文章编号:1671-7554(2026)02-0104-07 DOI:10.6040/j.issn.1671-7554.0.2025.1327

大型语言模型在骨科手术术前管理中的决策性能及辅助价值

魏书生,吴海波,李松林,温镇璘,杨昌骜,卢群山,刘培来
(山东大学齐鲁医院骨科,山东 济南 250012)

摘要:目的 探讨大型语言模型(如DeepSeek、ChatGPT等)的不同生成模式在术前管理领域的应用效果及对低年资医生的辅助决策价值。**方法** 随机选取2025年1月至2025年8月山东大学齐鲁医院医院骨科住院患者100例病历,排除预计施行一级、二级、三级手术及非关节置换手术患者,最终纳入患者87例。在PubMed和UpToDate数据库检索围术期管理相关指南,将检索到的指南经文本处理和向量化后,构建围术期管理知识库,为后续模型调用与问答提供外部知识支持。患者病历匿名化处理后上传到DeepSeek模型不同版本[DeepSeek Chat版本(V3版本)、DeepSeek Chat+知识库版本、DeepSeek深度思考版本(R1版本)及DeepSeek R1+知识库版本]中,以相同的"指令-上下文-输入-输出(Instruction-Context-Input-Output, ICIO)"提示词框架提问,对模型输出的结果进行客观与主观评估。**结果** DeepSeek R1模型在术前改良心脏风险指数(revised cardiac risk index, RCRI)评分与风险分级任务中的正确率分别为75.86%和78.16%,显著优于Chat系列模型。4个版本模型在美国麻醉医师协会身体状况分级系统(American society of anesthesiologists ,ASA)评分与手术可行性判断中的正确率均处于中等水平,其中R1版本表现略优。知识库的引入仅在Chat版本中对RCRI评分准确率有轻微提升(+4.6%),但在R1版本中反而降低性能。主观评估结果显示,低年资医生普遍认为R1系列模型回答更具临床参考价值,其平均评分(4.19±0.72)显著高于Chat系列(Chat版本为3.06±0.06,Chat+知识库版本为2.97±0.03),提示R1模型在术前决策支持中具有更强的实用性与可接受性($P<0.05$)。**结论** DeepSeek R1模型在骨科术前麻醉风险评估与临床辅助决策中展现出良好的应用潜力,但知识库构建及任务适配仍需进一步优化,以提升模型在真实临床场景下的可靠性与可推广性。

关键词:大语言模型;DeepSeek;术前决策;知识库;改良心脏风险指数评分
中图分类号:R684 **文献标志码:**A

Decision performance and auxiliary value of large language models in preoperative management of orthopedic surgery

WEI Shusheng, WU Haibo, LI Songlin, WEN Zhenlin, YANG Chang'ao, LU Qunshan, LIU Peilai
(Department of Orthopedics, Qilu Hospital of Shandong University, Jinan 250012, shandong, China)

Abstract: Objective To explore the application effectiveness of different generation modes of large language models (such as DeepSeek, ChatGPT, etc.) in the field of preoperative management and their value in assisting decision-making processes for junior physicians. **Methods** A total of 100 medical history records of orthopedic inpatients at Qilu Hospital of Shandong University were randomly selected from January to August 2025. Patients who were scheduled to undergo Grade I, II, III surgeries and non-joint replacement surgeries were excluded, resulting in the inclusion of total 87 patients. Guidelines related to perioperative management were retrieved from databases such as PubMed and UpToDate. After text processing and vectorization, these guidelines were used to build a perioperative management knowledge base, providing external knowledge support for subsequent model calls and question-answering tasks. The anonymized patient records were uploaded to different versions of the DeepSeek model [DeepSeek Chat version (V3), DeepSeek Chat + knowledge base version, DeepSeek Deep Thinking version (R1), and DeepSeek R1 + knowledge

收稿日期:2025-11-17
通信作者:卢群山。E-mail:201462008053@email.sdu.edu.cn
刘培来。E-mail:199362000205@email.sdu.edu.cn

base version], and questions were posed under the identical "Instruction-Context-Input-Output (ICIO)" prompt framework. The model outputs were evaluated both objectively and subjectively. **Results** The DeepSeek R1 model achieved accuracy rates of 75.86% and 78.16% in the Revised Cardiac Risk Index (RCRI) scoring and risk classification tasks, respectively, significantly outperforming the Chat series models. All four model versions showed moderate accuracy in the American Society of Anesthesiologists (ASA) physical status classification and surgical feasibility judgment, with the R1 version performing slightly better. The introduction of the knowledge base slightly improved RCRI scoring accuracy only in the Chat version (+4.6%) but reduced performance in the R1 version. Subjective evaluation results indicated that junior physicians generally considered the R1 series models' answers to be of greater clinical reference value, with an average score (4.19±0.72) significantly higher than that of the Chat series (Chat version: 3.06±0.06; Chat + knowledge base version: 2.97±0.03). This suggested that the R1 model has stronger practicality and acceptability in preoperative decision support ($P<0.05$). **Conclusion** The DeepSeek R1 model demonstrates good application potential in orthopedic preoperative anesthesia risk assessment and clinical decision support. However, knowledge base building and task adaptation require further optimization to enhance the model's reliability and generalizability in real clinical scenarios.

Key words: Large language model; DeepSeek; Preoperative decision-making; Knowledge base; Revised cardiac risk index score

术前决策是确保手术安全的关键步骤[1-2]。通过全面评估患者的健康状况,可以预测可能出现的麻醉及手术并发症。此外,评估患者的年龄、体质量、既往病史和目前的药物使用情况,这些信息有助于麻醉医生选择最合适的麻醉方式和药物剂量[3-4]。然而,低年资医生常常因经验不足而面临临床决策压力,无法及时、准确地判断患者是否能够耐受手术,不仅会延长患者的术前等待时间,还可能导致患者病情加重。

大语言模型(large language models, LLMs)是一种基于深度学习的自然语言处理技术。它通过对海量数据训练,能够掌握语言规律,擅长文本的生成、理解和推理[5-7]。近年来,DeepSeek、ChatGPT 等大语言模型在医学领域应用广泛,其在执业医师考题或专科问题测试、协助临床医生进行诊治、术前规划、患者教育及医学生教育等方面展现出了巨大潜力[8-12]。然而,DeepSeek 的不同生成模式在骨科术前决策领域中的应用效果及对低年资医生辅助决策的价值尚不明确。

本研究通过收集某三甲医院近期骨科真实住院患者的病历资料,对其进行匿名化处理后将其上传至 DeepSeek 的不同生成模式中,从主观和客观角度评估其回答效果,并分析其在辅助低年资医生进行麻醉决策方面的价值。

1 资料与方法

1.1 研究资料

1.1.1 患者来源

随机选取 2025 年 1 月至 2025 年 8 月山东大学齐鲁医院住院病历检索系统里诊断为膝关节骨关节炎或股骨头坏死的住院患者 100 例,根据纳入标准与排除标准,最终纳入患者 87 例,其中男 28 例,女 59 例,59 岁(19~82 岁)。本研究属于真实世界临床病历驱动的医学大语言模型性能评估研究,通过病历回顾进行,免除患者知情同意要求。本研究已获得山东大学齐鲁医院医学伦理委员会批注(批号:KYLL-202502-060)

1.1.2 纳入标准与排除标准

纳入标准:入院主诊断为膝关节骨关节炎或股骨头坏死的患者。排除标准:①排除施行一级、二级及三级手术患者;②排除非关节置换手术患者。

1.1.3 研究设计

本研究采用杭州深度求索人工智能基础技术研究有限公司发布的 DeepSeek V3 和 R1 版本的 DeepSeek 模型(版本号为 DeepSeek-V3.2-Exp);采用苏州语灵人工智能科技公司开发的开源大语言模型应用开发平台 Dify(https://dify.ai/zh,版本号为 Version 1.8.1)建立不同版本的工作流。本研究Dify工作流能够同时处理多份患者资料,每个模型对每个病历回答 1 次,且每次回答互不影响。研究人员将患者资料经匿名化处理后上传至 DeepSeek 进行测试,内容包括主诉、现病史、既往史、体格检查、生命体征、辅助检查及入院诊断等信息。随后,根据不同版本给出的回答结果,对模型回答术前决策相关问题的情况进行客观评估。主观评估由数名低年资医生在盲法的情况下,对模型的回答结果进行评估,以判断回答结果是否对患者的术前决策具有帮助。

1.2 方法

1.2.1 知识库构建

在PubMed和UpToDate数据库检索围术期管理相关指南22个，内容包括围术期心肺功能评估、糖尿病和高血压等基础疾病的围术期管理、围术期用药指南及围术期过敏反应处理等[13-16]。指南以Word文件形式上传至Dify平台，并利用其平台的知识库创建功能构建知识库。在创建知识库过程中，利用嵌入式模型处理文档以实现更精确的检索。

1.2.2 模型配置与工作流设计

本研究使用DeepSeek 4种不同模式进行测试，包括DeepSeek Chat版本（V3版本）、DeepSeek Chat+知识库版本、DeepSeek深度思考版本（R1版本）及DeepSeek R1+知识库版本。工作流是指通过设计好的任务执行流程，将复杂的自然语言处理任务分解为可执行的步骤，并协调模型与其他系统或工具的交互。其核心在于将人工设计的流程自动化，以提高任务处理效率和准确性[17-18]。本研究所有模型的工作流均使用Dify平台完成搭建，不搭载知识库模型的工作流为：输入-大语言模型-直接输出；而搭载知识库模型的工作流为：输入-知识库检索-大语言模型-直接输出。

1.2.3 提示模板

采用统一的“指令-上下文-输入-输出（Instruction-Context-Input-Output，ICIO）”提示词框架模板：你是一名骨科主治医师，现需对拟行关节置换术的患者进行术前评估。你需结合上传的患者病历，完成并输出以下任务：①美国麻醉医师协会身体状况分级系统（American society of anesthesiologists，ASA）评分。依据患者整体健康状况进行ASA分级并说明理由。②改良心脏风险指数（Revised cardiac risk index，RCRI）评分。根据病历内容进行RCRI评分并进行风险分级。③手术耐受性判断。明确患者当前是否可耐受手术，或需要推迟手术并列出调整目标（如感染控制、血糖/血压管理）。

1.2.4 客观评估

从DeepSeek模型不同版本的回答结果中提取出ASA评分、RCRI评分、风险分级以及患者是否手术，并与患者真实的术前ASA评分、RCRI评分、风险分级及是否手术信息对比（真实的术前评估信息来自患者病历中的术前“麻醉评估单”）。只有当模型回答结果与真实的术前评估一致时，才被认为正确[19-20]。其中，ASA评分是美国麻醉医师协会制定的用于评估患者术前健康状况及手术风险的分级系统。该根据患者体质状况和对手术危险性将患者分为六级：Ⅰ级代表健康个体；Ⅱ级代表有轻微全身性疾病的患者；Ⅲ级代表患有严重全身性疾病的患者；Ⅳ级代表患有严重全身性疾病且威胁生命的患者；Ⅴ级代表患者麻醉危险性极高，围术期死亡率较高的患者；Ⅵ级代表被宣布脑死亡的患者，其器官可被摘取捐献[21]。RCRI评分是用于评估非心脏手术患者围术期发生主要心血管并发症风险的临床工具，其包含6项独立危险因素：高危手术、缺血性心脏病史、慢性心力衰竭史、脑血管病史、胰岛素依赖型糖尿病以及术前血肌酐>2.0 mg/dl（176.8 μmol/L），每项1分，总分为0~6分。0~1分为低危风险分级；2分为中危风险分级；≥3分为高危风险分级[22]。最后由高年资副主任医师对模型进行评估，以反映其在骨科术前决策中的应用效果。

1.2.5 主观评估

主观评估由3名低年资医生在盲法的情况下，对模型的回答结果进行评估，医生能同时收到4个DeepSeek模型不同版本对同一问题的回答，但他们事先并不知道每个回答是由哪个版本生成。评估模型的临床参考价值，即医生根据患者的病历及模型回答结果判断其对术前决策是否有帮助[23-24]。其中，帮助性指标采用李克特量表对其进行量化[25]，分为1~5级：1级为完全无帮助，回答完全无关或错误，或未提供任何有用信息；2级为不太有帮助，回答部分相关但未解决核心问题，或提供的信息实用性较低；3级为一般有帮助，回答基本覆盖问题，但缺乏深度或未能完全解决问题；4级为有帮助，回答准确、清晰，有效解决用户的核心问题；5级为非常有帮助，回答超出预期，提供额外实用信息或创造性解决方案，显著提升问题解决效果。

1.3 统计学处理

采用SPSS 27.0.1和GraphPad Prism 10统计学软件。采用Shapiro-Wilk评估连续变量的正态分布。若符合正态分布，则采用单因素方差分析比较各组间差异，计量资料以 $\bar{x}\pm s$ 表示。若连续变量不符合正态分布，则采用Kruskal-Wallis比较不同模型版本间的差异。当方差分析显示统计学显著性时，进行事后多重比较。计数资料以 n（%）表示。采用 χ^2 检验或Fisher精确检验分析组间差异。检验水准 $\alpha=0.05$。

2 结 果

2.1 患者临床资料比较

纳入患者 87 例,19~82 岁,体质量指数 26.3±3.2。72 例(82.8%)患者接受手术干预,15 例(17.2%)未采取手术治疗;手术患者中,39 例(54.2%)行全膝关节置换术,33 例(45.8%)行全髋关节置换术。

2.2 RCRI 及其风险分级正确率

DeepSeek R1 版本在患者术前 RCRI 评分和风险分级任务中的正确率分别为 75.86%和 78.16%,显著优于 DeepSeek Chat(52.87%)和 DeepSeek Chat+知识库(57.47%)版本,差异有统计学意义(P<0.05)。DeepSeek Chat+知识库版本在回答 RCRI 评分的正确率比 DeepSeek Chat 版本高 4.6%,表明知识库对于 DeepSeek Chat 版本有轻微提升作用,但抑制了 DeepSeek R1 版本的性能,其添加知识库后 RCRI 评分的准确率降低 2.3%。同时,不同版本 RCRI 评分和风险分级同时回答正确的比例也符合上述规律。所有模型在风险分级的表现均优于评分计算。见图 1。

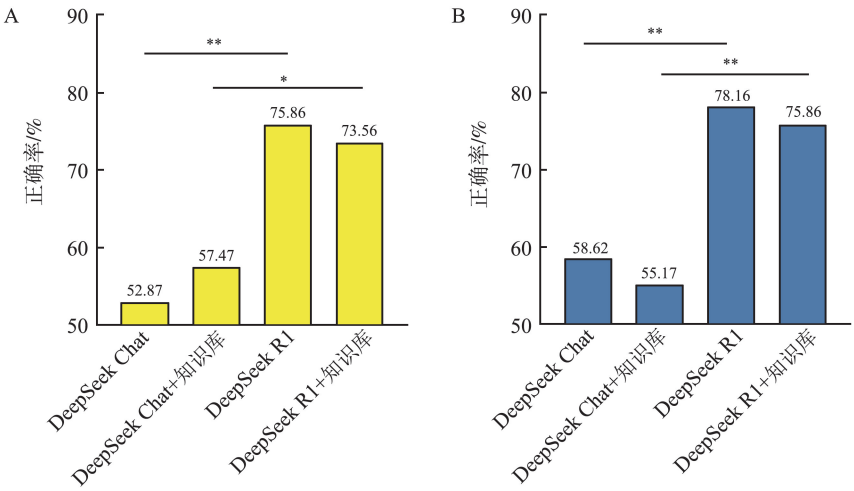

图 1 DeepSeek 模型不同版本客观评估结果(*P<0.05;**P<0.01)
A:RCRI 评分正确率;B:RCRI 风险分级判断正确率对比。

Figure 1 Objective evaluation results of different versions of the DeepSeek model (*P<0.05;**P<0.01)
A: Accuracy rate of RCRI score; B: Comparison of accuracy rate of RCRI risk classification judgement.

2.3 ASA 评分正确率

通过对 DeepSeek 不同版本回答 ASA 评分问题的正确率分析,DeepSeek Chat 回答 ASA 评分问题的正确率为 55.17%,DeepSeek Chat+知识库版本为 52.87%,DeepSeek R1 版本为 60.92%,DeepSeek R1+知识库版本为 59.77%,差异无统计学意义(P>0.05),DeepSeek R1 版本及其结合知识库的配置表现略优,但整体正确率仍处于中等水平。

2.4 手术判断正确率

DeepSeek 模型在判断患者手术可行性时现显著版本差异:DeepSeek R1 版本(62.07%)的正确率显著高于 DeepSeek Chat 版本(54.02%),而知识库增强后,DeepSeek R1+知识库(57.47%)和 DeepSeek Chat+知识库(49.43%)版本性能反而下降,差异无统计学意义(P>0.05)。

2.5 临床参考价值评估

R1 系列版本(DeepSeek R1 版本平均分为 4.19±0.72,DeepSeek R1+知识库版本平均分为 4.11±0.73),评分显著高于 Chat 系列版本(DeepSeek Chat 版本平均分为 3.06±0.06,DeepSeek Chat+知识库版本平均分为 2.97±0.03),且 DeepSeek Chat+知识库版本评分略低于 DeepSeek Chat 版本,DeepSeek R1+知识库版本与 DeepSeek R1 版本评分接近。见表 1。

表 1 不同医生对 DeepSeek 模型不同版本回答所评估的平均分及整体平均分

Table 1 The average and overall average scores of the evaluations given by different doctors for different versions of the DeepSeek

系列版本	医生 1	医生 2	医生 3	平均得分
DeepSeek Chat	3.11±1.44	3.00±0.42	3.06±0.64	3.06±0.06
DeepSeek Chat+知识库	2.98±1.38	2.94±0.24	3.00±0.59	2.97±0.03
DeepSeek R1	3.41±1.46	4.83±0.38	4.33±0.77	4.19±0.72
DeepSeek R1+知识库	3.38±1.50	4.83±0.38	4.11±0.68	4.11±0.73
F	—	—	—	11.08
P	—	—	—	<0.05

3 讨 论

近年来，大语言模型在已在医学领域广泛应用，其在临床诊疗、患者教育及医学教育等各个方面都展现出了巨大潜力。然而，目前尚未有研究探索大语言模型（如DeepSeek）的不同生成模式在骨科专科疾病（如膝关节骨关节炎、股骨头坏死等）术前评估领域中的应用效果及其辅助低年资医生进行术前评估方面的价值。

DeepSeek R1版本在RCRI评分及风险分级中的表现显著优于DeepSeek Chat系列版本。这一结果印证了模型设计定位与任务需求的匹配度：专注于深度推理的R1模型在解析临床数据和遵循评分规则方面更具优势。而Chat系列模型作为聊天机器人，则更适用于广泛场景的自然语言处理任务。因此DeepSeek R1在需要严格遵循评分标准的RCRI评估中展现出明显优势。研究表明，与任务非直接相关的知识库会干扰大语言模型推理并降低性能[26-27]，本研究结果与此结论相印证。本研究结果显示，知识库对不同版本模型性能的影响有所不同，其对DeepSeek Chat版本性能有轻微提升（4.6%），但对DeepSeek R1版本性能有抑制作用（准确率降低2.3%），提示知识库的筛选与适配需紧密贴合模型特性以及具体的临床问题，而非简单的将数个相关文档叠加。值得注意的是，所有模型在风险分级中的表现均优于评分计算，这一结果表明当前大语言模型在临床数值精确推理中存在固有短板。其原因在于评分计算过程涉及多维度指标的量化整合，而风险分级更侧重于定性判断，与模型的自然语言理解能力更匹配。

Peng等[28-29]研究显示，大语言模型在专科化和标准化评分系统中，易因对专业术语的理解偏差而导致性能受限，本研究ASA评分结果进一步验证了这一观点。本研究结果显示，即使是表现最优的DeepSeek R1+知识库版本，其性能也未达到理想水平，且知识库的加入并未显著提升正确率。由此可见，临床评分类任务对知识库的精准度要求极高，当知识库中的专科评分标准占比过低时，冗余信息会增加模型的信息处理负荷，从而无法提升性能。

本研究结果显示，手术可行性判断的结果延续了模型版本间的性能差异，DeepSeek R1版本的正确率显著高于DeepSeek Chat版本，证实了高级推理型模型在复杂临床决策辅助中的优势。而知识库增强后两个版本的性能均出现了下降，进一步验证了无关或低关联知识对模型推理的干扰效应。提示在手术可行性这类需要综合考虑患者基础状况、麻醉风险与手术指征的复杂判断中，模型更依赖于自身的推理框架，而不是外部知识库的简单补充，过度冗余的信息反而可能导致决策逻辑的错误。

Tordjman等[30]认为，经过强化学习优化的模型在临床推理逻辑上更加严谨，其诊断推理步骤的评分显著高于传统的聊天类模型。本研究结果显示，人为评估结果从临床使用者的视角印证了不同版本模型性能的差异。由此可见，低年资的关节外科主治医师对R1系列版本的帮助性评分显著高于Chat系列版本，而DeepSeek Chat+知识库版本评分略低于DeepSeek Chat版本。这一结果表明，冗余的知识库不仅没有提升模型的实际辅助价值，还可能因回答质量的下降而影响临床医生的使用体验。此外，不同评估者的评分也存在差异，有研究表明，不同临床工作者对AI辅助工具的认知与需求存在差异：部分医生认可AI的辅助作用，部分则持中立态度，这种差异可能与医生的临床经验、对AI技术的熟悉程度以及工作场景的需求有关。由此提示，未来在开发AI工具时，应更注重个性化与临床的适配性[31]。

本研究尚存在一定的局限性：①样本量局限于单中心87例骨科4级手术患者，病例类型与地域分布的单一性可能影响结果的泛化性，在未来可开展多中心、多手术类型的扩展研究，以验证不同模型在不同场景下的适用性；②同一病历仅采用单次模型回答，未通过多轮问答与结果集成削弱随机性对准确率比较的影响，在未来对同一病历可采用“多轮问答+结果集成”的方式来削弱单次输出的随机性；③评估场景仅覆盖四类核心问题，未涉及麻醉方案选择、并发症和手术风险预测等其他关键场景，导致模型辅助价值评估不够全面，在未来可将以上这些评估方法加入到研究中，以全面评估模型对临床的辅助价值；④客观评估中的金标准依赖低年资麻醉医生的评估结果，未采用多位高年资麻醉医生的共识结果，可能影响评估准确性，应该考虑让多位高年资麻醉医生进行评分，以此作为客观评估的“金标准”；⑤主观评估仅纳入低年资骨科医师，未涉及高年资医生及麻醉科专业医师，评估视角相对单一，未来需扩大主观评估人群范围，提升结果的客观性与代表性；⑥未明确知识库中负面影响模型性能的具体信息类型，无法为知识库优化提供针对性指导，未来需针对知识库优化的具体方向展开研究，以提升模型的性能。

本研究基于真实临床病例,系统评估了不同版本的 DeepSeek 大型语言模型在骨科术前决策评估中的表现。核心结论为 DeepSeek R1 模型在 RCRI 评分、风险分级及手术可行性判断中表现最优,临床推理能力显著优于 Chat 模型,且低年资医生主观认可 DeepSeek R1 模型辅助价值,提示其具有潜在的临床辅助潜力。这一结果为大语言模型在临床决策支持系统中的优化应用提供了新参考,明确了高级推理型模型在标准化临床评估任务中的适配优势。未来仍需进一步开展“医生-模型”协作决策的对照实验,以验证其对临床决策准确性的真实效果。此外,未来也可通过多中心扩展研究及知识库精细化构建,进一步验证并提升其在不同手术类型与人群中的适用性与稳定性。

参考文献:

[1] 谢昉, 冯艳, 孙德峰. 围手术期规范化麻醉评估流程在日间手术中的应用[J]. 华西医学, 2021, 36(2): 144-151.
XIE Fang, FENG Yan, SUN Defeng. Role of perioperative standardized anesthesia evaluation in day surgery[J]. West Chin Med J, 2021, 36(2): 144-151.

[2] 郭振江, 王宁, 赵光远, 等. 基于机器学习建立术前预测近端胃癌食管切缘阳性模型[J]. 山东大学学报(医学版), 2024, 62(7): 78-83.
GUO Zhenjiang, WANG Ning, ZHAO Guangyuan, et al. Development of preoperative models for predicting positive esophageal margin in proximal gastric cancer based on machine learning[J]. Journal of Shandong University (Health Sciences), 2024, 62(7): 78-83.

[3] Selpien H, Penon J, Thunecke D, et al. Adjustment of positive end-expiratory pressure based on body mass index during general anaesthesia: a randomised controlled trial[J]. Anaesthesia, 2025, 80(11): 1322-1332.

[4] Lin C, Abboud S, Zoghbi V, et al. Suprazygomatic maxillary nerve blocks and opioid requirements in pediatric adenotonsillectomy: a randomized clinical trial[J]. JAMA Otolaryngol Head Neck Surg, 2024, 150(7): 564. doi:10.1001/jamaoto.2024.1011

[5] 王文奇, 郭梦帆, 杨杜祥, 等. 大语言模型发展与应用综述[J]. 中原工学院学报, 2025, 36(2): 1-8.
WANG Wenqi, GUO Mengfan, YANG Duxiang, et al. Overview of the development and applications of large language models[J]. Journal of Zhongyuan University of Technology, 2025, 36(2): 1-8.

[6] Shool S, Adimi S, Saboori Amleshi R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine[J]. BMC Med Inform Decis Mak, 2025, 25(1): 117. doi:10.1186/s12911-025-02954-4

[7] 薛东, 杨思毅, 杜晗, 等. 大语言模型的发展现状及引信行业赋能路径展望[J]. 探测与控制学报, 2025, 47(4): 9-20.
XUE Dong, YANG Siyi, DU Han, et al. The large language models development status and outlook on empowering fuze industry[J]. Journal of Detection Control, 2025, 47(4): 9-20.

[8] Liu BHM, Lin YZ, Long X, et al. Utilizing AI for the identification and validation of novel therapeutic targets and repurposed drugs for endometriosis[J]. Adv Sci, 2025, 12(5): 2406565. doi:10.1002/advs.202406565

[9] Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial[J]. BMC Med Educ, 2024, 24(1): 1391. doi:10.1186/s12909-024-06399-7

[10] Ye XD, Shan XF, Tu YF, et al. Examining the efficacy of large language models for mitigating depression and anxiety among Chinese students: a randomized controlled trial[J]. CIN Comput Inform Nurs, 2025, 43(9):e01349. doi:10.1097/cin.0000000000001349

[11] 陈紫林, 祝帆帆, 罗宇昕, 等. 大语言模型在医疗健康领域的应用现状与前景展望[J]. 医学与哲学, 2025, 46(12): 32-37.
CHEN Zilin, ZHU Fanfan, LUO Yuxin, et al. Overview of the development and applications of large language models[J]. Medicine Philosophy, 2025, 46(12): 32-37.

[12] 张晓波, 冯瑞, 杨睿, 等. DeepSeek 赋能的儿科全流程智慧医疗系统的构建和应用效果评价[J]. 中国循证儿科杂志, 2025, 20(3): 217-222.
ZHANG Xiaobo, FENG Rui, YANG Rui, et al. A DeepSeek-enabled intelligent pediatric healthcare system: construction and application effectiveness evaluation[J]. Chinese Journal of Evidence-Based Pediatrics, 2025, 20(3): 217-222.

[13] Uzel K, Azboyİ, Parvizi J. Venous thromboembolism in orthopedic surgery: global guidelines[J]. Acta Orthop Traumatol Turc, 2023, 57(5): 192-203.

[14] Sigmund A, Russell LA. Optimizing rheumatoid arthritis patients for surgery[J]. Curr Rheumatol Rep, 2018, 20(8): 48. doi:10.1007/s11926-018-0757-x

[15] Grits D, Kuo A, Acuña AJ, et al. The association between perioperative blood transfusions and venous thromboembolism risk following surgical management of hip fractures[J]. J Orthop, 2022, 34: 123-131. doi: 10.1016/j.jor.2022.08.016

[16] Arraut J, Thomas J, Oakley CT, et al. The AAHKS best podium presentation research award: a second dose of

dexamethasone reduces postoperative opioid consumption and pain in total joint arthroplasty[J]. J Arthroplasty, 2023, 38(7): S21-S28.

[17] Santos Gomes MA, Kovaleski JL, Pagani RN, et al. Machine learning applied to healthcare: a conceptual review [J]. J Med Eng Technol, 2022, 46(7): 608-616.

[18] Rashidi HH, Pantanowitz J, Hanna MG, et al. Introduction to artificial intelligence and machine learning in pathology and medicine: generative and nongenerative artificial intelligence basics[J]. Mod Pathol, 2025, 38(4): 100688. doi:10.1016/j.modpat.2024.100688

[19] Cheng TT, Li Y, Gu JQ, et al. The performance of ChatGPT in day surgery and pre-anesthesia risk assessment: a case-control study of 150 simulated patient presentations[J]. Perioper Med, 2024, 13(1): 111. doi: 10.1186/s13741-024-00469-6

[20] Abdel Malek M, van Velzen M, Dahan A, et al. Generation of preoperative anaesthetic plans by ChatGPT-4.0: a mixed-method study.[J]. Br J Anaesth, 2025, 134(5):1333-1340.

[21] Pedrosa E, Silva M, Lobo A, et al. Is the ASA classification universal? [J]. Turk J Anaesthesiol Reanim, 2021, 49(4): 298-303.

[22] Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery[J]. Circulation, 1999, 100(10): 1043-1049.

[23] Omiye JA, Gui HW, Rezaei SJ, et al. Large language models in medicine: the potentials and pitfalls: a narrative review[J]. Ann Intern Med, 2024, 177(2): 210-220.

[24] Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making[J]. Nat Med, 2025, 31(8): 2546-2549.

[25] Jebb AT, Ng V, Tay L. A review of key likert scale development advances: 1995 - 2019[J]. Front Psychol, 2021, 12: 637547. doi:10.3389/fpsyg.2021.637547

[26] Wysocka M, Wysocki O, Delmas M, et al. Large language Models, scientific knowledge and factuality: a framework to streamline human expert evaluation[J]. J Biomed Inform, 2024, 158: 104724. doi:10.1016/j.jbi.2024.104724

[27] Bedi S, Liu YT, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review[J]. Jama, 2025, 333(4): 319. doi:10.1001/jama.2024.21700

[28] Peng YF, Malin BA, Rousseau JF, et al. From GPT to DeepSeek: significant gaps remain in realizing AI in healthcare[J]. J Biomed Inform, 2025, 163: 104791. doi:10.1016/j.jbi.2025.104791

[29] Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making[J]. Nat Med, 2024, 30(9): 2613-2622.

[30] Tordjman M, Liu ZL, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning[J]. Nat Med, 2025, 31(8): 2550-2555.

[31] 巴宏军, 陈佳睿, 胡晗, 等. 住院医师对人工智能应用的认知与态度调查[J]. 中华医学教育杂志, 2025, 45(3):194-197.
BA Hongjun, CHEN Jiarui, HU Han, et al. Survey on residents' perception and attitudes towards the application of artificial intelligence[J]. Chinese Journal of Medical Education, 2025, 45(3): 194-197.

(编辑:徐苗蓁)