分最高，其次是豆包、DeepSeek V3和KiMi。亚组分析表明，提问问题字数、烧伤部位数量和烧伤面积等变量均对模型得分产生显著影响，对于字数少、单部位、面积小的病历，Kimi模型得分显著高。总体及各亚组的统计检验均显示差异显著。本研究表明，LLMs在烧伤诊断中的平均得分虽然不是很高(2.9分)，但却显示出在烧伤救治乃至医学诊疗辅助决策支持中的潜力，特别是在大模型迭代速度快的今天，其性能必将在不远的未来进一步提升[6]。

多跳推理可以帮助医师更快速、准确地诊断和治疗患者，在临床决策中扮演重要作用。然而在医疗领域中多跳推理面临的一个主要问题是医学知识的模糊性和医学推理的复杂性，模型需要能够处理包含多种逻辑运算的知识并进行复杂的逻辑推理[21-22]，这可能是LLMs总体得分不高的主要原因。说明LLMs在该任务中还需要开展大量训练和改进，以提高此项能力。

本研究显示，不同LLMs在整合多源临床信息、进行复杂逻辑推理的能力存在差异。DeepSeek R1表现优异可能归功于捕捉临床语言的细微差别，使其能够以高精度区分具有重叠症状特征的疾病[23]，其深度思考与推理能力强，以及在训练过程中的领域自适应预训练、充分整合多模态医学语料库、分层注意力机制(症状-病理-治疗三级推理)、对抗性数据生成(罕见烧伤类型)和逻辑思维链推理，使其在处理信息量大、结构复杂的病历时具有更高的稳定性和准确性，尤其是在整体诊断准确性和置信水平方面表现更优。相反，DeepSeek V3和KiMi在面对多维临床信息时表现较弱，提示其模型架构或训练数据在专业性和复杂推理上的不足。有学者研究表明，大型预训练LLMs在上下文理解和推理能力上表现突出，但其在特定领域的应用效果仍然依赖于训练数据的专业性[4]。部分LLMs强调计算效率和可扩展性，使其特别适合于高容量或资源受限的临床环境中的快速诊断支持[24]。

研究表明，提问问题字数、烧伤部位数量和面积大小均会影响模型得分。其原因是，这些因素反映了输入信息的量级和复杂度[1-2]，在不同程度上影响模型的信息整合能力和推理性能。其中，问题字数反映了输入文本的信息量和复杂性，较长的提问可能提供更全面的背景信息和细节，从

而有助于模型进行更精确的推理；但同时，如果字数过多，则可能包含冗余或噪声信息，使得模型在注意力分配和信息整合时遇到困难。此外，较长的文本也可能超出模型的上下文处理能力，从而影响模型得分。同样，烧伤部位数量多、烧伤面积大也会增加临床信息的复杂度。单一部位的病历通常信息较为集中，而多部位、大面积烧伤往往伴随着更复杂的生理变化和并发症风险，要求模型整合和处理更多信息，从而可能导致得分下降。此外，多部位、大面积烧伤的信息冗余和交叉干扰也会加大模型的推理难度。而只有KiMi得分受问题字数、烧伤面积和部位数量的影响，其他三种模型并不受影响，说明其他LLMs的性能相对KiMi更加稳定。这些因素需要在实际临床辅助决策系统中予以充分考虑，以便在模型优化和选择过程中，更加针对性地提高多跳推理任务的准确性和稳定性。

本研究结果为急救中LLMs的应用提供了量化依据。针对多跳推理任务，优选性能稳定、能高效整合复杂信息的模型有助于提升紧急救治的决策效率和准确性。未来工作应在扩大样本量、引入客观评价指标的基础上，在模拟和实际环境中开展验证，同时针对不同临床变量开展模型优化，以构建更高效、可靠的智能临床决策支持系统。

本研究提示，LLMs在医疗领域的应用，需要结合专门知识库进行定制化训练，开展微调，使模型深入掌握垂直领域(包括烧伤救治领域)的专业知识，定制其输出能力；或者开展本地化部署，为其提供无法在公共网络中学习到的知识，提升其性能[25]。同时，另一条重要途径是引入临床医学专业大模型，包括医疗影像辅助诊断、中医药辅助诊疗、智能导诊、健康管理、医疗咨询、医疗资源配置，以及专病大模型等[26]。

本研究强调了LLMs引入临床救治时开展评估的重要性。LLMs有其固有缺陷，包括幻觉问题，如会生成看似合理但实际错误的内容；时效性问题，其训练数据通常截至于特定时间点，因此无法处理训练后发生的事件或更新的信息；其预训练通常是在通用数据上开展的，因此在对于某些垂直领域(如医疗)的专业知识和特定语境可能并不充分[27]，这导致了其存在医学知识不精准、回答逻辑不严谨等问题，回答正确率并不很高。不仅是通用模型需要评估，专业模型的性能也需要评估。在这方面还有很多工作要做，比如，首先要构建测评数据集。在通用领域包括医学领域已经建立了不少测评数据集，在模型性能测评中发挥了重要作用[28-29]，而在烧伤救治等专业领域的LLMs测评数据集尚未有报道。其次，测评方法和指标需要标准化，在当前LLMs不断涌现的情况下，模型性能的通用测评和专业测评对于社会公众和专业人员更加全面客观准确透明地掌握其对不同任务的性能，是正确开展LLMs应用的前提。这需要有关部门尽快建立相关规范性文件，对LLMs的应用加以引导和管理。

本研究存在一定局限性。首先，研究选取的LLMs数量少，还有很多没有涵盖其中；其次，样本量仅为30例，可能影响结果的外推性；再次，虽然专家评分的内部一致性较高，但评分中仍存在一定主观因素；最后，所使用的病历均来自医院救治，而非现场急救记录，实际条件可能更为复杂严苛。因此，研究结果在不同救治环境中的适用性仍需进一步验证。

综上，本研究首次对国内4种大语言模型在烧伤辅助诊疗多跳推理任务中的性能进行了比较，结果显示DeepSeek R1表现出显著优势，其在整合多源临床数据、生成初步诊断报告方面得分最高；而DeepSeek V3及KiMi在某些复杂任务中的准确性不足。不同亚组变量(问题字数、烧伤部位数量及面积)对个别LLMs的性能存在显著影响，为后续模型针对性优化提供了量化依据。本研究不仅为紧急救治中的临床决策支持提供了理论参考，也为构建高效、可靠的智能医疗辅助系统提供了有力支持。

作者贡献　张文一：模型盲评、论文撰写；李林：模型基于烧伤数据集生成答案、论文审读和修订，监督指导；郭九宫、郑金光：模型盲评；王庆梅：烧伤数据集设计。

利益冲突　所有作者声明无利益冲突。

数据共享声明　本论文相关数据可依据合理理由从作者处获取，Email：895374163@qq.com。

参考文献

1　Herndon DN. Total Burn Care［M］. 5th ed. Edunburgh：Elsevier，2018.
2　王正国. 外科学与野战外科学［M］. 北京：人民军医出版社，2007.
3　Yang S，Gribovskaya E，Kassner N，et al. Do large language models latently perform multi-hop reasoning?［EB/OL］. https：//arxiv.org/abs/2402.16837
4　Gupta GK，Pande P，Acharya N，et al. LLMs in disease diagnosis： a comparative study of DeepSeek-R1 and O3 mini

across chronic health conditions［EB/OL］. https：//arxiv. org/abs/2503.10486
5 Ríos-Hoyo A，Shan NL，Li AR，et al. Evaluation of large language models as a diagnostic aid for complex medical cases［J］. Front Med，2024，11：1380148.
6 Gupta GK，Singh A，Manikandan SV，et al. Digital diagnostics：the potential of large language models in recognizing symptoms of common illnesses［J］. AI，2025，6（1）：13.
7 Zhou S，Xu ZD，Zhang M，et al. Large language models for disease diagnosis：a scoping review［J］. NPJ Artif Intell，2025，1：9.
8 Gu BW，Desai RJ，Lin KJ，et al. Probabilistic medical predictions of large language models［J］. NPJ Digit Med，2024，7：367.
9 Bajwa J，Munir U，Nori A，et al. Artificial intelligence in healthcare：transforming the practice of medicine［J］. Future Healthc J，2021，8（2）：e188-e194.
10 Brown TB，Mann B，Ryder N，et al. Language models are few-shot learners［EB/OL］. https：//arxiv.org/abs/2005.14165
11 Esteva A，Robicquet A，Ramsundar B，et al. A guide to deep learning in healthcare［J］. Nat Med，2019，25（1）：24-29.
12 Topol EJ. High-performance medicine：the convergence of human and artificial intelligence［J］. Nat Med，2019，25（1）：44-56.
13 Xiong YT，Zhan ZZ，Zhong CL，et al. Evaluating the performance of large language models（LLMs）in answering and analysing the Chinese dental licensing examination［J］. Eur J Dent Educ，2025，29（2）：332-340.
14 马武仁，弓孟春，戴辉，等. 以ChatGPT为代表的大语言模型在临床医学中的应用综述［J］. 医学信息学杂志，2023，44（7）：9-17.
15 Wang HY，Wu WZ，Dou Z，et al. Performance and exploration of ChatGPT in medical examination，records and education in Chinese：Pave the way for medical AI［J］. Int J Med Inform，2023，177：105173.
16 Tavakol M，Dennick R. Making sense of cronbach’s alpha［J］. Int J Med Educ，2011，2：53-55.
17 Liu Y，Jain A，Eng C，et al. A deep learning system for differential diagnosis of skin diseases［J］. Nat Med，2020，26（6）：900-908.
18 Mei XY，Lee HC，Diao KY，et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19［J］. Nat Med，2020，26（8）：1224-1228.
19 Qiu SR，Joshi PS，Miller MI，et al. Development and validation of an interpretable deep learning framework for Alzheimer’s disease classification［J］. Brain，2020，143（6）：1920-1933.
20 Almubark I. Exploring the impact of large language models on disease diagnosis［J］. IEEE Access，2025，13：8225-8238.
21 徐寅鑫，杨宗保，林宇晨，等. 基于知识图谱和预训练语言模型深度融合的可解释生物医学推理［J］. 北京大学学报（自然科学版），2024，60（1）：62-70.
22 王昕瑞，陈涛，孙涛. 基于文本知识的多跳推理综述［J］. 计算机应用，2024，44（S1）：1-10.
23 DeepSeek-AI，Guo DY，Yang DJ，et al. DeepSeek-R1：incentivizing reasoning capability in LLMs via reinforcement learning［EB/OL］. https：//arxiv.org/abs/2501.12948
24 Ballon M，Algaba A，Ginis V. The relationship between reasoning and performance in large language models：o3（mini）thinks harder，not longer［EB/OL］. https：//arxiv.org/abs/2502.15631
25 Zhang C，Deng Y，Lin X，et al. 100 days after DeepSeek-R1：a survey on replication studies and more directions for reasoning language models［EB/OL］. https：//arxiv. org/abs/2505.00551
26 郑琰莉，韩福海，李舒玉，等. 人工智能大模型在医疗领域的应用现状与前景展望［J］. 医学信息学杂志，2024，45（6）：24-29.
27 Huang L，Yu WJ，Ma WT，et al. A survey on hallucination in large language models：principles，taxonomy，challenges，and open questions［J］. ACM Trans Inf Syst，2025，43（2）：1-55.
28 Bahak H，Taheri F，Zojaji Z，et al. Evaluating ChatGPT as a question answering system：a comprehensive analysis and comparison with existing models［EB/OL］. https：//arxiv. org/abs/2312.07592
29 Laskar MTR，Bari MS，Rahman M，et al. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets［EB/OL］. https：//arxiv.org/abs/2305.18486

（责任编辑：施晓亚，潘越）