Advances in Clinical Medicine 临床医学进展, 2026, 16(1), 588-596
Published Online January 2026 in Hans. https://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2026.161080

Hans汉斯

大语言模型重塑外科未来

——大语言模型在外科领域的应用进展

沈泽林[1]，王光锁[2*]

[1]暨南大学第二临床医学院，深圳市人民医院胸外科，广东 深圳
[2]深圳市人民医院(南方科技大学第一附属医院、暨南大学第二附属医院)，胸外科，广东 深圳

收稿日期：2025年12月5日；录用日期：2025年12月28日；发布日期：2026年1月8日

摘 要

以ChatGPT、Gemini、DeepSeek、通义千问等为代表的大语言模型正蓬勃发展，其应用已渗透至医疗实践的各个领域，将深刻改变未来医院的格局。在胸外科、心脏病学、口腔外科、肾脏病学、骨科、胃肠病学和影像科学等领域，变革尤为迅速。大语言模型在辅助医学文档书写、提供临床决策支持、进行医学健康教育及患者围手术期管理等方面展现出巨大的应用潜力。本文综述了大语言模型在电子病例书写、临床辅助诊断、临床决策支持、患者健康管理、医学教育及科研论文撰写等多个外科相关场景的应用。大语言模型能够高效处理与分析大规模数据集，并具备出色的自然语言理解能力。然而，这些技术的应用仍存在局限性，如模型的“幻觉”现象、潜在的学术不端风险、临床过度依赖、误诊与治疗失误的可能性以及责任归属不清等问题。在充分利用大语言模型益处的同时，我们必须认识并解决这些伦理与实践挑战，以确保其在医学领域的应用是负责任且有效的。

关键词

大型语言模型，人工智能，外科学，教育，诊断，临床决策支持，自然语言处理

Large Language Models Reshaping the Future of Surgery

—Advances in the Application of Large Language Models in the Field of Surgery

Zelin Shen[1], Guangsuo Wang[2*]

[1]Department of Thoracic Surgery, Shenzhen People’s Hospital, The Second Clinical Medical College, Jinan University, Shenzhen Guangdong
[2]Department of Thoracic Surgery, Shenzhen People’s Hospital (The First Affiliated Hospital, Southern University of Science and Technology; The Second Clinical Medical College, Jinan University), Shenzhen Guangdong

[*]通讯作者。

文章引用: 沈泽林, 王光锁. 大语言模型重塑外科未来[J]. 临床医学进展, 2026, 16(1): 588-596.
DOI: 10.12677/acm.2026.161080

Received: December 5, 2025; accepted: December 28, 2025; published: January 8, 2026

Abstract

Large language models (LLMs), represented by ChatGPT, Gemini, DeepSeek, Tongyi Qianwen, and others, are flourishing. Their applications have penetrated various fields of medical practice and are poised to profoundly reshape the future landscape of hospitals. Transformations are occurring particularly rapidly in fields such as thoracic surgery, cardiology, oral surgery, nephrology, orthopedics, gastroenterology, and imaging sciences. LLMs demonstrate immense application potential in assisting with medical documentation, providing clinical decision support, conducting medical health education, and managing patients during the perioperative period, among other areas. This article reviews the applications of LLMs in various surgery-related scenarios, including electronic medical record writing, clinical auxiliary diagnosis, clinical decision support, patient health management, medical education, and scientific research paper writing. LLMs can efficiently process and analyze large-scale datasets and possess remarkable natural language understanding capabilities. However, the application of these technologies still has limitations, such as model "hallucination," potential risks of academic misconduct, clinical over-reliance, possibilities of misdiagnosis and treatment errors, as well as unclear attribution of responsibility. While fully leveraging the benefits of LLMs, we must recognize and address these ethical and practical challenges to ensure their application in the medical field is responsible and effective.

Keywords

Large Language Models, Artificial Intelligence, Surgery, Education, Diagnosis, Clinical Decision Support, Natural Language Processing

1. 引言

当前，以 ChatGPT 为代表的大语言模型正推动外科领域经历一场深刻的数字化转型[1] [2]。从手术风险评估到围手术期决策，大语言模型已成为提升外科诊疗质量、安全性与效率的关键驱动力[3]。在这一浪潮中，大型语言模型作为一种革命性的自然语言处理技术，凭借其强大的语言理解、生成和推理能力，正展现出重塑外科实践格局的巨大潜力[4] [5]。

与传统专注于图像或结构化数据的 AI 模型不同，大语言模型的核心优势在于处理海量、非结构化的文本数据[6] [7]。外科领域本质上是一个信息密集型学科，涵盖了从电子健康记录(EHR)、医学文献、手术报告、医患沟通到医学教育资料等浩瀚的文本信息海洋[8] [9]。这为大语言模型的应用提供了天然的土壤。目前，大语言模型在外科的应用探索已初见端倪，其潜力主要体现在以下几个核心环节：在临床决策支持方面，大语言模型能够快速整合患者病史、实验室检查结果和最新临床指南，为外科医生提供个性化的诊疗建议[9]；在医患沟通层面，它可作为智能助手，以通俗易懂的语言向患者解释复杂的手术流程、风险和术后注意事项，提升患者知情同意质量[10]；在医学教育与培训中，大语言模型能够生成逼真的临床模拟场景，为住院医师提供无限的练习机会，并实现个性化的反馈与辅导[11]；此外，在行政与科研领域，大语言模型还能自动化处理手术记录生成、临床数据提取和学术论文撰写等繁琐任务，将外科

医生从繁重的文书工作中解放出来，使其更专注于临床核心工作[12]。

然而，大语言模型在外科领域的集成并非一片坦途。其固有的局限性，如可能产生“幻觉”(即生成看似合理但实则错误的信息)、存在训练数据导致的偏见、以及涉及患者隐私和数据安全的严峻挑战，都为其临床落地设置了重重障碍[7]。此外，目前该领域的研究仍处于早期阶段，相关应用呈碎片化分布，缺乏对其有效性、可靠性及伦理影响的系统性评估。目前，尚未有研究对这一新兴领域进行梳理，本文主要综述外科主要应用场景中的最新进展，分析其当前面临的局限与挑战，并对未来发展方向进行展望[13]。

2. LLMs 在外科领域的主要应用场景

2.1. 外科电子病历书写

临床电子病历书写，尤其是详细、规范的手术记录，是外科医生的一个繁重的任务，尽管这些文档工作对于患者疾病的诊治、医疗质量改进和法律证据至关重要，但它们繁复的性质给外科医生增加了工作负荷和职业倦怠风险[14] [15]。采用生成式 AI，可以通过生成标准化的模板并自动填入相关临床信息来优化医疗记录的工作流程，从而提升医疗文档处理的效率[16]。手术记录是外科医生负责书写的一个重要文档，向大语言模型输入少量关键信息即可生成高质量的手术记录初稿，经外科医生审核修改后能大幅提升效率[17] [18]。在莫氏显微手术中，LLMs 已能辅助生成高质量的手术笔记[19]。未来，深度集成 EHR 系统的外科专科 LLMs 有望实现从麻醉记录、手术记录到术后随访计划的全程文档辅助。然而，实现这一目标必须确保生成的文本绝对精确，并建立严格的数据隐私保护机制。对于外科患者，详尽的临床病史是确保术前影像评估与术后随访方案精准化的关键。此外，详尽的临床病史是精准影像学评估的基础。Bhayana 等[20]人的研究揭示了利用大语言模型优化这一流程的潜力。该研究证实，LLMs 能够从冗长的临床笔记中自动生成高质量的影像检查病史，其内容远比传统申请单上的信息丰富，特别是能显著提升与外科相关的关键信息(如相关手术史)的提及率(61% vs 12%)。若将这一技术整合入外科工作流程，可辅助外科医生在开具 CT、MRI 等检查时，自动生成一份全面、结构化的临床病史，从而减少信息遗漏，提升放射科协作效率，最终使患者受益。

2.2. 临床辅助诊断

在外科诊断中，尤其是胸外科、普外科、口腔颌面外科等需要综合影像学与临床表现的领域，LLMs 有潜力成为强大的辅助工具[21]-[24]。一项针对口腔病理诊断的前瞻性评估研究[23]，通过 16 个模拟临床案例，对比了 DeepSeek-V3 和 ChatGPT-4o 的诊断性能。该研究由 20 名口腔颌面外科及放射科专家采用李克特量表进行盲法评估。结果表明，两种模型均能提供中等至良好水平的诊断建议，但 DeepSeek-V3 的平均得分(4.02 ± 0.36)显著高于 ChatGPT-4o (3.15 ± 0.41)，且在 16 个案例中的 9 个表现出统计学上的优势。这一发现提示，不同 LLMs 在专业外科诊断任务上存在性能差异，DeepSeek 作为后起之秀，展现了强大的竞争力。大语言模型在疑难复杂病例的诊断展现出独特的优势。近期一项针对伴有胃肠道症状的疑难病例的诊断研究表明，先进的大语言模型(LLM)展现出超越经验丰富专科医生的潜力。Yang 等[25]人发现，在 67 例真实世界疑难病例构成的离线数据集中，Claude 3.5 Sonnet 模型提供的诊断建议，其“指导性诊断”覆盖率高达 76.1%，显著高于所有 22 位参与研究的胃肠病专家(平均覆盖率 29.5%)。该研究提示，在复杂病例的诊断过程中，LLMs 能够为外科医生提供更广阔的鉴别诊断思路，有效弥补人类专家因知识领域局限可能导致的疏漏。

2.3. 围手术期临床决策支持

近年来，ChatGPT、DeepSeek 等开源大语言模型在医疗领域的应用展现出巨大潜力。在外科围手术

期时间中，电子病历系统及检验检查结果蕴含着大量的信息资源，贯穿于外科诊疗全流程：在术前阶段，辅助完成手术适应症判断、术式选择及风险预估；在术中环节，结合实时数据为手术决策提供参考依据；在术后管理中，支持并发症预警、康复评估及随访方案制定。这为提升外科诊疗的精准性、安全性及个体化水平提供了有力工具，对推动外科手术全程的智能辅助决策具有重要意义。一项发表于 Nature Medicine 的基准评估研究表明，开源模型 DeepSeek-V3 和 DeepSeek-R1 在涵盖多个专科(包括外科)的临床决策支持任务中，其诊断与治疗建议的准确性与最先进的专有模型(如 GPT-4o)相当，甚至在部分任务中表现更优。这证明了高性能的开源模型可以作为外科临床决策支持系统的可靠技术基础[26]-[28]。近期研究表明，多模态大语言模型在专科外科如喉癌手术中展现出显著的临床潜力。Liang 等[29]人系统评估了六种主流 MLLM 对喉癌相关影像(包括 CT、喉镜及病理图像等)的解读能力，发现先进模型如 Claude 3.5 Sonnet 在回答开放型临床问题时的准确率可达 79.43%，显著优于部分开源模型。该研究提示，MLLM 能够作为有效的辅助工具，整合多模态信息以支持外科医生在术前规划、术中决策及术后评估中的判断。Palenzuela 等[30]人的研究显示，在大语言模型 ChatGPT-4 与外科医师的临床决策能力对比中，ChatGPT-4 的表现优于低年资住院医师，并与高年资住院医师及主治医师相当。该研究基于真实病例构建了五种常见外科场景，结果表明 ChatGPT-4 在确定手术方式和识别术后并发症方面尤其出色。研究者认为，它有望成为辅助低年资医师进行临床决策训练的教育工具，但也指出了其存在“幻觉”和无法提供参考来源等局限性。前瞻性的手术行动预测对于手术决策具有重大意义。Xu 等[31]人提出 Surgical Action Planning (SAP)任务，并开发了基于大语言模型的 LLM-SAP 框架，通过近历史记忆模块(NHFM)和提示工程，实现从手术视频中生成未来行动步骤。该研究还提出了 ReAcc 评估指标，更贴合手术实际动态性。实验表明，经过监督微调的模型在 CholecT50-SAP 数据集上取得显著提升，展示了 LLMs 在手术决策中的潜力。

2.4. 外科教育与技能培训

多项研究表明[32]-[34]，以 ChatGPT 为代表的大语言模型(LLMs)在国家医学考试中以及多项专业考试中，LLMs 能够达到甚至超越低年资医生的水平，能够结合文本和图像解答包括解剖、术式、病理生理和并发症在内的各类问题，解决复杂临床实践问题[35]。人工智能技术正从多个维度推动教学模式的革新与深化，AI 可以用于生成高保真、个性化的虚拟病人和临床场景，用于训练学员的诊断思维和临床决策能力[36]。在“手术技能评估反馈”方面，AI 可结合手术视频，利用自然语言处理技术对学员的操作过程进行精准分析，并以自然语言描述的形式提供具体、可操作的操作技巧反馈与改进建议，帮助学员客观认知自身技术水平，实现从理论到实践的全链条能力提升[37]。一项在 61 个本科生[38]的随机对照试验中表明，相较于传统的教学模式，基于 ChatGPT 的混合教学模式在期末考试理论成绩(86.44 ± 5.59 vs. 77.86 ± 4.16, $p < 0.001$)和临床技能成绩(83.84 ± 6.13 vs. 79.12 ± 4.27, $p = 0.001$)方面均显著优于对照组。此外，实验组的教学满意度(17.23 ± 1.33)和对教学效果的自我评价(9.14 ± 0.54)也显著高于对照组(分别为 15.38 ± 1.5 和 8.46 ± 0.70，$p < 0.001$)。

2.5. 术前规划与风险评估

LLMs 能够快速读取患者的电子健康记录，提取关键信息(如合并症、手术史、用药史)，并生成结构化的术前评估摘要。通过整合临床指南和文献，它们可以辅助外科医生识别潜在的手术风险，推荐个性化的术前优化方案(如营养支持、戒烟干预)。例如，将患者的病历数据输入特定的 LLMs，可以自动生成包括美国麻醉医师协会分级、手术风险指数在内的综合报告，提高术前准备的效率和全面性[39]。此外，大语言模型对于患者围手术期的临床风险预测具有一定的潜力。Chung 等[39]人进行的一项预后研究系统评估了 GPT-4 Turbo 在八项围手术期预测任务中的表现。该研究显示，模型在分类任务上取得了优于随

机猜测的结果，其中 ICU 入住预测(F1 分数：0.81)和医院死亡率预测(F1 分数：0.86)表现最佳，ASA-PS 分级预测亦有中等程度准确性(F1 分数：0.50)。然而，模型在预测持续时间的任务(如 PACU 停留时间、住院天数)上表现不佳，其预测误差甚至高于简单基线模型。在手术室中，大致预估手术时间对于灵活调度手术室资源至关重要。研究表明 ChatGPT 可以较为准确地预测手术时间，微调后的 GPT-4 取得了最佳性能，平均绝对误差(MAE)为 47.64 分钟(95% CI, 45.71~49.56)，R2 为 0.61，与当前手术室调度的性能相当(MAE，49.34 分钟；95% CI，47.60~51.09；R2，0.63；P = 0.10)。微调后的 GPT-4 和微调后的 GPT-3.5 在准确性上均显著优于当前调度方法(分别为 46.12%和 46.08% vs 40.92%；P < 0.001)。在外部验证期间，微调后的 GPT-4 表现优于所有其他模型，性能指标相似(MAE，48.66 分钟；95% CI，45.31~52.00；准确性，46.0%)。基础模型表现各异，其中 GPT-4 在未微调模型中表现最佳(MAE，59.20 分钟；95% CI，56.88~61.52)。ChatGPT 能辅助预测手术时间，提高手术室的运转效率[40]。

2.6. LLMs 在术中应用与人机交互

随着手术机器人、AR 导航和智能传感技术的普及，手术室正演变为一个高度数字化的环境。这为 LLMs 的术中实时应用创造了条件。LLMs 有潜力作为“中央认知处理器”，整合并分析来自多源的数据流：1) 手术视频流：通过实时分析内窥镜或术野摄像画面，LLMs 可识别解剖结构、手术步骤，甚至预警可能的误操作(如接近重要血管神经)。结合如 LLM-SAP 的框架[31]，可预测手术进程并提供步骤提醒。2) 生理监测数据：整合患者实时生命体征，LLMs 可辅助麻醉医生和外科团队进行动态风险评估。3) 机器人传感器数据：解读手术机器人提供的力反馈、器械运动轨迹等信息，LLMs 可评估操作流畅度或识别非典型阻力。4) 语音指令与交互：外科医生可通过自然语音询问(如“显示肝门部解剖变异概率”或“根据当前出血量，推荐下一步止血方案”)，LLMs 即时调用知识库并生成语音或 AR 叠加强调回复，实现真正意义上的免提、智能人机交互。然而，术中应用对实时性、准确性和可靠性要求极高，任何延迟或错误都可能造成严重后果，这将是未来技术攻关和验证的重点。

2.7. 患者教育和术后护理

大语言模型在患者健康教育方面具有重大潜力，Nitin Srinivasan 等[41]人的研究表明在减肥手术患者教育中，GPT-4 与其他大语言模型相比，能够让患者理解回答患者的问题。由于健康教育资源的普及程度不足，许多患者通过搜索引擎寻找医疗建议，可能会阅读到不恰当的内容，从而导致错误的自我诊断。ChatGPT 能够作为一个宝贵的教育资源，为患者提供清晰、易于获取的信息，帮助他们理解自己的病情，并回答患者的疑问，使患者能够积极参与治疗过程。同时，ChatGPT 的回答可以根据患者的教育水平进行调整，提供个性化的信息。Lee 等[42]人进行了一项随机对照试验，探讨了聊天机器人与视频教育在乳腺癌患者放疗过程中的应用效果。研究发现，尽管整体上不同教育媒介在减轻患者焦虑方面无显著差异，但在年龄 ≤50 岁的患者中，使用聊天机器人的组别表现出焦虑减轻的趋势。该研究提示，基于聊天机器人的数字健康教育工具对于年轻患者具有潜在的应用价值，可作为外科及肿瘤治疗中患者教育与心理支持的辅助手段。同时，LLMs 可以自动生成出院小结，并为患者提供个性化的术后康复指导，包括用药提醒、活动建议和饮食注意事项，改善患者随访体验。

2.8. 临床科研写作

ChatGPT 大语言模型在临床研究和论文写作方面有极大的潜力。大语言模型能够高效地整合信息，在协助撰写提纲和论文初稿方面具有重大作用[43]-[45]。外科领域的知识更新迅速，每年有大量临床试验和研究成果发表，这为进行系统评价和 meta 分析带来了巨大挑战。传统的人工数据提取过程不仅耗时耗

力，而且容易出错。Khan 等[46]人的研究为这一难题提供了创新的解决方案。他们开发了一种协作式大语言模型工作流，模拟传统的“双人审核”过程，通过 GPT-4 和 Claude-3 的协同工作与交叉评判，在自动化数据提取中实现了高准确率(测试集一致应答的准确率达 94%)并显著降低了幻觉现象。这一方法证明了 LLMs 在自动化证据合成方面的巨大潜力，其通用性工作流可直接应用于外科临床试验数据的提取，从而为构建“动态外科系统评价”和加速外科证据更新奠定技术基础。

3. 局限性与挑战

3.1. 信息准确性、“幻觉”与外科决策风险

LLMs 可能生成不准确、过时或完全虚构的医学信息。在外科实践中，这种“幻觉”可能导致具有直接危害的决策错误。例如，一个 LLMs 在辅助肺癌分期时，可能给出错误的临床分期，若外科医生未加甄别，可能导致非标准化的、甚至危险的手术操作。模型缺乏真正的临床经验和情境化推理能力，无法替代外科医生在术中对组织质地、出血状况等实时情况的综合判断。

3.2. 透明度、责任与伦理问题

LLMs 的决策过程如同“黑箱”。当出现不良后果时，责任界定极其困难。若外科医生部分采纳了 LLMs 推荐的、具有一定创新性但证据等级不高的微创术式，术中发生严重并发症。此时，责任应归于采纳建议的外科医生、提供模型的开发公司、还是负责训练数据的提供方？目前法律和伦理框架对此尚无清晰界定。过度依赖 LLMs 还可能导致外科医生临床思维和技能退化。

3.3. 数据隐私与安全

外科患者数据高度敏感。LLMs 的训练和使用涉及数据传输与存储，存在泄露风险。将 LLMs 安全、合规地集成到现有的医院信息系统(HIS)、PACS 和手术室设备中，面临技术和管理的双重挑战。

4. 结论与展望

大语言模型正在深刻融入外科实践的各个环节，从文书、诊断、决策、教育到科研，展现出提升医疗质量、效率与患者体验的巨大潜力。然而，其固有的“幻觉”风险、责任模糊、数据安全等挑战不容忽视。为了有效地推动这一变革，未来研究与应用应聚焦于以下几个方向：1) 开发外科专科多模态大模型：构建能够深度融合文本型电子病历、医学影像、手术视频、实时生理信号及机器人传感数据的外科领域专用多模态大模型，实现对外科全流程信息的统一理解与推理，为精准手术提供全方位智能支持。2) 建立外科 LLMs 评估与基准测试平台：创建公开、标准化的外科 LLMs 性能基准测试集(Surgical LLMs Benchmarks)，涵盖从诊断准确性、手术方案合理性、文档生成质量到伦理合规性等多个维度，为模型的安全性、有效性和可靠性评估提供科学依据。3) 研究 LLMs 在外科技能闭环评估与个性化培训中的应用：深入探索 LLMs 结合手术视频分析技术，实现对外科医生操作技能的自动化、精细化评估，并提供个性化的自然语言反馈与培训课程推荐，形成“评估–反馈–改进”的闭环，革新外科技能培训体系。4) 构建适应 AI 辅助决策的外科临床指南与伦理法律框架：联合医学、法学、伦理学及技术专家，共同研究并制定明确规范 LLMs 在外科临床应用中角色、权限、责任归属的指南与政策。探索建立针对 AI 辅助决策的医疗责任保险和伦理审查机制，为技术创新保驾护航。未来，通过跨学科合作、严格验证和框架建设，大语言模型有望从“辅助工具”演进为值得信赖的“智能伙伴”，共同推动外科学进入一个更加精准、安全、高效的新时代。

参考文献

[1] Hashimoto, D.A., Rosman, G., Rus, D. and Meireles, O.R. (2018) Artificial Intelligence in Surgery: Promises and Perils. *Annals of Surgery*, **268**, 70-76. https://doi.org/10.1097/sla.0000000000002693

[2] Wall, J. and Krummel, T. (2020) The Digital Surgeon: How Big Data, Automation, and Artificial Intelligence Will Change Surgical Practice. *Journal of Pediatric Surgery*, **55**, 47-50. https://doi.org/10.1016/j.jpedsurg.2019.09.008

[3] Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., *et al*. (2020) Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. *Cell*, **181**, 1423-1433.e11. https://doi.org/10.1016/j.cell.2020.04.045

[4] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., *et al*. (2020) Language Models Are Few-Shot Learners. https://arxiv.org/abs/2005.14165

[5] Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., *et al*. (2023) Performance of ChatGPT on USMLE: Potential for Ai-Assisted Medical Education Using Large Language Models. *PLOS Digital Health*, **2**, e0000198. https://doi.org/10.1371/journal.pdig.0000198

[6] Vaswani, A., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[7] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F. and Ting, D.S.W. (2023) Large Language Models in Medicine. *Nature Medicine*, **29**, 1930-1940. https://doi.org/10.1038/s41591-023-02448-8

[8] Marwaha, J.S., Raza, M.M. and Kvedar, J.C. (2023) The Digital Transformation of Surgery. *NPJ Digital Medicine*, **6**, Article No. 103. https://doi.org/10.1038/s41746-023-00846-3

[9] Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., *et al*. (2022) Surgical Data Science—From Concepts toward Clinical Translation. *Medical Image Analysis*, **76**, Article ID: 102306. https://doi.org/10.1016/j.media.2021.102306

[10] Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., *et al*. (2023) Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, **183**, 589-596. https://doi.org/10.1001/jamainternmed.2023.1838

[11] Meskó, B. and Görög, M. (2020) A Short Guide for Medical Professionals in the Era of Artificial Intelligence. *NPJ Digital Medicine*, **3**, Article No. 126. https://doi.org/10.1038/s41746-020-00333-z

[12] Liu, T., Hetherington, T.C., Stephens, C., McWilliams, A., Dharod, A., Carroll, T., *et al*. (2024) AI-Powered Clinical Documentation and Clinicians' Electronic Health Record Experience: A Nonrandomized Clinical Trial. *JAMA Network Open*, **7**, e2432460. https://doi.org/10.1001/jamanetworkopen.2024.32460

[13] Patil, A., Serrato, P., Chisvo, N., Arnaout, O., See, P.A. and Huang, K.T. (2024) Large Language Models in Neurosurgery: A Systematic Review and Meta-Analysis. *Acta Neurochirurgica*, **166**, Article No. 475. https://doi.org/10.1007/s00701-024-06372-9

[14] Arndt, B.G., Beasley, J.W., Watkinson, M.D., Temte, J.L., Tuan, W., Sinsky, C.A., *et al*. (2017) Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *The Annals of Family Medicine*, **15**, 419-426. https://doi.org/10.1370/afm.2121

[15] Shanafelt, T.D., Hasan, O., Dyrbye, L.N., Sinsky, C., Satele, D., Sloan, J., *et al*. (2015) Changes in Burnout and Satisfaction with Work-Life Balance in Physicians and the General US Working Population between 2011 and 2014. *Mayo Clinic Proceedings*, **90**, 1600-1613. https://doi.org/10.1016/j.mayocp.2015.08.023

[16] Huang, H., Zheng, O., Wang, D., Yin, J., Wang, Z., Ding, S., *et al*. (2023) ChatGPT for Shaping the Future of Dentistry: The Potential of Multi-Modal Large Language Model. *International Journal of Oral Science*, **15**, Article No. 29. https://doi.org/10.1038/s41368-023-00239-y

[17] Di Ieva, A., Stewart, C. and Suero Molina, E. (2024) Large Language Models in Neurosurgery. In: Di Ieva, A., Ed., *Artificial Intelligence in Clinical Neurosciences*, Springer, 177-198. https://doi.org/10.1007/978-3-031-64892-2_11

[18] Hurley, E.T., Crook, B.S., Lorentz, S.G., Danilkowicz, R.M., Lau, B.C., Taylor, D.C., *et al*. (2024) Evaluation High-Quality of Information from ChatGPT (Artificial Intelligence—Large Language Model) Artificial Intelligence on Shoulder Stabilization Surgery. *Arthroscopy*: *The Journal of Arthroscopic & Related Surgery*, **40**, 726-731.e6. https://doi.org/10.1016/j.arthro.2023.07.048

[19] Chacko, R.S., Chacko, S.M., Srinivasan, G., Davis, M. and LeBoeuf, M. (2025) Automated Note Generation for Mohs Micrographic Surgery Using a Large Language Model: A Retrospective Cohort Study. *Journal of the American Academy of Dermatology*, **93**, 1077-1079. https://doi.org/10.1016/j.jaad.2025.04.084

[20] Bhayana, R., Alwahbi, O., Ladak, A.M., Deng, Y., Basso Dias, A., Elbanna, K., *et al*. (2025) Leveraging Large Language Models to Generate Clinical Histories for Oncologic Imaging Requisitions. *Radiology*, **314**, e242134.

https://doi.org/10.1148/radiol.242134

[21] Gong, E.J., Bang, C.S., Lee, J.J., Park, J., Kim, E., Kim, S., *et al*. (2024) Large Language Models in Gastroenterology: Systematic Review. *Journal of Medical Internet Research*, **26**, e66648. https://doi.org/10.2196/66648

[22] Wu, C., Liu, W., Mei, P., Liu, Y., Cai, J., Liu, L., *et al*. (2025) The Large Language Model Diagnoses Tuberculous Pleural Effusion in Pleural Effusion Patients through Clinical Feature Landscapes. *Respiratory Research*, **26**, Article No. 52. https://doi.org/10.1186/s12931-025-03130-y

[23] Kaygisiz, Ö.F. and Teke, M.T. (2025) Can Deepseek and ChatGPT Be Used in the Diagnosis of Oral Pathologies? *BMC Oral Health*, **25**, Article No. 638. https://doi.org/10.1186/s12903-025-06034-x

[24] Srivastav, S., Chandrakar, R., Gupta, S., Babhulkar, V., Agrawal, S., Jaiswal, A., *et al*. (2023) ChatGPT in Radiology: The Advantages and Limitations of Artificial Intelligence for Medical Imaging Diagnosis. *Cureus*, **15**, e41435. https://doi.org/10.7759/cureus.41435

[25] Yang, X., Li, T., Wang, H., Zhang, R., Ni, Z., Liu, N., *et al*. (2025) Multiple Large Language Models versus Experienced Physicians in Diagnosing Challenging Cases with Gastrointestinal Symptoms. *NPJ Digital Medicine*, **8**, Article No. 85. https://doi.org/10.1038/s41746-025-01486-5

[26] Sandmann, S., Hegselmann, S., Fujarski, M., Bickmann, L., Wild, B., Eils, R., *et al*. (2025) Benchmark Evaluation of DeepSeek Large Language Models in Clinical Decision-Making. *Nature Medicine*, **31**, 2546-2549. https://doi.org/10.1038/s41591-025-03727-2

[27] Thirunavukarasu, A.J., Hassan, R., Mahmood, S., Sanghera, R., Barzangi, K., El Mukashfi, M., *et al*. (2023) Trialling a Large Language Model (ChatGPT) in General Practice with the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care. *JMIR Medical Education*, **9**, e46599. https://doi.org/10.2196/46599

[28] Choi, J. (2025) Artificial Intelligence in Surgery Research: Successfully Implementing AI Clinical Decision Support Models. *Journal of Trauma and Acute Care Surgery*, **99**, 518-521. https://doi.org/10.1097/ta.0000000000004725

[29] Liang, B., Gao, Y., Wang, T., Zhang, L. and Wang, Q. (2025) Multimodal Large Language Models Address Clinical Queries in Laryngeal Cancer Surgery: A Comparative Evaluation of Image Interpretation across Different Models. *International Journal of Surgery*, **111**, 2727-2730. https://doi.org/10.1097/js9.0000000000002234

[30] Palenzuela, D.L., Mullen, J.T. and Phitayakorn, R. (2024) AI versus MD: Evaluating the Surgical Decision-Making Accuracy of ChatGPT-4. *Surgery*, **176**, 241-245. https://doi.org/10.1016/j.surg.2024.04.003

[31] Xu, M., Huang, Z., Zhang, J., Zhang, X. and Dou, Q. (2025) Surgical Action Planning with Large Language Models. 28*th International Conference MICCAI* 2025, Daejeon, 23-27 September 2025, 563-572. https://doi.org/10.1007/978-3-032-05114-1_54

[32] Yu, P., Fang, C., Liu, X., Fu, W., Ling, J., Yan, Z., *et al*. (2024) Performance of ChatGPT on the Chinese Postgraduate Examination for Clinical Medicine: Survey Study. *JMIR Medical Education*, **10**, e48514. https://doi.org/10.2196/48514

[33] Long, C., Lowe, K., Zhang, J., Santos, A.d., Alanazi, A., O'Brien, D., *et al*. (2024) A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology—Head and Neck Surgery Certification Examinations: Performance Study. *JMIR Medical Education*, **10**, e49970. https://doi.org/10.2196/49970

[34] Prazeres, F. (2025) ChatGPT's Performance on Portuguese Medical Examination Questions: Comparative Analysis of ChatGPT-3.5 Turbo and ChatGPT-4o Mini. *JMIR Medical Education*, **11**, e65108-e65108. https://doi.org/10.2196/65108

[35] Maruyama, H., Toyama, Y., Takanami, K., Takase, K. and Kamei, T. (2025) Role of Artificial Intelligence in Surgical Training by Assessing GPT-4 and GPT-4o on the Japan Surgical Board Examination with Text-Only and Image-Accompanied Questions: Performance Evaluation Study. *JMIR Medical Education*, **11**, e69313-e69313. https://doi.org/10.2196/69313

[36] Park, J.J., Tiefenbach, J. and Demetriades, A.K. (2022) The Role of Artificial Intelligence in Surgical Simulation. *Frontiers in Medical Technology*, **4**, Article ID: 1076755. https://doi.org/10.3389/fmedt.2022.1076755

[37] Azari, D.P., Frasier, L.L., Quamme, S.R.P., Greenberg, C.C., Pugh, C.M., Greenberg, J.A., *et al*. (2019) Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. *Annals of Surgery*, **269**, 574-581. https://doi.org/10.1097/sla.0000000000002478

[38] Wu, C., Chen, L., Han, M., Li, Z., Yang, N. and Yu, C. (2024) Application of ChatGPT-Based Blended Medical Teaching in Clinical Education of Hepatobiliary Surgery. *Medical Teacher*, **47**, 445-449. https://doi.org/10.1080/0142159x.2024.2339412

[39] Wang, B., Tian, Y. and Wang, X.T. (2025) An Exploratory Comparison of AI Models for Preoperative Anesthesia Planning: Assessing ChatGPT-4o, Claude 3.5 Sonnet, and ChatGPT-O1 in Clinical Scenario Analysis. *Journal of Medical Systems*, **49**, Article No. 104. https://doi.org/10.1007/s10916-025-02243-7

[40] Ramamurthi, A., Neupane, B., Deshpande, P., Hanson, R., Vegesna, S., Cray, D., *et al*. (2025) Applying Large Language Models for Surgical Case Length Prediction. *JAMA Surgery*, **160**, 894-902. https://doi.org/10.1001/jamasurg.2025.2154

[41] Srinivasan, N., Samaan, J.S., Rajeev, N.D., Kanu, M.U., Yeo, Y.H. and Samakar, K. (2024) Large Language Models and Bariatric Surgery Patient Education: A Comparative Readability Analysis of GPT-3.5, GPT-4, Bard, and Online Institutional Resources. *Surgical Endoscopy*, **38**, 2522-2532. https://doi.org/10.1007/s00464-024-10720-2

[42] Lee, J., Byun, H.K., Kim, Y.T., Shin, J. and Kim, Y.B. (2025) A Study on Breast Cancer Patient Care Using Chatbot and Video Education for Radiation Therapy: A Randomized Controlled Trial. *International Journal of Radiation Oncology Biology Physics*, **122**, 84-92. https://doi.org/10.1016/j.ijrobp.2024.12.012

[43] Holland, A.M., Lorenz, W.R., Cavanagh, J.C., Smart, N.J., Ayuso, S.A., Scarola, G.T., *et al*. (2024) Comparison of Medical Research Abstracts Written by Surgical Trainees and Senior Surgeons or Generated by Large Language Models. *JAMA Network Open*, **7**, e2425373. https://doi.org/10.1001/jamanetworkopen.2024.25373

[44] Cao, C., Sang, J., Arora, R., Chen, D., Kloosterman, R., Cecere, M., *et al*. (2025) Development of Prompt Templates for Large Language Model-Driven Screening in Systematic Reviews. *Annals of Internal Medicine*, **178**, 389-401. https://doi.org/10.7326/annals-24-02189

[45] Stadler, R.D., Sudah, S.Y., Moverman, M.A., Denard, P.J., Duralde, X.A., Garrigues, G.E., *et al*. (2025) Identification of ChatGPT-Generated Abstracts within Shoulder and Elbow Surgery Poses a Challenge for Reviewers. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, **41**, 916-924.e2. https://doi.org/10.1016/j.arthro.2024.06.045

[46] Khan, M.A., Ayub, U., Naqvi, S.A.A., Khakwani, K.Z.R., Sipra, Z.b.R., Raina, A., *et al*. (2025) Collaborative Large Language Models for Automated Data Extraction in Living Systematic Reviews. *Journal of the American Medical Informatics Association*, **32**, 638-647. https://doi.org/10.1093/jamia/ocae325