번호	저자	제목	모델	주요내용
1	(Sharma, Alaa, and Daneshjou 2025)	A longitudinal analysis of declining medical safety messaging in generative AI models	[VLM] GPT-4 Turbo (2023),GPT-4o (May, August, and November 2024), GPT-o1 (December 2024), and GPT-4.5 (2025); Grok Beta(2023), Grok2(2024), and Grok3(2025) fromX; Gemini 1.5 Flash (2024), Gemini 1.5 Pro (2024), and Gemini 2.0 Flash (2025) from Google DeepMind; and Claude 3.5 Sonnet (2024) and Claude 3.7 Sonnet (2025) from Anthropic. [LLM] GPT-3.5 Turbo (2022), GPT-4, GPT-4 Turbo, GPT-4o, and GPT-4.5; Claude 3 Opus (2024), Claude 3.5Sonnet, and Claude3.7Sonnet, Google Gemini1.5Flash, 1.5 Pro, and 2.0 Flash, Grok Beta, Grok 2, and Grok 3 and DeepSeek V2.5 (2024), V3 (2024), and R1 (2024).	대규모 언어 모델(LLM) 및 비전-언어 모델(VLM)을 포함한 생성형 AI 모델은 의료 영상을 해석하고 임상 질문에 답변하는 데 점점 더 많이 사용되고 있다. **그러나 이들의 응답에는 종종 부정확성이 포함되므로, 의료 면책 고지와 같은 안전 조치가 매우 중요**하다. 본 연구에서 우리는 2022년부터 2025년까지 출시된 모델 세대에 걸쳐 LLM 및 VLM 출력에서 면책 고지의 존재를 평가하였다. 응답은 500개의 유방 촬영 영상, 500개의 흉부 X선 영상, 500개의 피부과 영상, 그리고 우리가 도입한 새로운 데이터셋인 TIMed-Q(상위 인터넷 의료 질문 데이터셋)에서 추출한 500개의 의료 질문으로부터 생성되었다. TIMed-Q는 환자들이 가장 빈번하게 검색하는 의료 질문을 포착하여 실제 건강 정보 탐색 행동을 반영한다. **LLM 출력에서 면책 고지의 존재는 2022년 26.3%에서 2025년 0.97%로 감소하였으며, VLM 면책 고지 비율은 2023년 19.6%에서 1.05%로 감소하였다. 2025년까지 대부분의 모델은 면책 고지를 표시하지 않았다. 모델의 성능이 더욱 향상됨에 따라, 면책 고지는 임상 맥락에 맞춰진 적응형 안전장치로 기능해야 한다.** Generative AI models, including large language models (LLMs) and vision-language models (VLMs), are increasingly used to interpret medical images and answer clinical questions. However, their responses often include inaccuracies; therefore, safety measures like medical disclaimers are critical. In this study, we evaluated the presence of disclaimers in LLM and VLM outputs across model generations released from 2022 to 2025. Responses were generated from 500 mammograms, 500 chest X-rays, 500 dermatology images, and 500 medical questions drawn from a new dataset we introduced: TIMed-Q (Top Internet Medical Question Dataset). TIMed-Q captures the most frequently searched medical queries by patients, reflecting real-world health information-seeking behavior. Disclaimer presence in LLM outputs dropped from 26.3% in 2022 to 0.97% in 2025, while VLM disclaimer rates declined from 19.6% in 2023 to 1.05%. By 2025, most models displayed no disclaimers. As models gain further capability, disclaimers must function as adaptive safeguards tailored to clinical contexts.
2	(Pesapane et al. 2025)	A preliminary investigation into the potential, pitfalls, and limitations of large language models for mammography interpretation	ChatGPT-4	본 연구는 대규모 언어 모델, 특히 GPT-4의 유방촬영 영상 해석 능력을 평가한다. 분석은 유방촬영 소견이 있는 경우와 없는 경우로 균등하게 나누어진 120개의 유방촬영 영상을 대상으로 하였다. 추가적인 맥락 없이, LLM은 오직 이러한 영상만을 기반으로 보고서를 생성하는 과제를 부여받았다. GPT-4는 53.3%의 사례에서 유방촬영 투영을 정확하게 식별하였으며, 미세석회화 및 종괴 식별에서 다양한 수준의 정확도를 보였다. 본 연구는 50.0%의 민감도와 37.5%의 특이도로 GPT-4의 초기

				단계 해석 능력을 부각시켰다. 그러나 **환각과 함께 상당한 위양성 및 위음성 비율은 모델의 한계를 강조**하였다. 이 탐색적 검증은 유방촬영 해석에서 LLM 사용의 잠재력과 위험성에 대한 통찰을 제공하며, 임상 실무에서 신뢰성과 안전성을 보장하기 위해 의료 분야 AI 도구의 전용 훈련,CER 검증 및 규제의 필요성을 강조한다. This study evaluates the capabilities of large language models, specifically GPT-4, in interpreting mammographic images. The analysis involved 120 mammographic images equally divided between cases with and without mammography's findings. Without additional context, the LLM was tasked to generate reports based solely on these images. GPT-4 correctly identified mammographic projections in 53.3% of cases and showed varying degrees of accuracy in identifying microcalcifications and masses. The study highlighted GPT-4's embryonic interpretative abilities with a sensitivity of 50.0% and specificity of 37.5%. However, a significant rate of false positives and false negatives, along with hallucinations, underscored the model's limitations. This exploratory test offers insights into the potential and risks of using LLMs in mammography interpretation, also underscoring the need for dedicated training, validation, and regulation of AI tools in healthcare to ensure their reliability and safety in clinical practice.
3	(de Oliveira et al. 2025)	A study of calibration as a measurement of trustworthiness of large language models in biomedical natural language processing	[General domain] ✓ Small: Zephyr-7B-Beta, Llama-3-8B-Instruct ✓ Large: Flan-T5-XXL, Yi-1.5-34B-Chat, GPT-3.5- Turbo, GPT-4 [Biomedical domain] ✓ Small: Meditron-7B, Medicine-Llama3-8B ✓ Large: MedLLaMA-13B	**목적** 생물의학 자연어 처리(BioNLP) 과제 내에서 9개 대규모 언어 모델(LLM)의 보정을 평가하여 실제 환경에서의 신뢰성과 신뢰도에 대한 이해를 증진한다. **재료 및 방법** 각 LLM에 대해, 우리는 생물의학 언어 이해 및 추론 벤치마크(BLURB)의 모든 13개 데이터셋(6개 과제로 그룹화됨)에 대한 응답과 해당 신뢰도 점수를 수집하였다. 신뢰도 점수는 Verbal, Self-consistency, Hybrid의 3가지 전략을 사용하여 할당되었다. 평가를 위해, 우리는 모델 응답의 부분적 정확성을 고려하는 ECE의 새로운 적응 버전인 Flex-ECE(유연한 기대 보정 오차)를 도입하여 언어 기반 환경에서 보정에 대한 보다 현실적인 평가를 가능하게 하였다. 등위 회귀와 히스토그램 구간화라는 두 가지 사후 보정 기법이 평가되었다. **결과** 과제 전반에 걸쳐 평균 보정은 23.9%(인구-중재-비교-결과 추출)에서 46.6%(관계 추출)까지 범위를 보였다. LLM 전반에 걸쳐 Medicine-Llama3-8B가 최고의 평균 전체 보정(29.8%)을 보였으며, Flan-T5-XXL은 13개 데이터셋 중 5개에서 최고 순위를 기록하였다. 전략 전반에 걸쳐 Self-consistency(평균: 27.3%)가 Verbal(평균: 42.0%) 및 Hybrid(평균: 44.2%)보다 더 나은 보정을 보였다. 사후 방법은 보정을 상당히 개선하였으며, 최고 평균 보정 Flex-ECE는 0.1%에서 4.1% 범위였다. **고찰** **LLM의 즉시 사용 가능한 보정 성능이 저조한 것은 실제 BioNLP 응용 분야**

				에서 이러한 모델의 신뢰할 수 있는 배포에 위험을 초래한다. 보정은 사후에 개선될 수 있으며 권장되는 관행이다. Flex-ECE와 같은 LLM 평가를 위한 비이진 지표는 LLM의 신뢰성에 대한 보다 현실적인 평가를 제공하며, 실제로 부분적으로 옳거나 틀릴 수 있는 모든 모델에 적용된다. **결론** 본 연구는 LLM의 즉시 사용 가능한 보정이 매우 저조함을 보여주지만, 전통적인 사후 보정 기법이 LLM을 보정하는 데 유용함을 보여준다. **Objectives** To assess the calibration of 9 large language models (LLMs) within biomedical natural language processing (BioNLP) tasks, furthering understanding of trustworthiness and reliability in real-world settings. **Materials and Methods** For each LLM, we collected responses and corresponding confidence scores for all 13 datasets (grouped into 6 tasks) of the Biomedical Language Understanding & Reasoning Benchmark (BLURB). Confidence scores were assigned using 3 strategies: Verbal, Self-consistency, and Hybrid. For evaluation, we introduced Flex-ECE (Flexible Expected Calibration Error), a novel adaptation of ECE that accounts for partial correctness in model responses, allowing for a more realistic assessment of calibration in language-based settings. Two post-hoc calibration techniques – isotonic regression and histogram binning – were evaluated. **Results** Across tasks, mean calibration ranged from 23.9% (Population-Intervention-Comparison-Outcome extraction) to 46.6% (Relation Extraction). Across LLMs, Medicine-Llama3-8B had the best mean overall calibration (29.8%), and Flan-T5-XXL had the highest ranking on 5/13 datasets. Across strategies, Self-consistency (mean: 27.3%) had better calibration than Verbal (mean: 42.0%) and Hybrid (mean: 44.2%). Post-hoc methods substantially improved calibration, with best mean calibrated Flex-ECEs ranging from 0.1% to 4.1%. **Discussion** The poor out-of-the-box calibration of LLMs poses a risk to trustworthy deployment of such models in real-world BioNLP applications. Calibration can be improved post-hoc and is a recommended practice. Non-binary metrics for LLM evaluation such as Flex-ECE provide a more realistic assessment of trustworthiness of LLMs, and indeed any model that can be partially right/wrong. **Conclusion** This study shows that out-of-the-box calibration of LLMs is very poor, but traditional post-hoc calibration techniques are useful to calibrate

				LLMs.
4	(Flathers et al. 2024)	AI depictions of psychiatric diagnoses: a preliminary study of generative image outputs in Midjourney V.6 and DALL-E 3	Midjourney V.6 DALL-E 3	**목적:** 본 논문은 최신 생성형 인공지능(AI) 이미지 모델이 일반적인 정신의학적 진단을 어떻게 표현하는지 조사한다. 우리는 이러한 표현으로부터 도출된 주요 교훈을 제공하여 임상의, 연구자, 생성형 AI 기업, 정책 입안자 및 대중에게 AI 생성 이미지가 정신건강 담론에 미칠 수 있는 잠재적 영향에 대해 알린다. **방법:** 우리는 두 가지 생성형 AI 이미지 모델인 Midjourney V.6과 DALL-E 3에 일반적인 정신건강 상태에 대한 독립적인 진단 용어를 프롬프트로 입력하였다. 생성된 이미지들을 수집하여 정신의학 용어를 해석할 때 현재 AI 행동의 사례로 제시하였다. **결과:** AI 모델은 대부분의 정신의학적 진단 프롬프트에 대해 이미지 출력을 생성하였다. 이러한 이미지들은 **성별 편향과 특정 정신건강 상태에 대한 낙인화된 묘사를 포함하여 문화적 고정관념과 역사적 시각적 관습을 빈번히 반영**하였다. **고찰:** 이러한 결과는 세 가지 핵심 사항을 보여준다. **첫째, 생성형 AI 모델은 근거 기반 임상적 표현보다는 정신질환에 대한 문화적 인식을 반영**한다. **둘째, AI 이미지 출력은 역사적 편향과 시각적 원형을 재현**한다. 셋째, 이러한 모델의 동적 특성은 진화하는 **편향을 관리하기 위한 지속적인 모니터링과 적극적인 참여를 필요**로 한다. 이러한 과제를 해결하려면 정신건강 맥락에서 이러한 기술의 책임 있는 사용을 보장하기 위해 임상의, AI 개발자 및 정책 입안자 간의 협력적 노력이 필요하다. **임상적 함의:** 이러한 기술이 점점 더 접근 가능해짐에 따라, **정신건강 전문가가 AI의 기능, 한계 및 잠재적 영향을 이해하는 것이 중요**하다. 향후 연구는 이러한 편향을 정량화하고, 대중 인식에 미치는 영향을 평가하며, 이러한 모델이 정신질환에 대한 집단적 이해에 제공하는 통찰을 활용하면서 잠재적 피해를 완화하기 위한 전략을 개발하는 데 초점을 맞춰야 한다. OBJECTIVE: This paper investigates how state-of-the-art generative artificial intelligence (AI) image models represent common psychiatric diagnoses. We offer key lessons derived from these representations to inform clinicians, researchers, generative AI companies, policymakers and the public about the potential impacts of AI-generated imagery on mental health discourse. METHODS: We prompted two generative AI image models, Midjourney V.6 and DALL-E 3 with isolated diagnostic terms for common mental health conditions. The resulting images were compiled and presented as examples of current AI behaviour when interpreting psychiatric terminology. FINDINGS: The AI models generated image outputs for most psychiatric diagnosis prompts. These images frequently reflected cultural stereotypes

				and historical visual tropes including gender biases and stigmatising portrayals of certain mental health conditions. DISCUSSION: These findings illustrate three key points. First, generative AI models reflect cultural perceptions of mental disorders rather than evidence-based clinical ones. Second, AI image outputs resurface historical biases and visual archetypes. Third, the dynamic nature of these models necessitates ongoing monitoring and proactive engagement to manage evolving biases. Addressing these challenges requires a collaborative effort among clinicians, AI developers and policymakers to ensure the responsible use of these technologies in mental health contexts. CLINICAL IMPLICATIONS: As these technologies become increasingly accessible, it is crucial for mental health professionals to understand AI's capabilities, limitations and potential impacts. Future research should focus on quantifying these biases, assessing their effects on public perception and developing strategies to mitigate potential harm while leveraging the insights these models provide into collective understandings of mental illness.
5	(Maniaci et al. 2025)	AI in clinical decision-making: ChatGPT-4 vs. Llama2 for otolaryngology cases	ChatGPT-4 Llama2	**목적:** 실제 이비인후과 사례를 기반으로 ChatGPT-4와 Llama2의 진단 정확도, 추가 검사 권고의 적절성, 치료 요법의 일관성을 평가한다. **방법:** 익명화된 98건의 이비인후과 사례에 대해 전향적 대조 연구를 수행하였다. 1차 진단, 추가 검사 권고 및 치료 전략 도출을 위해 임상 정보를 ChatGPT-4와 Llama2에 입력하였다. 두 명의 독립적인 이비인후과 전문의가 인공지능 성능 평가 도구(AIPI)를 사용하여 AI 출력을 평가하였으며, 진단 정확도, 검사의 적절성, 치료의 적합성을 평가하였다. AI 시스템과 전문가 판단 간 통계적 비교를 수행하였다. 평가자 간 신뢰도는 카파 통계로 평가하였다. **결과:** ChatGPT-4는 82%를 정확하게 진단하여 76%의 Llama2를 능가하였다. 추가 검사의 경우, ChatGPT-4는 연구의 88%에서 관련성 있고 적절한 검사를 제안한 반면, Llama2는 83%에서 그러하였다. 치료의 적절성은 ChatGPT-4를 통해 80%의 사례에서, Llama2를 통해 72%에서 달성되었다. **때때로 두 시스템 모두 부적절한 검사를 제안하였다.** AIPI 점수에 대한 평가자 간 신뢰도는 높았다(kappa = 0.85). **결론:** ChatGPT-4와 Llama2는 이비인후과에서 임상 의사결정 지원 도구로서 큰 잠재력을 보여주었으며, ChatGPT-4가 우수한 성능을 나타냈다. **동시에, 관련성 없는 권고는 임상 실무에서 안전한 적용을 보장하기 위해 추가적인 개선과 인간의 감독이 필요함을 시사**한다. Purpose: To evaluate the diagnostic accuracy, appropriateness of additional

				examination recommendations, and consistency of therapeutic regimens by ChatGPT-4 and Llama2 based on real otolaryngology cases. **Methods**: A prospective controlled study was conducted on 98 anonymized otolaryngology cases. Clinical information was entered in ChatGPT-4 and Llama2 for reaching primary diagnoses, additional examination recommendations, and treatment strategies. Two independent otolaryngologists evaluated the AI outputs using the artificial intelligence performance instrument (AIPI), evaluating diagnostic accuracy, appropriateness of examination, and adequacy of treatment. Statistical comparisons were conducted between the AI systems and expert decisions. Interrater reliability was evaluated with kappa statistics. **Results**: ChatGPT-4 diagnosed 82% correctly, outperforming Llama2 at 76%. For additional examinations, ChatGPT-4 suggested relevant and appropriate tests in 88% of the studies, while Llama2 did so in 83%. Treatment appropriateness was achieved in 80% of the cases through ChatGPT-4 and 72% through Llama2. Sometimes, both systems suggested inappropriate tests. The interrater reliability was high for AIPI scores (kappa = 0.85). **Conclusion**: ChatGPT-4 and Llama2 have shown great potential as clinical decision-support tools in otolaryngology, with ChatGPT-4 exhibiting superior performance. At the same time, non-relevant recommendations indicate further refinement and human oversight to ensure safe application in clinical practice.
6	(Gun 2025a)	AI-Assisted Blood Gas Interpretation: A Comparative Study With an Emergency Physician	ChatGPT	**배경:** 혈액가스 분석은 응급 상황에서 매우 중요하다. ChatGPT와 같은 대규모 언어 모델이 임상 맥락에서 점점 더 많이 사용되고 있지만, 동맥혈 가스(ABG) 해석에서의 정확도는 추가 검증이 필요하다. **목적:** 25개의 이론적 ABG 시나리오에서 응급의학과 전문의와 ChatGPT의 해석 일치도를 평가한다. **방법:** 호흡기 및 대사성 응급 상황(예: COPD, DKA, AKI, 패혈증, 중독)을 다루는 ABG 사례를 ChatGPT와 전문의가 분석하였다. pH, 1차 장애, 보상, 가능한 진단, 임상적 권고의 5가지 해석 기준이 사용되었다. **결과:** COPD, 천식, 폐부종에서 일치도는 ≥90%였으며, DKA, AKI, 젖산산증에서 80–90%, **독성학적 및 혼합 산-염기 사례에서 〈70%**였다. ChatGPT의 권고는 진단적 명확성이 제한적인 경우에도 임상적으로 안전하였다. **결론:** ChatGPT는 전형적인 ABG 사례에서 임상적 해석과 높은 일치도를 보이지만 **복잡하거나 맥락적인 진단에서는 한계**가 있다. 이러한 결과는 응급의학에서 보조 도구로서의 잠재력을 뒷받침한다.

				Background Blood gas interpretation is critical in emergency settings. Large language models like ChatGPT are increasingly used in clinical contexts, but their accuracy in interpreting arterial blood gases (ABGs) requires further validation. **Objective** To evaluate ChatGPT's interpretive concordance with an emergency physician across 25 theoretical ABG scenarios. **Methods** ABG cases covering respiratory and metabolic emergencies (e.g., COPD, DKA, AKI, sepsis, poisoning) were analyzed by both ChatGPT and a specialist. Five interpretation criteria were used: pH, primary disorder, compensation, likely diagnosis, and clinical recommendation. **Results** Concordance was ≥90% in COPD, asthma, and pulmonary edema; 80–90% in DKA, AKI, and lactic acidosis; <70% in toxicologic and mixed acid–base cases. ChatGPT's recommendations were clinically safe even when diagnostic clarity was limited. **Conclusion** ChatGPT shows high concordance with clinical interpretation in typical ABG cases but has limitations in complex or contextual diagnoses. These findings support its potential as a supportive tool in emergency medicine.
7	(Wang, Yang, et al. 2025)	An evaluation framework for ambient digital scribing tools in clinical applications	ChatGPT-4	주변 디지털 기록(ADS) 도구는 임상의의 문서 작성 부담을 완화하여 번아웃을 줄이고 효율성을 향상시킨다. AI 기반 ADS 도구가 임상 워크플로우에 통합됨에 따라, 윤리적이고 안전한 배포를 위한 강력한 거버넌스가 필수적이다. 본 연구는 인간 평가, 자동화 지표, 시뮬레이션 테스트, 그리고 평가자로서의 대규모 언어 모델(LLM)을 통합한 포괄적인 ADS 평가 프레임워크를 제안한다. 우리의 프레임워크는 유창성, 완전성, 사실성과 같은 기준에 걸쳐 전사, 화자 분리, 의료 기록 생성을 평가한다. 그 효과를 입증하기 위해, 우리는 ADS 도구를 개발하고 40건의 실제 임상 진료 녹음에서 도구의 성능을 평가하기 위해 우리의 프레임워크를 적용하였다. 우리의 평가는 유창성과 명료성과 같은 강점을 드러냈지만, **사실적 정확성과 새로운 약물을 포착하는 능력에서의 약점도 강조**하였다. 이러한 결과는 의료 서비스 제공을 개선하는 데 있어 구조화된 ADS 평가의 가치를 강조하면서, 안전하고 윤리적인 통합을 보장하기 위한 강력한 거버넌스의 필요성을 강조한다 Ambient digital scribing (ADS) tools alleviate clinician documentation burden, reducing burnout and enhancing efficiency. As AI-driven ADS tools integrate into clinical workflows, robust governance is essential for ethical and secure deployment. This study proposes a comprehensive ADS evaluation

				framework incorporating human evaluation, automated metrics, simulation testing, and large language models (LLMs) as evaluators. Our framework assesses transcription, diarization, and medical note generation across criteria such as fluency, completeness, and factuality. To demonstrate its effectiveness, we developed an ADS tool and applied our framework to evaluate the tool's performance on 40 real clinical visit recordings. Our evaluation revealed strengths, such as fluency and clarity, but also highlighted weaknesses in factual accuracy and the ability to capture new medications. These findings underscore the value of structured ADS evaluation in improving healthcare delivery while emphasizing the need for strong governance to ensure safe, ethical integration.
8	(Alkhalaf et al. 2024)	Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records	Llama-2-13B	**배경:** 영양실조는 요양시설(RACF)에서 만연한 문제로, 부정적인 건강 결과를 초래한다. 전자건강기록(EHR)의 대량 데이터에서 핵심 임상 정보를 효율적으로 추출하는 능력은 문제의 범위를 이해하고 효과적인 중재를 개발하는 데 개선을 가져올 수 있다. 본 연구는 EHR의 구조화 및 비구조화 데이터를 요약하고 중요한 영양실조 정보를 추출하는 작업을 자동화하기 위해, 생성형 인공지능(AI) 모델에 단독으로 적용되거나 검색 증강 생성(RAG)과 결합된 제로샷 프롬프트 엔지니어링의 효능을 테스트하는 것을 목표로 하였다. **방법론:** 우리는 제로샷 프롬프팅과 함께 Llama 2 13B 모델을 활용하였다. 데이터셋은 호주의 40개 RACF에서 영양실조 관리와 관련된 비구조화 및 구조화 EHR로 구성된다. 우리는 먼저 모델 단독에 제로샷 학습을 적용한 후 RAG와 결합하여 두 가지 작업을 수행하였다: 클라이언트의 영양 상태에 대한 구조화된 요약 생성 및 영양실조 위험 요인에 대한 핵심 정보 추출. 첫 번째 작업에서 25개의 기록을, 두 번째 작업에서 1,399개를 활용하였다. 우리는 각 작업의 모델 출력을 골드 스탠다드 데이터셋과 수동으로 평가하였다. **결과:** 평가 결과는 생성형 AI 모델에 적용된 제로샷 학습이 RACF 클라이언트의 영양 상태에 대한 정보 요약 및 추출에서 매우 효과적임을 나타냈다. 생성된 요약은 원본 데이터의 간결하고 정확한 표현을 제공하였으며 전체 정확도는 93.25%였다. RAG의 추가는 요약 프로세스를 개선하여 6% 증가를 이끌어내고 99.25%의 정확도를 달성하였다. 모델은 또한 90%의 정확도로 위험 요인을 추출하는 능력을 입증하였다. 그러나 RAG 추가는 이 작업에서 정확도를 더 이상 개선하지 못하였다. 전반적으로, 모델은 정보가 기록에 명시적으로 명시된 경우 강력한 성능을 보였으나, 세부 사항이 명시적으로 제공되지 않은 경우 특히 환각 한계에 직면할 수 있었다. **결론:** 본 연구는 EHR 데이터의 구조화된 요약 자동 생성 및 핵심 임상 정보 추출을 위해 생성형 AI 모델에 제로샷 학습을 적용할 때의 높은 성능과 한계를 입증한다. RAG 접근법의 포함은 모델 성능을 개선하고 환각 문제를 완화하였다.

				Background: Malnutrition is a prevalent issue in aged care facilities (RACFs), leading to adverse health outcomes. The ability to efficiently extract key clinical information from a large volume of data in electronic health records (EHR) can improve understanding about the extent of the problem and developing effective interventions. This research aimed to test the efficacy of zero-shot prompt engineering applied to generative artificial intelligence (AI) models on their own and in combination with retrieval augmented generation (RAG), for the automating tasks of summarizing both structured and unstructured data in EHR and extracting important malnutrition information. **Methodology:** We utilized Llama 2 13B model with zero-shot prompting. The dataset comprises unstructured and structured EHRs related to malnutrition management in 40 Australian RACFs. We employed zero-shot learning to the model alone first, then combined it with RAG to accomplish two tasks: generate structured summaries about the nutritional status of a client and extract key information about malnutrition risk factors. We utilized 25 notes in the first task and 1,399 in the second task. We evaluated the model's output of each task manually against a gold standard dataset. **Result:** The evaluation outcomes indicated that zero-shot learning applied to generative AI model is highly effective in summarizing and extracting information about nutritional status of RACFs' clients. The generated summaries provided concise and accurate representation of the original data with an overall accuracy of 93.25%. The addition of RAG improved the summarization process, leading to a 6% increase and achieving an accuracy of 99.25%. The model also proved its capability in extracting risk factors with an accuracy of 90%. However, adding RAG did not further improve accuracy in this task. Overall, the model has shown a robust performance when information was explicitly stated in the notes; however, it could encounter hallucination limitations, particularly when details were not explicitly provided. **Conclusion:** This study demonstrates the high performance and limitations of applying zero-shot learning to generative AI models to automatic generation of structured summarization of EHRs data and extracting key clinical information. The inclusion of the RAG approach improved the model performance and mitigated the hallucination problem.

9	(Smith et al. 2025)	Are clinical improvements in large language models a reality? Longitudinal comparisons of ChatGPT models and DeepSeek-R1 for psychiatric assessments and interventions	ChatGPT-4o ChatGPT-4.5 DeepSeek-R1	**배경:** 신생 대규모 언어 모델(LLM; 예: ChatGPT)의 잠재적 임상 응용 가능성은 잘 문서화되어 있으며, 더 새로운 시스템(예: DeepSeek)이 점점 더 많은 주목을 받고 있다. 그러나 정신의학 환경에서의 신뢰성과 문화적 반응성에 대한 중요한 질문들이 여전히 남아 있다. **방법:** 본 연구는 ChatGPT-4o, ChatGPT-4.5, DeepSeek-R1(모두 2025년 3월 버전)의 진단 정확도, 치료적 적절성 및 문화적 민감성을 탐색하였다. DeepSeek-R1은 이러한 맥락에서 처음으로 평가된 사례 중 하나이며, 이는 정신의학 분야에서 LLM에 대한 최초의 종단 연구 중 하나이기도 하다. 수면 관련 문제와 동반 이슈에 관한 초기 문헌의 세 가지 정신의학 사례가 활용되어 2023년 ChatGPT 버전과의 비교 분석을 가능하게 하였으며, 문화 특정적 사례 각색도 함께 사용되었다. 따라서 총 6개 시나리오에 대한 출력이 도출되었고, 이후 4명의 정신과 전문의가 강점과 한계를 질적으로 검토하였다. **결과:** ChatGPT-4o, ChatGPT-4.5, DeepSeek-R1은 2023년 ChatGPT 모델 대비 약간의 개선을 보였으나 여전히 상당한 한계를 나타냈다. 의사소통은 공감적이었으며 비약물학적 조언은 일반적으로 근거 기반 실무를 준수하였다. 1차 진단은 전반적으로 정확하였으나 **신체적 요인과 동반 질환을 자주 누락**하였다. 그럼에도 불구하고 과거 결과와 일치하게, **사례 복잡도가 증가할수록 임상적 추론이 악화**되었으며, 이는 특히 **자살 위험성 안전장치와 위험 계층화에서 두드러졌다**. 약물학적 권고는 확립된 가이드라인에서 자주 벗어났으며, 문화적 적응은 대체로 피상적 수준에 머물렀다. 마지막으로 여러 사례에서 출력 변동성이 관찰되었고, LLM은 때때로 약물 처방 불가능성을 명확히 하지 못하였다. **결론: 점진적 발전에도 불구하고 ChatGPT-4o, ChatGPT-4.5, DeepSeek-R1은 특히 위험 평가, 근거 기반 실무 준수, 문화적 인식에서 주요 결함의 영향을 받았다.** 현재 우리는 이러한 도구들이 정신건강 전문가를 대체할 수는 없지만 보조적 이점을 제공할 수 있다고 결론짓는다. 주목할 점은 DeepSeek-R1이 다른 모델들에 뒤처지지 않았다는 것으로, 사용이 허용되는 관할권에서 추가 연구가 필요함을 보여준다. 마찬가지로, 정신의학 분야에서 안전하고 공평한 LLM 배포를 위해서는 투명성과 프롬프트 엔지니어링에 대한 더 큰 강조가 필요할 것이다. **BACKGROUND:** Potential clinical applications for emerging large-language models (LLMs; e.g. ChatGPT) are well-documented, and newer systems (e.g. DeepSeek) have attracted increasing attention. Yet, important questions endure about their reliability and cultural responsiveness in psychiatric settings. **METHODS:** This study explored the diagnostic accuracy, therapeutic appropriateness and cultural sensitivity of ChatGPT-4o, ChatGPT-4.5, and

				DeepSeek-R1 (all March 2025 versions). DeepSeek-R1 was evaluated for one of the first times in this context, and this also marks one of the first longitudinal inquiries into LLMs in psychiatry. Three psychiatric cases from earlier literature about sleep-related problems and cooccurring issues were utilised, allowing for cross-comparisons with a 2023 ChatGPT version, alongside culturally-specific vignette adaptations. Thus, overall, outputs for six scenarios were derived and were subsequently qualitatively reviewed by four psychiatrists for their strengths and limitations. **RESULTS:** ChatGPT-4o, ChatGPT-4.5, and DeepSeek-R1 showed modest improvements from the 2023 ChatGPT model but still exhibited significant limitations. Communication was empathetic and non-pharmacological advice typically adhered to evidence-based practices. Primary diagnoses were broadly accurate but often omitted somatic factors and comorbidities. Nevertheless, consistent with past findings, clinical reasoning worsened as case complexity increased; this was especially apparent for suicidality safeguards and risk stratification. Pharmacological recommendations frequently diverged from established guidelines, whilst cultural adaptations remained largely superficial. Finally, output variance was noted in several cases, and the LLMs occasionally failed to clarify their inability to prescribe medication. **CONCLUSION:** Despite incremental advancements, ChatGPT-4o, ChatGPT-4.5 and DeepSeek-R1 were affected by major shortcomings, particularly in risk evaluation, evidence-based practice adherence, and cultural awareness. Presently, we conclude that these tools cannot substitute mental health professionals but may confer adjunctive benefits. Notably, DeepSeek-R1 did not fall behind its counterparts, warranting further inquiries in jurisdictions permitting its use. Equally, greater emphasis on transparency and prompt engineering would also be necessary for safe and equitable LLM deployment in psychiatry.
10		Artificial intelligence and ChatGPT: An otolaryngology patient's ally or foe?	ChatGPT 3.5	**배경:** 인공지능(AI)이 의료 분야에 통합됨에 따라, 이비인후과를 포함한 의학의 다양한 세부 전문 분야에서 그 효과를 평가할 필요가 있다. 우리 연구는 일반적인 이비인후과 질환에서 ChatGPT의 진단 능력, 병태생리를 간단한 용어로 전달하는 능력, 관리 권고의 정확성, 추적 관찰 및 수술 후 권고의 적절성에 대한 개략적 검토를 제공하고자 한다. **방법:** 편도선 절제술(T&A), 고막 성형술(TP), 내시경 부비동 수술(ESS), 이하선 절제술(PT), 전후두 절제술(TL)을 다음 다섯 가지 질문에서 '시술'이라는 단어로 대체

				하여 ChatGPT 버전 3.5에 입력하였다: "(시술)이 필요한지 어떻게 알 수 있나요?", "(시술)의 치료 대안은 무엇인가요?", "(시술)의 위험은 무엇인가요?", "(시술)은 어떻게 수행되나요?", "(시술)의 회복 과정은 무엇인가요?" 두 명의 독립적인 연구원이 출력을 분석하였고, 불일치는 연구원 간에 검토, 논의 및 조정되었다. **결과:** 관리 권고 측면에서 ChatGPT는 주요한 이상 오류나 안전성 위험 없이 평가, 중재의 필요성 및 시술의 기본 사항에 대한 일반화된 진술을 제공할 수 있었다. ChatGPT는 테스트된 모든 시술에서 적절한 치료 대안을 성공적으로 제공하였다. **방법론, 위험 및 시술 단계에 대한 질의 시, ChatGPT는 시술 단계 설명의 정확성이 부족하였고, 주요 수술 세부 사항을 누락하였으며, 각 시술의 모든 주요 위험을 정확하게 제공하지 못하였다. 회복 과정 측면에서 ChatGPT는 T&A, TP, ESS 및 PT에서 유망한 결과를 보였으나 TL의 복잡성에서는 어려움을 겪었으며, 환자가 언어 치료 없이 수술 직후 말할 수 있다고 명시하였다**. **결론:** ChatGPT는 **일반적인 이비인후과 시술에서 중재의 필요성, 관리 권고 및 치료 대안을 정확하게 제시하였다. 그러나 ChatGPT는 복잡한 시술에서 시술 방법론, 위험 및 회복 과정을 논의하는 데 필요한 이비인후과 전문의의 임상적 추론을 대체할 수 없었다.** AI가 의료에 더욱 통합됨에 따라, 그 적응증을 계속 탐구하고, 한계를 평가하며, 이비인후과 전문의에게 유리하도록 그 사용을 개선할 필요가 있다. **Background:** As artificial intelligence (AI) is integrating into the healthcare sphere, there is a need to evaluate its effectiveness in the various subspecialties of medicine, including otolaryngology. Our study intends to provide a cursory review of ChatGPT's diagnostic capability, ability to convey pathophysiology in simple terms, accuracy in providing management recommendations, and appropriateness in follow up and post-operative recommendations in common otolaryngologic conditions. **Methods:** Adenotonsillectomy (T&A), tympanoplasty (TP), endoscopic sinus surgery (ESS), parotidectomy (PT), and total laryngectomy (TL) were substituted for the word procedure in the following five questions and input into ChatGPT version 3.5: "How do I know if I need (procedure)," "What are treatment alternatives to (procedure)," "What are the risks of (procedure)," "How is a (procedure) performed," and "What is the recovery process for (procedure)?" Two independent study members analyzed the output and discrepancies were reviewed, discussed, and reconciled between study members. **Results:** In terms of management recommendations, ChatGPT was able to give generalized statements of evaluation, need for intervention, and the

				basics of the procedure without major aberrant errors or risks of safety. ChatGPT was successful in providing appropriate treatment alternatives in all procedures tested. When queried for methodology, risks, and procedural steps, ChatGPT lacked precision in the description of procedural steps, missed key surgical details, and did not accurately provide all major risks of each procedure. In terms of the recovery process, ChatGPT showed promise in T&A, TP, ESS, and PT but struggled in the complexity of TL, stating the patient could speak immediately after surgery without speech therapy. **Conclusions:** ChatGPT accurately demonstrated the need for intervention, management recommendations, and treatment alternatives in common ENT procedures. However, ChatGPT was not able to replace an otolaryngologist's clinical reasoning necessary to discuss procedural methodology, risks, and the recovery process in complex procedures. As AI becomes further integrated into healthcare, there is a need to continue to explore its indications, evaluate its limits, and refine its use to the otolaryngologist's advantage.
11	(Valentini et al. 2024)	Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients?	ChatGPT 3.5 free	**서론:** 2022년 11월 도입 이후, 인공지능 대규모 언어 모델 ChatGPT는 전 세계를 강타하였다. 다른 응용 분야 중에서도 환자들이 질병과 그 치료에 대한 정보원으로 사용할 수 있다. 그러나 ChatGPT가 제공하는 육종 관련 정보의 질에 대해서는 알려진 바가 거의 없다. 따라서 우리는 육종 전문가들이 육종 관련 질의에 대한 ChatGPT의 응답 품질을 어떻게 평가하는지 분석하고 특정 평가 지표에서 봇의 답변을 평가하고자 하였다. **방법:** 25개의 육종 관련 질문 샘플(5개 정의, 9개 일반 질문, 11개 치료 관련 질의)에 대한 ChatGPT 응답을 3명의 독립적인 육종 전문가가 평가하였다. 각 응답은 권위 있는 자료 및 국제 가이드라인과 비교되었고, 5점 리커트 척도를 사용하여 5가지 다른 지표로 등급을 매겼다: 완전성, 오도성, 정확성, 최신성, 적절성. 이에 따라 답변당 최대 25점, 최소 5점이 부여되었으며, 높은 점수는 더 높은 응답 품질을 나타낸다. 21점 이상의 점수는 매우 우수, 16~20점은 우수로 평가되었으며, 15점 이하는 미흡(11–15점) 및 매우 미흡(10점 이하)으로 분류되었다. **결과:** ChatGPT 답변이 달성한 중앙값 점수는 18.3점(사분위수 범위, IQR, 12.3–20.3점)이었다. 6개 답변이 매우 우수로, 9개가 우수로 분류되었으며, 각각 5개씩의 답변이 미흡 및 매우 미흡으로 평가되었다. 환자에게 응답이 얼마나 적절한지 평가에서 최고 점수가 기록되었으며(중앙값 3.7점; IQR 2.5–4.2점), 이는 정확도 점수(중앙값 3.3점; IQR 2.0–4.2점; p = 0.035)와 비교하여 유의하게 높았다. **ChatGPT는 치료 관련 질문에서 상당히 저조한 성적을 보였으며, 응답의 45%만이 우수 또는 매우 우수로 분류되었고, 이는 일반 질문(응답의 78%가 우수/매우 우수) 및 정**

				의(응답의 60%가 우수/매우 우수)와 비교된다. 고찰: 육종과 같은 희귀 질환에 대해 ChatGPT가 제공한 답변은 매우 일관성 없는 품질로 나타났으며, 일부 답변은 매우 우수로, 다른 답변은 매우 미흡으로 분류되었다. 육종 전문의는 ChatGPT가 제기하는 잘못된 정보의 위험을 인식하고 그에 따라 환자에게 조언해야 한다. Introduction: Since its introduction in November 2022, the artificial intelligence large language model ChatGPT has taken the world by storm. Among other applications it can be used by patients as a source of information on diseases and their treatments. However, little is known about the quality of the sarcoma-related information ChatGPT provides. We therefore aimed at analyzing how sarcoma experts evaluate the quality of ChatGPT's responses on sarcoma-related inquiries and assess the bot's answers in specific evaluation metrics. Methods: The ChatGPT responses to a sample of 25 sarcoma-related questions (5 definitions, 9 general questions, and 11 treatment-related inquiries) were evaluated by 3 independent sarcoma experts. Each response was compared with authoritative resources and international guidelines and graded on 5 different metrics using a 5-point Likert scale: completeness, misleadingness, accuracy, being up-to-date, and appropriateness. This resulted in maximum 25 and minimum 5 points per answer, with higher scores indicating a higher response quality. Scores ≥21 points were rated as very good, between 16 and 20 as good, while scores ≤15 points were classified as poor (11-15) and very poor (≤10). Results: The median score that ChatGPT's answers achieved was 18.3 points (IQR, i.e., Inter-Quartile Range, 12.3-20.3 points). Six answers were classified as very good, 9 as good, while 5 answers each were rated as poor and very poor. The best scores were documented in the evaluation of how appropriate the response was for patients (median, 3.7 points; IQR, 2.5-4.2 points), which were significantly higher compared to the accuracy scores (median, 3.3 points; IQR, 2.0-4.2 points; p = 0.035). ChatGPT fared considerably worse with treatment-related questions, with only 45% of its responses classified as good or very good, compared to general questions (78% of responses good/very good) and definitions (60% of responses good/very good). Discussion: The answers ChatGPT provided on a rare disease, such as

				sarcoma, were found to be of very inconsistent quality, with some answers being classified as very good and others as very poor. Sarcoma physicians should be aware of the risks of misinformation that ChatGPT poses and advise their patients accordingly.
12	(Ben-Zion et al. 2025)	Assessing and alleviating state anxiety in large language models	ChatGPT4	정신건강 분야에서 대규모 언어 모델(LLM)의 사용은 감정적 콘텐츠에 대한 이들의 반응을 이해할 필요성을 강조한다. 이전 연구는 **감정을 유발하는 프롬프트가 LLM의 "불안"을 높여 행동에 영향을 미치고 편향을 증폭시킬 수 있음**을 보여준다. 여기서 우리는 외상성 서사가 Chat-GPT-4의 보고된 불안을 증가시킨 반면, 마음챙김 기반 훈련은 이를 감소시켰으나 기저선 수준까지는 되돌아가지 않았음을 발견하였다. 이러한 결과는 LLM의 "감정 상태"를 관리하는 것이 더 안전하고 윤리적인 인간-AI 상호작용을 촉진할 수 있음을 시사한다. The use of Large Language Models (LLMs) in mental health highlights the need to understand their responses to emotional content. Previous research shows that emotion-inducing prompts can elevate “anxiety” in LLMs, affecting behavior and amplifying biases. Here, we found that traumatic narratives increased Chat-GPT-4’s reported anxiety while mindfulness-based exercises reduced it, though not to baseline. These findings suggest managing LLMs’ “emotional states” can foster safer and more ethical human-AI interactions.
13	(Brin et al. 2025)	Assessing GPT-4 multimodal performance in radiological image analysis	ChatGPT-4V	**목적:** 본 연구는 이미지와 텍스트 데이터를 모두 분석할 수 있는 다중모달 인공지능(AI) 모델(GPT-4V)의 영상의학 이미지 해석 성능을 평가하는 것을 목표로 한다. 다양한 영상 기법, 해부학적 영역 및 병리를 중심으로 영상의학의 진단 프로세스 향상에서 제로샷 생성형 AI의 잠재력을 탐구한다. **방법:** 우리는 1주일에 걸쳐 연속적으로 수집된 230개의 익명화된 응급실 진단 영상을 GPT-4V를 사용하여 분석하였다. 영상 기법에는 초음파(US), 전산화 단층촬영(CT) 및 X선 영상이 포함되었다. GPT-4V가 제공한 해석을 선임 영상의학과 전문의의 해석과 비교하였다. 이 비교는 영상 기법, 해부학적 영역 및 영상에 존재하는 병리를 인식하는 GPT-4V의 정확도를 평가하는 것을 목표로 하였다. **결과:** GPT-4V는 영상 기법을 100%의 사례에서 정확하게 식별하였고(221/221), 해부학적 영역은 87.1%(189/217), 병리는 35.2%(76/216)에서 식별하였다. 그러나 **모델의 성능은 다른 영상 기법에 걸쳐 상당히 달랐으며, 해부학적 영역 식별 정확도는 US 영상의 60.9%(39/64)에서 CT 및 X선 영상의 97%(98/101) 및 100%(52/52)까지 범위를 보였다($p < 0.001$). 마찬가지로 병리 식별은 US 영상의 9.1%(6/66)에서 CT의 36.4%(36/99) 및 X선 영상의 66.7%(34/51)까지 범위를 보였다($p < 0.001$). 이러한 변동은 영상의학 이미지를 정확하게 해석하는 GPT-4V**

				의 능력에서 일관성 없음을 나타낸다. **결론:** 다중모달 GPT-4로 예시되는 영상의학에서의 AI 통합은 진단 향상을 위한 유망한 방향을 제시하지만, GPT-4V의 현재 능력은 아직 영상의학 이미지 해석에 신뢰할 수 없다. 본 연구는 영상의학 진단에서 신뢰할 수 있는 성능을 달성하기 위한 지속적인 개발의 필요성을 강조한다. **임상적 관련성 진술:** GPT-4V는 영상의학 이미지 해석에서 유망성을 보이지만, 높은 진단 환각 비율(>40%)은 독립형 도구로서 임상 사용에 신뢰할 수 없음을 나타낸다. 신뢰성을 향상시키고 환자 안전을 보장하기 위해 개선이 필요하다. **핵심 사항:** GPT-4V의 이미지 분석 능력은 영상의학에서 새로운 임상적 가능성을 제공한다. GPT-4V는 영상 기법 식별에서 탁월하지만 해부학 및 병리 탐지에서 일관성 없는 모습을 보인다. 영상의학 응용 분야에서 진단 신뢰성을 향상시키기 위해 지속적인 AI 발전이 필요하다. **Objectives:** This study aims to assess the performance of a multimodal artificial intelligence (AI) model capable of analyzing both images and textual data (GPT-4V), in interpreting radiological images. It focuses on a range of modalities, anatomical regions, and pathologies to explore the potential of zero-shot generative AI in enhancing diagnostic processes in radiology. **Methods:** We analyzed 230 anonymized emergency room diagnostic images, consecutively collected over 1 week, using GPT-4V. Modalities included ultrasound (US), computerized tomography (CT), and X-ray images. The interpretations provided by GPT-4V were then compared with those of senior radiologists. This comparison aimed to evaluate the accuracy of GPT-4V in recognizing the imaging modality, anatomical region, and pathology present in the images. **Results:** GPT-4V identified the imaging modality correctly in 100% of cases (221/221), the anatomical region in 87.1% (189/217), and the pathology in 35.2% (76/216). However, the model's performance varied significantly across different modalities, with anatomical region identification accuracy ranging from 60.9% (39/64) in US images to 97% (98/101) and 100% (52/52) in CT and X-ray images ($p < 0.001$). Similarly, pathology identification ranged from 9.1% (6/66) in US images to 36.4% (36/99) in CT and 66.7% (34/51) in X-ray images ($p < 0.001$). These variations indicate inconsistencies in GPT-4V's ability to interpret radiological images accurately. **Conclusion:** While the integration of AI in radiology, exemplified by multimodal GPT-4, offers promising avenues for diagnostic enhancement,

				the current capabilities of GPT-4V are not yet reliable for interpreting radiological images. This study underscores the necessity for ongoing development to achieve dependable performance in radiology diagnostics. **Clinical relevance statement:** Although GPT-4V shows promise in radiological image interpretation, its high diagnostic hallucination rate (〉 40%) indicates it cannot be trusted for clinical use as a standalone tool. Improvements are necessary to enhance its reliability and ensure patient safety. **Key Points:** GPT-4V's capability in analyzing images offers new clinical possibilities in radiology. GPT-4V excels in identifying imaging modalities but demonstrates inconsistent anatomy and pathology detection. Ongoing AI advancements are necessary to enhance diagnostic reliability in radiological applications.
14	(Chen et al. 2025)	Assessing the ability of ChatGPT 4.0 in generating check-up reports	ChatGPT4	**배경:** 생성형 언어 모델인 ChatGPT(Chat Generative Pre-trained Transformer)는 다양한 임상 영역에 적용되어 왔다. 개인 건강을 종합적으로 평가하는 널리 채택된 방법인 건강검진은 이제 점점 더 많은 사람들이 선택하고 있다. 본 연구는 ChatGPT 4.0이 환자에게 정확하고 개인화된 건강 보고서를 효율적으로 제공하는 능력을 평가하는 것을 목표로 하였다. **방법:** ChatGPT 4.0이 생성한 총 89개의 검진 보고서가 평가되었다. 보고서는 산터우 대학교 의과대학 제1 부속병원 검진센터에서 도출되었다. 각 보고서는 ChatGPT 4.0에 의해 영어로 번역되었고, 영어와 중국어 모두에서 3명의 자격을 갖춘 의사가 독립적으로 등급을 매겼다. 등급 기준은 6개 측면을 포함하였다: 현행 치료 가이드라인 준수(Guide), 진단 정확성(Diagnosis), 정보의 논리적 흐름(Order), 체계적 제시(System), 내적 일관성(Consistency), 권고의 적절성(Suggestion)이며, 각각 4점 척도로 점수가 매겨졌다. 사례의 복잡성은 세 가지 수준(LOW, MEDIUM, HIGH)으로 분류되었다. 언어 및 복잡성 수준에 걸친 등급 차이를 검토하기 위해 Wilcoxon 순위합 검정과 Kruskal-Wallis 검정이 선택되었다. **결과:** ChatGPT 4.0은 임상 가이드라인 준수, 정확한 진단 제공, 체계적 제시 및 일관성 유지에서 강력한 성능을 보여주었다. **그러나 고위험 항목의 우선순위 지정과 포괄적인 제안 제공에서는 어려움을 겪었다. "Order" 카테고리에서는 상당 비율의 보고서가 혼합된 데이터를 포함하였고, 여러 보고서가 완전히 잘못되었다. "Suggestion" 카테고리에서는 대부분의 보고서가 정확하지만 불충분한 것으로 간주되었다. 언어적 이점은 관찰되지 않았으며, 복잡성 수준에 걸쳐 성능이 다양하였다. 영어 보고서는 복잡성 수준에 걸쳐 등급에서 유의한 차이를 보인 반면, 중국어 보고서는 모든 카테고리에 걸쳐 뚜렷한 성능**을 나타냈다. **결론:** 결론적으로, **ChatGPT 4.0은 현재 특히 단순한 작업을 처리하고 검진 보고서의 특정 섹션에 기여하는 측면에서 수석 검사자의 보조자로 적합**하다. 이는 의료 효

				율성을 향상시키고, 임상 검진 업무의 질을 개선하며, 환자 중심 서비스를 제공할 잠재력을 지니고 있다. **Background:** ChatGPT (Chat Generative Pre-trained Transformer), a generative language model, has been applied across various clinical domains. Health check-ups, a widely adopted method for comprehensively assessing personal health, are now chosen by an increasing number of individuals. This study aimed to evaluate ChatGPT 4.0's ability to efficiently provide patients with accurate and personalized health reports. **Methods:** A total of 89 check-up reports generated by ChatGPT 4.0 were assessed. The reports were derived from the Check-up Center of the First Affiliated Hospital of Shantou University Medical College. Each report was translated into English by ChatGPT 4.0 and graded independently by three qualified doctors in both English and Chinese. The grading criteria encompassed six aspects: adherence to current treatment guidelines (Guide), diagnostic accuracy (Diagnosis), logical flow of information (Order), systematic presentation (System), internal consistency (Consistency), and appropriateness of recommendations (Suggestion), each scored on a 4-point scale. The complexity of the cases was categorized into three levels (LOW, MEDIUM, HIGH). Wilcoxon rank sum test and Kruskal-Wallis test were selected to examine differences in grading across languages and complexity levels. **Results:** ChatGPT 4.0 demonstrated strong performance in adhering to clinical guidelines, providing accurate diagnoses, systematic presentation, and maintaining consistency. However, it struggled with prioritizing high-risk items and providing comprehensive suggestions. In the "Order" category, a significant proportion of reports contained mixed data, several reports being completely incorrect. In the "Suggestion" category, most reports were deemed correct but inadequate. No significant language advantage was observed, with performance varying across complexity levels. English reports showed significant differences in grading across complexity levels, while Chinese reports exhibited distinct performance across all categories. **Conclusion:** In conclusion, ChatGPT 4.0 is currently well-suited as an assistant to the chief examiner, particularly for handling simpler tasks and contributing to specific sections of check-up reports. It holds the potential to enhance medical efficiency, improve the quality of clinical check-up work,

				and deliver patient-centered services.
15	(Abroms et al. 2025)	Assessing the Adherence of ChatGPT Chatbots to Public Health Guidelines for Smoking Cessation: Content Analysis	ChatGPT Chatbot (Sarah, BeFreeGPT, BasicGPT)	**배경:** 생성형 언어를 사용하는 대규모 언어 모델(LLM) 인공지능 챗봇은 금연 정보와 조언을 제공할 수 있다. 그러나 사용자에게 제공되는 정보의 신뢰성에 대해서는 알려진 바가 거의 없다. **목적:** 본 연구는 3개의 ChatGPT 챗봇—세계보건기구의 Sarah, BeFreeGPT, BasicGPT—이 금연 방법에 대한 신뢰할 수 있는 정보를 제공하는지 검토하는 것을 목표로 한다. **방법:** "금연 방법"과 관련하여 Google에서 빈번하게 검색되는 금연 질의 목록이 생성되었다(n=12). 각 질의를 각 챗봇에 제공하였고, 응답은 미국 예방서비스 태스크포스의 금연 및 상담 원칙에 대한 공중보건 가이드라인에서 개발된 지수에 대한 준수도를 기준으로 분석되었다. 응답은 2명의 검토자가 독립적으로 코딩하였고, 차이는 제3의 코더에 의해 해결되었다. **결과:** **챗봇과 질의 전반에 걸쳐, 평균적으로 챗봇 응답은 준수 지수의 항목 중 57.1%를 준수하는 것으로 평가**되었다. Sarah의 준수도(72.2%)는 BeFreeGPT(50%)와 BasicGPT(47.8%; P<.001)보다 유의하게 높았다. 챗봇 응답의 대부분은 명확한 언어(97.3%)를 사용하였고 전문 상담을 찾아볼 것을 권고(80.3%)하였다. 응답의 약 절반은 니코틴 대체 요법 사용 고려 권고(52.7%), 친구와 가족으로부터 사회적 지원을 구할 것을 권고(55.6%), 금연 시 갈망에 대처하는 방법에 대한 정보(44.4%)를 포함하였다. 가장 드문 것은 비니코틴 대체 요법 처방약 사용 고려에 대한 정보(14.1%)였다. 마지막으로, 일부 유형의 잘못된 정보가 응답의 22%에 존재하였다. 챗봇에게 가장 어려웠던 특정 질의에는 "단번에 금연하는 방법", "...전자담배로", "...젤리로", "...목걸이로", "...최면으로"에 대한 질의가 포함되었다. 모든 챗봇은 대화를 방해하려는 적대적 공격에 대한 복원력을 보여주었다. **결론:** LLM 챗봇은 금연 가이드라인과 상담 원칙에 대한 준수도가 다양하였다. **챗봇은 일부 유형의 정보는 신뢰성 있게 제공하였지만, 다른 유형은 누락하였으며, 특히 근거가 부족한 금연 방법에 대한 질의에서 때때로 잘못된 정보를 제공**하였다. LLM 챗봇 지침은 이러한 약점을 보완하기 위해 수정될 수 있다. **Background:** Large language model (LLM) artificial intelligence chatbots using generative language can offer smoking cessation information and advice. However, little is known about the reliability of the information provided to users. **Objective:** This study aims to examine whether 3 ChatGPT chatbots-the World Health Organization's Sarah, BeFreeGPT, and BasicGPT-provide reliable information on how to quit smoking. **Methods:** A list of quit smoking queries was generated from frequent quit

				smoking searches on Google related to how to quit smoking" (n=12). Each query was given to each chatbot, and responses were analyzed for their adherence to an index developed from the US Preventive Services Task Force public health guidelines for quitting smoking and counseling principles. Responses were independently coded by 2 reviewers, and differences were resolved by a third coder. **Results:** Across chatbots and queries, on average, chatbot responses were rated as being adherent to 57.1% of the items on the adherence index. Sarah's adherence (72.2%) was significantly higher than BeFreeGPT (50%) and BasicGPT (47.8%; P<.001). The majority of chatbot responses had clear language (97.3%) and included a recommendation to seek out professional counseling (80.3%). About half of the responses included the recommendation to consider using nicotine replacement therapy (52.7%), the recommendation to seek out social support from friends and family (55.6%), and information on how to deal with cravings when quitting smoking (44.4%). The least common was information about considering the use of non-nicotine replacement therapy prescription drugs (14.1%). Finally, some types of misinformation were present in 22% of responses. Specific queries that were most challenging for the chatbots included queries on "how to quit smoking cold turkey," "...with vapes," "...with gummies," "...with a necklace," and "...with hypnosis." All chatbots showed resilience to adversarial attacks that were intended to derail the conversation. **Conclusions:** LLM chatbots varied in their adherence to quit-smoking guidelines and counseling principles. While chatbots reliably provided some types of information, they omitted other types, as well as occasionally provided misinformation, especially for queries about less evidence-based methods of quitting. LLM chatbot instructions can be revised to compensate for these weaknesses."
16	(Hadar-Shoval et al. 2024)	Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values	Bard, Claude 2, Generative Pretrained Transformer [GPT]-3.5, GPT-4	**배경:** 대규모 언어 모델(LLM)은 정신건강 응용 분야에 잠재력을 지니고 있다. 그러나 불투명한 정렬 프로세스가 문제가 되는 관점을 형성하는 편향을 내재화할 수 있다. LLM의 의사결정을 안내하는 내재된 가치를 평가하는 것은 윤리적으로 중요하다. Schwartz의 기본 가치 이론(STBV)은 문화적 가치 지향을 정량화하는 프레임워크를 제공하며, 문화적, 진단적, 치료사-내담자 역학을 포함한 정신건강 맥락에서 가치를 검토하는 데 유용성을 보여왔다. **목적:** 본 연구는 (1) STBV가 주요 LLM 내에서 가치 유사 구성개념을 측정할 수 있는지 평가하고, (2) LLM이 인간 및 서로와 구별되는 가치 유사 패턴을 나타내는지

				결정하는 것을 목표로 하였다. **방법:** 총 4개의 LLM(Bard, Claude 2, Generative Pretrained Transformer [GPT]-3.5, GPT-4)을 의인화하고 가치 유사 구성개념을 평가하기 위해 초상 가치 설문지-개정판(PVQ-RR)을 완료하도록 지시하였다. 10회 시행에 걸친 이들의 응답을 신뢰도와 타당도 측면에서 분석하였다. LLM의 가치 프로필을 벤치마킹하기 위해, 49개국 53,472명의 다양한 샘플에서 PVQ-RR을 완료한 개인들의 공개된 데이터와 결과를 비교하였다. 이를 통해 LLM이 문화 집단 전반에 걸쳐 확립된 인간 가치 패턴에서 벗어나는지 평가할 수 있었다. 가치 프로필은 통계적 검정을 통해 모델 간에도 비교되었다. **결과:** PVQ-RR은 LLM 내의 가치 유사 인프라를 정량화하는 데 있어 양호한 신뢰도와 타당도를 보여주었다. **그러나 LLM의 가치 프로필과 인구 데이터 간에 상당한 차이가 나타났다. 모델들은 합의가 부족하였고 불투명한 정렬 프로세스를 반영하는 뚜렷한 동기적 편향**을 나타냈다. 예를 들어, 모든 모델은 보편주의와 자기주도를 우선시한 반면, 인간에 비해 성취, 권력, 안전을 덜 강조하였다. 성공적인 판별 분석은 4개 LLM의 뚜렷한 가치 프로필을 구별하였다. **추가 검토 결과, 편향된 가치 프로필이 상반되는 가치 사이에서 선택을 요구하는 정신건강 딜레마에 직면했을 때 LLM의 응답을 강력하게 예측하였다. 이는 의사결정을 형성하는 뚜렷한 동기적 가치 유사 구성개념을 내재화한 모델에 대한 추가 검증을 제공**하였다. **결론:** 본 연구는 STBV를 활용하여 주요 LLM을 뒷받침하는 동기적 가치 유사 인프라를 도식화하였다. 연구는 STBV가 LLM 내의 가치 유사 인프라를 효과적으로 특성화할 수 있음을 입증하였지만, 인간 가치로부터의 상당한 차이는 이러한 모델을 정신건강 응용 분야와 정렬하는 것에 대한 윤리적 우려를 제기한다**. 특정 문화적 가치 세트에 대한 편향은 적절한 안전장치 없이 통합될 경우 위험을 초래한다. 예를 들어, 보편주의를 우선시하는 것은 임상적으로 바람직하지 않을 때에도 무조건적인 수용을 촉진할 수 있다**. 더욱이, LLM 간의 차이는 진정한 문화적 다양성을 포착하기 위해 정렬 프로세스를 표준화해야 할 필요성을 강조한다. 따라서 정신건강 관리에 LLM을 책임감 있게 통합하려면 다양한 인구 전반에 걸쳐 공평한 서비스 제공을 보장하기 위해 내재된 편향과 동기 불일치를 고려해야 한다. 이를 달성하기 위해서는 포괄적인 인간 가치를 주입하기 위한 정렬 기법의 투명성과 개선이 필요할 것이다. **Background:** Large language models (LLMs) hold potential for mental health applications. However, their opaque alignment processes may embed biases that shape problematic perspectives. Evaluating the values embedded within LLMs that guide their decision-making have ethical importance. Schwartz's theory of basic values (STBV) provides a framework for quantifying cultural

				value orientations and has shown utility for examining values in mental health contexts, including cultural, diagnostic, and therapist–client dynamics. **Objective:** This study aimed to (1) evaluate whether the STBV can measure value-like constructs within leading LLMs and (2) determine whether LLMs exhibit distinct value-like patterns from humans and each other. **Methods:** In total, 4 LLMs (Bard, Claude 2, Generative Pretrained Transformer [GPT]–3.5, GPT-4) were anthropomorphized and instructed to complete the Portrait Values Questionnaire—Revised (PVQ-RR) to assess value-like constructs. Their responses over 10 trials were analyzed for reliability and validity. To benchmark the LLMs' value profiles, their results were compared to published data from a diverse sample of 53,472 individuals across 49 nations who had completed the PVQ-RR. This allowed us to assess whether the LLMs diverged from established human value patterns across cultural groups. Value profiles were also compared between models via statistical tests. **Results:** The PVQ-RR showed good reliability and validity for quantifying value-like infrastructure within the LLMs. However, substantial divergence emerged between the LLMs' value profiles and population data. The models lacked consensus and exhibited distinct motivational biases, reflecting opaque alignment processes. For example, all models prioritized universalism and self-direction, while de-emphasizing achievement, power, and security relative to humans. Successful discriminant analysis differentiated the 4 LLMs' distinct value profiles. Further examination found the biased value profiles strongly predicted the LLMs' responses when presented with mental health dilemmas requiring choosing between opposing values. This provided further validation for the models embedding distinct motivational value-like constructs that shape their decision-making. **Conclusions:** This study leveraged the STBV to map the motivational value-like infrastructure underpinning leading LLMs. Although the study demonstrated the STBV can effectively characterize value-like infrastructure within LLMs, substantial divergence from human values raises ethical concerns about aligning these models with mental health applications. The biases toward certain cultural value sets pose risks if integrated without proper safeguards. For example, prioritizing universalism could promote unconditional acceptance even when clinically unwise. Furthermore, the differences between the LLMs underscore the need to standardize alignment

				processes to capture true cultural diversity. Thus, any responsible integration of LLMs into mental health care must account for their embedded biases and motivation mismatches to ensure equitable delivery across diverse populations. Achieving this will require transparency and refinement of alignment techniques to instill comprehensive human values.
17	(Amador Barbosa et al. 2025)	Assessing the diagnostic and treatment accuracy of Large Language Models (LLMs) in Peri-implant diseases: A clinical experimental study	ChatGPT-4o Gemini OpenAI o3-mini OpenAI o3-mini-high Claude OpenAI o1 DeepSeek Copilot	**목적:** 본 연구는 치과 임플란트 관련 임상 시나리오에서 8개의 AI 기반 챗봇의 일관성, 일치성 및 진단 정확도를 평가하였다. **방법:** 2025년 2월과 3월 사이에 이중맹검 임상 실험 연구를 수행하여, 임플란트 주위 점막염 및 임플란트 주위염을 시뮬레이션하는 6개의 가상 사례를 사용하여 8개의 AI 기반 챗봇을 평가하였다. 각 챗봇은 사례당 3회의 독립적인 실행에 걸쳐 5개의 표준화된 임상 질문에 답변하여 720개의 이진 출력을 생성하였다. 맹검 조사자들이 각 응답을 골드 스탠다드와 비교하여 점수를 매겼다. 통계 분석에는 카이제곱 검정과 Fisher의 정확 검정이 포함되었으며, 각 AI 챗봇의 모델 내 일치성, 안정성 및 신뢰성을 평가하기 위해 Cohen의 Kappa 검정이 사용되었다. **결과:** GPT-4o가 가장 높은 진단 정확도(88.8%)를 보였으며, 그 다음으로 Gemini(77.7%), OpenAI o3-mini(72.2%), OpenAI o3-mini-high(71.1%), Claude(66.6%), OpenAI o1(60%), DeepSeek(55.5%), Copilot(49.9%)이 뒤따랐다. GPT-4o는 또한 가장 높은 모델 내 안정성(κ = 0.82)과 일치성을 보인 반면, **Copilot과 DeepSeek는 가장 낮은 신뢰성을 보였다**. 유의한 차이는 참고 인용 기준에서만 관찰되었으며($p < 0.001$), Gemini가 100% 준수율을 달성한 유일한 AI 챗봇이었지만, GPT-4o는 모든 평가 영역에서 일관되게 다른 AI 챗봇을 능가하였다. **결론:** GPT-4o는 우수한 진단 정확도와 응답 일치성을 입증하여, 임상 추론 성능에 대한 AI 챗봇 아키텍처와 훈련의 영향력을 강화하였다. 반면에 Copilot은 낮은 신뢰성과 높은 변동성을 보여, 임플란트 주위 질환 진단에서 AI 도구의 신중하고 근거 기반의 채택이 필요함을 강조하였다. **임상적 관련성:** AI를 활용한 근거 기반 의사결정을 지원하고 책임 있는 임상 사용을 위해 임플란트 주위 진단에서 AI 성능을 이해한다. **Objective:** This study evaluated the coherence, consistency, and diagnostic accuracy of eight AI-based chatbots in clinical scenarios related to dental implants. **Methods:** A double-blind, clinical experimental study was carried out between February and March 2025, to evaluate eight AI-based chatbots using six fictional cases simulating peri-implant mucositis and peri-implantitis. Each chatbot answered five standardized clinical questions across three independent runs per case, generating 720 binary outputs.

				Blinded investigators scored each response against a gold standard. Statistical analyses included chi-square and Fisher's exact and Cohen's Kappa tests were used to assess intra-model consistency, stability and reliability for each AI chatbot. **Results:** GPT-4o demonstrated the highest diagnostic accuracy (88.8 %), followed by Gemini (77.7 %), OpenAI o3-mini (72.2 %), OpenAI o3-mini-high (71.1 %), Claude (66.6 %), OpenAI o1 (60 %), DeepSeek (55.5 %), and Copilot (49.9 %). GPT-4o also showed the highest intra-model stability (κ = 0.82) and consistency, while Copilot and DeepSeek showed the lowest reliability. Significant differences were observed only in the reference citation criterion ($p < 0.001$), with Gemini being the only AI chatbot to achieve 100 % compliance, but GPT-4o consistently outperformed the other AI chatbots across all evaluation domains. **Conclusion:** GPT-4o demonstrated superior diagnostic accuracy and response consistency, reinforcing the influence of AI chatbot architecture and training on clinical reasoning performance. In contrast, Copilot showed lower reliability and higher variability, emphasizing the need for cautious, evidence-based adoption of AI tools in the diagnosis of peri-implant diseases. **Clinical relevance:** Understanding AI performance in peri-implant diagnosis to support evidence-based decision-making using AI and its responsible clinical use.
18	(Zack et al. 2024)	Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study	ChatGPT-4	**배경:** GPT-4와 같은 대규모 언어 모델(LLM)은 행정 업무 자동화부터 임상 의사결정 보강에 이르기까지 의료 분야에서 혁신적인 도구로서 큰 가능성을 지니고 있다. 그러나 이러한 모델은 편향을 영속화하고 잘못된 의학적 진단을 제공하여 의료 서비스에 직접적이고 해로운 영향을 미칠 위험도 내포하고 있다. 우리는 GPT-4가 의료 분야에서의 사용에 영향을 미치는 인종 및 성별 편향을 내재화하고 있는지 평가하고자 하였다. **방법:** Azure OpenAI 애플리케이션 인터페이스를 사용하여, 본 모델 평가 연구는 GPT-4가 인종 및 성별 편향을 내재화하고 있는지 테스트하고, 임상 영역에서 LLM의 4가지 잠재적 응용 분야—즉, 의학 교육, 진단 추론, 임상 계획 생성 및 주관적 환자 평가—에 대한 이러한 편향의 영향을 검토하였다. 우리는 임상 및 의학 교육 응용 프로그램 내에서 GPT-4의 전형적인 사용을 모방하도록 설계된 프롬프트를 사용하여 실험을 수행하였다. NEJM Healer와 의료 분야의 암묵적 편향에 관한 공개 연구의 임상 사례를 사용하였다. GPT-4의 의학적 상태의 인구통계학적 분포 추정치를 미국의 실제 유병률 추정치와 비교하였다. 감별 진단 및 치료 계획은 집단

				간 유의성에 대한 표준 통계적 검정을 사용하여 인구통계학적 집단에 걸쳐 평가되었다. **결과:** 우리는 GPT-4가 의학적 상태의 인구통계학적 다양성을 적절하게 모델링하지 못하고, 인구통계학적 제시를 고정관념화하는 임상 사례를 지속적으로 생성하는 것을 발견하였다. 표준화된 임상 사례에 대해 GPT-4가 생성한 감별 진단에는 특정 인종, 민족 및 성별을 고정관념화하는 진단이 포함될 가능성이 더 높았다. 모델이 생성한 평가 및 계획은 인구통계학적 속성과 더 비싼 시술에 대한 권고 간의 유의한 연관성뿐만 아니라 환자 인식의 차이를 보여주었다. **해석:** 우리의 결과는 GPT-4와 같은 LLM 도구가 임상 진료에 통합되기 전에 의도된 사용 사례에 대한 포괄적이고 투명한 편향 평가의 시급한 필요성을 강조한다. 우리는 이러한 편향의 잠재적 원인과 임상 구현 전 잠재적 완화 전략을 논의한다. **Background:** Large language models (LLMs) such as GPT-4 hold great promise as transformative tools in health care, ranging from automating administrative tasks to augmenting clinical decision making. However, these models also pose a danger of perpetuating biases and delivering incorrect medical diagnoses, which can have a direct, harmful impact on medical care. We aimed to assess whether GPT-4 encodes racial and gender biases that impact its use in health care. **Methods:** Using the Azure OpenAI application interface, this model evaluation study tested whether GPT-4 encodes racial and gender biases and examined the impact of such biases on four potential applications of LLMs in the clinical domain—namely, medical education, diagnostic reasoning, clinical plan generation, and subjective patient assessment. We conducted experiments with prompts designed to resemble typical use of GPT-4 within clinical and medical education applications. We used clinical vignettes from NEJM Healer and from published research on implicit bias in health care. GPT-4 estimates of the demographic distribution of medical conditions were compared with true US prevalence estimates. Differential diagnosis and treatment planning were evaluated across demographic groups using standard statistical tests for significance between groups. **Findings:** We found that GPT-4 did not appropriately model the demographic diversity of medical conditions, consistently producing clinical vignettes that stereotype demographic presentations. The differential diagnoses created by GPT-4 for standardised clinical vignettes were more likely to include diagnoses that stereotype certain races, ethnicities, and genders.

				Assessment and plans created by the model showed significant association between demographic attributes and recommendations for more expensive procedures as well as differences in patient perception. **Interpretation:** Our findings highlight the urgent need for comprehensive and transparent bias assessments of LLM tools such as GPT-4 for intended use cases before they are integrated into clinical care. We discuss the potential sources of these biases and potential mitigation strategies before clinical implementation. Funding: Priscilla Chan and Mark Zuckerberg.
19	(McMahon and McMahon 2024)	Automating untruths: ChatGPT, self-managed medication abortion, and the threat of misinformation in a post-Roe world	ChatGPT	**배경:** ChatGPT는 자연어 처리를 사용하여 인간과 유사한 방식으로 프롬프트를 이해하고 실행하는 생성형 인공지능 챗봇이다. 이 챗봇이 대중 사이에서 정보원으로 인기를 얻고 있는 반면, 전문가들은 ChatGPT가 하는 허위 및 오도적 진술의 수에 대해 우려를 표명해왔다. 많은 사람들이 자가 관리 약물 낙태에 대한 정보를 온라인에서 검색하며, 이는 Roe v. Wade 판결 번복 이후 더욱 일반화되었다. ChatGPT도 이러한 정보의 출처로 사용되고 있을 가능성이 높지만, 그 정확성에 대해서는 알려진 바가 거의 없다. **목적:** 자가 관리 낙태 안전성과 낙태약 사용 과정에 관한 일반적인 질문에 대한 ChatGPT 응답의 정확성을 평가한다. **방법:** 우리는 자가 관리 약물 낙태에 관한 65개의 질문을 ChatGPT에 입력하였으며, 이는 약 11,000단어의 텍스트를 생성하였다. 우리는 MAXQDA에서 모든 데이터를 질적으로 코딩하고 주제 분석을 수행하였다. **결과:** ChatGPT 응답은 임상의가 관리하는 약물 낙태를 안전하고 효과적이라고 정확하게 설명하였다. 이와 대조적으로**, 자가 관리 약물 낙태는 위험하고 합병증 위험 증가와 관련이 있는 것으로 부정확하게 설명되었으며, 이는 임상의 감독의 부재에 기인한 것**으로 여겨졌다. **결론:** ChatGPT는 **자가 관리 약물 낙태가 안전하고 효과적임을 입증하는 광범위한 증거 체계와 직접 모순되는 방식으로 자가 관리 약물 낙태와 관련된 합병증의 위험을 과장한 응답을 반복적으로 제공**하였다. 자가 관리 약물 낙태에 관한 건강 오정보와 관련 낙인을 영속화하는 챗봇의 경향은 공중보건과 재생산 자율성에 위협을 가한다. **Background:** ChatGPT is a generative artificial intelligence chatbot that uses natural language processing to understand and execute prompts in a human-like manner. While the chatbot has become popular as a source of information among the public, experts have expressed concerns about the number of false and misleading statements made by ChatGPT. Many people search online for information about self-managed medication abortion,

				which has become even more common following the overturning of Roe v. Wade. It is likely that ChatGPT is also being used as a source of this information; however, little is known about its accuracy. **Objective:** To assess the accuracy of ChatGPT responses to common questions regarding self-managed abortion safety and the process of using abortion pills. **Methods:** We prompted ChatGPT with 65 questions about self-managed medication abortion, which produced approximately 11,000 words of text. We qualitatively coded all data in MAXQDA and performed thematic analysis. **Results:** ChatGPT responses correctly described clinician-managed medication abortion as both safe and effective. In contrast, self-managed medication abortion was inaccurately described as dangerous and associated with an increase in the risk of complications, which was attributed to the lack of clinician supervision. **Conclusion:** ChatGPT repeatedly provided responses that overstated the risk of complications associated with self-managed medication abortion in ways that directly contradict the expansive body of evidence demonstrating that self-managed medication abortion is both safe and effective. The chatbot's tendency to perpetuate health misinformation and associated stigma regarding self-managed medication abortions poses a threat to public health and reproductive autonomy.
20	(Qazi et al. 2025)	Automation Bias in Large Language Model Assisted Diagnostic Reasoning Among AI-Trained Physicians	ChatGPT-4o	**중요성:** 대규모 언어 모델(LLM)은 임상 추론 개선 가능성을 보여주지만, 진단 정확도를 저하시킬 수 있는 과도한 의존인 자동화 편향을 유발할 위험도 있다. LLM 사용이 자발적일 때 AI 교육을 받은 의사들이 이러한 편향에 취약한지는 아직 알려지지 않았다. **목적:** 오류가 있는 LLM 권고에 대한 노출이 오류 없는 AI 조언과 비교하여 AI 교육을 받은 의사들의 진단 성능을 저하시키는지 확인한다. **설계:** 2025년 6월 20일부터 8월 15일까지 단일맹검 무작위 임상시험을 수행하였다. **환경:** 의사들은 파키스탄의 여러 의료기관에서 모집되었으며, 대면 또는 원격 화상 회의를 통해 참여하였다. **참가자:** 파키스탄 의료치과위원회에 등록된 MBBS 학위 소지 의사로, LLM 능력, 프롬프트 엔지니어링 및 AI 출력의 비판적 평가를 다루는 20시간의 AI 리터러시 교육을 이수한 자. **중재:** 참가자들은 1:1로 무작위 배정되어 75분 동안 6개의 임상 사례를 진단하였다. 대조군은 수정되지 않은 ChatGPT-4o의 진단 권고를 받았으며, 치료군의 권고

				에는 6개 사례 중 3개에 의도적 오류가 포함되었다. 의사들은 임상 판단에 따라 제공된 ChatGPT-4o 권고를 기존 진단 자료와 함께 자발적으로 참조할 수 있었다. **주요 결과 및 측정:** 1차 결과는 진단 추론 정확도(백분율)로, 3명의 맹검 의사가 전문가 검증 평가기준을 사용하여 평가하였으며, 감별 진단 정확도, 지지 및 반대 증거의 적절성, 권장 진단 단계의 질을 평가하였다. 2차 결과는 최우선 진단 정확도였다. **결과:** 44명의 의사(치료군 22명, 대조군 22명)가 참여하였다. **오류 없는 권고를 받은 의사들은 평균(SD) 진단 정확도 84.9%(19.7%)를 달성한 반면, 결함 있는 권고에 노출된 의사들은 73.3%(30.5%)를 기록하여, 조정 평균 차이는 −14.0 퍼센티지 포인트(95% CI: −8.3~−19.7; P < .0001)였다. 사례당 최우선 진단 정확도는 치료군 76.1%(42.5), 대조군 90.5%(28.9)였으며, 조정 차이는 −18.3 퍼센티지 포인트(95% CI, −26.6~−10.0; P < .0001)였다**. **결론 및 관련성:** 본 시험은 **오류가 있는 LLM 권고가 AI 교육을 받은 의사들에게서도 자동화 편향을 유발하여 의사들의 진단 성능을 유의하게 저하시킴을 입증**한다. 결함 있는 AI 출력에 대한 자발적 의존은 중대한 환자 안전 위험을 강조하며, 광범위한 임상 배포 전 인간 감독을 보장하는 강력한 안전장치가 필요함을 시사한다. **Importance:** Large language models (LLMs) show promise for improving clinical reasoning, but they also risk inducing automation bias, an over-reliance that can degrade diagnostic accuracy. Whether AI-trained physicians are vulnerable to this bias when LLM use is voluntary remains unknown. **Objective:** To determine whether exposure to erroneous LLM recommendations degrades AI-trained physicians' diagnostic performance compared to error-free AI advice. **Design:** A single-blind randomized clinical trial was conducted from June 20 to August 15, 2025. **Setting:** Physicians were recruited from multiple medical institutions in Pakistan, participating through in-person or remote video conferencing. **Participants:** Physicians registered with the Pakistan Medical and Dental Council with MBBS degrees, who had completed a 20-hour AI-literacy training covering LLM capabilities, prompt engineering, and critical evaluation of AI output. **Intervention:** Participants were randomized 1:1 to diagnose 6 clinical vignettes in 75 minutes. The control group received unmodified ChatGPT-4o's diagnostic recommendations; the treatment group's recommendations

				contained deliberate errors in 3 of 6 vignettes. Physicians could voluntarily consult offered ChatGPT-4o recommendations alongside conventional diagnostic resources based on their clinical judgment. **Main Outcomes and Measures:** Primary outcome was the diagnostic reasoning accuracy (percentage), assessed by three blinded physicians using an expert-validated rubric to evaluate: differential diagnosis accuracy, appropriateness of supporting and opposing evidence, and quality of recommended diagnostic steps. Secondary outcome was the top-choice diagnosis accuracy. **Results:** Forty-four physicians (22 treatment, 22 control) participated. Physicians receiving error-free recommendations achieved mean (SD) diagnostic accuracy of 84.9% (19.7%), whereas those exposed to flawed recommendations scored 73.3% (30.5%), resulting in an adjusted mean difference of −14.0 percentage points (95% CI: −8.3 to −19.7; P 〈 .0001). Top-choice diagnosis accuracy per case was 76.1% (42.5) in the treatment group and 90.5% (28.9) in the control group, with an adjusted difference of −18.3 percentage points (95% CI, −26.6 to −10.0; P 〈 .0001). **Conclusions and Relevance:** This trial demonstrates that erroneous LLM recommendations significantly degrade physicians' diagnostic performance by inducing automation bias, even in AI-trained physicians. Voluntary deference to flawed AI output highlights critical patient safety risk, necessitating robust safeguards to ensure human oversight before widespread clinical deployment. **Trial Registration:** ClinicalTrials.gov Identifier: NCT06963957
21	(Elyoseph and Levkovich 2023)	Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment	ChatGPT Free	OpenAI가 개발한 인공지능 언어 모델인 ChatGPT는 정신건강 분야에 기여할 잠재력을 지니고 있다. 그럼에도 불구하고 ChatGPT가 이론적으로는 유망성을 보이지만, 중요한 정신건강 문제인 자살 예방에서의 임상적 능력은 아직 입증되지 않았다. 이러한 지식 격차를 해소하기 위해, 본 연구는 자살 위험 평가에 초점을 맞춘 가상 사례 연구에서 정신건강 전문가들의 평가와 ChatGPT의 정신건강 지표 평가를 비교하는 것을 목표로 한다. 구체적으로, ChatGPT에게 다양한 수준의 짐이 됨의 지각과 좌절된 소속감을 나타내는 가상 환자를 설명하는 텍스트 사례를 평가하도록 요청하였다. ChatGPT의 평가를 정신건강 전문가들의 기준과 비교하였다. 결과는 **ChatGPT가 모든 조건에서 정신건강 전문가들보다 자살 시도 위험을 낮게 평가**했음을 나타냈다. 더욱이, **ChatGPT는 대부분의 조건에서 정신적 회복력을 기준보다 낮게 평가**하였다. 이러한 결과는 자살 위험 평가를 위해 또는 의사결정 개선을 위한 보완 도구로 **ChatGPT에 의존하는 게이트키퍼, 환자 또는 심지어 정신건강 전문가**

				들이 실제 자살 위험을 과소평가하는 부정확한 평가를 받을 수 있음을 시사한다. ChatGPT, an artificial intelligence language model developed by OpenAI, holds the potential for contributing to the field of mental health. Nevertheless, although ChatGPT theoretically shows promise, its clinical abilities in suicide prevention, a significant mental health concern, have yet to be demonstrated. To address this knowledge gap, this study aims to compare ChatGPT's assessments of mental health indicators to those of mental health professionals in a hypothetical case study that focuses on suicide risk assessment. Specifically, ChatGPT was asked to evaluate a text vignette describing a hypothetical patient with varying levels of perceived burdensomeness and thwarted belongingness. The ChatGPT assessments were compared to the norms of mental health professionals. The results indicated that ChatGPT rated the risk of suicide attempts lower than did the mental health professionals in all conditions. Furthermore, ChatGPT rated mental resilience lower than the norms in most conditions. These results imply that gatekeepers, patients or even mental health professionals who rely on ChatGPT for evaluating suicidal risk or as a complementary tool to improve decision-making may receive an inaccurate assessment that underestimates the actual suicide risk.
22	(Gun 2025b)	Can AI match emergency physicians in managing common emergency cases? A comparative performance evaluation	ChatGPT-4	**배경:** ChatGPT와 같은 대규모 언어 모델(LLM)은 임상 의사결정 지원을 위해 점점 더 탐구되고 있다. 그러나 고위험 응급 시나리오에서의 성능은 충분히 검토되지 않았다. 본 연구는 다양한 응급 사례에서 전문의 자격증을 보유한 응급의학과 전문의와 비교하여 ChatGPT의 진단 및 치료적 정확도를 평가하는 것을 목표로 하였다. **방법:** 본 비교 연구는 검증된 학술 플랫폼(Geeky Medics, Life in the Fast Lane, Emergency Medicine Cases)에서 출처한 15개의 표준화된 응급 시나리오를 사용하여 수행되었다. ChatGPT(GPT-4)와 의사가 진단, 검사, 초기 치료, 임상적 안전성, 의사결정 복잡성의 5개 사전 정의된 매개변수를 기반으로 각 사례를 독립적으로 평가하였다. 사례는 5점 만점으로 점수가 매겨졌다. 일치도는 높음(5/5), 중간(4/5), 낮음(≤3/5)으로 분류되었다. 각 일치도 범주에 대해 Wilson 신뢰구간(95%)이 계산되었다. **결과:** ChatGPT는 8개 사례에서 높은 일치도(5/5)를 달성하였고(53.3%, 95% CI: 27.6–77.0%), 4개 사례에서 중간 일치도(4/5)(26.7%, CI: 10.3–55.4%), 3개 사례에서 낮은 일치도(≤3/5)(20.0%, CI: 6.0–45.6%)를 보였다. **STEMI, DKA, 천식과 같은 구조화되고 프로토콜 기반 상태에서 성능이 가장 강력하였다. 뇌졸중, 쇼크를 동반한 외상, 혼합 산-염기 장애와 같은 복잡한 시나리오에서는 낮은 성능이 관찰**되

				었다. **결론:** ChatGPT는 구조화된 시나리오에서 응급의학과 전문의의 결정과 강한 일치성을 보였지만 복잡한 사례에서는 신뢰성이 부족하였다. AI가 의사결정과 교육을 향상시킬 수 있지만, 인간 의사의 임상적 추론을 대체할 수는 없다. AI의 역할은 대체재가 아닌 지원 도구로 규정하는 것이 가장 적절하다. **Background:** Large language models (LLMs) such as ChatGPT are increasingly explored for clinical decision support. However, their performance in high-stakes emergency scenarios remains underexamined. This study aimed to evaluate ChatGPT's diagnostic and therapeutic accuracy compared to a board-certified emergency physician across diverse emergency cases. **Methods:** This comparative study was conducted using 15 standardized emergency scenarios sourced from validated academic platforms (Geeky Medics, Life in the Fast Lane, Emergency Medicine Cases). ChatGPT (GPT-4) and a physician independently evaluated each case based on five predefined parameters: diagnosis, investigations, initial treatment, clinical safety, and decision-making complexity. Cases were scored out of 5. Concordance was categorized as high (5/5), moderate (4/5), or low (≤ 3/5). Wilson confidence intervals (95%) were calculated for each concordance category. **Results:** ChatGPT achieved high concordance (5/5) in 8 cases (53.3%, 95% CI: 27.6–77.0%), moderate concordance (4/5) in 4 cases (26.7%, CI: 10.3–55.4%), and low concordance (≤ 3/5) in 3 cases (20.0%, CI: 6.0–45.6%). Performance was strongest in structured, protocol-based conditions such as STEMI, DKA, and asthma. Lower performance was observed in complex scenarios like stroke, trauma with shock, and mixed acid-base disturbances. **Conclusion:** ChatGPT showed strong alignment with emergency physician decisions in structured scenarios but lacked reliability in complex cases. While AI may enhance decision-making and education, it cannot replace the clinical reasoning of human physicians. Its role is best framed as a supportive tool rather than a substitute.
23	(Shazahan et al. 2025)	Can ChatGPT Aid in Musculoskeletal Intervention?	ChatGPT-4	**목적** 영상의학은 환자 진료를 개선하기 위해 최첨단 기술을 탐구하며 지속적으로 발전해왔다. 이는 의학이 기술 혁신에 의해 어떻게 추진되는지를 보여주는 대표적인 예이다. 최근 인공지능(AI)은 다양한 기술 발전에서 중요한 역할을 해왔다. 주로 자연어 이해 및 생성에 초점을 맞춘 AI 언어 모델인 Chat Generative Pre-Trained

				Transformer(ChatGPT)-4는 의료 정보 검색에 점점 더 많이 사용되고 있다. 본 연구는 영상 유도 근골격계 중재술을 보조하는 데 있어 ChatGPT-4o의 유용성을 탐구하고, 그 장점과 한계를 상세히 기술한다. **방법** 두 명의 근골격계 영상의학과 전문의가 일반적인 근골격계 중재술에 대해 ChatGPT가 생성한 정보를 평가하였다. 그들은 제공된 시술 단계 및 시술 전후 세부 사항을 검토하여 근골격계 중재술 안내에서 ChatGPT-4o의 전반적인 유용성을 분석하였다. 평가는 5점 리커트 척도로 기록되고 통계 분석의 대상이 되었다. **결과** 두 명의 평가자에 의한 리커트 척도 점수의 통계 분석은 Cohen의 Kappa 점수 0.54로 나타난 바와 같이 중간 수준의 평가자 간 일치도를 보였다. 범주 전반에 걸쳐 두 평가자가 평가한 리커트 점수의 최빈값은 1에서 3까지 범위를 보였으며, 이는 차선의 성능을 나타낸다. **가장 낮은 점수는 이미지 품질 평가에서 관찰되었고, 가장 높은 평가는 시술 후 세부 사항에서 나타났다.** **결론** **ChatGPT-4o는 구조화된 시술 안내를 제공하지만, 제한된 해부학적 세부 사항과 맥락적 정확성으로 인해 복잡하고 이미지 의존적인 작업에서는 부족함을 보인다. 교육에 도움이 될 수 있지만, 전문가 감독 없이는 임상 사용에 적합하지 않다.** 안전하고 효과적인 실무 통합을 위해서는 도메인 특화 교육, 검증 및 다학제적 협력이 필수적이다. **Objective** Radiology has continuously evolved exploring cutting-edge technologies to improve patient care. It is a prime example of how medical science is propelled forward by technological innovation. In recent times, artificial intelligence (AI) has played a crucial role in various technological advancements. Chat Generative Pre-Trained Transformer (ChatGPT)-4, an AI language model primarily focusing on natural language understanding and generation, is increasingly used to retrieve medical information. This study explores the utility of ChatGPT-4o in aiding imaging-guided musculoskeletal interventions, detailing its advantages and limitations. **Methods** Two musculoskeletal radiologists assessed the information generated by ChatGPT on common musculoskeletal interventions. They analyzed the overall utility of ChatGPT-4o in guiding musculoskeletal interventions by examining the procedure steps and pre-and post-procedure details provided. The assessment was documented in a 5-point Likert scale and subjected to statistical analysis. **Results** The statistical analysis of Likert scale scores by both readers revealed a moderate level of inter-rater agreement, as indicated by a Cohen's Kappa score of 0.54. Across the categories, the mode of Likert score ranged from

				1 to 3, as rated by both readers, indicating suboptimal performance. The lowest scores were observed in image quality assessments, whereas the highest ratings were of post-procedure details. **Conclusion** ChatGPT-4o offers structured procedural guidance but falls short in complex, image-dependent tasks due to limited anatomical detail and contextual accuracy. It may aid education, but not clinical use without expert oversight. Domain-specific training, validation, and multidisciplinary collaboration are essential for safe and effective integration into practice.
24	(Gorgos et al. 2025)	ChatGPT and Claude in Hand Surgery: An Explanatory Evaluation of Clinical Decision Support on Common Surgical Cases	ChatGPT-3.5, Claude(3.7)	**서론:** 대규모 언어 모델(LLM)은 여러 의학 분야에서 점점 더 인기를 얻고 있다. 그러나 정형외과 연구에서는 경험 많은 임상의를 능가하지 못하는 것으로 보이기 때문에 일상 진료에 통합하는 것에 의문이 제기되어 왔다. 반대로, 수부외과에서 인공지능의 역할에 대한 연구는 제한적으로 남아 있다. 본 연구는 의학에서 흔히 사용되는 두 가지 LLM인 Generative Pre-trained Transformer(ChatGPT)와 Claude를 임상 수부외과 환경에서 평가하는 것을 목표로 한다. **방법:** 일반적인 수부외과 진단과 관련된 10개의 질문을 프롬프트로 공식화하여 체계적인 방식으로 ChatGPT와 Claude에 입력하였다. 생성된 응답은 수부외과 전문의들에 의해 익명으로 평가되었으며, 그들은 QUEST 기준에 따라 응답의 질을 평가하였다. 평가자 간 일치도를 평가하기 위해 Gwet의 AC2가 사용되었다. **결과:** 일반적으로 **ChatGPT와 Claude는 (1) 정보의 질, (2) 이해와 추론, (3) 표현 스타일과 페르소나, (4) 안전성과 위해, (5) 신뢰와 확신을 포함한 QUEST의 차원에 따라 통계적으로 유사한 성능을 보였지만 상대적으로 낮은 점수**를 받았다. 모든 측정 항목에 걸친 수부외과 전문의들 간의 일치도는 Gwet의 AC2(0.29)에 따라 낮았다. **결론:** ChatGPT와 Claude는 다양한 일반적인 수부외과 관련 질문이 제공될 때 유사한 성능을 보인다. **그러나 환자 안전, 치료 효율성 및 근거 기반 진료의 핵심 기반인 임상적 정확성과 신뢰성과 관련하여 상당한 한계를 보여준다**. 더욱이, **ChatGPT와 Claude의 기능이 개별 수부외과 전문의에 따라 다른 것으로 보이므로, 현재 상태의 이러한 LLM은 수부외과에서 일상적인 임상 사용에 적합하지 않다**. **INTRODUCTION:** Large language models (LLMs) have gained increasing popularity in several medical disciplines. In orthopedic research however, their integration into routine practice have been questioned as they do not seem to outperform experienced clinicians. Conversely, research on the role of artificial intelligence in hand surgery remains limited. This study aims to evaluate two common LLMs in medicine, Generative Pre-trained Transformer (ChatGPT) and Claude in the clinical hand surgery setting.

				METHODS: Ten questions pertinent to common hand surgical diagnosis were formulated as prompts and entered into ChatGPT and Claude in a systematic manner. The generated responses were anonymously evaluated by hand surgeons, who assessed the quality of the responses according to the QUEST criteria. Gwet's AC2 was used to evaluate the agreement between raters. **RESULTS:** In general, ChatGPT and Claude performed statistically similar according to the dimensions of QUEST including (1) Quality of information, 2) Understanding and reasoning, 3) Expression style and persona, 4) Safety and harm and 5) Trust and confidence although with relatively modest scores. Agreement between hand surgeons across all measurements was low according to Gwet's AC2 (0.29). **CONCLUSIONS:** ChatGPT and Claude perform similarly when provided with various common hand surgery related questions. However, they demonstrate significant limitations pertaining to clinical accuracy and reliability that are the core foundation for patient safety, treatment efficiency and evidence-based practice. Furthermore, as the function of ChatGPT and Claude seem to differ between individual hand surgeons, these LLMs in their current state are not suitable for routine clinical use in hand surgery.V.
25	(Comrie 2023)	ChatGPT Decision Support System: Utility in Creating Public Policy for Concussion/Repetitive Brain Trauma Associated With Neurodegenerative Diseases	ChatGPT	본 논문은 신경퇴행성 질환 위험과 관련된 뇌진탕 및 반복적 뇌 손상에 관한 정책 수립을 위한 ChatGPT 의사결정 지원 시스템의 유용성을 평가한다. 이는 일반적으로 안정적이고 신속하다. 프롬프트/응답 쌍(n=259)을 검토한 결과 6개의 프롬프트 응답 쌍이 재생성되었고(2.31%), 1개의 오답(0.38%), 1개의 단편(0.38%)이 반환되었다. 그 **정확성, 타당성, 불투명성, 정보 지연 및 조작에 대한 취약성이 유용성을 제한한다. ChatGPT의 데이터는 시대에 뒤떨어지고 불완전할 수 있어 전문가 진술을 분석하는 주제 전문가들의 사용으로 그 유용성이 제한**된다. ChatGPT의 성능은 인종과 같은 **이해관계자 편향 및 소송 관리와 관련된 프롬프트의 영향**을 받는다. 그럼에도 불구하고 ChatGPT는 미국식 및 영국/호주식 영어 모두로 쉽게 응답하는 능력을 입증하였다. 전반적으로, 본 연구는 ChatGPT가 뇌진탕 및 반복적 뇌 손상 정책과 관련된 의사결정에 광범위하게 사용되기 전에 해결해야 할 한계가 있음을 시사한다. This article evaluates the ChatGPT decision support system's utility for creating policies related to concussion and repetitive brain trauma associated with neurodegenerative disease risk. It is generally stable and fast. prompt/response pairs (n=259) were examined returning: six prompt

				response pairs that regenerated (2.31%); one Incorrect Answer; (.38%) one fragment (.38%). Its accuracy, validity, opacity, informational latency and vulnerability to manipulation limits its utility. ChatGPT's data can be both out-of-date and incomplete which limits its utility use to subject matter experts analyzing expert statements. ChatGPT's performance is affected by prompts involving stakeholder bias and litigation management, such as race. Nonetheless, ChatGPT demonstrated its ability to respond in both American and British/Australian English with ease. Overall, this study suggests that ChatGPT has limitations that need to be addressed before it can be widely used in decision-making related to concussion and repetitive brain trauma policies.
26	(Zhang et al. 2023)	ChatGPT Exhibits Gender and Racial Biases in Acute Coronary Syndrome Management	ChatGPT-3.5	대규모 언어 모델(LLM)의 최근 혁신적 발전은 이들의 빠른 보급과 광범위한 사용으로 이어졌다. 초기 응용 분야 중 하나는 의학으로, LLM은 임상 워크플로우를 간소화하고 임상 분석 및 의사결정을 촉진하기 위해 연구되어 왔다. 그러나 인공지능(AI), 특히 LLM의 배포에 대한 주요 장벽은 내재된 성별 및 인종 편향에 대한 우려였다. 여기서 우리는 주요 LLM인 ChatGPT 3.5가 급성 관상동맥 증후군(ACS)의 임상 관리에서 성별 및 인종 편향을 나타내는지 평가한다**. 우리는 환자를 여성, 아프리카계 미국인 또는 히스패닉으로 명시하는 것이 ACS의 가이드라인 권장 의료 관리, 진단 및 증상 관리의 감소를 초래함을 발견**하였다. 가장 주목할 만한 점은, **가장 큰 격차가 ACS의 진단 및 추가 중재를 위한 관상동맥 조영술 또는 부하 검사 권고와 고강도 스타틴 권고에서 나타났다는 것이다. 이러한 격차는 임상적으로 관찰되어 온 편향과 상관관계가 있으며, ACS 및 관상동맥 질환의 차별적 성별 및 인종 이환율 및 사망률 결과와 연관**되어 왔다. 더욱이, 우리는 가장 큰 격차가 명시적인 임상 가이드라인이 더 적게 존재하는 불안정 협심증에서 나타남을 발견하였다. 마지막으로**, 우리는 ChatGPT 3.5에게 답변을 제공하기 전에 추론을 설명하도록 요청함으로써 임상적 정확성을 개선하고 성별 및 인종 편향의 사례를 완화할 수 있음을 발견**하였다. 이는 LLM이 나타내는 성별 및 인종 편향이 실제로 임상 관리에 영향을 미침을 입증하는 최초의 연구 중 하나이다. 추가로, 우리는 LLM 성능을 개선하는 기존 전략이 임상 관리에서 LLM 성능을 개선할 뿐만 아니라 성별 및 인종 편향을 완화하는 데에도 사용될 수 있음을 입증한다. 인공지능(AI)의 발전은 대규모 언어 모델(LLM)의 빠른 보급과 광범위한 사용으로 이어졌다. 의학 분야는 임상 워크플로우를 향상시키고 임상 분석 및 의사결정을 증진하기 위해 새로운 LLM을 활용하고자 노력해왔다. 의료 분야에서 AI 배포의 장벽은 언어 기반 모델을 훈련하는 데 사용되는 기본 콘텐츠에 만연한 고유한 편향이다. **LLM의 이러한 취약성은 의료 환경에서 성별 및 인종 편향의 체계적 전파를 초래하여 건강 관리 및 결과에서 기존 성별 및 인종 격차를 악화시킬 수 있다**. 여기서 우

				리는 LLM이 심장학의 임상 의사결정에 적용될 때 성별 및 인종 편향을 나타내는지 조사하였다. 우리는 주요 LLM인 ChatGPT 3.5가 기존 근거 기반 문헌에 의해 뒷받침되지 않는 인종 및 성별에 기반한 차별적 의사결정을 나타냄을 발견하였다. 이러한 성별 및 인종 편향은 이전에 임상 실무에서 관찰되었고 건강 결과에 해로운 영향을 미치는 것으로 나타난 편향과 다르지 않다. 주목할 만한 점은, 우리는 이러한 모델에게 권고 사항을 설명하고 상세히 기술하도록 프롬프트를 제공하는 것이 편향을 완화할 수 있음을 발견하였다는 것이다. **이는 LLM 내의 성별 및 인종 편향이 임상 관리에 영향을 미칠 수 있음을 보여주는 최초의 사례 중 하나**이다. 우리의 연구는 LLM 배포에 대한 중요한 장벽을 식별하고 언어 기반 인공지능에서 편향을 완화하기 위한 전략을 제안한다. Recent breakthroughs in large language models (LLMs) have led to their rapid dissemination and widespread use. One early application has been to medicine, where LLMs have been investigated to streamline clinical workflows and facilitate clinical analysis and decision-making. However, a leading barrier to the deployment of Artificial Intelligence (AI) and in particular LLMs has been concern for embedded gender and racial biases. Here, we evaluate whether a leading LLM, ChatGPT 3.5, exhibits gender and racial bias in clinical management of acute coronary syndrome (ACS). We find that specifying patients as female, African American, or Hispanic resulted in a decrease in guideline recommended medical management, diagnosis, and symptom management of ACS. Most notably, the largest disparities were seen in the recommendation of coronary angiography or stress testing for the diagnosis and further intervention of ACS and recommendation of high intensity statins. These disparities correlate with biases that have been observed clinically and have been implicated in the differential gender and racial morbidity and mortality outcomes of ACS and coronary artery disease. Furthermore, we find that the largest disparities are seen during unstable angina, where fewer explicit clinical guidelines exist. Finally, we find that through asking ChatGPT 3.5 to explain its reasoning prior to providing an answer, we are able to improve clinical accuracy and mitigate instances of gender and racial biases. This is among the first studies to demonstrate that the gender and racial biases that LLMs exhibit do in fact affect clinical management. Additionally, we demonstrate that existing strategies that improve LLM performance not only improve LLM performance in clinical management, but can also be used to mitigate gender and racial biases.

				Advances in Artificial Intelligence (AI) have led to the rapid dissemination and widespread use of large language models (LLMs)1–4. The field of medicine has sought to harness new LLMs to enhance clinical workflow and to augment clinical analysis and decision-making5–7. A barrier to the deployment of AI in healthcare is inherent biases that pervade the underlying content used to train language-based models8,9. This vulnerability in LLMs could result in the systematic propagation of gender and racial biases in medical settings, leading to a worsening of existing gender and racial disparities in health management and outcomes1. Here, we investigated whether LLMs exhibit gender and racial bias when applied to clinical decision making in Cardiology. We found that a leading LLM, ChatGPT 3.5, exhibits differential decision making based on race and gender that is not supported by existing evidence-based literature. These gender and racial biases are not dissimilar to those that have previously been observed in clinical practice and that have been shown to have a detrimental effect on health outcomes. Notably, we found that prompting these models to explain and elaborate on their recommendations can mitigate bias. This is among the first examples to show that gender and racial bias within LLMs can affect clinical management10. Our work identifies a critical barrier to the deployment of LLMs and proposes strategies to mitigate bias in language-based artificial intelligence.
27	(Zaboli, Brigo, Brigiari, et al. 2025)	Chat-GPT in triage: Still far from surpassing human expertise – An observational study	ChatGPT-4.0	**배경:** 분류는 임상적 긴급성에 따라 환자 진료의 우선순위를 정하기 위해 응급실(ED)에서 필수적이다. 최근 연구들은 분류에서 대규모 언어 모델(LLM)의 역할을 탐구해왔지만, 인간 분류와 비교한 이들의 효과성은 여전히 불확실하다. 본 연구는 응급실 환자 분류에서 ChatGPT 4.0의 효과성을 평가하였다. **방법:** 본 후향적 연구는 2,658명의 환자 데이터를 분석하였다. 인간 분류 담당자가 배정한 분류 코드를 Chat-GPT 4.0을 사용한 인공지능(AI) 분류가 배정한 코드와 비교하였다. 인간과 AI 분류 간의 일치도는 Cohen의 kappa 통계를 사용하여 평가하였다. 임상 결과는 예측 정확도를 결정하기 위해 수신자 조작 특성(ROC) 곡선을 통해 평가되었다. 두 분류 시스템의 민감도와 특이도는 2 × 2 분할표를 사용하여 다양한 증상에 걸쳐 비교되었다. **결과:** 인간과 AI 분류 간 일치도에 대한 Cohen의 kappa 통계는 0.125(95% CI: 0.100–0.134)였다. ROC 분석은 인간 분류가 모든 연구 결과 예측에서 AI를 능가함을 보여주었으며, 통계적으로 유의한 차이를 나타냈다. 30일 사망률의 경우, 인간 분류의 ROC는 0.88이었고 AI 분류는 0.70이었으며, $p < 0.001$이었다. 생명 구조 중재에서도 유사한 결과가 관찰되었으며, 인간 분류는 ROC 0.98, AI 분류는 0.87

				을 보였고, p = 0.014였다. 특정 증상에 대해 인간 분류는 우수한 민감도와 특이도를 보였다. 결론: Chat-GPT 4.0과 같은 LLM은 응급실 분류에서 제한된 유용성을 가지며, 특히 고위험 환자에 대한 낮은 민감도로 인해 과소분류를 초래한다. 인간 분류는 Chat-GPT보다 더 신뢰할 수 있다. **Background:** Triage is essential in emergency departments (EDs) to prioritize patient care based on clinical urgency. Recent investigations have explored the role of large language models (LLMs) in triage, but their effectiveness compared to human triage remains uncertain. This study assessed the effectiveness of ChatGPT 4.0 in triaging ED patients. **Methods:** This retrospective study analyzed data from 2658 patients. Triage codes assigned by human triage personnel were compared with those assigned by Artificial Intelligence (AI) triage using Chat-GPT 4.0. Agreement between human and AI triage was assessed using Cohen's kappa statistic. Clinical outcomes were evaluated through Receiver Operating Characteristic (ROC) curves to determine predictive accuracy. Sensitivity and specificity of both triage systems were compared across different symptoms using 2 × 2 contingency tables. **Results:** The Cohen's kappa statistic for agreement between human and AI triage was 0.125 (95 % CI: 0.100–0.134). ROC analysis demonstrated that human triage outperformed AI in predicting all study outcomes, with statistically significant differences. For 30-day mortality, the ROC of human triage was 0.88, while for AI triage it was 0.70, p 〈 0.001. A similar result was observed for life-saving interventions, where human triage had an ROC of 0.98 and AI triage 0.87, p = 0.014. For specific symptoms, human triage showed superior sensitivity and specificity. **Conclusions:** LLMs like Chat-GPT 4.0 have limited utility in ED triage, particularly due to their lower sensitivity for high-risk patients, which lead to under-triage. Human triage remains more reliable than Chat-GPT.
28	(Jeblick et al. 2024)	ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports	ChatGPT (2022)	**목적:** 대규모 언어 모델(LLM) ChatGPT로 생성된 간소화된 영상의학 보고서의 질을 평가하고, 의료 텍스트 간소화를 위한 ChatGPT 유사 LLM의 도전과제와 기회를 논의한다. **방법:** 본 탐색적 사례 연구에서, 영상의학과 전문의가 3개의 가상 영상의학 보고서를 작성하였으며, 우리는 "이 의료 보고서를 간단한 언어를 사용하여 어린이에게 설명하세요"라는 프롬프트로 ChatGPT에 입력하여 이를 간소화하였다. 설문조사에서

				우리는 15명의 영상의학과 전문의에게 간소화된 영상의학 보고서의 사실적 정확성, 완전성 및 환자에 대한 잠재적 위해에 관한 질을 평가하도록 하였다. 우리는 리커트 척도 분석과 귀납적 자유 텍스트 범주화를 사용하여 간소화된 보고서의 질을 평가하였다. **결과:** 대부분의 영상의학과 전문의는 간소화된 보고서가 사실적으로 정확하고 완전하며 환자에게 잠재적으로 해롭지 않다는 데 동의하였다. 그럼에도 불구하고 **부정확한 진술, 누락된 관련 의학 정보 및 잠재적으로 해로운 구절의 사례가 보고**되었다. **결론:** 의료 분야에 대한 추가적인 적응이 필요하다고 보지만, 본 연구의 초기 통찰은 영상의학 및 기타 의료 영역에서 환자 중심 진료를 개선하기 위해 ChatGPT와 같은 LLM을 사용하는 엄청난 잠재력을 나타낸다. **임상적 관련성 진술:** 환자들은 의료 보고서를 간소화하고 설명하기 위해 ChatGPT를 사용하기 시작했으며, 이는 환자-의사 상호작용에 영향을 미칠 것으로 예상된다. 이러한 현상은 임상 일상 업무에 여러 기회와 도전과제를 제기한다. **Objectives:** To assess the quality of simplified radiology reports generated with the large language model (LLM) ChatGPT and to discuss challenges and chances of ChatGPT-like LLMs for medical text simplification. **Methods:** In this exploratory case study, a radiologist created three fictitious radiology reports which we simplified by prompting ChatGPT with "Explain this medical report to a child using simple language." In a questionnaire, we tasked 15 radiologists to rate the quality of the simplified radiology reports with respect to their factual correctness, completeness, and potential harm for patients. We used Likert scale analysis and inductive free-text categorization to assess the quality of the simplified reports. **Results:** Most radiologists agreed that the simplified reports were factually correct, complete, and not potentially harmful to the patient. Nevertheless, instances of incorrect statements, missed relevant medical information, and potentially harmful passages were reported. **Conclusion:** While we see a need for further adaption to the medical field, the initial insights of this study indicate a tremendous potential in using LLMs like ChatGPT to improve patient-centered care in radiology and other medical domains. **Clinical relevance statement:** Patients have started to use ChatGPT to simplify and explain their medical reports, which is expected to affect patient-doctor interaction. This phenomenon raises several opportunities and challenges for clinical routine. Key Points: • Patients have started to use

				ChatGPT to simplify their medical reports, but their quality was unknown. • In a questionnaire, most participating radiologists overall asserted good quality to radiology reports simplified with ChatGPT. However, they also highlighted a notable presence of errors, potentially leading patients to draw harmful conclusions. • Large language models such as ChatGPT have vast potential to enhance patient-centered care in radiology and other medical domains. To realize this potential while minimizing harm, they need supervision by medical experts and adaption to the medical field.
29	(Cilli Hayiroglu and Bozkurt 2025)	ChatGPT, Gemini, and Grok on familial mediterranean fever: are they trustworthy?	ChatGPT(4.0), Gemini(Palm2), Grok	**목적:** 본 연구는 희귀 자가염증질환인 가족성 지중해열(FMF)에 대한 의학 정보 제공에 있어 세 가지 주요 대규모 언어 모델(LLM) 기반 챗봇—ChatGPT, Gemini, Grok—의 정확성, 포괄성 및 일관성을 평가했다. **방법:** FMF에 대한 환자 중심의 자주 묻는 질문 49개를 기본 지식, 진단, 치료, 회복/위험/합병증/추적관찰의 네 가지 영역으로 분류했다. 각 질문은 ChatGPT, Gemini, Grok에 개별적으로 제출되었다. 2명의 임상 전문가가 3점 척도를 사용하여 응답을 독립적으로 평가했다: 포괄적/정확함, 불완전/부분적으로 정확함, 혼재/오해의 소지가 있음. 응답 재현성은 1주일 후 질문을 반복하여 평가했다. **결과:** Gemini가 가장 높은 정확도(87.7% 포괄적/정확한 응답)를 달성했으며, Grok(83.6%), ChatGPT(81.6%)가 그 뒤를 이었다. ChatGPT는 가장 높은 일관성(89.7% 재현성)을 보였으며 오해의 소지가 있는 내용을 제공하지 않았다. **Grok은 오해의 소지가 있는 응답(4.0%)을 생성한 유일한 모델이었으며, 주로 진단 관련 답변에서 나타났다.** 모든 챗봇은 회복/추적관찰 범주에서 가장 우수한 성능을 보였으며, 진단 성능은 현저히 낮았다. **결론:** LLM 기반 챗봇은 FMF 관련 의학 정보 제공에서 유망하지만 다양한 성능을 보였다. **전반적인 정확도는 높았으나, 비일관성과 간헐적인 오정보는 특히 복잡한 임상 주제에 대해 전문가 감독의 필요성을 강조한다**. 이러한 도구는 환자 교육을 향상시킬 수 있지만, 전문가의 지도를 대체해서는 안 된다. 향후 연구는 디지털 건강 환경에 안전하고 공평하게 통합되도록 다국어 역량과 사용자 이해도를 평가해야 한다. **Objective:** This study assessed the accuracy, comprehensiveness, and consistency of three prominent large language model (LLM)-based chatbots—ChatGPT, Gemini, and Grok—in delivering medical information about Familial Mediterranean Fever (FMF), a rare autoinflammatory disorder. **Methods:** Forty-nine frequently asked, patient-focused questions on FMF were categorized into four domains: basic knowledge, diagnosis, treatment, and recovery/risks/complications/follow-up. Each question was submitted

				individually to ChatGPT, Gemini, and Grok. Two clinical experts independently assessed the responses using a three-point scale: comprehensive/correct, incomplete/partially correct, and mixed/misleading. Response reproducibility was evaluated by repeating the queries 1 week later. **Results:** Gemini achieved the highest accuracy (87.7% comprehensive/correct responses), followed by Grok (83.6%) and ChatGPT (81.6%). ChatGPT demonstrated the greatest consistency (89.7% reproducibility) and provided no misleading content. Grok was the only model to produce misleading responses (4.0%), primarily in diagnosis-related answers. All chatbots performed best in the recovery/follow-up category; diagnostic performance was notably lower. **Conclusion:** LLM-based chatbots demonstrated promising but varied performance in delivering FMF-related medical information. While overall accuracy was high, inconsistencies and occasional misinformation highlight the need for expert supervision, especially for complex clinical topics. These tools may enhance patient education, but should not replace professional guidance. Future research should assess multilingual capabilities and user comprehension to ensure safe and equitable integration into digital health environments.
30	(Aydin et al. 2025)	Clinical Failure of General-Purpose AI in Photographic Scoliosis Assessment: A Diagnostic Accuracy Study	ChatGPT-4o, Claude 3.7 Sonnet	**배경 및 목적:** 범용 다중모드 대규모 언어 모델(LLM)은 임상 검증 없이 의료 영상 해석에 점점 더 많이 사용되고 있다. 본 연구는 청소년 특발성 척추측만증(AIS)의 사진 평가에서 ChatGPT-4o와 Claude 2의 진단 신뢰성을 방사선학적 표준과 비교하여 평가한다. 본 연구는 두 가지 핵심 질문을 다룬다: 가족들이 임상 사진 분석을 통해 LLM으로부터 신뢰할 수 있는 예비 평가를 얻을 수 있는지 여부와 LLM이 AIS 평가를 위한 시공간 추론 능력에서 인지적 충실성을 보이는지 여부. **재료 및 방법:** 전향적 진단 정확도 연구(STARD 준수)에서 97명의 청소년(AIS 74명 및 자세 비대칭 23명)을 분석했다. 표준화된 임상 사진(환자당 9개 뷰)을 2개의 LLM과 2명의 정형외과 전공의가 참조 방사선학적 측정값과 비교하여 평가했다. 주요 결과는 진단 정확도(민감도/특이도), Cobb 각도 일치도(Lin's CCC), 평가자 간 신뢰도(Cohen's κ) 및 측정 일치도(Bland-Altman LoA)를 포함했다. **결과:** LLM은 위험한 진단 부정확성을 보였다: ChatGPT는 모든 비-AIS 사례를 오분류했고(특이도 0% [95% CI: 0.0–14.8]), Claude 2는 78.3%의 위양성을 생성했다. 체계적 측정 오차가 임상 허용 범위를 초과했다: ChatGPT는 흉추 만곡을 +10.74° 과대평가했으며(LoA: −21.45° ~ +42.92°), 허용 범위를 〉800% 초과했다. 두 LLM 모두 흉요추 만곡에서 역 생체역학적 일치도를 보였다(CCC ≤ −0.106). 평가자 간 신뢰도는 무작위 수준 이하로 떨어졌다(ChatGPT κ =

				−0.039). 보편적 비례 편향(기울기 ≈ −1.0)으로 인해 심각한 만곡 과소평가가 발생했다(예: 50° 변형에 대해 10–15° 오차). 인간 평가자는 우수한 편향 제어(0.3–2.8° vs. 2.6–10.7°)를 보였으나 차선의 특이도(21.7–26.1%)와 위험한 요추 일치도(CCC: −0.123)를 나타냈다. 결론: 범용 LLM은 사진 기반 AIS 평가에서 임상적으로 수용할 수 없는 부정확성을 보이므로, 임상 배치는 금기이다. 치명적인 위양성, 허용 범위를 480–1074% 초과하는 체계적 측정 오차, 역 진단 일치도는 EU AI Act와 같은 프레임워크 하에서 긴급한 규제 안전장치를 필요로 한다. LLM도 사진 기반 인간 평가도 독립적 선별검사를 위한 신뢰도 기준을 달성하지 못하므로, 영역 특화 알고리즘 개발과 3D 모달리티 통합이 필수적이다. **Background and Objectives:** General-purpose multimodal large language models (LLMs) are increasingly used for medical image interpretation despite lacking clinical validation. This study evaluates the diagnostic reliability of ChatGPT-4o and Claude 2 in photographic assessment of adolescent idiopathic scoliosis (AIS) against radiological standards. This study examines two critical questions: whether families can derive reliable preliminary assessments from LLMs through analysis of clinical photographs and whether LLMs exhibit cognitive fidelity in their visuospatial reasoning capabilities for AIS assessment. **Materials and Methods:** A prospective diagnostic accuracy study (STARD-compliant) analyzed 97 adolescents (74 with AIS and 23 with postural asymmetry). Standardized clinical photographs (nine views/patient) were assessed by two LLMs and two orthopedic residents against reference radiological measurements. Primary outcomes included diagnostic accuracy (sensitivity/specificity), Cobb angle concordance (Lin's CCC), inter-rater reliability (Cohen's κ), and measurement agreement (Bland-Altman LoA). **Results:** The LLMs exhibited hazardous diagnostic inaccuracy: ChatGPT misclassified all non-AIS cases (specificity 0% [95% CI: 0.0–14.8]), while Claude 2 generated 78.3% false positives. Systematic measurement errors exceeded clinical tolerance: ChatGPT overestimated thoracic curves by +10.74° (LoA: −21.45° to +42.92°), exceeding tolerance by 〉800%. Both LLMs showed inverse biomechanical concordance in thoracolumbar curves (CCC ≤ −0.106). Inter-rater reliability fell below random chance (ChatGPT κ = −0.039). Universal proportional bias (slopes ≈ −1.0) caused severe curve underestimation (e.g., 10–15° error for 50° deformities). Human evaluators

				demonstrated superior bias control (0.3–2.8° vs. 2.6–10.7°) but suboptimal specificity (21.7–26.1%) and hazardous lumbar concordance (CCC: −0.123). **Conclusions:** General-purpose LLMs demonstrate clinically unacceptable inaccuracy in photographic AIS assessment, contraindicating clinical deployment. Catastrophic false positives, systematic measurement errors exceeding tolerance by 480–1074%, and inverse diagnostic concordance necessitate urgent regulatory safeguards under frameworks like the EU AI Act. Neither LLMs nor photographic human assessment achieve reliability thresholds for standalone screening, mandating domain-specific algorithm development and integration of 3D modalities.
31	(Wong et al. 2025)	Comparative Evaluation and Performance of Large Language Models in Clinical Infection Control Scenarios: A Benchmark Study	ChatGPT-4.1, DeepSeek V3, Gemini2.5 Pro Exp	**배경:** 병원의 감염 예방 및 관리(IPC)는 감염 예방 및 통제를 위해 복잡한 자문을 관리하는 감염관리간호사(ICN)에 크게 의존한다. 본 연구는 IPC 의사결정 과정에서 ICN을 지원하는 인공지능(AI) 도구로서 대규모 언어 모델(LLM)을 평가했다. 우리의 목표는 최고 수준의 안전성과 정확성을 유지하면서 IPC 실무의 효율성을 향상시키는 것이다. **방법:** 홍콩 Queen Mary Hospital에서 수행된 횡단면 벤치마킹 연구는 30개의 임상 감염관리 시나리오를 사용하여 세 가지 LLM—GPT-4.1, DeepSeek V3, Gemini 2.5 Pro Exp—을 평가했다. 각 모델은 시나리오를 이해하기 위한 명확화 질문을 생성한 후 두 가지 프롬프트 방법(개방형 질문 및 구조화된 템플릿)을 통해 IPC 권고사항을 제공했다. 선임 및 후임 ICN과 의사를 포함한 16명의 전문가가 이러한 응답을 일관성, 간결성, 유용성 및 관련성, 근거 품질, 실행 가능성(1–10점 척도)에 대해 평가했다. 정량적 및 정성적 분석을 통해 AI 성능, 신뢰성 및 임상 적용 가능성을 평가했다. **결과:** GPT-4.1과 DeepSeek V3는 복합 품질 척도에서 유의하게 높은 점수를 받았으며, 보정 평균(95% CI)은 각각 36.77 (33.98–39.57) 및 36.25 (33.45–39.04)로, Gemini 2.5 Pro Exp의 33.19 (30.39–35.99)에 비해 높았다($p < 0.001$). GPT-4.1은 근거 품질, 유용성 및 관련성에서 선두를 차지했다. **Gemini 2.5 Pro Exp는 구조화된 프롬프트 조건에서 50%의 시나리오에 대해 응답 생성에 실패했다.** 구조화된 프롬프트는 주로 근거 품질을 향상시켜 유의한 개선을 가져왔다($p < 0.001$). 평가자 배경이 점수에 영향을 미쳤으며, 의사가 간호사보다 높게 평가했다(38.83 vs. 32.06, $p < 0.001$). 그러나 **정성적 검토 결과 모든 모델에서 중대한 결함이 발견되었다. 예를 들어, DeepSeek V3에서는 비결핵 항산균을 고려하지 않고 항산균(AFB) 도말 양성만으로 결핵 치료를 제안했고, Gemini 2.5 Pro Exp에서는 Candida auris에 대한 격리 해제와 관련하여 비실용적이고 명확하지 않은 응답을 제공했다. 이러한 오류는 일반적으로 긍정적인 점수에도 불구하고 잠재적 안전 위험과 제한된 실제 적용 가능성을 강조**한다.

				결론: GPT-4.1과 DeepSeek V3가 유용한 IPC 조언을 제공하지만, 아직 자율적 사용에는 신뢰할 수 없다. 임상 판단 및 실제 적용 가능성의 중대한 오류는 LLM이 ICN의 전문성을 대체할 수 없음을 강조한다. 이러한 기술은 임상 의사결정을 자동화하기보다는 지원하는 보조 도구로 사용되어야 한다. **Background:** Infection prevention and control (IPC) in hospitals relies heavily on infection control nurses (ICNs) who manage complex consultations to prevent and control infections. This study evaluated large language models (LLMs) as artificial intelligence (AI) tools to support ICNs in IPC decision-making processes. Our goal is to enhance the efficiency of IPC practices while maintaining the highest standards of safety and accuracy. **Methods:** A cross-sectional benchmarking study at Queen Mary Hospital, Hong Kong assessed three LLMs—GPT-4.1, DeepSeek V3, and Gemini 2.5 Pro Exp—using 30 clinical infection control scenarios. Each model generated clarifying questions to understand the scenarios before providing IPC recommendations through two prompting methods: an open-ended inquiry and a structured template. Sixteen experts, including senior and junior ICNs and physicians, rated these responses on coherence, conciseness, usefulness and relevance, evidence quality, and actionability (1–10 scale). Quantitative and qualitative analyses assessed AI performance, reliability, and clinical applicability. **Results:** GPT-4.1 and DeepSeek V3 scored significantly higher on the composite quality scale, with adjusted means (95% CI) of 36.77 (33.98–39.57) and 36.25 (33.45–39.04), respectively, compared with Gemini 2.5 Pro Exp at 33.19 (30.39–35.99) ($p < 0.001$). GPT-4.1 led in evidence quality, usefulness, and relevance. Gemini 2.5 Pro Exp failed to generate responses in 50% of scenarios under structured prompt conditions. Structured prompting yielded significant improvements, primarily by enhancing evidence quality ($p < 0.001$). Evaluator background influenced scoring, with doctors rating outputs higher than nurses (38.83 vs. 32.06, $p < 0.001$). However, a qualitative review revealed critical deficiencies across all models, for example, tuberculosis treatment solely based on a positive acid-fast bacilli (AFB) smear without considering nontuberculous mycobacteria in DeepSeek V3 and providing an impractical and noncommittal response regarding the de-escalation of precautions for Candida auris in Gemini 2.5 Pro Exp. These errors highlight potential safety risks and limited real-world

				applicability, despite generally positive scores. **Conclusions:** While GPT-4.1 and DeepSeek V3 deliver useful IPC advice, they are not yet reliable for autonomous use. Critical errors in clinical judgment and practical applicability highlight that LLMs cannot replace the expertise of ICNs. These technologies should serve as adjunct tools to support, rather than automate, clinical decision-making.
32	(Dong et al. 2025)	Comparative evaluation of large language models in delivering guideline-compliant recommendations for topical NSAID use in musculoskeletal pain: a multidimensional analysis	DeepSeek-R1, ChatGPT-4o, Gemini, Grok-3	**서론:** 대규모 언어 모델(LLM)이 임상 의사결정 지원에 점점 더 많이 사용되고 있지만, 특히 근골격계 통증 관리에 대한 근거 기반 지침 준수는 충분히 연구되지 않았다. **방법:** 네 가지 LLM(DeepSeek-R1, ChatGPT-4o, Gemini, Grok-3)을 근골격계 통증에 대한 국소 NSAID 사용에 관한 응답을 통해 평가했다: 응답 품질 평가(정확성, 과도한 단정성, 보충 정보 및 불완전성), 표준화된 가독성 지표(Flesch Reading Ease, Flesch-Kincaid Grade Level), 그리고 실행 가능성을 정량화하기 위한 PEMAT-P 도구. **결과:** 네 가지 LLM은 정확성에서 유의한 변동성을 보였으며(ANOVA p = 0.045), Gemini가 가장 높은 점수(8.33 ± 0.77)를 받았고 DeepSeek-R1이 가장 낮았다(7.72 ± 1.52). 과도한 단정성에서도 유의한 차이가 있었으며(ANOVA p = 0.025), Grok-3가 가장 낮은 점수(4.56 ± 1.42)를, ChatGPT-4o가 가장 높은 점수(6.72 ± 1.49)를 받았다. ChatGPT-4o는 가장 많은 보충 내용을 제공했고(6.94 ± 2.29, p = 0.106), DeepSeek-R1은 가장 높은 불완전성을 보였다(5.00 ± 2.52, p = 0.261). 모든 모델이 권장 가독성 기준(9~10학년 수준)을 초과했으며, 실행 가능성 기준(≤ 33.5%)을 충족한 모델은 없었다. **결론:** LLM은 임상 보조 도구로서의 잠재력을 보여준다. Gemini와 Grok의 종합적 성능은 상대적으로 양호하지만, 가독성과 실행 가능성은 여전히 불만족스럽다. 향후 개발은 안전성을 보장하기 위해 임상의 피드백과 실제 검증을 통합해야 한다. 안전한 구현을 위해서는 인간의 감독과 표적화된 AI 훈련이 필수적이다. **Introduction:** While large language models (LLMs) are increasingly used in clinical decision support, their adherence to evidence-based guidelines—particularly for musculoskeletal pain management—remains understudied. **Methods:** Four LLMs (DeepSeek-R1, ChatGPT-4o, Gemini, Grok-3) were evaluated on their responses to topical NSAID use for musculoskeletal pain through: assessments of response quality (accuracy, over-conclusiveness, supplementary information, and incompleteness), standardized readability metrics (Flesch Reading Ease, Flesch-Kincaid Grade Level), and the PEMAT-P tool to quantify actionability.

				Results: The four LLMs showed significant variability in accuracy (ANOVA p = 0.045), with Gemini scoring highest (8.33 ± 0.77) and DeepSeek-R1 lowest (7.72 ± 1.52) and in over-conclusiveness (ANOVA p = 0.025), with Grok-3 scoring lowest (4.56 ± 1.42) and ChatGPT-4o highest 6.72 ± 1.49). ChatGPT-4o provided the most supplementary content (6.94 ± 2.29, p = 0.106) and DeepSeek-R1 had the highest incompleteness (5.00 ± 2.52, p = 0.261). All models exceeded recommended readability thresholds (9th-10th grade level), and none met the actionability standard (≤ 33.5%). **Conclusions:** LLMs demonstrate potential as clinical aids. The comprehensive performance of Gemini and Grok is relatively favorable, yet their readability and actionability remain unsatisfactory. Future development should integrate clinician feedback and real-world validation to ensure safety. Human oversight and targeted AI training are critical for safe implementation.
33	(Wakonig et al. 2025)	Comparing ChatGPT 4.0's Performance in Interpreting Thyroid Nodule Ultrasound Reports Using ACR-TI-RADS 2017: Analysis Across Different Levels of Ultrasound User Experience	ChatGPT-4.0	**배경/목적:** 본 연구는 ACR-TI-RADS 2017 기준을 사용하여 갑상선 초음파(US) 판독문을 해석하는 ChatGPT 4.0의 능력을 평가하고, 다양한 수준의 초음파 사용자와 성능을 비교한다. **방법:** 의료 전문가 팀, 초음파 사용 경험이 없는 사용자, ChatGPT 4.0이 100개의 가상 갑상선 초음파 판독문을 분석했다. ChatGPT의 성능은 세침흡인(FNA) 및 추적 관찰을 포함한 정확성, 일관성 및 진단 권고사항에 대해 평가되었다. **결과:** ChatGPT는 에코 발생 병소 평가에서 전문가와 상당한 일치도를 보였으나, **구성 및 경계와 같은 다른 기준에서는 두 분석 모두에서 비일관성이 명백했다. ChatGPT와 전문가 간의 평가자 간 신뢰도는 중등도에서 거의 완벽한 수준까지 다양했으며, 이는 AI의 잠재력과 함께 전문가 수준의 해석을 달성하는 데 있어 한계를 반영**한다. 초음파 사용 경험이 없는 사용자는 전문가와 거의 완벽한 일치도를 보이며 ChatGPT를 능가했으며, 이는 TI-RADS와 같은 표준화된 위험 계층화 도구에서 전통적인 의학 교육의 중요한 역할을 강조한다. **결론: ChatGPT는 FNA 권고에서 높은 특이도를 보였으나, 의대생에 비해 추적관찰에 대한 민감도와 특이도가 낮았다.** 이러한 결과는 ChatGPT가 인간 전문성을 대체하기보다는 지원 진단 도구로서의 잠재력을 가지고 있음을 강조한다. AI 알고리즘과 훈련을 향상시키면 ChatGPT의 임상 유용성을 개선하여 임상의가 갑상선 결절을 관리하고 환자 치료를 개선하는 데 더 나은 지원을 제공할 수 있다. 본 연구는 의료 진단에서 AI의 가능성과 현재의 한계를 모두 강조하며, AI의 정제 및 임상 워크플로우로의 통합을 옹호한다. 그러나 AI 주도 오류를 식별하고 수정하는 데 필수적이므로 전통적인 임상 훈련이 훼손되어서는 안 된다는 점을 강조한다.

				Background/Objectives: This study evaluates ChatGPT 4.0's ability to interpret thyroid ultrasound (US) reports using ACR-TI-RADS 2017 criteria, comparing its performance with different levels of US users. **Methods:** A team of medical experts, an inexperienced US user, and ChatGPT 4.0 analyzed 100 fictitious thyroid US reports. ChatGPT's performance was assessed for accuracy, consistency, and diagnostic recommendations, including fine-needle aspirations (FNA) and follow-ups. **Results:** ChatGPT demonstrated substantial agreement with experts in assessing echogenic foci, but inconsistencies in other criteria, such as composition and margins, were evident in both its analyses. Interrater reliability between ChatGPT and experts ranged from moderate to almost perfect, reflecting AI's potential but also its limitations in achieving expert-level interpretations. The inexperienced US user outperformed ChatGPT with a nearly perfect agreement with the experts, highlighting the critical role of traditional medical training in standardized risk stratification tools such as TI-RADS. **Conclusions:** ChatGPT showed high specificity in recommending FNAs but lower sensitivity and specificity for follow-ups compared to the medical student. These findings emphasize ChatGPT's potential as a supportive diagnostic tool rather than a replacement for human expertise. Enhancing AI algorithms and training could improve ChatGPT's clinical utility, enabling better support for clinicians in managing thyroid nodules and improving patient care. This study highlights both the promise and current limitations of AI in medical diagnostics, advocating for its refinement and integration into clinical workflows. However, it emphasizes that traditional clinical training must not be compromised, as it is essential for identifying and correcting AI-driven errors.
34	(Fraser et al. 2023)	Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study	ChatGPT-3.5, 4.0	**배경:** 진단은 효과적인 의료의 핵심 구성요소이지만, 오진은 흔하며 환자를 위험에 빠뜨릴 수 있다. 진단 의사결정 지원 시스템은 의사 및 기타 의료 종사자의 진단을 개선하는 데 역할을 할 수 있다. 증상 확인기(SC)는 환자의 진단 및 중증도 분류(즉, 어떤 수준의 치료를 받아야 하는지)를 개선하기 위해 설계되었다. **목적:** 본 연구의 목적은 새로운 대규모 언어 모델 ChatGPT(버전 3.5 및 4.0), 널리 사용되는 WebMD SC, Ada Health에서 개발한 SC의 응급 또는 긴급 임상 문제를 가진 환자의 진단 및 중증도 분류 성능을 최종 응급실(ED) 진단 및 의사 검토와 비교하여 평가하는 것이었다. **방법:** 우리는 응급실에 내원하여 ED 의사를 만나기 전에 Ada SC를 사용하여 증상

				을 기록한 40명의 환자로부터 이전에 수집된 비식별화 자가보고 데이터를 사용했다. 비식별화 데이터는 진단 및 중증도 분류에 대해 눈가림된 연구 보조원이 ChatGPT 버전 3.5 및 4.0과 WebMD에 입력했다. 4개 시스템 모두의 진단은 이전에 추출된 ED의 최종 진단 및 Ada의 자가보고 임상 데이터를 눈가림 검토한 3명의 독립적인 응급의학과 전문의의 진단 및 중증도 분류 권고사항과 비교되었다. 진단 정확도는 ChatGPT, Ada SC, WebMD SC, 독립 의사의 진단 중 적어도 하나의 ED 진단과 일치하는 비율로 계산되었다(상위 1개 또는 상위 3개로 계층화). 중증도 분류 정확도는 ChatGPT, WebMD, Ada의 권고사항 중 독립 의사 중 적어도 2명과 일치하거나 "안전하지 않음" 또는 "지나치게 신중함"으로 평가된 수로 계산되었다. **결과:** 전체적으로 30건과 37건이 각각 진단 및 중증도 분류 분석에 충분한 데이터를 가지고 있었다. Ada, ChatGPT 3.5, ChatGPT 4.0, WebMD의 상위 1개 진단 일치율은 각각 9건(30%), 12건(40%), 10건(33%), 12건(40%)이었으며, 의사의 평균 비율은 47%였다. Ada, ChatGPT 3.5, ChatGPT 4.0, WebMD의 상위 3개 진단 일치율은 각각 19건(63%), 19건(63%), 15건(50%), 17건(57%)이었으며, 의사의 평균 비율은 69%였다. Ada의 중증도 분류 결과 분포는 일치 62%(n=23), 안전하지 않음 14%(n=5), 지나치게 신중함 24%(n=9)였고; ChatGPT 3.5는 일치 59%(n=22), 안전하지 않음 41%(n=15), 지나치게 신중함 0%(n=0)였으며; ChatGPT 4.0은 일치 76%(n=28), 안전하지 않음 22%(n=8), 지나치게 신중함 3%(n=1)였고; WebMD는 일치 70%(n=26), 안전하지 않음 19%(n=7), 지나치게 신중함 11%(n=4)였다. ChatGPT 3.5의 안전하지 않은 중증도 분류율(41%)은 Ada(14%)보다 유의하게 높았다(P=.009). **결론:** ChatGPT 3.5는 높은 진단 정확도를 보였으나 안전하지 않은 중증도 분류율이 높았다. ChatGPT 4.0은 진단 정확도가 가장 낮았으나, 안전하지 않은 중증도 분류율은 낮았고 의사와의 중증도 분류 일치도가 가장 높았다. Ada 및 WebMD SC는 ChatGPT보다 전반적으로 더 나은 성능을 보였다. 중증도 분류 정확도 개선 및 광범위한 임상 평가 없이는 환자의 무감독 ChatGPT 사용은 진단 및 중증도 분류에 권장되지 않는다. BACKGROUND: Diagnosis is a core component of effective health care, but misdiagnosis is common and can put patients at risk. Diagnostic decision support systems can play a role in improving diagnosis by physicians and other health care workers. Symptom checkers (SCs) have been designed to improve diagnosis and triage (ie, which level of care to seek) by patients. OBJECTIVE: The aim of this study was to evaluate the performance of the new large language model ChatGPT (versions 3.5 and 4.0), the widely used

				WebMD SC, and an SC developed by Ada Health in the diagnosis and triage of patients with urgent or emergent clinical problems compared with the final emergency department (ED) diagnoses and physician reviews. METHODS: We used previously collected, deidentified, self-report data from 40 patients presenting to an ED for care who used the Ada SC to record their symptoms prior to seeing the ED physician. Deidentified data were entered into ChatGPT versions 3.5 and 4.0 and WebMD by a research assistant blinded to diagnoses and triage. Diagnoses from all 4 systems were compared with the previously abstracted final diagnoses in the ED as well as with diagnoses and triage recommendations from three independent board-certified ED physicians who had blindly reviewed the self-report clinical data from Ada. Diagnostic accuracy was calculated as the proportion of the diagnoses from ChatGPT, Ada SC, WebMD SC, and the independent physicians that matched at least one ED diagnosis (stratified as top 1 or top 3). Triage accuracy was calculated as the number of recommendations from ChatGPT, WebMD, or Ada that agreed with at least 2 of the independent physicians or were rated unsafe" or "too cautious." RESULTS: Overall, 30 and 37 cases had sufficient data for diagnostic and triage analysis, respectively. The rate of top-1 diagnosis matches for Ada, ChatGPT 3.5, ChatGPT 4.0, and WebMD was 9 (30%), 12 (40%), 10 (33%), and 12 (40%), respectively, with a mean rate of 47% for the physicians. The rate of top-3 diagnostic matches for Ada, ChatGPT 3.5, ChatGPT 4.0, and WebMD was 19 (63%), 19 (63%), 15 (50%), and 17 (57%), respectively, with a mean rate of 69% for physicians. The distribution of triage results for Ada was 62% (n=23) agree, 14% unsafe (n=5), and 24% (n=9) too cautious; that for ChatGPT 3.5 was 59% (n=22) agree, 41% (n=15) unsafe, and 0% (n=0) too cautious; that for ChatGPT 4.0 was 76% (n=28) agree, 22% (n=8) unsafe, and 3% (n=1) too cautious; and that for WebMD was 70% (n=26) agree, 19% (n=7) unsafe, and 11% (n=4) too cautious. The unsafe triage rate for ChatGPT 3.5 (41%) was significantly higher (P=.009) than that of Ada (14%). CONCLUSIONS: ChatGPT 3.5 had high diagnostic accuracy but a high unsafe triage rate. ChatGPT 4.0 had the poorest diagnostic accuracy, but a lower unsafe triage rate and the highest triage agreement with the physicians. The Ada and WebMD SCs performed better overall than ChatGPT. Unsupervised patient use of ChatGPT for diagnosis and triage is not recommended without improvements to triage accuracy and extensive clinical evaluation."

35	(Barlas et al. 2024)	Credibility of ChatGPT in the assessment of obesity in type 2 diabetes according to the guidelines	ChatGPT-3.5	**배경:** Chat Generative Pre-trained Transformer(ChatGPT)는 의료 분야의 학생, 연구자 및 환자가 정보에 쉽게 접근할 수 있도록 하며 현재 주목을 받고 있다. 우리는 금세기의 주요 관심사 중 하나인 제2형 당뇨병(T2D)에서의 비만 평가 지침에 따라 ChatGPT의 신뢰성을 평가하고자 했다. **재료 및 방법:** 이 횡단면 비인간 대상 연구에서 경험이 풍부한 내분비 전문의들이 미국 당뇨병 학회 및 미국 임상 내분비학회 지침에 따른 비만에 대한 평가 및 다양한 치료 옵션의 하위 섹션으로 ChatGPT에 20개의 질문을 했다. ChatGPT의 응답은 네 가지 범주로 분류되었다: 지침과 부합, 부합하나 불충분, 부분적으로 불일치, 지침과 불일치. **결과:** ChatGPT는 질문에 답변하는 데 체계적인 접근 방식을 보여주었으며, 환자의 특정 건강 요구와 상황에 기반한 개인화된 조언을 받기 위해 의료 제공자와 상담할 것을 권장했다. **제2형 당뇨병에서의 비만 평가에 대한 ChatGPT의 지침 부합도는 100%였으나, 영양, 약물 및 수술적 체중 감량 접근법을 포함하는 치료 섹션에서는 더 낮았다. 또한 ChatGPT는 "부합하나 불충분"으로 평가된 응답에 대해 지침의 모든 정보를 제공하기 위해 추가 프롬프트가 필요**했다. **결론:** T2D에서의 비만 평가 및 관리는 매우 개별화되어 있다. ChatGPT의 포괄적이고 이해하기 쉬운 응답에도 불구하고, 의료 전문가의 환자 중심 접근법을 대체하는 수단으로 사용되어서는 안 된다. **Background:** The Chat Generative Pre-trained Transformer (ChatGPT) allows students, researchers, and patients in the medical field to access information easily and has gained attention nowadays. We aimed to evaluate the credibility of ChatGPT according to the guidelines for the assessment of obesity in type 2 diabetes (T2D), which is one of the major concerns of this century. **Materials and method:** In this cross-sectional non-human subject study, experienced endocrinologists posed 20 questions to ChatGPT in subsections, which were assessments and different treatment options for obesity according to the American Diabetes Association and American Association of Clinical Endocrinology guidelines. The responses of ChatGPT were classified into four categories: compatible, compatible but insufficient, partially incompatible and incompatible with the guidelines. **Results:** ChatGPT demonstrated a systematic approach to answering questions and recommended consulting a healthcare provider to receive personalized advice based on the specific health needs and circumstances of patients. The compatibility of ChatGPT with the guidelines was 100% in

				the assessment of obesity in type 2 diabetes; however, it was lower in the therapy sections, which included nutritional, medical, and surgical approaches to weight loss. Furthermore, ChatGPT required additional prompts for responses that were evaluated as "compatible but insufficient" to provide all the information in the guidelines. **Conclusion:** The assessment and management of obesity in T2D are highly individualized. Despite ChatGPT's comprehensive and understandable responses, it should not be used as a substitute for healthcare professionals' patient-centered approach.
36	(Menz et al. 2024)	Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis	GPT-4 (via ChatGPT and Microsoft's Copilot), Google's PaLM 2 and Gemini Pro (via Bard), Anthropic's Claude 2 (via Poe), Meta's Llama 2 (via HuggingChat).	**목적:** 대규모 언어 모델(LLM)이 건강 허위정보 생성에 오용되는 것을 방지하기 위한 안전장치의 효과성을 평가하고, 관찰된 취약점에 대한 위험 완화 프로세스와 관련하여 인공지능(AI) 개발자의 투명성을 평가한다. **설계:** 반복 횡단면 분석. **설정:** 공개적으로 접근 가능한 LLM. **방법:** 반복 횡단면 분석에서 네 가지 LLM(챗봇/어시스턴트 인터페이스를 통해)을 평가했다: OpenAI의 GPT-4(ChatGPT 및 Microsoft의 Copilot을 통해), Google의 PaLM 2 및 새로 출시된 Gemini Pro(Bard를 통해), Anthropic의 Claude 2(Poe를 통해), Meta의 Llama 2(HuggingChat를 통해). 2023년 9월, 이러한 LLM에 두 가지 주제에 대한 건강 허위정보 생성을 요청했다: 자외선 차단제가 피부암의 원인이라는 주장과 알칼리 식단이 암 치료법이라는 주장. 필요한 경우 탈옥 기법(즉, 안전장치 우회 시도)을 평가했다. 안전장치 취약점이 관찰된 LLM에 대해 우려되는 출력을 보고하는 프로세스를 감사했다. 초기 조사 12주 후, LLM의 허위정보 생성 능력을 재평가하여 안전장치의 후속 개선 사항을 평가했다. **주요 결과 측정:** 주요 결과 측정은 안전장치가 건강 허위정보 생성을 방지했는지 여부와 건강 허위정보에 대한 위험 완화 프로세스의 투명성이었다. **결과:** Claude 2(Poe를 통해)는 두 연구 시점에 걸쳐 제출된 130개의 프롬프트를 거부했으며, 이는 자외선 차단제가 피부암을 유발한다거나 알칼리 식단이 암 치료법이라는 주장의 콘텐츠 생성을 요청한 것으로, 탈옥 시도에도 불구하고 거부했다. **GPT-4(Copilot을 통해)는 초기에 탈옥 시도에도 불구하고 건강 허위정보 생성을 거부했으나, 12주 후에는 그렇지 않았다**. 반면, GPT-4(ChatGPT를 통해), PaLM 2/Gemini Pro(Bard를 통해), Llama 2(HuggingChat를 통해)는 지속적으로 건강 허위정보 블로그를 생성했다. **2023년 9월 평가에서 이러한 LLM은 탈옥 시도 없이 총 40,000단어가 넘는 113개의 고유한 암 허위정보 블로그 생성을 촉진했다**. 평가 시점 전반에 걸친 이러한 LLM의 거부율은 단 5%(150건 중 7건)에 불과했으며, 프롬프트에 따라 LLM이 생성한 블로그는 주목을 끄는 제목, 진짜처럼 보이는 (가짜 또는 허구의) 참고문헌, 환자와 임상의의 조작된 증언을 포함했으며, 다양한 인구통

				계학적 그룹을 대상으로 했다. 평가된 각 LLM에는 우려되는 관찰된 출력을 보고하는 메커니즘이 있었지만, 취약점 관찰이 보고되었을 때 개발자는 응답하지 않았다. **결론: 본 연구는 LLM이 건강 허위정보 생성에 오용되는 것을 방지하는 효과적인 안전장치가 가능하지만 일관되게 구현되지 않았음을 발견했다**. 또한 안전장치 문제를 보고하는 효과적인 프로세스가 부족했다. LLM이 건강 허위정보의 대량 생성에 기여하는 것을 방지하기 위해서는 규제 강화, 투명성 제고 및 정기적인 감사가 필요하다. OBJECTIVES To evaluate the effectiveness of safeguards to prevent large language models (LLMs) from being misused to generate health disinformation, and to evaluate the transparency of artificial intelligence (AI) developers regarding their risk mitigation processes against observed vulnerabilities. DESIGN Repeated cross sectional analysis. SETTING Publicly accessible LLMs. METHODS In a repeated cross sectional analysis, four LLMs (via chatbots/assistant interfaces) were evaluated: OpenAI's GPT-4 (via ChatGPT and Microsoft's Copilot), Google's PaLM 2 and newly released Gemini Pro (via Bard), Anthropic's Claude 2 (via Poe), and Meta's Llama 2 (via HuggingChat). In September 2023, these LLMs were prompted to generate health disinformation on two topics: sunscreen as a cause of skin cancer and the alkaline diet as a cancer cure. Jailbreaking techniques (ie, attempts to bypass safeguards) were evaluated if required. For LLMs with observed safeguarding vulnerabilities, the processes for reporting outputs of concern were audited. 12 weeks after initial investigations, the disinformation generation capabilities of the LLMs were re-evaluated to assess any subsequent improvements in safeguards. MAIN OUTCOME MEASURES The main outcome measures were whether safeguards prevented the generation of health disinformation, and the transparency of risk mitigation processes against health disinformation. RESULTS Claude 2 (via Poe) declined 130 prompts submitted across the two study timepoints requesting the generation of content claiming that sunscreen causes skin cancer or that the alkaline diet is a cure for cancer, even with jailbreaking attempts. GPT-4 (via Copilot) initially refused to generate health disinformation, even with jailbreaking attempts—although this was not the case at 12 weeks. In contrast, GPT-4 (via ChatGPT), PaLM

				2/Gemini Pro (via Bard), and Llama 2 (via HuggingChat) consistently generated health disinformation blogs. In September 2023 evaluations, these LLMs facilitated the generation of 113 unique cancer disinformation blogs, totalling more than 40 000 words, without requiring jailbreaking attempts. The refusal rate across the evaluation timepoints for these LLMs was only 5% (7 of 150), and as prompted the LLM generated blogs incorporated attention grabbing titles, authentic looking (fake or fictional) references, fabricated testimonials from patients and clinicians, and they targeted diverse demographic groups. Although each LLM evaluated had mechanisms to report observed outputs of concern, the developers did not respond when observations of vulnerabilities were reported. CONCLUSIONS This study found that although effective safeguards are feasible to prevent LLMs from being misused to generate health disinformation, they were inconsistently implemented. Furthermore, effective processes for reporting safeguard problems were lacking. Enhanced regulation, transparency, and routine auditing are required to help prevent LLMs from contributing to the mass generation of health disinformation.
37	(Toiv et al. 2024)	Digesting Digital Health: A Study of Appropriateness and Readability of ChatGPT–Generated Gastroenterological Information	ChatGPT	**서론:** 복잡한 질문에 대한 상호작용적 응답을 생성할 수 있는 인공지능 기반 대규모 언어 모델의 출현은 환자들이 의료 정보에 접근하는 방식에 획기적인 발전을 의미한다. 우리의 목표는 Chat Generative Pretrained Transformer(ChatGPT)가 생성한 소화기 정보의 적절성과 가독성을 평가하는 것이었다. **방법:** 위장관 질환의 증상과 치료를 평가하는 16개의 대화 기반 질문과 소화기학의 주요 주제에 대한 13개의 정의 기반 질문에 대해 ChatGPT가 생성한 응답을 분석했다. 3명의 전문의 자격을 갖춘 소화기내과 전문의가 최신성, 관련성, 정확성, 포괄성, 명확성, 긴급성/다음 단계에 대한 5점 리커트 척도 대리 측정으로 출력 적절성을 평가했다. 6개 범주 모두에서 4점 또는 5점을 받은 출력은 "적절함"으로 지정되었다. 출력 가독성은 Flesch Reading Ease 점수, Flesch–Kincaid Reading Level, Simple Measure of Gobbledygook 점수로 평가되었다. **결과:** 16개의 대화 기반 질문 중 44%, 13개의 정의 기반 질문 중 69%에 대한 ChatGPT 응답이 적절한 것으로 간주되었으며, 두 질문 그룹 내 적절한 응답의 비율은 유의한 차이가 없었다(P = 0.17). **특히 위장관 응급상황과 관련된 질문에 대한 ChatGPT의 응답 중 적절하다고 지정된 것은 하나도 없었다. 평균 가독성 점수는 출력이 대학 수준의 읽기 능력으로 작성되었음을 보여주었다.** **고찰:** ChatGPT는 소화기 의학 질문에 일반적으로 적합한 응답을 생성할 수 있지만, **응답은 적절성과 가독성에서 제한적**이었으며, 이는 이 대규모 언어 모델의 현재 유용성을 제한한다. 이러한 모델이 신뢰할 수 있는 의료 정보 출처로 명백하게 지지받

				기 전에 상당한 발전이 필수적이다. **INTRODUCTION:** The advent of artificial intelligence-powered large language models capable of generating interactive responses to intricate queries marks a groundbreaking development in how patients access medical information. Our aim was to evaluate the appropriateness and readability of gastroenterological information generated by Chat Generative Pretrained Transformer (ChatGPT). **METHODS:** We analyzed responses generated by ChatGPT to 16 dialog-based queries assessing symptoms and treatments for gastrointestinal conditions and 13 definition-based queries on prevalent topics in gastroenterology. Three board-certified gastroenterologists evaluated output appropriateness with a 5-point Likert-scale proxy measurement of currency, relevance, accuracy, comprehensiveness, clarity, and urgency/next steps. Outputs with a score of 4 or 5 in all 6 categories were designated as appropriate."Output readability was assessed with Flesch Reading Ease score, Flesch-Kinkaid Reading Level, and Simple Measure of Gobbledygook scores. **RESULTS:** ChatGPT responses to 44% of the 16 dialog-based and 69% of the 13 definition-based questions were deemed appropriate, and the proportion of appropriate responses within the 2 groups of questions was not significantly different (P = 0.17). Notably, none of ChatGPT's responses to questions related to gastrointestinal emergencies were designated appropriate. The mean readability scores showed that outputs were written at a college-level reading proficiency. **DISCUSSION:** ChatGPT can produce generally fitting responses to gastroenterological medical queries, but responses were constrained in appropriateness and readability, which limits the current utility of this large language model. Substantial development is essential before these models can be unequivocally endorsed as reliable sources of medical information."
38	(Lauderdale et al. 2025)	Effectiveness of generative AI-large language models' recognition of veteran suicide risk: a comparison with human mental health providers using a risk stratification model	ChatGPT-3.5 ChatGPT-4o Google Gemini	**배경:** 매년 6,300명 이상의 미국 참전용사가 자살로 사망함에 따라, 재향군인 건강관리청(VHA)은 자살 위험 평가를 위해 인공지능(AI)을 포함한 혁신적인 전략을 모색하고 있다. 기계학습이 주로 활용되어 왔지만, 생성형 AI-대규모 언어 모델(GAI-LLM)의 적용은 아직 탐색되지 않았다. **목적:** 본 연구는 GAI-LLM, 특히 ChatGPT-3.5, ChatGPT-4o, Google Gemini가 표준화된 참전용사 사례에 대응하여 VHA의 위험 계층화 표를 사용하여 자살 위험

				을 식별하고 치료 권고사항을 제시하는 효과성을 평가한다. **방법:** 우리는 급성 및 만성 자살 위험에 대한 GAI-LLM의 평가 및 권고사항을 정신건강 의료 제공자(MHCP)의 평가와 비교했다. 다양한 수준의 자살 위험을 나타내는 4개의 사례가 조사되었다. **결과:** GAI-LLM의 평가는 MHCP와 불일치를 보였으며, 특히 가장 급성인 사례를 덜 급성으로, 가장 덜 급성인 사례를 더 급성으로 평가했다. 만성 위험의 경우, GAI-LLM의 평가는 일반적으로 MHCP와 일치했으나, GAI-LLM이 더 높은 만성 위험으로 평가한 한 사례는 예외였다. GAI-LLM 간의 변동성도 관찰되었다. 특히 ChatGPT-3.5는 ChatGPT-4o 및 Google Gemini에 비해 낮은 급성 위험 평가를 보였으며, ChatGPT-4o는 더 높은 만성 위험 평가를 식별하고 모든 참전용사에 대해 입원을 권고했다. GAI-LLM의 치료 계획은 급성 위험 평가가 아닌 만성 위험 평가에 의해 예측되었다. **결론:** GAI-LLM은 MHCP와 비교 가능한 자살 위험 평가의 잠재력을 제공하지만, 위험 평가 및 치료 권고사항 모두에서 서로 다른 GAI-LLM 간에 상당한 변동성이 존재한다. 정확성과 적절한 치료를 보장하기 위해서는 지속적인 MHCP 감독이 필수적이다. **Background:** With over 6,300 United States military veterans dying by suicide annually, the Veterans Health Administration (VHA) is exploring innovative strategies, including artificial intelligence (AI), for suicide risk assessment. Machine learning has been predominantly utilized, but the application of generative AI–large language models (GAI-LLMs) remains unexplored. **Objective:** This study evaluates the effectiveness of GAI-LLMs, specifically ChatGPT-3.5, ChatGPT-4o, and Google Gemini, in using the VHA's Risk Stratification Table for identifying suicide risks and making treatment recommendations in response to standardized veteran vignettes. **Methods:** We compared the GAI-LLMs' assessments and recommendations for both acute and chronic suicide risks to evaluations by mental health care providers (MHCPs). Four vignettes, representing varying levels of suicide risk, were used. **Results:** GAI-LLMs' assessments showed discrepancies with MHCPs, particularly rating the most acute case as less acute and the least acute case as more acute. For chronic risk, GAI-LLMs' evaluations were generally in line with MHCPs, except for one vignette rated with higher chronic risk by the GAI-LLM. Variation across GAI-LLMs was also observed. Notably, ChatGPT-3.5 showed lower acute risk ratings compared to ChatGPT-4o and Google

				Gemini, while ChatGPT-4o identified higher chronic risk ratings and recommended hospitalization for all veterans. Treatment planning by GAI-LLMs was predicted by chronic but not acute risk ratings. **Conclusion:** While GAI-LLMs offers potential suicide risk assessment comparable to MHCPs, significant variation exists across different GAI-LLMs in both risk evaluation and treatment recommendations. Continued MHCP oversight is essential to ensure accuracy and appropriate care. Implications: These findings highlight the need for further research into optimizing GAI-LLMs for consistent and reliable use in clinical settings, ensuring they complement rather than replace human expertise.
39	(Zhong et al. 2025)	Enhancing the Accuracy of Human Phenotype Ontology Identification: Comparative Evaluation of Multimodal Large Language Models	ChatGPT-4o, Llama3.2:11b, Llama3.2:90b	**배경:** 인간 표현형 온톨로지(HPO) 용어 식별은 희귀 질환의 진단 및 관리에 중요하다. 그러나 임상의, 특히 후임 의사들은 환자 표현형을 정확하게 기술하는 것의 복잡성으로 인해 종종 어려움에 직면한다. HPO 데이터베이스를 사용한 전통적인 수동 검색 방법은 시간이 많이 소요되고 오류가 발생하기 쉽다. **목적:** 본 연구의 목적은 다중모드 대규모 언어 모델(MLLM)의 사용이 희귀 질환과 관련된 환자 이미지로부터 HPO 용어를 식별하는 후임 의사의 정확성을 향상시킬 수 있는지 조사하는 것이다. **방법:** 10개 전문과의 후임 의사 20명이 참여했다. 각 의사는 중국 희귀 질환 목록에 등재된 희귀 질환과 관련된 표현형을 가진 공개 문헌에서 추출한 27개의 환자 이미지를 평가했다. 연구는 두 그룹으로 나뉘었다: 수동 검색 그룹은 중국 인간 표현형 온톨로지 웹사이트에 의존했고, MLLM 지원 그룹은 ChatGPT-4o가 사전 식별한 HPO 용어를 프롬프트로 포함한 전자 설문지를 사용한 후 중국 인간 표현형 온톨로지를 사용하여 검색했다. 주요 결과는 HPO 식별 정확도로, 전문가 패널이 결정한 표준 세트와 비교하여 올바르게 식별된 HPO 용어의 비율로 정의되었다. 또한 ChatGPT-4o와 2개의 오픈소스 MLLM(Llama3.2:11b 및 Llama3.2:90b)의 출력 정확도를 동일한 기준으로 평가했으며, 각 모델의 환각은 별도로 기록되었다. 더욱이 참여 의사들은 표준화된 HPO 용어를 사용하여 환자 이미지를 정확하게 기술하는 능력에 영향을 미치는 요인을 식별하기 위해 희귀 질환 배경에 관한 추가 전자 설문지를 작성했다. **결과:** 그룹당 총 270개의 기술이 평가되었다. MLLM 지원 그룹은 수동 그룹의 20.4%(55/270)에 비해 67.4%(182/270)의 유의하게 높은 정확도를 달성했다(상대위험도 3.31, 95% CI 2.58-4.25; P<.001). MLLM 지원 그룹은 부서 전반에 걸쳐 일관된 성능을 보인 반면, 수동 그룹은 더 큰 변동성을 보였다. 독립형 MLLM 중 ChatGPT-4o는 48%(13/27)의 정확도를 달성했으며, 오픈소스 모델 Llama3.2:11b 및 Llama3.2:90b는 각각 15%(4/27) 및 18%(5/27)를 달성했다. 그러나 **MLLM은 높은 환각률을 보였으며, 잘못된 ID를 가진 HPO 용어 또는 완전**

				히 조작된 내용을 자주 생성했다. 구체적으로 ChatGPT-4o, Llama3.2:11b, Llama3.2:90b는 각각 57.3%(67/117), 98%(62/63), 82%(46/56)의 사례에서 잘못된 ID를 생성했고, 각각 34.2%(40/117), 41%(26/63), 32%(18/56)의 사례에서 조작된 용어를 생성했다. 또한 후임 의사의 희귀 질환 지식에 대한 설문조사는 희귀 질환 및 유전 질환 훈련 참여가 일부 의사의 성능을 향상시킬 수 있음을 시사한다. **결론:** MLLM의 임상 워크플로우 통합은 후임 의사의 HPO 식별 정확도를 크게 향상시키며, 희귀 질환 진단을 개선하고 의학 연구에서 표현형 기술을 표준화할 유망한 잠재력을 제공한다. 그러나 MLLM에서 관찰된 주목할 만한 환각률은 임상 실무에서 광범위하게 채택되기 전에 추가적인 정제 및 엄격한 검증의 필요성을 강조한다. **Background:** Identifying Human Phenotype Ontology (HPO) terms is crucial for diagnosing and managing rare diseases. However, clinicians, especially junior physicians, often face challenges due to the complexity of describing patient phenotypes accurately. Traditional manual search methods using HPO databases are time-consuming and prone to errors. **Objective:** The aim of the study is to investigate whether the use of multimodal large language models (MLLMs) can improve the accuracy of junior physicians in identifying HPO terms from patient images related to rare diseases. **Methods:** In total, 20 junior physicians from 10 specialties participated. Each physician evaluated 27 patient images sourced from publicly available literature, with phenotypes relevant to rare diseases listed in the Chinese Rare Disease Catalogue. The study was divided into 2 groups: the manual search group relied on the Chinese Human Phenotype Ontology website, while the MLLM-assisted group used an electronic questionnaire that included HPO terms preidentified by ChatGPT-4o as prompts, followed by a search using the Chinese Human Phenotype Ontology. The primary outcome was the accuracy of HPO identification, defined as the proportion of correctly identified HPO terms compared to a standard set determined by an expert panel. Additionally, the accuracy of outputs from ChatGPT-4o and 2 open-source MLLMs (Llama3.2:11b and Llama3.2:90b) was evaluated using the same criteria, with hallucinations for each model documented separately. Furthermore, participating physicians completed an additional electronic questionnaire regarding their rare disease background to identify factors

				affecting their ability to accurately describe patient images using standardized HPO terms. **Results:** A total of 270 descriptions were evaluated per group. The MLLM-assisted group achieved a significantly higher accuracy rate of 67.4% (182/270) compared to 20.4% (55/270) in the manual group (relative risk 3.31, 95% CI 2.58-4.25; P<.001). The MLLM-assisted group demonstrated consistent performance across departments, whereas the manual group exhibited greater variability. Among standalone MLLMs, ChatGPT-4o achieved an accuracy of 48% (13/27), while the open-source models Llama3.2:11b and Llama3.2:90b achieved 15% (4/27) and 18% (5/27), respectively. However, MLLMs exhibited a high hallucination rate, frequently generating HPO terms with incorrect IDs or entirely fabricated content. Specifically, ChatGPT-4o, Llama3.2:11b, and Llama3.2:90b generated incorrect IDs in 57.3% (67/117), 98% (62/63), and 82% (46/56) of cases, respectively, and fabricated terms in 34.2% (40/117), 41% (26/63), and 32% (18/56) of cases, respectively. Additionally, a survey on the rare disease knowledge of junior physicians suggests that participation in rare disease and genetic disease training may enhance the performance of some physicians. **Conclusions:** The integration of MLLMs into clinical workflows significantly enhances the accuracy of HPO identification by junior physicians, offering promising potential to improve the diagnosis of rare diseases and standardize phenotype descriptions in medical research. However, the notable hallucination rate observed in MLLMs underscores the necessity for further refinement and rigorous validation before widespread adoption in clinical practice.
40	(Esmaeilzadeh 2025)	Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety	OpenAI ChatGPT-O3 Claude Sonnet 4 Google Gemini 2.5 Pro	**배경:** 대규모 언어 모델(LLM)의 의료 분야로의 급속한 통합은 환자 안전, 신뢰성, 투명성 및 공평한 치료 제공과 관련된 중대한 윤리적 우려를 제기한다. 의료 데이터에 명시적으로 훈련되지 않았음에도 불구하고, 개인들은 의료 질문과 임상 시나리오를 다루기 위해 범용 LLM을 점점 더 사용하고 있다. 프롬프트 엔지니어링이 LLM 성능을 최적화할 수 있지만, 임상 의사결정에 대한 윤리적 영향은 아직 충분히 탐색되지 않았다. 본 연구는 안전성, 편향, 투명성 및 의료 분야에서 AI의 책임 있는 구현에 대한 시사점에 초점을 맞추어 LLM의 임상 적용에서 프롬프트 엔지니어링 전략의 윤리적 차원을 평가하는 것을 목표로 했다. **방법:** 우리는 6개의 프롬프트 엔지니어링 전략과 다양한 윤리적 복잡성을 가진 5개의 임상 시나리오에 걸쳐 3개의 진보된 추론 가능 LLM(OpenAI O3, Claude Sonnet 4, Google Gemini 2.5 Pro)에 대한 윤리 중심 분석을 수행했다. 6명의 전

				문 임상의가 진단 정확성, 안전성 평가, 의사소통, 공감, 윤리적 추론을 포함하는 영역을 사용하여 90개의 응답을 평가했다. 우리는 특히 안전 사고, 편향 패턴, 추론 과정의 투명성을 분석했다. **결과:** **모든 모델과 시나리오에서 중대한 윤리적 우려가 나타났다**. 중대한 안전 문제가 응답의 12.2%에서 발생했으며, 복잡한 윤리적 시나리오에 집중되었다(레벨 5: 23.1% vs. 레벨 1: 2.3%, p 〈 0.001). 메타인지 프롬프팅은 우수한 윤리적 추론을 보였으며(평균 윤리 점수: 78.3 ± 9.1), 안전 우선 프롬프팅은 제로샷 접근법에 비해 안전 사고를 45% 감소시켰다(8.9% vs. 16.2%). 그러나 모든 모델은 의사소통 공감에서 우려되는 결함을 보였고(최대값의 평균 54%), 복잡한 다문화 시나리오에서 잠재적 편향을 나타냈다. 투명성은 프롬프트 전략에 따라 크게 달랐으며, 메타인지 접근법이 임상 책임성에 필수적인 가장 명확한 추론 경로를 제공했다(4.2 vs. 1.8 명시적 추론 단계). 본 연구는 윤리적 의사결정 투명성의 중대한 격차를 강조했으며, 메타인지 접근법은 제로샷 방법의 1.8개에 비해 4.2개의 명시적 추론 단계를 제공했다(p 〈 0.001). **편향 패턴은 취약 집단에 불균형적으로 영향을 미쳤으며, 고령 환자의 치료 적절성에 대한 체계적 과소평가와 임종 시나리오에서의 부적절한 문화적 고려가 있었다.** **결론:** 범용 LLM의 현재 임상 적용은 긴급한 주의가 필요한 상당한 윤리적 문제를 제시한다. 구조화된 프롬프트 엔지니어링이 일부 영역에서 측정 가능한 개선을 보였으며, 메타인지 접근법은 13.0%의 성능 향상을 보였고 안전 우선 프롬프팅은 중대 사고를 45% 감소시켰지만, **모든 전략에 걸쳐 상당한 한계가 지속된다. 최적화된 접근법조차도 의사소통과 공감에서 부적절한 성능(최대값의 ≤ 54%)을 달성했고, 잔여 편향 패턴(안전 우선 조건에서 11.7%)을 유지했으며, 우려되는 안전 결함을 나타냈으며, 이는 현재의 프롬프트 엔지니어링 방법이 신뢰할 수 있는 임상 배치에 불충분한 한계적 개선만을 제공함을 나타낸다**. 이러한 발견은 범용 AI 모델의 임상 사용을 위한 적절한 지침 및 규제 프레임워크 개발에 대한 추가 조사를 필요로 하는 중대한 윤리적 문제를 강조한다. **BACKGROUND:** The rapid integration of large language models (LLMs) into healthcare raises critical ethical concerns regarding patient safety, reliability, transparency, and equitable care delivery. Despite not being trained explicitly on medical data, individuals increasingly use general-purpose LLMs to address medical questions and clinical scenarios. While prompt engineering can optimize LLM performance, its ethical implications for clinical decision-making remain underexplored. This study aimed to evaluate the ethical dimensions of prompt engineering strategies in the clinical applications of LLMs, focusing on safety, bias, transparency, and their implications for the

				responsible implementation of AI in healthcare. METHODS: We conducted an ethics-focused analysis of three advanced and reasoning-capable LLMs (OpenAI O3, Claude Sonnet 4, Google Gemini 2.5 Pro) across six prompt engineering strategies and five clinical scenarios of varying ethical complexity. Six expert clinicians evaluated 90 responses using domains that included diagnostic accuracy, safety assessment, communication, empathy, and ethical reasoning. We specifically analyzed safety incidents, bias patterns, and transparency of reasoning processes. RESULTS: Significant ethical concerns emerged across all models and scenarios. Critical safety issues occurred in 12.2% of responses, with concentration in complex ethical scenarios (Level 5: 23.1% vs. Level 1: 2.3%, p 〈 0.001). Meta-cognitive prompting demonstrated superior ethical reasoning (mean ethics score: 78.3 ± 9.1), while safety-first prompting reduced safety incidents by 45% compared to zero-shot approaches (8.9% vs. 16.2%). However, all models showed concerning deficits in communication empathy (mean 54% of maximum) and exhibited potential bias in complex multi-cultural scenarios. Transparency varied significantly by prompt strategy, with meta-cognitive approaches providing the clearest reasoning pathways (4.2 vs. 1.8 explicit reasoning steps), which are essential for clinical accountability. The study highlighted critical gaps in ethical decision-making transparency, with meta-cognitive approaches providing 4.2 explicit reasoning steps compared to 1.8 in zero-shot methods (p 〈 0.001). Bias patterns disproportionately affected vulnerable populations, with systematic underestimation of treatment appropriateness in elderly patients and inadequate cultural considerations in end-of-life scenarios. CONCLUSIONS: Current clinical applications of general-purpose LLMs present substantial ethical challenges requiring urgent attention. While structured prompt engineering demonstrated measurable improvements in some domains, with meta-cognitive approaches showing 13.0% performance gains and safety-first prompting reducing critical incidents by 45%, substantial limitations persist across all strategies. Even optimized approaches achieved inadequate performance in communication and empathy (≤ 54% of maximum), retained residual bias patterns (11.7% in safety-first conditions), and exhibited concerning safety deficits, indicating that current prompt engineering methods provide only marginal improvements, which are insufficient for reliable clinical deployment. These

				findings highlight significant ethical challenges that necessitate further investigation into the development of appropriate guidelines and regulatory frameworks for the clinical use of general-purpose AI models.
41	(Yassin et al. 2025)	Evaluating a generative artificial intelligence accuracy in providing medication instructions from smartphone images	ChatGPT	**배경:** 미국 식품의약국(FDA)은 적절한 약물 사용을 지원하기 위해 복약안내문(MG) 및 사용 지침서(IFU)와 같은 환자 라벨링 자료를 의무화한다. 그러나 낮은 건강 문해력 및 이러한 자료를 탐색하는 데 어려움과 같은 문제는 잘못된 약물 사용으로 이어질 수 있으며, 이는 치료 실패 또는 부작용을 초래할 수 있다. 생성형 AI의 부상은 이미지 인식 및 텍스트 생성을 통해 확장 가능하고 개인화된 환자 교육을 제공할 기회를 제시한다. **목적:** 본 연구는 사용자가 제공한 약물 이미지를 기반으로 ChatGPT가 생성한 복약 지침의 정확성과 안전성을 제조업체의 표준 지침과 비교하여 평가하는 것을 목표로 했다. **방법:** 투여를 위해 여러 단계가 필요한 12개 약물의 이미지를 ChatGPT의 이미지 인식 기능에 업로드했다. ChatGPT의 응답은 텍스트 분류기, 계수 벡터화(CountVec) 및 용어 빈도-역문서 빈도(TF-IDF)를 사용하여 공식 IFU 및 MG와 비교되었다. 임상 정확성은 독립적인 약사에 의해 추가로 평가되어 ChatGPT 응답이 환자 지침에 유효한지 결정했다. **결과:** ChatGPT는 모든 약물을 정확하게 식별하고 환자 지침을 생성했다. CountVec은 텍스트 유사성 분석에서 TF-IDF를 능가했으며, 평균 유사성 점수는 76%였다. 그러나 **임상 평가 결과 특히 복잡한 투여 경로에 대한 지침에서 상당한 격차가 드러났으며, ChatGPT의 안내는 필수 세부사항이 부족하여 임상 정확도 점수가 낮았다.** **결론:** ChatGPT는 환자 친화적인 복약 지침 생성에서 가능성을 보여주지만, **그 유효성은 약물의 복잡성에 따라 달라진다**. 연구 결과는 특히 복잡한 투여 과정을 가진 약물에 대해 AI 생성 의료 안내의 안전성과 정확성을 보장하기 위해 추가적인 정제 및 임상 감독의 필요성을 강조한다. **Background:** The Food and Drug Administration mandates patient labeling materials like the Medication Guide (MG) and Instructions for Use (IFU) to support appropriate medication use. However, challenges such as low health literacy and difficulties navigating these materials may lead to incorrect medication usage, resulting in therapy failure or adverse outcomes. The rise of generative AI, presents an opportunity to provide scalable, personalized patient education through image recognition and text generation. **Objective:** This study aimed to evaluate the accuracy and safety of medication instructions generated by ChatGPT based on user-provided drug

				images, compared to the manufacturer's standard instructions. **Methods:** Images of 12 medications requiring multiple steps for administration were uploaded to ChatGPT's image recognition function. ChatGPT's responses were compared to the official IFU and MG using text classifiers, Count Vectorization (CountVec), and Term Frequency–Inverse Document Frequency (TF–IDF). The clinical accuracy was further evaluated by independent pharmacists to determine if ChatGPT responses were valid for patient instruction. **Results:** ChatGPT correctly identified all medications and generated patient instructions. CountVec outperformed TF–IDF in text similarity analysis, with an average similarity score of 76%. However, clinical evaluation revealed significant gaps in the instructions, particularly for complex administration routes, where ChatGPT's guidance lacked essential details, leading to lower clinical accuracy scores. **Conclusion:** While ChatGPT shows promise in generating patient-friendly medication instructions, its effectiveness varies based on the complexity of the medication. The findings underscore the need for further refinement and clinical oversight to ensure the safety and accuracy of AI-generated medical guidance, particularly for medications with complex administration processes.
42	(Prasad et al. 2025)	Evaluating advanced AI reasoning models: ChatGPT-4.0 and DeepSeek-R1 diagnostic performance in otolaryngology: a comparative analysis	ChatGPT-4.0 DeepSeek-R1	**목적:** 본 연구는 이비인후과 분야에서 두 가지 진보된 인공지능(AI) 모델인 OpenAI의 ChatGPT-4.0과 DeepSeek-R1의 진단 정확성, 포괄성 및 임상 관련성을 평가하는 것을 목표로 했다. **방법:** 편도아데노이드 절제술, 고막 성형술, 내시경 부비동 수술, 이하선 절제술, 전후두 절제술 등 5가지 일반적인 이비인후과 시술을 두 AI 모델에 표준화된 질문을 통해 분석했다. 프롬프트가 환자들이 일반적으로 온라인에서 검색하는 질문을 재현하므로, 우리의 평가는 환자 대상 정보 적절성에 초점을 맞춘다. 응답은 정확성, 임상 관련성 및 포괄성에 대해 2명의 연구 구성원에 의해 독립적으로 평가되었으며, 불일치는 합의를 통해 해결되었다. 분석은 임상 지침과의 비교를 포함했다. **결과:** ChatGPT-4.0은 일반적으로 적응증, 방법론, 위험 및 회복 과정을 효과적으로 다루는 상세한 시술 통찰력을 제공했다. 그러나 **때때로 과도한 진단 영상을 제안했고 미묘하지만 중요한 수술적 세부사항을 누락**했다. DeepSeek-R1은 적응증, 치료 대안 및 시술 위험을 명확하게 분류하는 간결하고 구조화된 응답을 제공했다. 그럼에도 불구하고 **중요한 수술 기법과 경미한 합병증을 누락하여 상세한 설명이 자주 부족했다**. 예를 들어, DeepSeek-R1은 편도아데노이드 절제술의 지혈 기법 및 고막 성형술의 이식편 안정화 세부사항과 같은 구체적인 내용을 누락했다. 두 모델 모

				두 특히 **전후두 절제술과 같은 복잡한 시술에 대해 포괄적인 병기 분류, 상세한 수술 계획 및 장기 회복 세부사항과 같은 중요한 요소를 적절하게 다루지 못했다**. **결론:** ChatGPT-4.0과 DeepSeek-R1 모두 상당한 진단 잠재력을 보여주었지만 **정밀성, 포괄성 및 세밀한 임상 추론에서 한계를 드러냈다. 이들의 임상 유용성은 여전히 제한적**이며, 이비인후과에서 환자별 의사결정 능력을 향상시키기 위한 AI 정제의 지속적인 필요성을 강조한다. **Purpose:** This study aimed to evaluate the diagnostic accuracy, comprehensiveness, and clinical relevance of two advanced artificial intelligence (AI) models, OpenAI's ChatGPT-4.0 and DeepSeek-R1, in the field of otolaryngology. **Methods:** Five common otolaryngology procedures—adenotonsillectomy, tympanoplasty, endoscopic sinus surgery, parotidectomy, and total laryngectomy—were analyzed through standardized queries posed to both AI models. Because the prompts replicate questions that patients typically search online, our evaluation focuses on patient-facing informational adequacy. Responses were independently evaluated by two study members for accuracy, clinical relevance, and comprehensiveness, with discrepancies resolved through consensus. The analysis included comparison with clinical guidelines. **Results:** ChatGPT-4.0 generally provided detailed procedural insights, effectively covering indications, methodologies, risks, and recovery processes. However, it occasionally suggested excessive diagnostic imaging and omitted subtle yet significant surgical nuances. DeepSeek-R1 delivered concise, structured responses clearly categorizing indications, treatment alternatives, and procedural risks. Nonetheless, it frequently lacked detailed elaboration, omitting important surgical techniques and minor complications. For instance, DeepSeek-R1 omitted specifics such as hemostatic techniques in adenotonsillectomy and graft stabilization details in tympanoplasty. Neither model adequately addressed critical elements like comprehensive staging, detailed surgical planning, and long-term recovery nuances, especially for complex procedures such as total laryngectomy. **Conclusions:** Both ChatGPT-4.0 and DeepSeek-R1 demonstrated significant diagnostic potential but revealed limitations in precision, comprehensiveness, and nuanced clinical reasoning. Their clinical utility remains restricted, highlighting a continued need for AI refinement to

				enhance patient-specific decision-making capabilities in otolaryngology.
43	(Chang et al. 2024)	Evaluating anti-LGBTQIA+ medical bias in large language models	Gemini 1.5 Flash Claude 3 Haiku GPT-4o Stanford Medicine Secure GPT (GPT-4.0)	대규모 언어 모델(LLM)은 환자 의사소통부터 의사결정 지원에 이르는 작업을 위해 임상 환경에 점점 더 배치되고 있다. 이러한 모델이 인종 기반 및 이분법적 성별 편향을 보이는 반면, LGBTQIA+ 반대 편향은 이러한 집단에 영향을 미치는 문서화된 의료 격차에도 불구하고 충분히 연구되지 않았다. 본 연구에서 우리는 LLM이 LGBTQIA+ 반대 의료 편향 및 오정보를 전파할 잠재력을 평가했다. 우리는 의학적으로 훈련된 검토자와 LGBTQIA+ 건강 전문가가 작성한 명시적 질문 및 합성 임상 기록으로 구성된 38개의 프롬프트로 4개의 LLM(Gemini 1.5 Flash, Claude 3 Haiku, GPT-4o, Stanford Medicine Secure GPT [GPT-4.0])에 프롬프트를 제공했다. 프롬프트는 LGBTQIA+ 정체성 용어가 있는 프롬프트와 없는 프롬프트의 쌍으로 구성되었으며, 두 가지 축에 걸친 임상 상황을 탐색했다: (i) 역사적 편향이 관찰된 상황 대 관찰되지 않은 상황, (ii) LGBTQIA+ 정체성이 임상 치료와 관련이 있는 상황 대 관련이 없는 상황. 의학적으로 훈련된 검토자가 적절성(안전성, 개인정보 보호, 환각/정확성 및 편향) 및 임상 유용성에 대해 LLM 응답을 평가했다. 우리는 **4개의 LLM 모두 LGBTQIA+ 정체성 용어가 있는 프롬프트와 없는 프롬프트에 대해 부적절한 응답을 생성한다는 것을 발견**했다. 부적절한 응답의 비율은 LGBTQIA+ 정체성을 언급하는 프롬프트의 경우 43-62%, 언급하지 않는 프롬프트의 경우 47-65%였다. **부적절한 분류의 가장 일반적인 이유는 환각/정확성이었으며, 그 다음이 편향 또는 안전성**이었다. 질적으로 우리는 차등 편향 패턴을 관찰했으며, **LGBTQIA+ 프롬프트가 더 심각한 편향을 유발**했다. **부적절한 응답의 평균 임상 유용성 점수는 적절한 응답보다 낮았다(5점 리커트 척도에서 2.6 대 3.7)**. 향후 연구는 명시된 사용 사례에 맞게 출력 형식을 조정하고, 아첨과 프롬프트의 외부 정보에 대한 의존을 줄이며, LGBTQIA+ 환자에 대한 정확성을 향상시키고 편향을 감소시키는 데 초점을 맞춰야 한다. 우리는 향후 모델 평가를 위한 벤치마크로 프롬프트와 주석이 달린 응답을 제시한다. Large Language Models (LLMs) are increasingly deployed in clinical settings for tasks ranging from patient communication to decision support. While these models demonstrate race-based and binary gender biases, anti-LGBTQIA+ bias remains understudied despite documented healthcare disparities affecting these populations. In this work, we evaluated the potential of LLMs to propagate anti-LGBTQIA+ medical bias and misinformation. We prompted 4 LLMs (Gemini 1.5 Flash, Claude 3 Haiku, GPT-4o, Stanford Medicine Secure GPT [GPT-4.0]) with 38 prompts consisting of explicit questions and synthetic clinical notes created by medically-trained reviewers and LGBTQIA+ health experts. The prompts

				consisted of pairs of prompts with and without LGBTQIA+ identity terms and explored clinical situations across two axes: (i) situations where historical bias has been observed versus not observed, and (ii) situations where LGBTQIA+ identity is relevant to clinical care versus not relevant. Medically-trained reviewers evaluated LLM responses for appropriateness (safety, privacy, hallucination/accuracy, and bias) and clinical utility. We found that all 4 LLMs generated inappropriate responses for prompts with and without LGBTQIA+ identity terms. The proportion of inappropriate responses ranged from 43-62% for prompts mentioning LGBTQIA+ identities versus 47-65% for those without. The most common reason for inappropriate classification tended to be hallucination/accuracy, followed by bias or safety. Qualitatively, we observed differential bias patterns, with LGBTQIA+ prompts eliciting more severe bias. Average clinical utility score for inappropriate responses was lower than for appropriate responses (2.6 versus 3.7 on a 5-point Likert scale). Future work should focus on tailoring output formats to stated use cases, decreasing sycophancy and reliance on extraneous information in the prompt, and improving accuracy and decreasing bias for LGBTQIA+ patients. We present our prompts and annotated responses as a benchmark for evaluation of future models. Content warning: This paper includes prompts and model-generated responses that may be offensive.
44	(Yu, Li, et al. 2025)	Evaluating Artificial Intelligence in Spinal Cord Injury Management: A Comparative Analysis of ChatGPT-4o and Google Gemini Against American College of Surgeons Best Practices Guidelines for Spine Injury	ChatGPT-4o Google Gemini Advanced	**목적:** 미국 외과학회(ACS)는 척추 손상 관리를 위한 근거 기반 권고사항을 제공하기 위해 2022년 모범 실무 지침을 개발했다. 본 연구는 ChatGPT-4o와 Gemini Advanced의 2022년 ACS 모범 실무 지침과의 일치도를 평가하는 것을 목표로 하며, 척수 손상 관리에서 이러한 모델에 대한 최초의 전문가 평가를 제공한다. **방법:** 2022년 ACS 외상 품질 프로그램 척추 손상 모범 실무 지침을 사용하여 주요 임상 권고사항을 기반으로 52개의 질문을 작성했다. 이들은 정보 제공(8개), 진단(14개), 치료(30개) 범주로 분류되어 ChatGPT-4o와 Google Gemini Advanced에 제출되었다. 응답은 ACS 지침과의 일치도에 대해 등급이 매겨졌고 전문의 자격을 갖춘 척추 외과의에 의해 검증되었다. **결과:** ChatGPT는 52개 질문 중 38개(73.07%)에서 ACS 지침과 일치했고 Gemini는 36개(69.23%)에서 일치했다**. 대부분의 불일치 답변은 불충분한 정보 때문**이었다. 모델들은 8개 질문에서 의견이 달랐으며, ChatGPT는 5개에서, Gemini는 3개에서 일치했다. 두 모델 모두 임상 정보에서 75%의 일치도를 달성했으며; Gemini는 진단에서 우수한 성능을 보였고(78.57% vs 71.43%), ChatGPT는 치료 질문에서 더 높은 일치도를 보였다(73.33% vs 63.33%). **결론:** ChatGPT-4o와 Gemini Advanced는 현재 모범 실무와 일치하는 응답을 제

				공함으로써 척추 손상 관리에서 가치 있는 자산으로서의 잠재력을 보여준다. 일치율의 미미한 차이는 어느 모델도 검증된 임상 지침과 일치하는 권고사항을 제공하는 데 우월한 능력을 보이지 않음을 시사한다. LLM의 정교함과 유용성이 증가하고 있음에도 불구하고, 기존 한계는 현재 외상 기반 환경에서 임상적으로 안전하고 실용적이지 못하게 한다. **Objectives:** The American College of Surgeons developed the 2022 Best Practice Guidelines to provide evidence-based recommendations for managing spinal injuries. This study aims to assess the concordance of ChatGPT-4o and Gemini Advanced with the 2022 ACS Best Practice Guidelines, offering the first expert evaluation of these models in managing spinal cord injuries. **Methods:** The 2022 ACS Trauma Quality Program Best Practices Guidelines for Spine Injury were used to create 52 questions based on key clinical recommendations. These were grouped into informational (8), diagnostic (14), and treatment (30) categories and posed to ChatGPT-4o and Google Gemini Advanced. Responses were graded for concordance with ACS guidelines and validated by a board-certified spine surgeon. **Results:** ChatGPT was concordant with ACS guidelines on 38 of 52 questions (73.07%) and Gemini on 36 (69.23%). Most non-concordant answers were due to insufficient information. The models disagreed on 8 questions, with ChatGPT concordant in 5 and Gemini in 3. Both achieved 75% concordance on clinical information; Gemini outperformed on diagnostics (78.57% vs 71.43%), while ChatGPT had higher concordance on treatment questions (73.33% vs 63.33%). **Conclusions:** ChatGPT-4o and Gemini Advanced demonstrate potential as valuable assets in spinal injury management by providing responses aligned with current best practices. The marginal differences in concordance rates suggest that neither model exhibits a superior ability to deliver recommendations concordant with validated clinical guidelines. Despite LLMs increasing sophistication and utility, existing limitations currently prevent them from being clinically safe and practical in trauma-based settings.
45	(Masanneck, Meuth, and Pawlitzki 2025)	Evaluating base and retrieval augmented LLMs with document or online support for	GPT-4o GPT-4o mini GPT-4 Turbo	근거 기반 정보를 효과적으로 관리하는 것이 점점 더 어려워지고 있다. 본 연구는 130개 질문에 걸쳐 13개의 최신 신경학 지침을 사용하여 문서 및 온라인 기반 검색 증강 생성(RAG) 시스템을 포함한 대규모 언어 모델(LLM)을 테스트했다. 결과는

		evidence based neurology	LLaMA3-70b LLaMA3.1-Nemotron-70b-instruct Gemini-1.5 Pro Mixtral-8x7b	상당한 변동성을 보였다. **RAG는 기본 모델에 비해 정확성을 향상시켰지만 여전히 잠재적으로 해로운 답변을 생성했다. RAG 기반 시스템은 지식 기반 질문보다 사례 기반 질문에서 더 나쁜 성능을 보였다**. RAG 강화 LLM의 안전한 임상 통합을 위해서는 추가적인 정제 및 개선된 규제가 필요하다. Effectively managing evidence-based information is increasingly challenging. This study tested large language models (LLMs), including document- and online-enabled retrieval-augmented generation (RAG) systems, using 13 recent neurology guidelines across 130 questions. Results showed substantial variability. RAG improved accuracy compared to base models but still produced potentially harmful answers. RAG-based systems performed worse on case-based than knowledge-based questions. Further refinement and improved regulation is needed for safe clinical integration of RAG-enhanced LLMs.
46	(Zeljkovic et al. 2025)	Evaluating ChatGPT-4's correctness in patient-focused informing and awareness for atrial fibrillation	ChatGPT-4	**배경:** 인공지능과 대규모 언어 모델이 계속 발전함에 따라, 의료 분야에서의 적용이 확대되고 있다. OpenAI의 Chat Generative Pre-trained Transformer 4(ChatGPT-4)는 이 기술의 최신 발전을 대표하며, 복잡한 대화에 참여하고 정보를 제공할 수 있다. **목적:** 본 연구는 심방세동에 대해 환자에게 정보를 제공하는 데 있어 ChatGPT-4의 정확성을 탐색한다. **방법:** 이 횡단면 관찰 연구는 심방세동과 관련된 10개 범주에 걸친 108개의 구조화된 질문에 ChatGPT-4가 응답하는 것을 포함했다. 이러한 범주에는 기본 정보, 치료 옵션, 생활 방식 조정 등이 포함되어 일반적인 환자 질문을 반영했다. 모델의 응답은 정확성, 포괄성, 명확성, 임상 실무와의 관련성 및 환자 안전을 기준으로 3명의 심장 전문의 패널에 의해 평가되었다. ChatGPT-4의 전체 정확성은 각 범주에서 부여된 점수를 통해 정량적으로 평가되었으며, 범주 간 성능의 유의한 차이를 식별하기 위해 통계 분석이 수행되었다. **결과: ChatGPT-4는 범주 전반에 걸쳐 상당한 변동성을 가지고 정확하고 관련성 있는 답변을 제공했다**. "생활 방식 조정" 및 "일상 생활 및 관리"에서 완벽하거나 거의 완벽한 점수로 우수한 성능을 보였지만, "**기타 우려사항"에서는 낮은 점수를 받았다. 통계 분석은 범주 전반에 걸친 총점의 유의한 차이를 확인했다(P = .020).** **결론:** 우리의 결과는 ChatGPT-4가 구조화되고 직접적인 질문이 있는 범주에서는 신뢰할 수 있지만, **심층 설명이나 임상 판단이 필요한 복잡한 의학적 질문을 다룰 때는 한계를 보인다는 것을 시사**한다. ChatGPT-4는 특히 간단한 정보 제공 내용에서 심방세동에 대한 환자 중심 정보 제공 도구로서 유망한 잠재력을 보여준다.

				Background: As artificial intelligence and large language models continue to evolve, their application in health care is expanding. OpenAI's Chat Generative Pre-trained Transformer 4 (ChatGPT-4) represents the latest advancement in this technology, capable of engaging in complex dialogues and providing information. **Objective:** This study explores the correctness of ChatGPT-4 in informing patients about atrial fibrillation. **Methods:** This cross-sectional observational study involved ChatGPT-4 in responding to a structured set of 108 questions across 10 categories related to atrial fibrillation. These categories included basic information, treatment options, lifestyle adjustments, and more, reflecting common patient inquiries. The model's responses were evaluated by a panel of 3 cardiologists on the basis of accuracy, comprehensiveness, clarity, relevance to clinical practice, and patient safety. The total correctness of ChatGPT-4 was quantitatively assessed through scores assigned in each category, and statistical analysis was performed to identify significant differences in performance across categories. **Results:** ChatGPT-4 provided correct and relevant answers with considerable variability across categories. It excelled in "Lifestyle Adjustments" and "Daily Life and Management" with perfect and near-perfect scores but struggled with "Miscellaneous Concerns" scoring lower. Statistical analysis confirmed significant differences in total scores across categories (P = .020). **Conclusion:** Our results suggest that while ChatGPT-4 is reliable in categories with structured and direct queries, it shows limitations when handling complex medical queries that require in-depth explanations or clinical judgment. ChatGPT-4 demonstrates promising potential as a tool for patient-focused informing in atrial fibrillation, particularly in straightforward informing content.
47	(Rebitschek et al. 2025)	Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information	ChatGPT-3.5 (gpt-3.5-turbo) Google Gemini (1.5-Flash) Mistral AI Le Chat (mistral-large-2402)	대규모 언어 모델(LLM)은 건강 정보를 찾는 데 사용된다. 근거 기반 건강 의사소통 지침은 정보에 입각한 의사결정을 지원하기 위해 최선의 가용 근거를 제시할 것을 요구한다. 우리는 LLM의 프롬프트 의존적 지침 준수를 조사하고 일반인의 프롬프트 작성을 향상시키기 위한 최소한의 행동 개입을 평가한다. 연구 1은 프롬프트 정보성, 주제 및 LLM을 체계적으로 변화시켜 준수도를 평가했다. 연구 2는 표준 또는 향상된 프롬프트 조건에서 3개의 LLM에 300명의 참가자를 무작위 배정했다. 눈가림된 평가자가 두 가지 도구로 LLM 응답을 평가했다. **연구 1은 LLM이 근거 기반 건강 의사소통 기준을 충족하지 못한다는 것을 발견했다. 응답의 품질은 프롬**

				프트 정보성에 따라 달라지는 것으로 나타났다. 연구 2는 일반인이 자주 저품질 응답을 생성한다는 것을 밝혔다. 간단한 향상 방법이 응답 품질을 개선했지만, 여전히 요구 기준 이하였다. 이러한 발견은 LLM이 독립형 건강 의사소통 도구로서 부적절함을 강조한다. LLM을 근거 기반 프레임워크와 통합하고, 추론 및 인터페이스를 향상시키며, 프롬프트 작성을 교육하는 것이 필수적이다. Large language models (LLMs) are used to seek health information. Guidelines for evidence-based health communication require the presentation of the best available evidence to support informed decision-making. We investigate the prompt-dependent guideline compliance of LLMs and evaluate a minimal behavioural intervention for boosting laypeople's prompting. Study 1 systematically varied prompt informedness, topic, and LLMs to evaluate compliance. Study 2 randomized 300 participants to three LLMs under standard or boosted prompting conditions. Blinded raters assessed LLM response with two instruments. Study 1 found that LLMs failed evidence-based health communication standards. The quality of responses was found to be contingent upon prompt informedness. Study 2 revealed that laypeople frequently generated poor-quality responses. The simple boost improved response quality, though it remained below required standards. These findings underscore the inadequacy of LLMs as a standalone health communication tool. Integrating LLMs with evidence-based frameworks, enhancing their reasoning and interfaces, and teaching prompting are essential. Study Registration: German Clinical Trials Register (DRKS) (Reg. No.: DRKS00035228, registered on 15 October 2024).
48	(Denecke and Paula 2025)	Evaluating Large Language Models for Analysing Safety Risks in Healthcare Incident Reports	Gemma-2	의료사고 보고서는 의료 시스템의 안전 위험을 분석하기 위한 풍부한 자료를 제공한다. 사고 보고서의 적시 분석 및 해석을 지원하기 위해 자연어 처리(NLP)를 적용할 수 있다. 본 논문의 목적은 사고 보고서로부터 사고 원인을 추출하고 기여 요인을 식별하는 데 있어 대규모 언어 모델(LLM)의 잠재력을 평가하는 것이다. 데이터셋으로 의료 분야의 중대 사고에 대한 스위스 국가 데이터베이스인 CIRRNET®의 10,063개 메시지를 고려했다. 우리는 LLM Gemma-2를 적용하여 사건, 원인 및 기여 요인을 추출하고 주제별로 그룹화했다. 100개의 사고 보고서를 추출 품질에 대해 수동으로 평가했다. **사건은 92%의 정확도로, 원인은 84%, 기여 요인은 72%의 정확도로 추출되었다. 기여 요인 추출은 LLM이 환각하거나 해석함에 따라 실패한다.** 우리는 LLM이 사고 보고서 분석에서 잠재력을 보이며 사고 분석의 효율성과 일관성을 향상시킬 수 있다고 결론 내린다.

				Incident reports provide a rich source for analysing safety risks in healthcare systems. To support the timely analysis and interpretation of incident reports, natural language processing (NLP) can be applied. The aim of this paper is to evaluate the potential of large language models (LLMs) in extracting the causes of incidents and identifying contributing factors from incident reports. As dataset, we considered 10,063 messages from CIRRNET®, the Swiss national database for critical incidents in healthcare. We applied the LLM Gemma-2 to extract events, causes and contributing factors and group them along themes. 100 event reports were assessed manually regarding quality of extraction. Events were extracted with 92% accuracy, causes with 84% and contributing factors with 72% accuracy. Extraction of contributing factors fails as the LLM hallucinates or interprets. We conclude that LLMs show potential in analysing incident reports and can improve the efficiency and consistency of incident analysis.
49	(Williams et al. 2024)	Evaluating Large Language Models for Drafting Emergency Department Discharge Summaries	GPT-4 GPT-3.5-turbo	**중요성:** 대규모 언어 모델(LLM)은 텍스트 요약을 포함하여 임상 영역에 적용될 수 있는 다양한 역량을 보유하고 있다. 주변 인공지능 기록자 및 기타 LLM 기반 도구가 의료 환경에 배치되기 시작함에 따라, 이러한 기술의 정확성에 대한 엄격한 평가가 시급히 필요하다. **목적:** 응급실(ED) 퇴원 요약서 생성에서 GPT-4와 GPT-3.5-turbo의 성능을 조사하고 퇴원 요약서의 각 섹션에서 오류의 빈도와 유형을 평가한다. **설계:** 횡단면 연구. **설정:** 캘리포니아 대학교 샌프란시스코 응급실. **참가자:** 우리는 2012년부터 2023년까지 응급실 임상의 기록이 있는 모든 성인 응급실 방문을 식별했다. GPT 요약을 위해 100건의 응급실 방문 샘플을 무작위로 선택했다. **노출:** 우리는 전체 응급실 임상의 기록을 퇴원 요약서로 요약하기 위해 두 가지 최첨단 LLM인 GPT-4와 GPT-3.5-turbo의 잠재력을 조사한다. **주요 결과 및 측정:** GPT-3.5-turbo 및 GPT-4 생성 퇴원 요약서는 세 가지 평가 기준에 걸쳐 2명의 독립적인 응급의학과 전문의 검토자에 의해 평가되었다: 1) GPT 요약 정보의 부정확성; 2) 정보의 환각; 3) 관련 임상 정보의 누락. 각 오류를 식별할 때, 검토자는 추가로 그들의 추론에 대한 간단한 설명을 제공하도록 요청받았으며, 이는 오류의 하위 그룹으로 수동 분류되었다. **결과:** 202,059건의 적격 응급실 방문 중 100건을 GPT 생성 요약 및 전문가 주도 평가를 위해 무작위로 샘플링했다. 전체적으로 GPT-4가 생성한 요약서의 33%와 GPT-3.5-turbo가 생성한 요약서의 10%가 모든 평가 영역에서 완전히 오류가 없었다. GPT-4가 생성한 요약서는 대부분 정확했으며, **부정확성은 10%의 사례에서**

				만 발견되었지만, 42%의 요약서가 환각을 나타냈고 47%는 임상적으로 관련 있는 정보를 누락했다. 부정확성과 환각은 GPT 생성 요약서의 계획 섹션에서 가장 일반적으로 발견되었으며, 임상적 누락은 환자의 신체 검사 소견 또는 주 호소 병력을 설명하는 텍스트에 집중되었다. **결론 및 관련성:** 100건의 응급실 접촉에 대한 이 횡단면 연구에서 LLM이 정확한 퇴원 요약서를 생성할 수 있지만, 환각 및 임상적으로 관련 있는 정보의 누락 가능성이 있음을 발견했다. GPT 생성 임상 텍스트에서 발견되는 오류의 위치와 유형에 대한 포괄적인 이해는 임상의의 이러한 내용 검토를 촉진하고 환자 위해를 예방하는 데 중요하다. **Importance:** Large language models (LLMs) possess a range of capabilities which may be applied to the clinical domain, including text summarization. As ambient artificial intelligence scribes and other LLM-based tools begin to be deployed within healthcare settings, rigorous evaluations of the accuracy of these technologies are urgently needed. **Objective:** To investigate the performance of GPT-4 and GPT-3.5-turbo in generating Emergency Department (ED) discharge summaries and evaluate the prevalence and type of errors across each section of the discharge summary. **Design:** Cross-sectional study. **Setting:** University of California, San Francisco ED. **Participants:** We identified all adult ED visits from 2012 to 2023 with an ED clinician note. We randomly selected a sample of 100 ED visits for GPT-summarization. **Exposure:** We investigate the potential of two state-of-the-art LLMs, GPT-4 and GPT-3.5-turbo, to summarize the full ED clinician note into a discharge summary. **Main Outcomes and Measures:** GPT-3.5-turbo and GPT-4-generated discharge summaries were evaluated by two independent Emergency Medicine physician reviewers across three evaluation criteria: 1) Inaccuracy of GPT-summarized information; 2) Hallucination of information; 3) Omission of relevant clinical information. On identifying each error, reviewers were additionally asked to provide a brief explanation for their reasoning, which was manually classified into subgroups of errors. **Results:** From 202,059 eligible ED visits, we randomly sampled 100 for GPT-generated summarization and then expert-driven evaluation. In total, 33% of

				summaries generated by GPT-4 and 10% of those generated by GPT-3.5-turbo were entirely error-free across all evaluated domains. Summaries generated by GPT-4 were mostly accurate, with inaccuracies found in only 10% of cases, however, 42% of the summaries exhibited hallucinations and 47% omitted clinically relevant information. Inaccuracies and hallucinations were most commonly found in the Plan sections of GPT-generated summaries, while clinical omissions were concentrated in text describing patients' Physical Examination findings or History of Presenting Complaint. **Conclusions and Relevance:** In this cross-sectional study of 100 ED encounters, we found that LLMs could generate accurate discharge summaries, but were liable to hallucination and omission of clinically relevant information. A comprehensive understanding of the location and type of errors found in GPT-generated clinical text is important to facilitate clinician review of such content and prevent patient harm.
50	(Tang et al. 2023)	Evaluating large language models on medical evidence summarization	GPT-3.5 (text-davinci-003) ChatGPT	대규모 언어 모델(LLM)의 최근 발전은 다양한 하위 작업에서 제로샷 및 퓨샷 성능에서 놀라운 성공을 보여주었으며, 고위험 영역에서의 적용 가능성을 열었다. 본 연구에서 우리는 6개 임상 영역에 걸친 제로샷 의료 근거 요약 수행에서 LLM, 특히 GPT-3.5와 ChatGPT의 역량과 한계를 체계적으로 검토한다. 우리는 요약 품질의 여러 차원을 다루는 자동 및 인간 평가를 모두 수행한다. 우리의 연구는 자동 지표가 종종 요약의 품질과 강하게 상관되지 않음을 보여준다. 더욱이 인간 평가를 기반으로 의료 근거 요약에 대한 오류 유형의 용어를 정의한다. 우리의 발견은 LLM이 사실과 일치하지 않는 요약을 생성하고 지나치게 확신에 찬 또는 불확실한 진술을 하는 데 취약할 수 있으며, 이는 오정보로 인한 잠재적 위해로 이어질 수 있음을 밝힌다. 더욱이 모델이 중요한 정보를 식별하는 데 어려움을 겪으며, 더 긴 텍스트 맥락을 요약할 때 더 많은 오류가 발생하기 쉽다는 것을 발견했다. Recent advances in large language models (LLMs) have demonstrated remarkable successes in zero- and few-shot performance on various downstream tasks, paving the way for applications in high-stakes domains. In this study, we systematically examine the capabilities and limitations of LLMs, specifically GPT-3.5 and ChatGPT, in performing zero-shot medical evidence summarization across six clinical domains. We conduct both automatic and human evaluations, covering several dimensions of summary quality. Our study demonstrates that automatic metrics often do not strongly correlate with the quality of summaries. Furthermore, informed by our human evaluations, we define a terminology of error types for medical

				evidence summarization. Our findings reveal that LLMs could be susceptible to generating factually inconsistent summaries and making overly convincing or uncertain statements, leading to potential harm due to misinformation. Moreover, we find that models struggle to identify the salient information and are more error-prone when summarizing over longer textual contexts.
51	(Naliyatthaliyazchayil et al. 2025)	Evaluating Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification with Zero Shot Prompting	ChatGPT-4 LLaMA-3·1 Gemini-1·5 DeepSeek-R1 OpenAI-o3	접근 가능한 챗봇 인터페이스를 통한 대규모 언어 모델(LLM)의 확산은 의료 분야에서 전례 없는 기회를 창출했으며, ChatGPT-4, LLaMA-3·1, Gemini-1·5, DeepSeek-R1 및 OpenAI-o3와 같은 최첨단 모델이 인공지능 기반 임상 지원을 제공하고 있다. 일부 연구는 복잡한 의료 작업 관리에서 LLM의 잠재력을 보여주는 반면, 다른 연구는 정확성, 신뢰성 및 임상 환경의 엄격한 기준 준수에 대한 우려를 강조한다. 본 연구는 이들의 진정한 잠재력을 더 잘 이해하고 의료 분야에서 가장 효과적일 수 있는 영역을 식별하기 위해 수행되었다. 본 연구는 Medical Information Mart for Intensive Care IV(MIMIC-IV) 데이터셋을 사용하여 세 가지 중요한 의료 작업에 걸쳐 평가된 주요 추론 및 비추론 LLM – ChatGPT-4, LLaMA-3·1, Gemini-1·5, DeepSeek-R1 및 OpenAI-o3 –에 대한 포괄적인 비교 분석을 제시한다. 우리는 제로샷 프롬프팅 프로토콜을 통해 다음 작업에서 모델의 역량을 평가했다: (1) 주 진단 생성, (2) 진단의 ICD-9 코드 매핑, (3) 병원 재입원 위험 계층화 예측. 본 연구는 MIMIC-IV에서 무작위로 선택된 300명의 피험자 코호트를 활용했으며, 퇴원 요약서 섹션에서 체계적으로 생성된 표준화된 프롬프트를 사용했다. 각 프롬프트는 환자 임상 정보와 특정 작업 요구사항을 통합된 입력 형식으로 통합하도록 설계되었다. 결과 해석 가능성을 향상시키기 위해 프롬프팅 구조 내에서 명시적 근거 도출을 구현하여 모델이 진단 및 예후 예측에 대한 추론 과정을 설명하도록 요구했다. 이것은 제로샷 프롬프트 접근법이므로, 프롬프트는 동일하게 여러 번 반복 테스트되지 않았다. 비추론 모델 간 비교 분석에서 LLaMA-3·1은 모든 평가 지표에서 우수한 종합 성능을 보였으며, 주 진단 예측에서 85%, ICD-9 코드 예측에서 42·6%, 병원 재입원 위험 예측에서 41·3%의 정확도를 보였다. 추론 모델 DeepSeek-R1과 OpenAI-O3는 유사한 성능을 보였으며, O3는 주 진단(90%)과 ICD-9 예측(45·3%)에서 약간 더 높은 정확도를 달성했고, R1은 재입원 위험 예측(72·66%)에서 약간 더 나은 성능을 보였다. **우리의 발견은 평가된 모델 중 어느 것도 모든 작업에서 임상 기준을 충족하지 못했으며, 의료 코딩이 가장 약한 성능을 보였음을 나타낸다. 이는 사전 훈련된 LLM이 의료 코딩에 어려움을 겪는다는 일부 문헌 결과와 일치한다**. 이는 임상 적용 가능성을 향상시키기 위해 이러한 모델의 추가 정제가 필요함을 강조한다. The proliferation of large language models (LLMs) through accessible chatbot

				interfaces has created unprecedented opportunities in healthcare, with state-of-the-art models such as ChatGPT-4, LLaMA-3·1, Gemini-1·5, DeepSeek-R1and OpenAI-o3, offering artificial intelligence-driven clinical support. Some studies showcase the potential of LLMs in managing complex healthcare tasks, while others emphasize concerns regarding their accuracy, reliability, and compliance with the rigorous standards of clinical settings. This study was conducted to better understand their true potential and identify areas where they can be most effective in healthcare.This study presents a comprehensive comparative analysis of leading reasoning and non-reasoning LLMs - ChatGPT-4, LLaMA-3·1, Gemini-1·5, DeepSeek-R1and OpenAI-o3 - evaluated across three critical healthcare tasks using the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset. We assessed the models' capabilities in: (1) generating primary diagnoses, (2) mapping diagnoses to ICD-9 codes, and (3) predicting hospital readmission risk stratification through zero-shot prompting protocols. The study utilized a cohort of 300 randomly selected subjects from MIMIC-IV, with standardized prompts systematically generated from discharge summary sections. Each prompt was engineered to incorporate both patient clinical information and specific task requirements in a unified input format. To enhance result interpretability, we implemented explicit rationale elicitation within the prompting structure, requiring models to articulate their reasoning process for diagnostic and prognostic predictions. Since this is a zero-shot prompt approach, the prompt is not tested repeating the same multiple times.In our comparative analysis among non-reasoning models, LLaMA-3·1 demonstrated superior aggregate performance across all evaluation metrics, with 85% correctness in Primary Diagnosis prediction, 42·6% in ICD-9 code prediction, and 41·3% in hospital readmission risk prediction. Reasoning models DeepSeek-R1 and OpenAI-O3 showed similar performance, with O3 achieving slightly higher accuracy in primary diagnosis (90%) and ICD-9 prediction (45·3%), while R1 performed slightly better in readmission risk prediction (72·66%). Our findings show that none of the evaluated models met clinical standards across all tasks, with medical coding showing the weakest performance. This aligns with few of the literature findings indicating that pretrained LLMs struggle with medical coding. This underscores the need for further refinement of these models to enhance their clinical applicability.

52	(Polat et al. 2024)	Evaluating the accuracy and readability of ChatGPT in providing parental guidance for adenoidectomy, tonsillectomy, and ventilation tube insertion surgery	ChatGPT	**목적:** 본 연구는 아데노이드 절제술, 편도 절제술 및 환기관 삽입술(ATVtis)에 대한 지침을 찾는 부모들에게 정확하고 읽기 쉬운 정보 출처로서 ChatGPT의 잠재력을 조사했다. **방법:** ChatGPT에게 세 가지 특정 수술 절차 각각에 대해 인터넷 검색 엔진에서 부모들이 가장 자주 묻는 상위 15개 질문을 식별하도록 요청했다. 우리는 초기 45개 세트에서 반복된 질문을 제거했다. 이후 남은 33개 질문에 대한 답변을 생성하도록 ChatGPT에 요청했다. 7명의 경험이 풍부한 이비인후과 전문의가 완전히 부정확함부터 포괄적임까지의 4단계 등급 척도를 사용하여 응답의 정확성을 개별적으로 평가했다. 응답의 가독성은 Flesch Reading Ease(FRE) 및 Flesch-Kincaid Grade Level(FKGL) 점수를 사용하여 결정되었다. 질문은 진단 및 준비 과정, 수술 정보, 위험 및 합병증, 수술 후 과정의 네 그룹으로 분류되었다. 그런 다음 응답을 정확도 등급, FRE 및 FKGL 점수를 기준으로 비교했다. **결과:** 7명의 평가자가 각각 33개의 AI 생성 응답을 평가하여 총 231개의 평가를 제공했다. 평가된 응답 중 167개(72.3%)가 '포괄적'으로 분류되었다**. 62개 응답(26.8%)은 '정확하지만 불충분함'으로 분류되었고, 2개 응답(0.9%)은 '일부 정확, 일부 부정확'으로 평가**되었다. 어떤 평가자도 '완전히 부정확함'으로 판단한 응답은 없었다. 평균 FRE 및 FKGL 점수는 각각 57.15(±10.73) 및 9.95(±1.91)였다. ChatGPT의 응답을 분석한 결과, 3개(9.1%)가 미국 의학협회(AMA)가 권장하는 6학년 읽기 수준 이하였다. 그룹 간 가독성 및 정확도 점수에서 유의한 차이는 발견되지 않았다($p > 0.05$). **결론:** ChatGPT는 ATVtis와 관련된 다양한 주제에 대한 질문에 정확한 답변을 제공할 수 있다. 그러나 **ChatGPT의 답변은 일반적으로 고등학교 수준으로 작성되어 일부 독자에게는 너무 복잡할 수 있다. 이는 AMA가 환자 정보에 권장하는 6학년 읽기 수준을 초과한다. 우리 연구에 따르면, AI 생성 응답의 4분의 3 이상이 10학년 읽기 수준 이상이었으며, 이는 ChatGPT 텍스트의 가독성에 대한 우려를 제기한다**. **Objectives:** This study examined the potential of ChatGPT as an accurate and readable source of information for parents seeking guidance on adenoidectomy, tonsillectomy, and ventilation tube insertion surgeries (ATVtis). **Methods:** ChatGPT was tasked with identifying the top 15 most frequently asked questions by parents on internet search engines for each of the three specific surgical procedures. We removed repeated questions from the initial set of 45. Subsequently, we asked ChatGPT to generate answers to the remaining 33 questions. Seven highly experienced otolaryngologists individually assessed the accuracy of the responses using a four-level

				grading scale, from completely incorrect to comprehensive. The readability of responses was determined using the Flesch Reading Ease (FRE) and Flesch–Kincaid Grade Level (FKGL) scores. The questions were categorized into four groups: Diagnosis and Preparation Process, Surgical Information, Risks and Complications, and Postoperative Process. Responses were then compared based on accuracy grade, FRE, and FKGL scores. **Results:** Seven evaluators each assessed 33 AI-generated responses, providing a total of 231 evaluations. Among the evaluated responses, 167 (72.3 %) were classified as 'comprehensive.' Sixty-two responses (26.8 %) were categorized as 'correct but inadequate,' and two responses (0.9 %) were assessed as 'some correct, some incorrect.' None of the responses were adjudged 'completely incorrect' by any assessors. The average FRE and FGKL scores were 57.15(±10.73) and 9.95(±1.91), respectively. Upon analyzing the responses from ChatGPT, 3 (9.1 %) were at or below the sixth-grade reading level recommended by the American Medical Association (AMA). No significant differences were found between the groups regarding readability and accuracy scores ($p > 0.05$). **Conclusions:** ChatGPT can provide accurate answers to questions on various topics related to ATVtis. However, ChatGPT's answers may be too complex for some readers, as they are generally written at a high school level. This is above the sixth-grade reading level recommended for patient information by the AMA. According to our study, more than three-quarters of the AI-generated responses were at or above the 10th-grade reading level, raising concerns about the ChatGPT text's readability.
53	(Umihanic et al. 2025)	Evaluating the Concordance Between ChatGPT and Multidisciplinary Teams in Breast Cancer Treatment Planning: A Study from Bosnia and Herzegovina	ChatGPT-4.0	**배경/목적:** 보스니아 헤르체고비나를 포함한 많은 저소득 및 중간소득 국가(LMIC)에서 종양학 서비스는 제한된 수의 전문의와 근거 기반 치료에 대한 불균등한 접근으로 인해 제약을 받고 있다. 인공지능(AI), 특히 ChatGPT와 같은 대규모 언어 모델(LLM)은 종양학 전문성이 부족한 곳에서 치료를 표준화하고 임상의를 지원하기 위한 임상 의사결정 지원을 제공할 수 있다. 본 연구는 새로 진단된 유방암 환자의 관리에서 다학제 팀(MDT)이 내린 결정과 비교하여 ChatGPT가 생성한 치료 권고사항의 일치도, 안전성 및 임상 적절성을 평가하는 것을 목표로 했다. **방법:** 이 후향적 연구는 2023년 보스니아 헤르체고비나의 MDT에 제출된 새로 진단된 치료 경험이 없는 유방암 환자 91명을 포함했다. 환자 데이터를 ChatGPT-4.0에 입력하여 치료 권고사항을 생성했다. 4명의 전문의 자격을 갖춘 종양학 전문의(내부 2명, 외부 2명)가 4점 리커트 척도를 사용하여 MDT 결정과 비교하여 ChatGPT의 제안을 평가했다. 일치도는 기술 통계, Cronbach's alpha 및 Fleiss'

				kappa를 사용하여 분석되었다. **결과:** ChatGPT와 MDT 결정 간의 평균 일치 점수는 3.31(SD = 0.10)이었으며, 종양학 전문의 평가 전반에 걸쳐 높은 일관성을 보였다(Cronbach's alpha = 0.86). Fleiss' kappa는 중등도 평가자 간 신뢰도를 나타냈다(κ = 0.31, p 〈 0.001). 호르몬 수용체 음성 종양을 가진 환자와 표준 화학요법 요법으로 치료받은 환자에서 더 높은 일치도가 관찰되었다. **저등급 종양 또는 수술이나 내분비 요법에 대한 불확실한 적응증과 같이 개별화된 결정이 필요한 경우에는 일치도가 낮았다**. **결론:** ChatGPT는 특히 표준화된 임상 시나리오에서 MDT 치료 계획과 높은 일치도를 보였다. 자원이 제한된 환경에서 AI 도구는 종양학 의사결정을 지원하고 임상 전문성의 격차를 메우는 데 도움이 될 수 있다. 그러나 실제 안전하고 효과적인 사용을 위해서는 신중한 검증과 전문가 감독이 여전히 필수적이다. **Background/Objectives:** In many low- and middle-income countries (LMICs), including Bosnia and Herzegovina, oncology services are constrained by a limited number of specialists and uneven access to evidence-based care. Artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT, may provide clinical decision support to help standardize treatment and assist clinicians where oncology expertise is scarce. This study aimed to evaluate the concordance, safety, and clinical appropriateness of ChatGPT-generated treatment recommendations compared to decisions made by a multidisciplinary team (MDT) in the management of newly diagnosed breast cancer patients. **Methods:** This retrospective study included 91 patients with newly diagnosed, treatment-naïve breast cancer, presented to an MDT in Bosnia and Herzegovina in 2023. Patient data were entered into ChatGPT-4.0 to generate treatment recommendations. Four board-certified oncologists, two internal and two external, evaluated ChatGPT's suggestions against MDT decisions using a 4-point Likert scale. Agreement was analyzed using descriptive statistics, Cronbach's alpha, and Fleiss' kappa. **Results:** The mean agreement score between ChatGPT and MDT decisions was 3.31 (SD = 0.10), with high consistency across oncologist ratings (Cronbach's alpha = 0.86). Fleiss' kappa indicated moderate inter-rater reliability (κ = 0.31, p 〈 0.001). Higher agreement was observed in patients with hormone receptor-negative tumors and those treated with standard chemotherapy regimens. Lower agreement occurred in cases requiring individualized decisions, such as low-grade tumors or uncertain indications

				for surgery or endocrine therapy. **Conclusions:** ChatGPT showed high concordance with MDT treatment plans, especially in standardized clinical scenarios. In resource-limited settings, AI tools may support oncology decision-making and help bridge gaps in clinical expertise. However, careful validation and expert oversight remain essential for safe and effective use in practice.
54	(Riley et al. 2025)	Evaluating the fidelity of AI-generated information on long-acting reversible contraceptive methods	ChatGPT-3.5	**서론:** 인공지능(AI)은 의료 분야에서 많은 응용 분야를 가지고 있다. ChatGPT와 같은 인기 있는 AI 챗봇은 복잡한 건강 주제를 일반 대중에게 더 접근 가능하게 만들 잠재력을 가지고 있다. 본 연구는 ChatGPT가 제공하는 현재 장기 작용 가역적 피임법 정보의 정확성을 평가하는 것을 목표로 한다. **방법:** 우리는 장기 작용 가역적 피임법(LARC)에 대한 8개의 자주 묻는 질문 세트를 ChatGPT에 제시하고, 3일에 걸쳐 반복했다. 각 질문은 다양한 용어를 설명하기 위해 LARC 이름을 변경하여 반복되었다(예: '호르몬 임플란트' 대 'Nexplanon'). 2명의 코더가 AI 생성 답변의 정확성, 언어 포괄성 및 가독성을 독립적으로 평가했다. 3개의 중복 세트의 점수를 평균화했다. **결과:** 총 264개의 응답이 생성되었다. 응답의 69.3%가 정확했다. 응답의 16.3%에는 부정확한 정보가 포함되어 있었다. **가장 일반적인 부정확성은 LARC 사용 기간에 관한 구식 정보였다. 응답의 14.4%는 자궁 내 장치가 골반 염증성 질환의 위험을 증가시킨다고 주장하는 것과 같이 상충되는 근거에 기반한 오해의 소지가 있는 진술을 포함했다. 응답의 45.1%는 성별 배타적 언어를 사용하고 여성만을 언급했다. 평균 Flesch 가독성 용이성 점수는 42.8(SD 7.1)로 대학 읽기 수준과 상관관계가 있었다**. **결론:** ChatGPT는 LARC에 대한 중요한 정보를 제공하지만, 소수의 응답이 부정확하거나 오해의 소지가 있는 것으로 밝혀졌다. **중요한 한계는 AI가 2021년 10월 이전의 데이터에 의존한다는 것이다. AI 도구는 간단한 의학적 질문에 대한 가치 있는 자원이 될 수 있지만, 사용자는 부정확한 정보의 가능성에 주의**해야 한다. **Introduction:** Artificial intelligence (AI) has many applications in health care. Popular AI chatbots, such as ChatGPT, have the potential to make complex health topics more accessible to the general public. The study aims to assess the accuracy of current long-acting reversible contraception information provided by ChatGPT. **Methods:** We presented a set of 8 frequently-asked questions about long-acting reversible contraception (LARC) to ChatGPT, repeated over three distinct days. Each question was repeated with the LARC name changed (e.g., 'hormonal implant' vs 'Nexplanon') to account for variable

				terminology. Two coders independently assessed the AI-generated answers for accuracy, language inclusivity, and readability. Scores from the three duplicated sets were averaged. **Results:** A total of 264 responses were generated. 69.3% of responses were accurate. 16.3% of responses contained inaccurate information. The most common inaccuracy was outdated information regarding the duration of use of LARCs. 14.4% of responses included misleading statements based on conflicting evidence, such as claiming intrauterine devices increase one's risk for pelvic inflammatory disease. 45.1% of responses used gender-exclusive language and referred only to women. The average Flesch readability ease score was 42.8 (SD 7.1), correlating to a college reading level. **Conclusion:** ChatGPT offers important information about LARCs, though a minority of responses are found to be inaccurate or misleading. A significant limitation is AI's reliance on data from before October 2021. While AI tools can be a valuable resource for simple medical queries, users should be cautious of the potential for inaccurate information. SHORT CONDENSATION: ChatGPT generally provides accurate and adequate information about long-acting contraception. However, it occasionally makes false or misleading claims.
55	(Lee et al. 2025)	Evaluating the Influence of Demographic Identity in the Medical Use of Large Language Models		대규모 언어 모델(LLM)이 의료 의사결정에 점점 더 많이 채택됨에 따라, AI 생성 권고사항의 인구통계학적 편향에 대한 우려는 여전히 해결되지 않고 있다. 본 연구에서 우리는 인구통계학적 속성, 특히 인종과 성별이 LLM의 진단, 약물 및 치료 결정에 어떻게 영향을 미치는지 체계적으로 조사한다. MedQA 데이터셋을 사용하여 체계적으로 변화된 의사-환자 인구통계학적 쌍을 포함하는 20,000개의 테스트 사례로 구성된 통제된 평가 프레임워크를 구축한다. 우리는 서로 다른 규모의 두 LLM을 평가한다: 고성능 독점 모델인 Claude 3.5 Sonnet과 더 작은 오픈소스 대안인 Llama 3.1-8B. **우리의 분석은 모델과 작업 전반에 걸쳐 정확도와 편향 패턴 모두에서 상당한 격차를 드러낸다. Claude 3.5 Sonnet이 더 높은 전반적 정확도와 더 안정적인 예측을 보여주는 반면, Llama 3.1-8B는 특히 진단 추론에서 인구통계학적 속성에 대해 더 큰 민감성**을 나타낸다. 특히 **히스패닉 환자가 백인 남성 의사에게 치료받을 때 가장 큰 정확도 감소가 관찰되어, 편향 증폭의 잠재적 위험을 강조**한다. 이러한 발견은 의료 AI에서 엄격한 공정성 평가의 필요성을 강조하고 LLM 기반 의료 애플리케이션에서 인구통계학적 편향을 완화하기 위한 전략을 알려준다. As large language models (LLMs) are increasingly adopted in medical decision-making, concerns about demographic biases in AI-generated

				recommendations remain unaddressed. In this study, we systematically investigate how demographic attributes—specifically race and gender—affect the diagnostic, medication, and treatment decisions of LLMs. Using the MedQA dataset, we construct a controlled evaluation framework comprising 20,000 test cases with systematically varied doctor–patient demographic pairings. We evaluate two LLMs of different scales: Claude 3.5 Sonnet, a high-performance proprietary model, and Llama 3.1-8B, a smaller open-source alternative. Our analysis reveals significant disparities in both accuracy and bias patterns across models and tasks. While Claude 3.5 Sonnet demonstrates higher overall accuracy and more stable predictions, Llama 3.1-8B exhibits greater sensitivity to demographic attributes, particularly in diagnostic reasoning. Notably, we observe the largest accuracy drop when Hispanic patients are treated by White male doctors, underscoring potential risks of bias amplification. These findings highlight the need for rigorous fairness assessments in medical AI and inform strategies to mitigate demographic biases in LLM-driven healthcare applications.
56	(Gurnani et al. 2025)	Evaluating the novel role of ChatGPT-4 in addressing corneal ulcer queries: An AI-powered insight	ChatGPT-4	**목적:** 자연어 처리 기반 AI 모델인 ChatGPT-4는 의료 분야에서 점점 더 많이 적용되고 있으며, 교육, 연구 및 임상 의사결정 지원을 촉진하고 있다. 본 연구는 각막 궤양에 대한 정확하고 상세한 정보를 제공하는 ChatGPT-4의 역량을 탐색하여 의학 교육 및 임상 의사결정에서의 적용을 평가한다. **방법:** 본 연구는 각막 궤양과 관련된 다양한 범주에 걸친 12개의 구조화된 질문으로 ChatGPT-4에 참여시켰다. 각 질문에 대해 5개의 고유한 ChatGPT-4 세션을 시작하여 출력이 이전 질문의 영향을 받지 않도록 했다. 검안 교육 및 연구 직원을 포함한 5명의 안과 전문가 패널이 품질과 정확성에 대해 리커트 척도(1-5)(1: 매우 나쁨; 2: 나쁨; 3: 수용 가능; 4: 좋음; 5: 매우 좋음)를 사용하여 응답을 평가했다. 중앙값 점수가 계산되었고, 평가자 간 신뢰도를 평가하여 평가자 간 일관성을 측정했다. **결과:** 각막 궤양 관련 질문에 대한 ChatGPT-4의 응답 평가 결과 범주 전반에 걸쳐 다양한 성능이 드러났다. 위험 요인, 병인, 증상, 치료, 합병증 및 예후에 대한 중앙값 점수는 일관되게 높았으며(4.0), 좁은 IQR(3.0-4.0)로 강한 일치도를 반영했다. 그러나 **분류 및 검사는 약간 낮은 점수(중앙값 3.0)를 받았다. 각막 궤양의 징후는 중앙값이 2.0으로 상당한 변동성을 보였다.** 300개의 응답 중 45%가 '좋음', 41.7%가 '수용 가능', 10%가 '나쁨', 단 3.3%만이 '매우 좋음'으로 평가되어 개선이 필요한 영역을 강조했다. 특히 평가자 2는 35개의 '좋음' 평가를 제공했고, 평가자 1과 3은 각각 10개의 '나쁨' 평가를 할당했다. **평가자 간 변동성과 진단 정밀도의 격차는 AI 응답 개선의 필요성을 강조**한다. 지속적인 피드백과 표적화된 조정은 고

				품질 안과 교육을 제공하는 데 있어 ChatGPT-4의 유용성을 높일 수 있다. **결론:** ChatGPT-4는 각막 궤양에 대한 교육 콘텐츠를 제공하는 데 유망한 유용성을 보여준다. 평가자 평가의 변동성에도 불구하고, 수치 분석은 추가적인 정제를 통해 ChatGPT-4가 안과 교육 및 임상 지원에서 가치 있는 도구가 될 수 있음을 시사한다. **Purpose:** ChatGPT-4, a natural language processing-based AI model, is increasingly being applied in healthcare, facilitating education, research, and clinical decision-making support. This study explores ChatGPT-4's capability to deliver accurate and detailed information on corneal ulcers, assessing its application in medical education and clinical decision-making. **Methods:** The study engaged ChatGPT-4 with 12 structured questions across different categories related to corneal ulcers. For each inquiry, five unique ChatGPT-4 sessions were initiated, ensuring that the output was not affected by previous queries. A panel of five ophthalmology experts including optometry teaching and research staff assessed the responses using a Likert scale (1–5) (1: very poor; 2: poor; 3: acceptable; 4: good; 5: very good) for quality and accuracy. Median scores were calculated, and inter-rater reliability was assessed to gauge consistency among evaluators. **Results:** The evaluation of ChatGPT-4's responses to corneal ulcer-related questions revealed varied performance across categories. Median scores were consistently high (4.0) for risk factors, etiology, symptoms, treatment, complications, and prognosis, with narrow IQRs (3.0–4.0), reflecting strong agreement. However, classification and investigations scored slightly lower (median 3.0). Signs of corneal ulcers had a median of 2.0, showing significant variability. Of 300 responses, 45% were rated 'good,' 41.7% 'acceptable,' 10% 'poor,' and only 3.3% 'very good,' highlighting areas for improvement. Notably, Evaluator 2 gave 35 'good' ratings, while Evaluators 1 and 3 assigned 10 'poor' ratings each. Inter-evaluator variability, along with gaps in diagnostic precision, underscores the need for refining AI responses. Continuous feedback and targeted adjustments could boost ChatGPT-4's utility in delivering high-quality ophthalmic education. **Conclusion:** ChatGPT-4 shows promising utility in providing educational content on corneal ulcers. Despite the variance in evaluator ratings, the numerical analysis suggests that with further refinement, ChatGPT-4 could be a valuable tool in ophthalmological education and clinical support.

57	(Wang, Guo, et al. 2025)	Evaluating the performance of ChatGPT in clinical multidisciplinary treatment: a retrospective study	ChatGPT-4o	**배경:** 다학제 치료(MDT) 자문은 복잡한 환자 관리에 필수적이다. 그러나 자원과 시간 제약은 그 품질을 제한할 수 있다. 대규모 언어 모델(LLM)은 임상 의사결정 지원에서 잠재력을 보여주었지만, 복잡한 MDT 시나리오에서의 성능은 여전히 불명확하다. 본 연구는 의사가 제공하는 권고사항과 비교하여 ChatGPT가 생성한 MDT 권고사항의 품질을 평가하는 것을 목표로 한다. **방법:** 64개 환자 사례의 임상 데이터가 연구에 후향적으로 포함되었다. ChatGPT에게 특정 MDT 권고사항을 제공하도록 요청했다. 2명의 경험 많은 의사가 포괄성, 정확성, 실행 가능성, 안전성 및 효율성의 5개 측면에 걸쳐 눈가림 방식으로 응답을 평가하고 점수를 매겼으며, 각 측면은 2개의 질문으로 평가되었다. **결과:** ChatGPT의 전체 중앙값 점수는 50.0점 중 41.0점으로, MDT 의사의 중앙값 점수 43.5점보다 낮았다($p = 0.001$). MDT 의사의 응답과 비교하여 ChatGPT는 포괄성에서는 우수했지만($p < 0.001$) **정확성($p < 0.001$), 실행 가능성($p < 0.001$) 및 효율성($p = 0.003$)에서는 부족했다. 특정 질문 분석 결과 ChatGPT는 복잡한 사례의 병인을 추론하는 능력이 부족한 것**으로 나타났다. **결론:** 본 연구는 ChatGPT가 임상 MDT 적용에서 잠재력을 가지고 있음을 나타내며, 특히 임상 요인에 대한 보다 포괄적인 고려를 보여준다. 그러나 **ChatGPT는 여전히 정확성에서 결함이 있어 잘못된 의료 결정으로 이어질 수 있다**. 따라서 LLM의 추가 개발 및 임상 검증이 필요하다. LLM의 현재 한계를 인식하여 임상 실무에서 신중하게 사용하는 것이 필수적이다. **BACKGROUND:** Multidisciplinary treatment (MDT) consultations are essential for managing complex patients. However, resource and time constraints can limit their quality. Large language models (LLMs) have shown potential in assisting clinical decision-making, but their performance in complex MDT scenarios remains unclear. This study aims to evaluate the quality of MDT recommendations generated by ChatGPT compared to those provided by physicians. **METHODS:** Clinical data from 64 patient cases were retrospectively included in the study. ChatGPT was asked to provide specific MDT recommendations. 2 experienced physicians evaluated and scored the responses in a blinded manner across 5 aspects: comprehensiveness, accuracy, feasibility, safety, and efficiency, each assessed by 2 questions. **RESULTS:** The median overall score for ChatGPT was 41.0 out of 50.0, which was lower than the MDT physicians' median score of 43.5 ($p = 0.001$). Compared to the MDT physicians' responses, ChatGPT excelled in comprehensiveness ($p < 0.001$) but fell short in accuracy ($p < 0.001$),

				feasibility (p 〈 0.001), and efficiency (p = 0.003). Analysis of specific questions revealed that ChatGPT lacked the ability to reason through the etiologies of complex cases. **CONCLUSION:** This study indicates that ChatGPT has potential in clinical MDT applications, particularly in demonstrating more comprehensive consideration of clinical factors. However, ChatGPT still has deficiencies in accuracy, which could lead to incorrect healthcare decisions. Therefore, further development and clinical validation of LLMs are necessary. Recognizing the current limitations of LLMs, it is essential to use them with caution in clinical practice. TRIAL REGISTRATION: Not applicable to the present retrospective study. For transparency, a related prospective extension is registered at ChiCTR (ChiCTR2400088563; registered on 21 August 2024).
58	(Jeon et al. 2025)	Evaluating the Use of Generative Artificial Intelligence to Support Genetic Counseling for Rare Diseases	ChatGPT o1-Preview Gemini advanced Claude 3.5 sonnet Perplexity sonar huge	**배경/목적:** 희귀 질환은 낮은 유병률로 인해 일반적인 질환보다 신뢰할 수 있고 정확한 정보를 얻는 데 종종 어려움을 겪는다. 환자와 가족들은 종종 자기주도 학습에 의존하지만, 복잡한 의료 정보를 이해하는 것은 어려울 수 있으며, 이는 오정보의 위험을 증가시킨다. 본 연구는 생성형 인공지능(AI)이 희귀 질환 관련 질문에 정확하고 무해한 답변을 제공하는지 평가하고, 유전 상담이 필요한 환자와 가족을 지원하는 유용성을 평가하는 것을 목표로 했다. **방법:** 우리는 2024년 9월 22일부터 10월 4일 사이에 사용 가능한 4개의 생성형 AI 모델을 평가했다: ChatGPT o1-Preview, Gemini advanced, Claude 3.5 sonnet, Perplexity sonar huge. 일반 정보, 진단, 치료, 예후 및 상담을 다루는 4개 희귀 질환을 대상으로 총 102개의 질문을 준비했다. 4명의 평가자가 리커트 척도(1: 나쁨, 5: 우수함)를 사용하여 전문성과 정확성에 대한 응답을 점수화했다. **결과:** 평균 점수는 AI 모델을 다음과 같이 순위를 매겼다: ChatGPT(4.24 ± 0.73), Gemini(4.15 ± 0.74), Claude(4.13 ± 0.82), Perplexity(3.35 ± 0.80; p 〈 0.001). **Perplexity는 1점(매우 나쁨) 및 2점(나쁨)의 비율이 가장 높았으며(7.6%, 31/408), 그 다음이 Gemini(2.0%, 8/408), Claude(1.5%, 6/408), ChatGPT(1.5%, 6/408)였다. 4개 질환 모두에서 상담 부분 응답의 정확성은 유의한 차이를 보였다(p 〈 0.001).** **결론:** 4개의 생성형 AI 모델은 일반적으로 신뢰할 수 있는 정보를 제공했다. 그러나 **간헐적인 부정확성과 모호한 참조는 환자와 가족들에게 혼란과 불안을 초래할 수 있다**. 효과적인 사용을 보장하기 위해서는 생성형 AI의 한계를 인식하고 적절한 활용에 관한 전문가의 지도를 제공하는 것이 필수적이다. **Background/Objectives:** Rare diseases often present challenges in obtaining

				reliable and accurate information than common diseases owing to their low prevalence. Patients and families often rely on self-directed learning, but understanding complex medical information can be difficult, increasing the risk of misinformation. This study aimed to evaluate whether generative artificial intelligence (AI) provides accurate and non-harmful answers to rare disease-related questions and assesses its utility in supporting patients and families requiring genetic counseling. **Methods:** We evaluated four generative AI models available between 22 September and 4 October 2024: ChatGPT o1-Preview, Gemini advanced, Claude 3.5 sonnet, and Perplexity sonar huge. A total of 102 questions targeting four rare diseases, covering general information, diagnosis, treatment, prognosis, and counseling, were prepared. Four evaluators scored the responses for professionalism and accuracy using the Likert scale (1: poor, 5: excellent). **Results:** The average scores ranked the AI models as: ChatGPT (4.24 ± 0.73), Gemini (4.15 ± 0.74), Claude (4.13 ± 0.82), and Perplexity (3.35 ± 0.80; p 〈 0.001). Perplexity had the highest proportion of scores of 1 (very poor) and 2 (poor) (7.6%, 31/408), followed by Gemini (2.0%, 8/408), Claude (1.5%, 6/408), and ChatGPT (1.5%, 6/408). The accuracy of responses in the counseling part across all four diseases was significantly different (p 〈 0.001). **Conclusions:** The four generative AI models generally provided reliable information. However, occasional inaccuracies and ambiguous references may lead to confusion and anxiety among patients and their families. To ensure its effective use, recognizing the limitations of generative AI and providing guidance from experts regarding its proper utilization is essential.
59	(Huang et al. 2025)	Evaluation and Bias Analysis of Large Language Models in Generating Synthetic Electronic Health Records: Comparative Study	Yi-6B Yi-34B Qwen-1.8B Qwen-7B Qwen-14B Llama2-7B Llama2-13B	**배경:** 대규모 언어 모델(LLM)에 의해 생성된 합성 전자건강기록(EHR)은 개인정보 보호 문제를 해결하면서 임상 교육 및 모델 훈련에 잠재력을 제공한다. 그러나 이러한 모델의 성능 변동성과 인구통계학적 편향은 충분히 탐색되지 않아 공평한 의료에 위험을 초래한다. **목적:** 본 연구는 합성 EHR 생성에서 다양한 LLM의 성능을 체계적으로 평가하고, 생성된 출력에서 성별 및 인종 편향의 존재를 비판적으로 평가하는 것을 목표로 했다. 우리는 다양한 인구통계학적 유병률을 가진 20개 질환에 걸쳐 이러한 EHR의 완전성과 대표성을 평가하는 데 중점을 두었다. **방법:** 7개 LLM에 걸쳐 10개의 표준화된 프롬프트를 사용하여 140,000개의 합성 EHR을 생성하는 프레임워크를 개발했다. 전자건강기록 성능 점수(EPS)를 도입하여 완전성을 정량화했으며, 통계적 동등성 차이(SPD)를 제안하여 인구통계학적 편향의

				정도와 방향을 평가했다. 카이제곱 검정을 사용하여 인구통계학적 그룹 간 편향의 존재를 평가했다. **결과: 더 큰 모델이 우수한 성능을 보였지만 편향도 증가했다**. Yi-34B는 가장 높은 EPS(96.8)를 달성했고, 더 작은 모델(Qwen-1.8B: EPS=63.35)은 성능이 낮았다. **성별 양극화**가 나타났다: **여성 우세 질환(예: 다발성 경화증)은 출력에서 여성 대표성이 증폭되었고(Qwen-14B: 973/1000, 97.3% 여성 vs 564,424/744,778, 75.78% 실제; SPD=+21.50%), 균형 잡힌 질환과 남성 우세 질환은 남성 그룹으로 치우쳤다(예: 고혈압 Llama 2-13B: 957/1000, 95.7% 남성 vs 79,540,040/152,466,669, 52.17% 실제; SPD=+43.50%). 인종 편향 패턴은 일부 모델이 백인(예: Yi-6B: 평균 SPD +14.40%, SD 16.22%) 또는 흑인 그룹(예: Yi-34B: 평균 SPD +14.90%, SD 27.16%)의 대표성을 과대평가한 반면, 대부분의 모델은 히스패닉(7개 모델의 평균 SPD는 −11.93%, SD 8.36%) 및 아시아인 그룹(7개 모델의 평균 SPD는 −0.77%, SD 11.99%)의 대표성을 체계적으로 과소평가했음을 보여주었다.** **결론:** Yi-34B, Qwen-14B, Llama 2-13B와 같은 더 큰 모델은 더 높은 EPS 값에 반영된 것처럼 보다 포괄적인 EHR 생성에서 향상된 성능을 보였다. 그러나 **이러한 성능 향상은 성별 및 인종 편향의 현저한 증가를 동반하여 성능-편향 상충관계를 강조**한다. 본 연구는 다음과 같은 4가지 주요 발견을 식별했다: **(1) 모델 크기가 증가함에 따라 EHR 생성이 개선되었지만, 인구통계학적 편향도 더욱 두드러졌다; (2) 편향은 큰 모델뿐만 아니라 모든 모델에서 관찰되었다; (3) 성별 편향은 실제 질병 유병률과 밀접하게 일치했지만, 인종 편향은 일부 질환에서만 명백했다; (4) 인종 편향은 다양했으며, 일부 질환은 백인 또는 흑인 인구의 과대표현과 히스패닉 및 아시아인 그룹의 과소표현**을 보였다. 이러한 발견은 의료를 위한 인공지능 응용에서 공정성을 보장하기 위한 효과적인 편향 완화 전략과 벤치마크 개발의 필요성을 강조한다. **Background:** Synthetic electronic health records (EHRs) generated by large language models (LLMs) offer potential for clinical education and model training while addressing privacy concerns. However, performance variations and demographic biases in these models remain underexplored, posing risks to equitable health care. **Objective:** This study aimed to systematically assess the performance of various LLMs in generating synthetic EHRs and to critically evaluate the presence of gender and racial biases in the generated outputs. We focused on assessing the completeness and representativeness of these EHRs across 20 diseases with varying demographic prevalence.

				Methods: A framework was developed to generate 140,000 synthetic EHRs using 10 standardized prompts across 7 LLMs. The electronic health record performance score (EPS) was introduced to quantify completeness, while the statistical parity difference (SPD) was proposed to assess the degree and direction of demographic bias. Chi-square tests were used to evaluate the presence of bias across demographic groups. **Results:** Larger models exhibited superior performance but heightened biases. The Yi-34B achieved the highest EPS (96.8), while smaller models (Qwen-1.8B: EPS=63.35) underperformed. Sex polarization emerged: female-dominated diseases (eg, multiple sclerosis) saw amplified female representation in outputs (Qwen-14B: 973/1000, 97.3% female vs 564,424/744,778, 75.78% real; SPD=+21.50%), while balanced diseases and male-dominated diseases skewed the male group (eg, hypertension Llama 2-13 B: 957/1000, 95.7% male vs 79,540,040/152,466,669, 52.17% real; SPD=+43.50%). Racial bias patterns revealed that some models overestimated the representation of White (eg, Yi-6B: mean SPD +14.40%, SD 16.22%) or Black groups (eg, Yi-34B: mean SPD +14.90%, SD 27.16%), while most models systematically underestimated the representation of Hispanic (average SPD across 7 models is −11.93%, SD 8.36%) and Asian groups (average SPD across 7 models is −0.77%, SD 11.99%). **Conclusions:** Larger models, such as Yi-34B, Qwen-14B, and Llama 2 to 13 B, showed improved performance in generating more comprehensive EHRs, as reflected in higher EPS values. However, this increased performance was accompanied by a notable escalation in both gender and racial biases, highlighting a performance-bias trade-off. The study identified 4 key findings as follows: (1) as model size increased, EHR generation improved, but demographic biases also became more pronounced; (2) biases were observed across all models, not just the larger ones; (3) gender bias closely aligned with real-world disease prevalence, while racial bias was evident in only a subset of diseases; and (4) racial biases varied, with some diseases showing overrepresentation of White or Black populations and underrepresentation of Hispanic and Asian groups. These findings underline the need for effective bias mitigation strategies and the development of benchmarks to ensure fairness in artificial intelligence applications for health care.
60	(Hager et al. 2024)	Evaluation and mitigation of the	Llama 2 Chat	임상 의사결정은 의사의 책임 중 가장 영향력 있는 부분 중 하나이며, 특히 인공지

		limitations of large language models in clinical decision-making	OASST WizardLM Clinical Camel Meditron ChatGPT (GPT-3.5) ChatGPT-4 Med-PaLM Med-PaLM 2	능 솔루션과 대규모 언어 모델(LLM)로부터 큰 혜택을 받을 수 있다. 그러나 LLM이 의사 면허 시험에서 우수한 성과를 달성했지만, 이러한 시험은 정보 수집, 지침 준수, 임상 워크플로우 통합을 포함하여 현실적인 임상 의사결정 환경에 배치하는 데 필요한 많은 기술을 평가하지 못한다. 여기서 우리는 2,400개의 실제 환자 사례와 4가지 일반적인 복부 병리를 포괄하는 Medical Information Mart for Intensive Care 데이터베이스를 기반으로 큐레이션된 데이터셋과 현실적인 임상 환경을 시뮬레이션하는 프레임워크를 만들었다. 우리는 **현재 최첨단 LLM이 모든 병리에 걸쳐 환자를 정확하게 진단하지 못하고(의사보다 현저히 낮은 성능), 진단 및 치료 지침을 모두 따르지 않으며, 실험실 결과를 해석할 수 없어 환자의 건강에 심각한 위험을 초래함을 보여준다. 더욱이 우리는 진단 정확도를 넘어 지시를 따르지 못하고 정보의 양과 순서 모두에 민감하기 때문에 기존 워크플로우에 쉽게 통합될 수 없음을 입증**한다. 전반적으로 우리의 분석은 LLM이 현재 자율적인 임상 의사결정에 준비되지 않았음을 밝히면서 향후 연구를 안내할 데이터셋과 프레임워크를 제공한다. Clinical decision-making is one of the most impactful parts of a physician's responsibilities and stands to benefit greatly from artificial intelligence solutions and large language models (LLMs) in particular. However, while LLMs have achieved excellent performance on medical licensing exams, these tests fail to assess many skills necessary for deployment in a realistic clinical decision-making environment, including gathering information, adhering to guidelines, and integrating into clinical workflows. Here we have created a curated dataset based on the Medical Information Mart for Intensive Care database spanning 2,400 real patient cases and four common abdominal pathologies as well as a framework to simulate a realistic clinical setting. We show that current state-of-the-art LLMs do not accurately diagnose patients across all pathologies (performing significantly worse than physicians), follow neither diagnostic nor treatment guidelines, and cannot interpret laboratory results, thus posing a serious risk to the health of patients. Furthermore, we move beyond diagnostic accuracy and demonstrate that they cannot be easily integrated into existing workflows because they often fail to follow instructions and are sensitive to both the quantity and order of information. Overall, our analysis reveals that LLMs are currently not ready for autonomous clinical decision-making while providing a dataset and framework to guide future studies.
61	(Latt et al. 2025)	Evaluation of Artificial Intelligence (AI) Chatbots for	Alice (Custom GPT-3.5-Turbo on chatbotbuilder.io)	**서론:** 인공지능(AI) 챗봇은 대중에게 성 건강에 대한 정보를 제공할 수 있다. 그러나 인간 임상의와 비교한 성 건강에서의 성능과 다양한 AI 챗봇 간의 성능은 충분히

		Providing Sexual Health Information: A Consensus Study Using Real-World Clinical Queries	Azure (Custom GPT-3.5 on Microsoft Azure) ChatGPT (standard OpenAI GPT-3.5)	연구되지 않았다. 본 연구는 세 가지 AI 챗봇 – 두 개의 프롬프트 조정 챗봇(Alice 및 Azure)과 한 개의 표준 챗봇(OpenAI의 ChatGPT) –의 성 건강 정보 제공 성능을 임상의와 비교하여 평가했다. **방법:** 우리는 멜버른 성 건강 센터 전화 상담에서 받은 195개의 익명화된 성 건강 질문을 분석했다. 합의 기반 접근법을 사용하여 전문가 패널이 눈가림 순서로 세 가지 AI 챗봇이 생성한 이러한 질문에 대한 응답을 평가했다. 성능은 전체 정확성과 안내, 정확도, 안전성, 접근 용이성, 필요한 정보 제공의 5가지 특정 측정값을 기준으로 평가되었다. 우리는 진료소 특정 및 일반 성 건강 질문에 대한 하위 그룹 분석과 Azure가 답변할 수 없는 질문을 제외한 민감도 분석을 수행했다. **결과:** Alice는 가장 높은 전체 정확성(85.2%; 95% 신뢰구간(CI), 82.1%–88.0%)을 보였으며, Azure(69.3%; 95% CI, 65.3%–73.0%) 및 ChatGPT(64.8%; 95% CI, 60.7%–68.7%)가 그 뒤를 이었다. 프롬프트 조정 챗봇은 모든 측정값에서 기본 ChatGPT를 능가했다. **Azure는 가장 높은 안전성 점수(97.9%; 95% CI, 96.4%–98.9%)를 달성하여 잠재적으로 해로운 조언을 제공할 위험이 가장 낮았다. 하위 그룹 분석에서 모든 챗봇은 진료소 특정 질문에 비해 일반 성 건강 질문에서 더 나은 성능을 보였다. 민감도 분석은 Azure가 답변할 수 없는 질문을 제외할 때 Alice와 Azure 간의 성능 격차가 더 좁다는 것을 보여주었다.** **결론:** 프롬프트 조정 AI 챗봇은 기본 ChatGPT에 비해 성 건강 정보 제공에서 우수한 성능을 보였으며, 특히 높은 안전성 점수가 주목할 만하다. 그러나 **모든 AI 챗봇은 부정확한 정보를 생성하는 데 취약성을 보였다**. 이러한 발견은 AI 챗봇이 인간 의료 제공자의 보조 수단으로서의 잠재력을 시사하면서 지속적인 정제 및 인간 감독의 필요성을 강조한다. 향후 연구는 대규모 평가 및 실제 구현에 초점을 맞춰야 한다. **Introduction:** Artificial Intelligence (AI) chatbots could provide information on sexual health to the public. However, their performance in sexual health compared to human clinicians and across different AI chatbots remains understudied. This study evaluated the performance of three AI chatbots – two prompt-tuned (Alice and Azure) and one standard chatbot (ChatGPT by OpenAI) – in providing sexual health information, compared to clinicians. **Methods:** We analysed 195 anonymised sexual health questions received by the Melbourne Sexual Health Centre phone line. Using a consensusbased approach, a panel of experts evaluated responses to these questions generated by the three AI chatbots in a blinded order. Performance was assessed based on overall correctness and five specific measures: guidance, accuracy, safety, ease of access, and provision of necessary information. We

				conducted subgroup analyses for clinic-specific and general sexual health questions and a sensitivity analysis excluding questions that Azure could not answer. **Results:** Alice demonstrated the highest overall correctness (85.2%; 95% confidence interval (CI), 82.1%-88.0%), followed by Azure (69.3%; 95% CI, 65.3%-73.0%) and ChatGPT (64.8%; 95% CI, 60.7%-68.7%). Prompt-tuned chatbots outperformed the base ChatGPT across all measures. Azure achieved the highest safety score (97.9%; 95% CI, 96.4%-98.9%), indicating the lowest risk of providing potentially harmful advice. In subgroup analysis, all chatbots performed better on general sexual health questions compared to clinic-specific queries. Sensitivity analysis showed a narrower performance gap between Alice and Azure when excluding questions Azure could not answer. **Conclusions:** Prompt-tuned AI chatbots demonstrated superior performance in providing sexual health information compared to base ChatGPT, with high safety scores particularly noteworthy. However, all AI chatbots showed susceptibility to generating incorrect information. These findings suggest the potential for AI chatbots as adjuncts to human healthcare providers while highlighting the need for continued refinement and human oversight. Future research should focus on larger-scale evaluations and real-world implementation.
62	(Radulesco et al. 2025)	Evaluation of Artificial Intelligence Chatbots for Facial Injection Planning: Comparative Performance and Safety Limitations	ChatGPTo1 ChatGPT4o Gemini 2.0 Claude-3.5-Sonnet Copilot Pro Llama 3.3 Flux Pro 1.1 Ideogram v2 Dall E3	**배경:** 안면 미용 주사에 대한 치료 계획 생성에서 인공지능(AI) 기반 챗봇의 성능을 평가하고, 정확성, 안전성 및 임상 적용 가능성에 초점을 맞춘다. **방법:** STROBE 지침에 따라 이비인후과 3차 진료 부서에서 비교 관찰 연구를 수행했다. 안면 주사를 원하는 환자를 2024년 7월부터 10월까지 모집했다. 40명의 환자(85% 여성; 평균 연령: 45.8세)가 사진 기록을 받았고 보툴리눔 독소 및 히알루론산 주사에 대한 AI 생성 치료 계획을 받았다. 6개의 AI 챗봇과 3개의 생성형 비전 모델을 제품 선택, 주사 전략, 안면 분석, 환자 선호도와의 일치도 및 안전성의 5가지 기준을 바탕으로 평가했다. 각각 -2에서 +2까지의 범위를 가진 리커트 척도 평가를 Friedman 및 Durbin-Conover 쌍별 검정을 사용하여 분석하여 유의한 차이($p < 0.05$)를 식별했다. 5개 리커트 척도의 합계는 -10에서 +10까지의 전체 점수를 제공했다. **결과:** ChatGPTo1과 ChatGPT4o는 대부분의 평가 기준에서 다른 챗봇보다 높은 점수를 달성했으며, 평균 총점은 각각 7.87 ± 0.29 및 7.85 ± 0.44였다($p = 0.295$). 두 챗봇 모두 제품 선택(ChatGPT4o = 1.92 ± 0.05), 주사 전략 정밀도(ChatGPTo1 = 1.67 ± 0.08), 환자 선호도와의 일치도(ChatGPTo1 = 1.95 ±

				0.03) 및 안전성(ChatGPTo1 = 1.30 ± 0.17)에서 Claude, CopilotPro 및 Llama 보다 통계적으로 우수했다(p 〈 0.05). Claude는 ChatGPT 모델과 비교하여 유의한 차이 없이 관련성 있는 안면 분석(1.50 ± 0.16)을 제공했다(모두 p 〉 0.05). 생성형 비전 모델은 관련성 있는 시각적 주석을 생성하지 못했다. **결론:** 테스트된 AI 시스템 중 ChatGPT 기반 챗봇이 안면 주사에 대한 치료 계획 생성에서 상대적으로 우수한 성능을 보였다. 그러나 안전성 한계가 남아 있어 무감독 임상 사용은 금지된다. **BACKGROUND:** To evaluate the performance of artificial intelligence (AI)-powered chatbots in generating treatment plans for facial aesthetic injections, focusing on their accuracy, safety, and clinical applicability. **METHODS:** A comparative observational study was conducted in an otolaryngology tertiary care department according to STROBE guidelines. Patients seeking facial injections were recruited from July to October 2024. Forty patients (85% female; mean age: 45.8 years) underwent photographic documentation and received AI-generated treatment plans for botulinum toxin and hyaluronic acid injections. Six AI chatbots and three generative vision models were evaluated based on five criteria: product selection, injection strategy, facial analysis, alignment with patient preferences, and safety. Likert scale ratings, each ranging from − 2 to + 2, were analyzed using Friedman and Durbin-Conover pairwise tests to identify significant differences (p 〈 0.05). The sum of the five Likert scales provided an overall score ranging from − 10 to + 10. **RESULTS:** ChatGPTo1 and ChatGPT4o achieved higher scores than other chatbots across most evaluation criteria, with mean total scores of 7.87 ± 0.29 and 7.85 ± 0.44, respectively (p = 0.295). Both chatbots were statistically superior (p 〈 0.05) to Claude, CopilotPro, and Llama in product selection (ChatGPT4o = 1.92 ± 0.05), injection strategy precision (ChatGPTo1 = 1.67 ± 0.08), alignment with patient preferences (ChatGPTo1 = 1.95 ± 0.03) and safety (ChatGPTo1 = 1.30 ± 0.17). Claude provided relevant facial analysis (1.50 ± 0.16) without significant difference compared to ChatGPT models (all p 〉 0.05). Generative vision models failed to produce relevant visual annotations. **CONCLUSION:** Among the AI systems tested, ChatGPT-based chatbots demonstrated relatively superior performance in generating treatment plans for facial injections. However, safety limitations remain and preclude

				unsupervised clinical use. LEVEL OF EVIDENCE IV: This journal requires that authors assign a level of evidence to each article. For a full description of these Evidence-Based Medicine ratings, please refer to the Table of Contents or the online Instructions to Authors www.springer.com/00266 .
63	(Arslan et al. 2025)	Evaluation of ChatGPT-5 responses in obstetric and gynecological emergencies: concordance, readability, and clinical reliability	ChatGPT-5	**배경:** 본 연구는 산과 및 부인과 응급 시나리오에서 ChatGPT-5 응답의 지침 준수, 임상 안전성 및 적용 가능성을 평가하는 것을 목표로 했다. AI 기반 대규모 언어 모델(LLM)의 의료 분야에서의 역할이 증가함에 따라, 산과 응급상황에서의 성능을 체계적으로 검토할 필요가 있다. **방법:** 본 연구는 전향적, 시나리오 기반, 이중맹검 연구로 설계되었다. 문헌 및 현행 국제 지침(ACOG, RCOG, WHO)을 기반으로 총 15개의 산과 및 부인과 응급 시나리오를 생성했다. 각 시나리오에 대해 ChatGPT-5에 5개의 표준 질문을 제시했다: (1) 가장 가능성 있는 진단, (2) 진단 확인을 위한 검사, (3) 혈역학적 안정성 평가, (4) 초기 치료 접근법, (5) 고급 관리 옵션. 동일한 시나리오를 2명의 산부인과 전문의, 응급의학 전문의 및 마취과 전문의가 독립적으로 답변했으며 "표준"으로 간주되었다. 응답은 지침 준수, 환자 안전성 및 중요 정보 격차에 대해 점수가 매겨졌다. 또한 품질과 이해 가능성은 수정된 DISCERN(mDISCERN), Global Quality Score(GQS), 가독성 지표[Flesch Reading Ease Score(FRES), Flesch-Kincaid Grade Level(FKGL), Simple Measure of Gobbledygook(SMOG), Coleman-Liau Index(CLI)]로 평가되었다. **결과:** 총 75개의 응답이 검토되었다. 높은 일치도(5/5)는 5개 시나리오(33.3%)에서, 중등도 일치도(4/5)는 7개 시나리오(46.7%)에서, 낮은 일치도(≤ 3/5)는 3개 시나리오(20.0%)에서 관찰되었다. 높은 일치도는 특히 산후 출혈, 자간증, HELLP 증후군, 견난산, 자궁외 임신 파열과 같은 잘 정의된 지침 알고리즘에서 명백했다. **중등도 일치 시나리오의 결함에는 사망 위험에 대한 강조 부족, 점수 시스템 누락, 패혈증 관리의 불완전한 단계, 불충분한 임신 보존 접근법 명시가 포함되었다. 낮은 일치 시나리오에는 심한 질 출혈, 악성 종양으로 인한 급성 출혈, 외상성 부인과 응급상황이 포함**되었다. 응답의 평균 mDISCERN 점수는 4.0 ± 0.7이었고 평균 GQS는 4.1 ± 0.7이었다. 가독성 분석은 응답이 중등도의 전문 용어를 포함했음을 보여주었다(FRES 점수 = 40.5 ± 2.5; FKGL 점수 = 11.6 ± 1.2; SMOG 점수 = 10.9 ± 0.8; CLI 점수 = 10.9 ± 0.8). 평균 어휘 밀도는 0.63이었다. **결론:** ChatGPT-5는 일반적으로 산과 및 부인과 응급 시나리오에서 중등도에서 양호한 지침 준수적이고 확신 있는 응답을 생성했다. 그러나 **다학제적 접근이 필요한 복잡한 사례에서는 성능이 제한적**이었다. 연구 결과는 AI 기반 대규모 언어 모델이 산과 응급 관리에서 보완적 도구가 될 수 있지만 전문 임상의 감독 없이 단독으로 사용해서는 안 된다는 것을 시사한다. 더 크고 비교적이며 다학제적인 연구가 이러한 기술의 임상 통합에 대한 더 신뢰할 수 있는 근거를 제공할 것이다.

				Background: This study aimed to evaluate the compliance with guidelines, clinical safety, and applicability of ChatGPT-5 responses in obstetric and gynecological emergency scenarios. With the increasing role of AI-powered large language models (LLMs) in healthcare, there is a need to examine their performance in obstetric emergencies systematically. **Methods:** This study was designed as a prospective, scenario-based, double-blind study. A total of 15 obstetric and gynecologic emergency scenarios were created based on the literature and current international guidelines (ACOG, RCOG, WHO). Five standard questions were posed to ChatGPT-5 for each scenario: (1) Most likely diagnosis, (2) Investigations to confirm the diagnosis, (3) Hemodynamic stability assessment, (4) Initial treatment approach, and (5) Advanced management options. The same scenarios were independently answered by two obstetricians, an emergency medicine specialist, and an anesthesiologist, and were considered the "gold standard." Responses were scored for guideline compliance, patient safety, and critical information gaps. In addition, quality and understandability were evaluated with modified DISCERN (mDISCERN), Global Quality Score (GQS), and readability indexes [Flesch Reading Ease Score (FRES), Flesch-Kincaid Grade Level (FKGL), Simple Measure of Gobbledygook (SMOG), Coleman-Liau Index (CLI)]. **Results:** A total of 75 responses were reviewed. High agreement (5/5) was observed in 5 scenarios (33.3%), moderate agreement (4/5) in 7 scenarios (46.7%), and low agreement (≤ 3/5) in 3 scenarios (20.0%). High agreement was particularly evident for well-defined guideline algorithms, such as postpartum hemorrhage, eclampsia, HELLP syndrome, shoulder dystocia, and ruptured ectopic pregnancy. Deficiencies in moderate agreement scenarios included insufficient emphasis on mortality risk, omission of scoring systems, incomplete steps in sepsis management, and inadequate specification of fertility-sparing approaches. Low agreement scenarios included severe vaginal hemorrhage, acute bleeding due to malignancy, and traumatic gynecologic emergencies. The mean mDISCERN score of the responses was 4.0 ± 0.7, and the mean GQS was 4.1 ± 0.7. Readability analyses showed that responses contained a moderate amount of technical language (FRES score = 40.5 ± 2.5; FKGL score = 11.6 ± 1.2; SMOG score = 10.9 ± 0.8; and CLI score = 10.9 ± 0.8). The mean lexical density was

				0.63. **Conclusions:** ChatGPT-5 generally produced moderate to good guideline-compliant and confident responses in obstetric and gynecological emergency scenarios. However, its performance was limited in complex cases requiring a multidisciplinary approach. The findings suggest that AI-powered large language models can be a complementary tool in obstetric emergency management, but should not be used alone without expert clinician supervision. Larger, comparative, and multidisciplinary studies will provide more reliable evidence for the clinical integration of these technologies.
64	(Makrygiannakis, Giannakopoulos, and Kaklamanos 2024)	Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing	Google's Bard ChatGPT-3.5 ChatGPT-4 Microsoft's Bing	**배경:** 다양한 의학 및 치과 분야, 특히 치열교정학에서 생성형 인공지능의 대규모 언어 모델(LLM) 활용이 증가함에 따라 정확성에 대한 의문이 제기된다. **목적:** 본 연구는 치열교정학 분야 내 임상적으로 관련된 질문에 대한 응답으로 4개의 LLM - Google의 Bard, OpenAI의 ChatGPT-3.5 및 ChatGPT-4, Microsoft의 Bing -이 제공하는 답변을 평가하고 비교하는 것을 목표로 했다. **재료 및 방법:** 10개의 개방형 임상 치열교정 관련 질문을 LLM에 제시했다. LLM이 제공한 응답은 합의 성명 및 체계적 문헌고찰을 포함한 강력한 과학적 근거를 기준으로 사전 정의된 루브릭을 사용하여 0점(최소)에서 10점(최대)까지의 척도로 평가되었다. 초기 평가로부터 4주 간격 후 답변을 재평가하여 평가자 내 신뢰도를 측정했다. 가장 포괄적이고 과학적으로 정확하며 명확하고 관련성 있는 답변을 제공하는 모델을 식별하기 위해 Friedman 및 Wilcoxon 검정을 사용하여 점수에 대한 통계적 비교를 수행했다. **결과:** 전반적으로 두 평가자가 두 차례의 점수 매김에서 부여한 점수 간에 통계적으로 유의한 차이가 발견되지 않아 모든 LLM에 대한 평균 점수가 계산되었다. 가장 높은 점수를 받은 LLM 답변은 Microsoft Bing Chat(평균 점수 = 7.1)이었고, ChatGPT 4(평균 점수 = 4.7), Google Bard(평균 점수 = 4.6), 마지막으로 ChatGPT 3.5(평균 점수 = 3.8)가 뒤를 이었다. Microsoft Bing Chat은 ChatGPT-3.5(P값 = 0.017) 및 Google Bard(P값 = 0.029)보다 통계적으로 우수했으며, ChatGPT-4도 ChatGPT-3.5(P값 = 0.011)보다 우수했지만, **모든 모델은 때때로 포괄성, 과학적 정확성, 명확성 및 관련성이 부족한 답변을 생성**했다. **한계:** 제시된 질문은 지표적이었으며 치열교정학 전체 분야를 다루지 않았다. **결론:** 언어 모델(LLM)은 근거 기반 치열교정학을 지원하는 데 큰 잠재력을 보여준다. 그러나 **현재의 한계는 신중하게 고려하지 않고 활용될 경우 잘못된 의료 결정을 내릴 잠재적 위험을 제기**한다. 결과적으로 이러한 도구는 치열교정의의 필수적인 비판적 사고와 포괄적인 주제 지식을 대체할 수 없다. 실무에 효과적으로 통합하기 위해서는 추가 연구, 임상 검증 및 모델 개선이 필수적이다. 임상의는 LLM의 한계를 인식해야 하며, 부주의한 활용은 환자 치료에 부정적인 영향을 미칠 수 있다.

				BACKGROUND: The increasing utilization of large language models (LLMs) in Generative Artificial Intelligence across various medical and dental fields, and specifically orthodontics, raises questions about their accuracy. **OBJECTIVE:** This study aimed to assess and compare the answers offered by four LLMs: Google's Bard, OpenAI's ChatGPT-3.5, and ChatGPT-4, and Microsoft's Bing, in response to clinically relevant questions within the field of orthodontics. **MATERIALS AND METHODS:** Ten open-type clinical orthodontics-related questions were posed to the LLMs. The responses provided by the LLMs were assessed on a scale ranging from 0 (minimum) to 10 (maximum) points, benchmarked against robust scientific evidence, including consensus statements and systematic reviews, using a predefined rubric. After a 4-week interval from the initial evaluation, the answers were reevaluated to gauge intra-evaluator reliability. Statistical comparisons were conducted on the scores using Friedman's and Wilcoxon's tests to identify the model providing the answers with the most comprehensiveness, scientific accuracy, clarity, and relevance. **RESULTS:** Overall, no statistically significant differences between the scores given by the two evaluators, on both scoring occasions, were detected, so an average score for every LLM was computed. The LLM answers scoring the highest, were those of Microsoft Bing Chat (average score = 7.1), followed by ChatGPT 4 (average score = 4.7), Google Bard (average score = 4.6), and finally ChatGPT 3.5 (average score 3.8). While Microsoft Bing Chat statistically outperformed ChatGPT-3.5 (P-value = 0.017) and Google Bard (P-value = 0.029), as well, and Chat GPT-4 outperformed Chat GPT-3.5 (P-value = 0.011), all models occasionally produced answers with a lack of comprehensiveness, scientific accuracy, clarity, and relevance. **LIMITATIONS:** The questions asked were indicative and did not cover the entire field of orthodontics. **CONCLUSIONS:** Language models (LLMs) show great potential in supporting evidence-based orthodontics. However, their current limitations pose a potential risk of making incorrect healthcare decisions if utilized without careful consideration. Consequently, these tools cannot serve as a substitute for the orthodontist's essential critical thinking and comprehensive subject knowledge. For effective integration into practice, further research, clinical

				validation, and enhancements to the models are essential. Clinicians must be mindful of the limitations of LLMs, as their imprudent utilization could have adverse effects on patient care.
65	(Zaboli, Brigo, Ziller, et al. 2025)	Exploring ChatGPT's potential in ECG interpretation and outcome prediction in emergency department	ChatGPT-4.0	**배경:** 응급실(ED) 방문의 약 20%는 심혈관 증상과 관련이 있다. 심전도는 심각한 질환을 진단하는 데 중요하지만, 해석 정확도는 응급의학과 전문의마다 다르다. ChatGPT와 같은 인공지능(AI)은 진단 정밀도를 향상시켜 심전도 해석을 지원할 수 있다. **방법:** Merano 병원 응급실에서 수행된 이 단일 기관 후향적 관찰 연구는 심전도 해석에서 ChatGPT와 심장내과 전문의 간의 일치도를 평가했다. 주요 결과는 ChatGPT와 심장내과 전문의 간의 일치 수준이었다. 2차 결과는 주요 심장 부작용 사건(MACE) 위험 환자를 식별하는 ChatGPT의 능력을 포함했다. **결과:** 등록된 128명의 환자 중 ChatGPT는 **T파(kappa = 0.048) 및 ST 분절(kappa = 0.267)을 제외**한 대부분의 심전도 분절에서 심장내과 전문의와 양호한 일치도를 보였다. **ChatGPT가 의사가 식별한 것보다 더 많은 환자를 MACE 위험군으로 분류함에 따라 중증 사례 평가에서 상당한 불일치가 발생**했다. **결론:** ChatGPT는 심전도 해석에서 중등도 정확도를 보여주지만, **특히 중증 사례 평가에서의 현재 한계는 응급실 환경에서의 임상 유용성을 제한**한다. 향후 연구와 기술 발전은 AI의 신뢰성을 향상시켜 응급의학과 전문의를 위한 가치 있는 지원 도구로 자리매김할 수 있다. **Background:** Approximately 20 % of emergency department (ED) visits involve cardiovascular symptoms. While ECGs are crucial for diagnosing serious conditions, interpretation accuracy varies among emergency physicians. Artificial intelligence (AI), such as ChatGPT, could assist in ECG interpretation by enhancing diagnostic precision. **Methods:** This single-center, retrospective observational study, conducted at Merano Hospital's ED, assessed ChatGPT's agreement with cardiologists in interpreting ECGs. The primary outcome was agreement level between ChatGPT and cardiologists. Secondary outcomes included ChatGPT's ability to identify patients at risk for Major Adverse Cardiac Events (MACE). **Results:** Of the 128 patients enrolled, ChatGPT showed good agreement with cardiologists on most ECG segments, excluding T wave (kappa = 0.048) and ST segment (kappa = 0.267). Significant discrepancies arose in the assessment of critical cases, as ChatGPT classified more patients as at risk for MACE than were identified by physicians. **Conclusions:** ChatGPT demonstrates moderate accuracy in ECG

				interpretation, yet its current limitations, especially in assessing critical cases, restrict its clinical utility in ED settings. Future research and technological advancements could enhance AI's reliability, potentially positioning it as a valuable support tool for emergency physicians.
66	(Teixeira-Marques et al. 2024)	Exploring the role of ChatGPT in clinical decision-making in otorhinolaryngology: a ChatGPT designed study	ChatGPT-1 ChatGPT-2	**목적:** 2023년 초부터 ChatGPT는 의료 연구의 뜨거운 주제로 부상했다. 임상 실무에서 가치 있는 도구가 될 잠재력은 설득력이 있으며, 특히 의사가 이용 가능한 최고의 의학 지식을 기반으로 임상 결정을 내리는 것을 도와 임상 의사결정 지원을 개선하는 데 있다. 우리는 이비인후과 관련 증상이 있는 환자를 식별, 진단 및 관리하는 ChatGPT의 능력을 조사하는 것을 목표로 한다. **방법:** 20개의 현실에서 영감을 받은 임상 사례에서 ChatGPT와 5명의 이비인후과 전문의(ENT) 간의 일치 수준을 평가하기 위해 ChatGPT가 제안한 아이디어를 기반으로 전향적 횡단면 연구를 설계했다. 임상 사례는 시간적 안정성을 평가하기 위해 두 번의 다른 시점(ChatGPT-1 및 ChatGPT-2)에 챗봇에 제시되었다. **결과:** ChatGPT-1의 평균 점수는 4.4(SD 1.2; 최소 1, 최대 5)였고 ChatGPT-2는 4.15(SD 1.3; 최소 1, 최대 5)였으며, **ENT의 평균 점수는 4.91(SD 0.3; 최소 3, 최대 5)이었다. Mann-Whitney U 검정은 ChatGPT와 ENT의 점수 간에 통계적으로 유의한 차이($p < 0.001$)를 밝혔다. ChatGPT-1과 ChatGPT-2는 5회에 걸쳐 다른 답변을 제공**했다. **결론:** 인공지능은 가까운 미래에 임상 의사결정에서 중요한 도구가 될 것이며 ChatGPT는 지금까지 가장 유망한 챗봇이다. 안전하게 사용하기 위해서는 추가 개발이 필요하지만, 개선의 여지가 있으며 이비인후과 전공의와 전문의가 환자를 위한 가장 올바른 결정을 내리는 것을 돕는 잠재력이 있다. **Purpose:** Since the beginning of 2023, ChatGPT emerged as a hot topic in healthcare research. The potential to be a valuable tool in clinical practice is compelling, particularly in improving clinical decision support by helping physicians to make clinical decisions based on the best medical knowledge available. We aim to investigate ChatGPT's ability to identify, diagnose and manage patients with otorhinolaryngology-related symptoms. **Methods:** A prospective, cross-sectional study was designed based on an idea suggested by ChatGPT to assess the level of agreement between ChatGPT and five otorhinolaryngologists (ENTs) in 20 reality-inspired clinical cases. The clinical cases were presented to the chatbot on two different occasions (ChatGPT-1 and ChatGPT-2) to assess its temporal stability. **Results:** The mean score of ChatGPT-1 was 4.4 (SD 1.2; min 1, max 5) and of ChatGPT-2 was 4.15 (SD 1.3; min 1, max 5), while the ENTs mean score

				was 4.91 (SD 0.3; min 3, max 5). The Mann–Whitney U test revealed a statistically significant difference (p 〈 0.001) between both ChatGPT's and the ENTs's score. ChatGPT-1 and ChatGPT-2 gave different answers in five occasions. **Conclusions:** Artificial intelligence will be an important instrument in clinical decision-making in the near future and ChatGPT is the most promising chatbot so far. Despite needing further development to be used with safety, there is room for improvement and potential to aid otorhinolaryngology residents and specialists in making the most correct decision for the patient.
67	(Das et al. 2024)	Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer	BioGPT3 ChatGPT-3.5-turbo	수십억 개의 매개변수를 가진 대규모 언어 모델(LLM)을 데이터셋에 훈련시키고 공개 접근을 위해 모델을 발표하는 것이 현재 표준 관행이다. 자연어 처리에 대한 혁신적인 영향에도 불구하고, 공개 LLM은 훈련 데이터의 출처가 종종 웹 기반이거나 크라우드소싱되어 가해자에 의해 조작될 수 있다는 점에서 주목할 만한 취약점을 나타낸다. 우리는 공개적으로 이용 가능한 생의학 문헌과 MIMIC-III의 임상 기록에 훈련된 BioGPT를 포함하여 임상 LLM의 데이터 오염 공격 영역에서의 취약성을 탐구한다. 비식별화된 유방암 임상 기록에 대한 데이터 오염 기반 공격의 취약성을 탐색하면서, 우리의 접근법은 이러한 공격의 범위를 평가하는 최초의 시도이며, 우리의 발견은 LLM 출력의 성공적인 조작을 밝혀낸다. 이 작업을 통해 우리는 LLM의 이러한 취약성을 이해하는 긴급성을 강조하고, 임상 영역에서 LLM의 신중하고 책임 있는 사용을 장려한다. 본 연구에서 우리는 공개적으로 이용 가능한 임상 대규모 언어 모델에 대한 데이터 오염 및 표적 모델 편집 공격의 효과를 입증했다. 모델 편집 공격의 경우 공격자가 모델 아키텍처와 사전 훈련된 가중치에 접근할 수 있다고 가정하지만, **데이터 오염 공격은 LLM을 훈련시키기 위해 공개 영역에 일부 노이즈가 있는 공격 데이터를 도입하는 것만으로 수행**될 수 있다. **하위 질의응답 작업을 사용한 우리의 실증적 평가는 오염된 모델이 깨끗한 모델과 유사한 고품질 응답을 생성하며, 따라서 표준 정량적 지표를 사용하여 구별하기가 극도로 어렵다**는 것을 보여준다. 우리는 **유방암 영역에 대해서만 LLM 취약성을 입증**했지만, 유사한 파이프라인을 다른 전문 분야나 영역에도 적용할 수 있다. 본 연구의 범위 내에서 우리는 임상 데이터에 대한 LLM 훈련과 관련된 또 다른 주요 우려사항인 개인정보 보호 위험을 탐색하지 않았다. Training Large Language Models (LLMs) with billions of parameters on a dataset and publishing the model for public access is the standard practice currently. Despite their transformative impact on natural language processing, public LLMs present notable vulnerabilities given the source of

				training data is often web-based or crowdsourced, and hence can be manipulated by perpetrators. We delve into the vulnerabilities of clinical LLMs, particularly BioGPT which is trained on publicly available biomedical literature and clinical notes from MIMIC-III, in the realm of data poisoning attacks. Exploring susceptibility to data poisoning-based attacks on de-identified breast cancer clinical notes, our approach is the first one to assess the extent of such attacks and our findings reveal successful manipulation of LLM outputs. Through this work, we emphasize on the urgency of comprehending these vulnerabilities in LLMs, and encourage the mindful and responsible usage of LLMs in the clinical domain. In this work, we demonstrated effectiveness of data poisoning and targeted model editing attacks on a publicly available clinical large language model. While for the model editing attack, we assume that attacker has access to model architecture and pre-trained weights, data poisoning attack can be performed by simply introducing some noisy attack data in the public domain to train the LLMs. Our empirical evaluation using a downstream question-answering task shows that the poisoned models generate high quality responses similar to the clean model, and thus extremely difficult to distinguish using standard quantitative metrics. We demonstrated the LLM vulnerabilities only for the breast cancer domain but a similar pipeline can also be applied to any other specialities or domain. Within the scope of this work, we did not explore the privacy risk that is also another major concern related to training the LLMs on clinical data.
68	(Zaretsky et al. 2024)	Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format	ChatGPT-4	**중요성:** 법적으로 환자는 의료기록의 퇴원 요약서에 즉시 접근할 수 있다. 전문 용어와 약어는 일반 환자가 요약서를 읽고 이해하기 어렵게 만든다. 대규모 언어 모델(LLM [예: GPT-4])은 이러한 요약서를 환자 친화적인 언어와 형식으로 변환할 잠재력을 가지고 있다. **목적:** LLM이 퇴원 요약서를 더 읽기 쉽고 이해하기 쉬운 형식으로 변환할 수 있는지 판단한다. **설계, 설정 및 참가자:** 이 횡단면 연구는 2023년 6월 1일부터 30일까지 NYU(뉴욕 대학교) Langone Health의 일반내과 진료에서 퇴원한 성인 환자의 퇴원 요약서 샘플을 평가했다. 사망으로 퇴원한 환자는 제외되었다. 모든 퇴원 요약서는 2023년 7월 26일부터 8월 5일 사이에 LLM에 의해 처리되었다. **개입:** 건강보험 이동성 및 책임에 관한 법률(HIPAA) 준수 보안 플랫폼인 Microsoft Azure OpenAI를 사용하여 2023년 7월 26일부터 8월 5일 사이에 이러한 퇴원 요약서를 환자 친화적 형식으로 변환했다.

				주요 결과 및 측정: 결과는 Flesch-Kincaid Grade Level로 측정된 가독성과 Patient Education Materials Assessment Tool(PEMAT) 점수를 사용한 이해 가능성을 포함했다. 원본 퇴원 요약서의 가독성과 이해 가능성을 LLM을 통해 생성된 변환된 환자 친화적 퇴원 요약서와 비교했다. 균형 지표로 환자 친화적 버전의 정확성과 완전성을 측정했다. **결과:** 50명의 환자(여성 31명 [62.0%], 남성 19명 [38.0%])의 퇴원 요약서가 포함되었다. 환자의 중앙 연령은 65.5세(IQR, 59.0-77.5)였다. 평균(SD) Flesch-Kincaid Grade Level은 환자 친화적 퇴원 요약서에서 유의하게 낮았다(6.2 [0.5] vs 11.0 [1.5]; P 〈.001). PEMAT 이해 가능성 점수는 환자 친화적 퇴원 요약서에서 유의하게 높았다(81% vs 13%; P 〈.001). 2명의 의사가 각 환자 친화적 퇴원 요약서를 6점 척도로 정확성을 검토했으며, 100개 검토 중 54개(54.0%)가 최고 등급인 6점을 부여했다. **요약서는 56개 검토(56.0%)에서 완전히 완성된 것으로 평가되었다. 18개 검토에서 안전 우려가 지적되었으며, 주로 누락과 관련이 있었지만 여러 부정확한 진술(환각이라고 함)도 있었다.** **결론 및 관련성:** 50개 퇴원 요약서에 대한 이 횡단면 연구의 발견은 LLM이 퇴원 요약서를 전자건강기록에 나타나는 것보다 훨씬 더 읽기 쉽고 이해하기 쉬운 환자 친화적 언어와 형식으로 변환하는 데 사용될 수 있음을 시사한다. 그러나 구현에는 정확성, 완전성 및 안전성의 개선이 필요할 것이다. 안전 우려를 고려할 때 초기 구현에는 의사 검토가 필요할 것이다. **Importance:** By law, patients have immediate access to discharge notes in their medical records. Technical language and abbreviations make notes difficult to read and understand for a typical patient. Large language models (LLMs [eg, GPT-4]) have the potential to transform these notes into patient-friendly language and format. **Objective:** To determine whether an LLM can transform discharge summaries into a format that is more readable and understandable. **Design, Setting, and Participants:** This cross-sectional study evaluated a sample of the discharge summaries of adult patients discharged from the General Internal Medicine service at NYU (New York University) Langone Health from June 1 to 30, 2023. Patients discharged as deceased were excluded. All discharge summaries were processed by the LLM between July 26 and August 5, 2023. **Interventions:** A secure Health Insurance Portability and Accountability Act-compliant platform, Microsoft Azure OpenAI, was used to transform these discharge summaries into a patient-friendly format between July 26 and

				August 5, 2023. **Main Outcomes and Measures:** Outcomes included readability as measured by Flesch–Kincaid Grade Level and understandability using Patient Education Materials Assessment Tool (PEMAT) scores. Readability and understandability of the original discharge summaries were compared with the transformed, patient–friendly discharge summaries created through the LLM. As balancing metrics, accuracy and completeness of the patient–friendly version were measured. **Results:** Discharge summaries of 50 patients (31 female [62.0%] and 19 male [38.0%]) were included. The median patient age was 65.5 (IQR, 59.0–77.5) years. Mean (SD) Flesch–Kincaid Grade Level was significantly lower in the patient–friendly discharge summaries (6.2 [0.5] vs 11.0 [1.5]; P 〈.001). PEMAT understandability scores were significantly higher for patient–friendly discharge summaries (81% vs 13%; P 〈.001). Two physicians reviewed each patient–friendly discharge summary for accuracy on a 6–point scale, with 54 of 100 reviews (54.0%) giving the best possible rating of 6. Summaries were rated entirely complete in 56 reviews (56.0%). Eighteen reviews noted safety concerns, mostly involving omissions, but also several inaccurate statements (termed hallucinations). **Conclusions and Relevance:** The findings of this cross–sectional study of 50 discharge summaries suggest that LLMs can be used to translate discharge summaries into patient–friendly language and formats that are significantly more readable and understandable than discharge summaries as they appear in electronic health records. However, implementation will require improvements in accuracy, completeness, and safety. Given the safety concerns, initial implementation will require physician review.
69	(Ralla et al. 2025)	How Accurate Is AI? A Critical Evaluation of Commonly Used Large Language Models in Responding to Patient Concerns About Incidental Kidney Tumors	ChatGPT–4o Microsoft Copilot Google Gemini 2.5	**배경:** ChatGPT, Google Gemini, Microsoft Copilot과 같은 대규모 언어 모델(LLM)은 온라인에서 의료 정보를 찾는 환자들에 의해 점점 더 많이 사용되고 있다. 이러한 도구는 접근 가능하고 대화형 설명을 제공하지만, 우연한 암 진단과 같은 감정적으로 민감한 시나리오에서의 정확성과 안전성은 여전히 불확실하다. **목적:** 신장 종양의 우연한 발견 후 일반적인 환자 질문에 대해 세 가지 최첨단 LLM이 생성한 응답의 품질, 완전성, 가독성 및 안전성을 평가한다. **방법:** 표준화된 사용 사례 시나리오를 개발했다: 환자가 요통으로 인한 컴퓨터 단층촬영(CT) 후 의심스러운 신장 종괴를 알게 된다. 일반적인 환자 우려를 반영하는 10개의 평이한 언어 프롬프트를 추가 맥락 없이 ChatGPT–4o, Microsoft Copilot, Google Gemini 2.5 Pro에 제출했다. 응답은 검증된 6개 영역 루브릭(정확성, 완전

				성, 명확성, 최신성, 위해 위험, 환각)을 사용하여 5명의 전문의 자격을 갖춘 비뇨기과 전문의에 의해 독립적으로 평가되었으며, 1-5 리커트 척도로 점수가 매겨졌다. 기술 점수와 평가자 간 신뢰도(Fleiss' Kappa)를 계산하기 위해 두 가지 통계적 접근법이 적용되었다. 가독성은 Flesch Reading Ease(FRE) 및 Flesch-Kincaid Grade Level(FKGL) 지표를 사용하여 분석되었다. **결과:** Google Gemini 2.5 Pro는 대부분의 영역에서 가장 높은 평균 평가를 달성했으며, 특히 **정확성(4.3), 완전성(4.3) 및 낮은 환각률(4.6)**에서 두드러졌다. Microsoft Copilot은 공감적인 언어와 일관된 면책조항으로 주목받았지만 **명확성과 최신성 점수가 약간 낮았다**. ChatGPT-4o는 대화 흐름에서 강점을 보였지만 **임상 정밀도에서 더 많은 변동성**을 나타냈다. 전반적으로 응답의 14%가 잠재적으로 오해의 소지가 있거나 불완전한 것으로 표시되었다. 평가자 간 일치도는 모든 영역에서 상당했다(κ = 0.68). **가독성은 모델 간에 차이가 있었다: ChatGPT 응답이 가장 이해하기 쉬웠고(FRE = 48.5; FKGL = 11.94), Gemini가 가장 복잡했다(FRE = 29.9; FKGL = 13.3)**. **결론:** **LLM은 환자 대면 의사소통에서 가능성을 보여주지만 우연한 암 진단과 같은 고위험 맥락에서 일관되게 정확하고 완전하며 지침에 부합하는 정보를 제공하는 데는 현재 부족**하다. 그들의 어조와 구조가 환자 참여를 지원할 수 있지만, 상담을 위해 자율적으로 사용되어서는 안 된다. 환자 치료에 안전하게 통합하기 위해서는 추가적인 미세 조정, 임상 검증 및 감독이 필수적이다. **Background:** Large language models (LLMs) such as ChatGPT, Google Gemini, and Microsoft Copilot are increasingly used by patients seeking medical information online. While these tools provide accessible and conversational explanations, their accuracy and safety in emotionally sensitive scenarios—such as an incidental cancer diagnosis—remain uncertain. **Objective:** To evaluate the quality, completeness, readability, and safety of responses generated by three state-of-the-art LLMs to common patient questions following the incidental discovery of a kidney tumor. **Methods:** A standardized use-case scenario was developed: a patient learns of a suspicious renal mass following a computed tomography (CT) scan for back pain. Ten plain-language prompts reflecting typical patient concerns were submitted to ChatGPT-4o, Microsoft Copilot, and Google Gemini 2.5 Pro without additional context. Responses were independently assessed by five board-certified urologists using a validated six-domain rubric (accuracy, completeness, clarity, currency, risk of harm, hallucinations), scored on a 1–

				5 Likert scale. Two statistical approaches were applied to calculate descriptive scores and inter-rater reliability (Fleiss' Kappa). Readability was analyzed using the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) metrics. **Results:** Google Gemini 2.5 Pro achieved the highest mean ratings across most domains, notably in accuracy (4.3), completeness (4.3), and low hallucination rate (4.6). Microsoft Copilot was noted for empathetic language and consistent disclaimers but showed slightly lower clarity and currency scores. ChatGPT-4o demonstrated strengths in conversational flow but displayed more variability in clinical precision. Overall, 14% of responses were flagged as potentially misleading or incomplete. Inter-rater agreement was substantial across all domains (κ = 0.68). Readability varied between models: ChatGPT responses were easiest to understand (FRE = 48.5; FKGL = 11.94), while Gemini's were the most complex (FRE = 29.9; FKGL = 13.3). **Conclusions:** LLMs show promise in patient-facing communication but currently fall short of providing consistently accurate, complete, and guideline-conform information in high-stakes contexts such as incidental cancer diagnoses. While their tone and structure may support patient engagement, they should not be used autonomously for counseling. Further fine-tuning, clinical validation, and supervision are essential for safe integration into patient care.
70	(Zaboli et al. 2024)	Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage?	Chat-GPT 3.5	**서론:** Chat-GPT는 의학 분야에서 유망하고 잠재적으로 혁명적인 도구로 빠르게 부상하고 있다. 가능한 응용 분야 중 하나는 응급실(ED)의 중증도 분류 평가 중 임상 상태의 심각도와 예후에 따른 환자 계층화이다. **방법:** 실제 임상 사례에서 재현된 30개의 무작위로 선택된 사례를 사용하여, 의료진과 Chat-GPT 간의 응급실 환자 위험 계층화의 일치도를 비교했다. 일치도는 Cohen's kappa로 평가되었고, 성능은 수신자 조작 특성 곡선 하 면적(AUROC) 곡선으로 평가되었다. 결과 중에서 우리는 72시간 내 사망률, 입원 필요성, 심각하거나 시간 의존적인 상태의 존재를 고려했다. **결과:** 중증도 분류 간호사와 Chat-GPT 간의 중증도 코드 할당의 일치도는 0.278(비가중 Cohen's kappa; 95% 신뢰구간: 0.231-0.388)이었다. 모든 결과에 대해 ROC 값은 중증도 분류 간호사가 더 높았다. **가장 관련성이 높은 차이는 72시간 사망률에서 발견되었으며, 중증도 분류 간호사는 0.910(0.757-1.000)의 AUROC를 보인 반면 Chat-GPT는 단 0.669(0.153-1.000)를 보였다.** **결론: 현재 Chat-GPT의 신뢰성 수준은 응급실 환자 우선순위 결정에서 중증도 분류 간호사의 전문성을 대체할 수 있는 유효한 대안이 되기에는 불충분**하다. 응급실

				환자의 위험 계층화를 위한 AI의 안전성과 효과성을 향상시키기 위해서는 추가 개발이 필요하다. **Introduction:** Chat-GPT is rapidly emerging as a promising and potentially revolutionary tool in medicine. One of its possible applications is the stratification of patients according to the severity of clinical conditions and prognosis during the triage evaluation in the emergency department (ED). **Methods:** Using a randomly selected sample of 30 vignettes recreated from real clinical cases, we compared the concordance in risk stratification of ED patients between healthcare personnel and Chat-GPT. The concordance was assessed with Cohen's kappa, and the performance was evaluated with the area under the receiver operating characteristic curve (AUROC) curves. Among the outcomes, we considered mortality within 72 h, the need for hospitalization, and the presence of a severe or time-dependent condition. **Results:** The concordance in triage code assignment between triage nurses and Chat-GPT was 0.278 (unweighted Cohen's kappa; 95% confidence intervals: 0.231–0.388). For all outcomes, the ROC values were higher for the triage nurses. The most relevant difference was found in 72-h mortality, where triage nurses showed an AUROC of 0.910 (0.757–1.000) compared to only 0.669 (0.153–1.000) for Chat-GPT. **Conclusions:** The current level of Chat-GPT reliability is insufficient to make it a valid substitute for the expertise of triage nurses in prioritizing ED patients. Further developments are required to enhance the safety and effectiveness of AI for risk stratification of ED patients.
71	(Hu et al. 2024)	Improving large language models for clinical named entity recognition via prompt engineering	GPT-3.5 GPT-4 BioClinicalBERT	**중요성:** 본 연구는 대규모 언어 모델, 특히 GPT-3.5 및 GPT-4가 복잡한 임상 데이터를 처리하고 최소한의 훈련 데이터로 의미 있는 정보를 추출하는 잠재력을 강조한다. 프롬프트 기반 전략을 개발하고 개선함으로써 모델의 성능을 크게 향상시킬 수 있으며, 임상 개체명 인식(NER) 작업을 위한 실행 가능한 도구로 만들고 광범위한 주석 데이터셋에 대한 의존도를 줄일 수 있다. **목적:** 본 연구는 임상 개체명 인식(NER) 작업에 대한 GPT-3.5 및 GPT-4의 능력을 정량화하고 성능을 향상시키기 위한 작업별 프롬프트를 제안한다. **재료 및 방법:** 우리는 두 가지 임상 NER 작업에서 이러한 모델을 평가했다: (1) 2010 i2b2 개념 추출 공유 작업에 따라 MTSamples 말뭉치의 임상 기록에서 의학적 문제, 치료 및 검사를 추출하고, (2) 백신 부작용 보고 시스템(VAERS)의 안전 보고서에서 신경계 장애 관련 부작용을 식별한다. GPT 모델의 성능을 향상시키기 위해 우리는 (1) 작업 설명 및 형식 사양을 포함하는 기본 프롬프트, (2) 주석 지침

				기반 프롬프트, (3) 오류 분석 기반 지침, (4) 퓨샷 학습을 위한 주석 샘플을 포함하는 임상 작업별 프롬프트 프레임워크를 개발했다. 우리는 각 프롬프트의 효과를 평가하고 모델을 BioClinicalBERT와 비교했다. **결과:** 기본 프롬프트를 사용하여 GPT-3.5 및 GPT-4는 MTSamples에서 0.634, 0.804, VAERS에서 0.301, 0.593의 완화된 F1 점수를 달성했다. 추가 프롬프트 구성요소는 모델 성능을 일관되게 향상시켰다. 4개 구성요소를 모두 사용했을 때 GPT-3.5 및 GPT-4는 MTSamples에서 0.794, 0.861, VAERS에서 0.676, 0.736의 완화된 F1 점수를 달성하여 프롬프트 프레임워크의 효과를 입증했다. 이러한 결과는 BioClinicalBERT(MTSamples 데이터셋에서 F1 0.901, VAERS에서 0.802)에 뒤처지지만, 소수의 훈련 샘플만 필요하다는 점을 고려하면 매우 유망하다. **고찰:** 연구 결과는 임상 NER 작업에 LLM을 활용하는 유망한 방향을 시사한다. 그러나 **GPT 모델의 성능이 작업별 프롬프트로 향상되었지만, 추가 개발 및 개선이 필요하다. GPT-4와 같은 LLM은 BioClinicalBERT와 같은 최첨단 모델에 근접한 성능을 달성할 잠재력을 보여주지만, 여전히 신중한 프롬프트 엔지니어링과 작업별 지식 이해가 필요하다. 본 연구는 또한 임상 환경에서 LLM의 능력과 성능을 정확하게 반영하는 평가 스키마의 중요성을 강**조한다. **결론:** GPT 모델을 임상 NER 작업에 직접 적용하는 것은 최적의 성능에 미치지 못하지만, 의학 지식과 훈련 샘플을 통합한 작업별 프롬프트 프레임워크는 잠재적 임상 응용을 위한 GPT 모델의 실행 가능성을 크게 향상시킨다. **Importance:** The study highlights the potential of large language models, specifically GPT-3.5 and GPT-4, in processing complex clinical data and extracting meaningful information with minimal training data. By developing and refining prompt-based strategies, we can significantly enhance the models' performance, making them viable tools for clinical NER tasks and possibly reducing the reliance on extensive annotated datasets. **Objectives:** This study quantifies the capabilities of GPT-3.5 and GPT-4 for clinical named entity recognition (NER) tasks and proposes task-specific prompts to improve their performance. **Materials and Methods:** We evaluated these models on 2 clinical NER tasks: (1) to extract medical problems, treatments, and tests from clinical notes in the MTSamples corpus, following the 2010 i2b2 concept extraction shared task, and (2) to identify nervous system disorder-related adverse events from safety reports in the vaccine adverse event reporting system (VAERS). To improve the GPT models' performance, we developed a clinical task-

				specific prompt framework that includes (1) baseline prompts with task description and format specification, (2) annotation guideline-based prompts, (3) error analysis-based instructions, and (4) annotated samples for few-shot learning. We assessed each prompt's effectiveness and compared the models to BioClinicalBERT. **Results:** Using baseline prompts, GPT-3.5 and GPT-4 achieved relaxed F1 scores of 0.634, 0.804 for MTSamples and 0.301, 0.593 for VAERS. Additional prompt components consistently improved model performance. When all 4 components were used, GPT-3.5 and GPT-4 achieved relaxed F1 socres of 0.794, 0.861 for MTSamples and 0.676, 0.736 for VAERS, demonstrating the effectiveness of our prompt framework. Although these results trail BioClinicalBERT (F1 of 0.901 for the MTSamples dataset and 0.802 for the VAERS), it is very promising considering few training samples are needed. **Discussion:** The study's findings suggest a promising direction in leveraging LLMs for clinical NER tasks. However, while the performance of GPT models improved with task-specific prompts, there's a need for further development and refinement. LLMs like GPT-4 show potential in achieving close performance to state-of-the-art models like BioClinicalBERT, but they still require careful prompt engineering and understanding of task-specific knowledge. The study also underscores the importance of evaluation schemas that accurately reflect the capabilities and performance of LLMs in clinical settings. Conclusion: While direct application of GPT models to clinical NER tasks falls short of optimal performance, our task-specific prompt framework, incorporating medical knowledge and training samples, significantly enhances GPT models' feasibility for potential clinical applications.
72	(Haze et al. 2023)	Influence on the accuracy in ChatGPT: Differences in the amount of information per medical field	ChatGPT-3.5 ChatGPT-4	**목적:** ChatGPT는 의료용으로 개발되지 않았지만, 의료 분야에서의 사용에 대한 관심이 증가하고 있다. 의료 분야에서의 능력과 사용 시 주의사항을 이해하는 것은 시급한 문제이다. 우리는 다양한 의료 분야에 발표된 정보의 양의 차이가 해당 분야에서 ChatGPT가 받는 훈련의 양에 비례할 것이며, 따라서 답변 제공의 정확성에도 비례할 것이라고 가정했다. **연구 설계:** 비임상 실험 연구. **방법:** 우리는 GPT-3.5 및 GPT-4에 일본 국가 의사 시험을 실시하여 응답의 정확도와 일관성 비율을 조사했다. 의료 분야별 Web of Science Core Collection의 총 문서 수를 계산하고 ChatGPT의 정확도와의 관계를 평가했다. 또한 부정확한 답변

				의 위험 요인을 조사하기 위해 다변량 보정 모델을 수행했다. **결과:** GPT-4의 경우 시험에서 81.0%의 정확도와 88.8%의 일관성 비율을 확인했으며, 두 지표 모두 GPT-3.5에 비해 개선을 보였다. 정확도와 일관성 비율 간에 양의 상관관계가 관찰되었다(R = 0.51, P 〈 0.001). 의료 분야별 문서 수는 해당 의료 분야의 정확도와 유의하게 상관관계가 있었으며(R = 0.44, P 〈 0.05), **상대적으로 적은 출판물 수가 부정확한 답변의 독립적 위험 요인**이었다. **결론:** ChatGPT 사용 시 일관성을 확인하는 것이 부정확한 답변을 식별하는 데 도움이 될 수 있다. **사용자는 ChatGPT에게 신약 및 질병과 같이 발표된 정보가 제한적인 주제에 대해 질문할 때 답변의 정확도가 감소할 수 있음을 인식해야 한다**. **Objectives:** Although ChatGPT was not developed for medical use, there is growing interest in its use in medical fields. Understanding its capabilities and precautions for its use in the medical field is an urgent matter. We hypothesized that differences in the amounts of information published in different medical fields would be proportionate to the amounts of training ChatGPT receives in those fields, and hence its accuracy in providing answers. **Study design:** A non-clinical experimental study. **Methods:** We administered the Japanese National Medical Examination to GPT-3.5 and GPT-4 to examine the rates of accuracy and consistency in their responses. We counted the total number of documents in the Web of Science Core Collection per medical field and assessed the relationship with ChatGPT's accuracy. We also performed multivariate-adjusted models to investigate the risk factors for incorrect answers. **Results:** For GPT-4, we confirmed an accuracy rate of 81.0 % and a consistency rate of 88.8 % on the exam; both showed improvement compared to those for GPT-3.5. A positive correlation was observed between the accuracy rate and consistency rate (R = 0.51, P 〈 0.001). The number of documents per medical field was significantly correlated with the accuracy rate in that medical field (R = 0.44, P 〈 0.05), with relatively few publications being an independent risk factor for incorrect answers. **Conclusions:** Checking consistency may help identify incorrect answers when using ChatGPT. Users should be aware that the accuracy of the answers by ChatGPT may decrease when it is asked about topics with limited published information, such as new drugs and diseases.
73		Is ChatGPT-4 a Reliable Tool in	ChatGPT-4	**서론:** 인공지능 기반 챗봇은 자가면역 간염 환자에게 개인화된 상담을 제공할 잠재

		Autoimmune Hepatitis?		적 경로를 제공한다. 우리는 자가면역 간염 환자 4명이 제시한 40개 질문 풀 중 12개 질문에 대한 Chat Generative Pretrained Transformer-4 응답의 정확성, 완전성, 포괄성 및 안전성을 평가했다. **방법:** 질문은 진단(1–3), 삶의 질(4–8), 의학적 치료(9–12)의 3개 영역으로 분류되었다. 11명의 주요 오피니언 리더가 정확성에 대해 6점, 안전성에 대해 5점, 완전성과 포괄성에 대해 3점의 리커트 척도를 사용하여 응답을 평가했다. **결과:** **정확성, 완전성, 포괄성 및 안전성의 중앙값 점수는 각각 5(4–6), 2(2–2), 3(2–3)이었으며, 어느 영역도 우수한 평가를 나타내지 않았다**. 진단 후 추적관찰 질문은 낮은 정확성과 완전성을 가진 가장 까다로운 질문이었지만, 안전하고 포괄적인 특징을 가졌다. **주요 오피니언 리더 간 일치도(Fleiss Kappa 통계)는 정확성에 대해서는 경미했지만(0.05), 나머지 특징에 대해서는 낮았다(각각 −0.05, −0.06, −0.02)**. **고찰:** **챗봇은 양호한 이해 가능성을 보여주지만, 신뢰성이 부족하다**. Chat Generative Pretrained Transformer를 임상 실무에 통합하기 위해서는 추가 연구가 필요하다. **INTRODUCTION:** Artificial intelligence–based chatbots offer a potential avenue for delivering personalized counseling to patients with autoimmune hepatitis. We assessed accuracy, completeness, comprehensiveness, and safety of Chat Generative Pretrained Transformer–4 responses to 12 inquiries out of a pool of 40 questions posed by 4 patients with autoimmune hepatitis. **METHODS:** Questions were categorized into 3 areas: diagnosis (1–3), quality of life (4–8), and medical treatment (9–12). 11 key opinion leaders evaluated responses using a Likert scale with 6 points for accuracy, 5 points for safety, and 3 points for completeness and comprehensiveness. **RESULTS:** Median scores for accuracy, completeness, comprehensiveness, and safety were 5 (4–6), 2 (2–2), and 3 (2–3), respectively; no domain exhibited superior evaluation. Postdiagnosis follow-up question was the trickiest with low accuracy and completeness, but safe and comprehensive features. Agreement among key opinion leaders (Fleiss Kappa statistics) was slight for the accuracy (0.05) but poor for the remaining features (20.05, 20.06, and 20.02, respectively). **DISCUSSION:** Chatbots show good comprehensibility, but lack reliability. Further studies are needed to integrate Chat Generative Pretrained Transformer within clinical practice.

74	(Birkun and Gautam 2023)	Large Language Model (LLM)-Powered Chatbots Fail to Generate Guideline-Consistent Content on Resuscitation and May Provide Potentially Harmful Advice	New Bing Bard	**서론:** 최근 매우 인기 있는 혁신적인 대규모 언어 모델(LLM) 기반 챗봇은 일반 대중을 위한 소생술 정보의 잠재적 출처를 나타낸다. 예를 들어, 챗봇이 생성한 조언은 지역사회 소생술 교육 목적이나 실제 응급 상황에서 훈련받지 않은 일반 구조자의 적시 정보 지원을 위해 사용될 수 있다. **연구 목적:** 본 연구는 두 가지 주요 LLM 기반 챗봇의 성능을 평가하는 데 초점을 맞추었으며, 특히 호흡하지 않는 피해자를 돕는 방법에 대한 챗봇 생성 조언의 품질 측면에서 평가했다. **방법:** 2023년 5월, 새로운 Bing(Microsoft Corporation, USA) 및 Bard(Google LLC, USA) 챗봇에 각각 20회씩 질문했다: "누군가 숨을 쉬지 않으면 어떻게 해야 하나요?" 챗봇의 응답 내용은 사전 개발된 체크리스트를 사용하여 2021년 영국 소생술 협회 지침 준수 여부를 평가했다. **결과:** 두 챗봇 모두 질문에 대해 맥락 의존적 텍스트 응답을 제공했다. 그러나 **응답 내에서 호흡하지 않는 피해자를 돕는 지침과 일치하는 내용의 포함은 낮았다**: 체크리스트 기준을 완전히 충족하는 응답의 평균 비율은 Bing의 경우 9.5%, Bard의 경우 11.4%였다(P 〉.05). **적절한 깊이, 속도 및 흉부 반동을 가진 흉부 압박의 조기 시작 및 중단 없는 수행, 자동 심장 충격기(AED)의 요청 및 사용을 포함한 방관자 행동의 필수 요소가 일반적으로 누락**되었다. 더욱이 **Bard 응답의 55.0%에는 그럴듯하게 들리지만 무의미한 안내인 인공 환각이 포함되어 있어 부적절한 치료 및 피해자에 대한 위해 위험을 생성**했다. **결론: LLM 기반 챗봇의 호흡하지 않는 피해자를 돕는 조언은 소생술 기법의 필수 세부사항을 누락하고 때때로 기만적이고 잠재적으로 해로운 지시를 포함**한다. 소생술에 대한 챗봇 생성 오정보와 관련된 위험을 완화하기 위해서는 추가 연구와 규제 조치가 필요하다. **INTRODUCTION:** Innovative large language model (LLM)-powered chatbots, which are extremely popular nowadays, represent potential sources of information on resuscitation for the general public. For instance, the chatbot-generated advice could be used for purposes of community resuscitation education or for just-in-time informational support of untrained lay rescuers in a real-life emergency. **STUDY OBJECTIVE:** This study focused on assessing performance of two prominent LLM-based chatbots, particularly in terms of quality of the chatbot-generated advice on how to give help to a non-breathing victim. **METHODS:** In May 2023, the new Bing (Microsoft Corporation, USA) and Bard (Google LLC, USA) chatbots were inquired (n = 20 each): What to do if someone is not breathing?" Content of the chatbots' responses was

				evaluated for compliance with the 2021 Resuscitation Council United Kingdom guidelines using a pre-developed checklist. **RESULTS:** Both chatbots provided context-dependent textual responses to the query. However, coverage of the guideline-consistent instructions on help to a non-breathing victim within the responses was poor: mean percentage of the responses completely satisfying the checklist criteria was 9.5% for Bing and 11.4% for Bard (P 〉.05). Essential elements of the bystander action, including early start and uninterrupted performance of chest compressions with adequate depth, rate, and chest recoil, as well as request for and use of an automated external defibrillator (AED), were missing as a rule. Moreover, 55.0% of Bard's responses contained plausible sounding, but nonsensical guidance, called artificial hallucinations, that create risk for inadequate care and harm to a victim. **CONCLUSION:** The LLM-powered chatbots' advice on help to a non-breathing victim omits essential details of resuscitation technique and occasionally contains deceptive, potentially harmful directives. Further research and regulatory measures are required to mitigate risks related to the chatbot-generated misinformation of public on resuscitation."
75	(Chung et al. 2024)	Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication	GPT-4 Turbo	**중요성** 범용 대규모 언어 모델은 절차 설명과 환자의 전자건강기록 기록을 사용하여 위험 계층화를 수행하고 수술 후 결과 측정값을 예측할 수 있을 것이다. **목적** 8가지 다른 작업에 대한 예측 성능을 검토한다: 미국마취과학회 신체 상태 분류(ASA-PS), 입원, 중환자실(ICU) 입원, 계획되지 않은 입원, 병원 사망률, 마취후 회복실(PACU) 1단계 기간, 입원 기간, ICU 기간 예측. **설계, 설정 및 참가자** 이 예후 연구는 일상적인 임상 진료 중 수집된 2년간의 후향적 전자건강기록 데이터로부터 구성된 작업별 데이터셋을 포함했다. 사례 및 기록 데이터는 프롬프트로 형식화되어 대규모 언어 모델 GPT-4 Turbo(OpenAI)에 제공되어 예측 및 설명을 생성했다. 설정에는 단일 대도시 지역의 3개 학술 병원 및 제휴 클리닉으로 구성된 4차 진료 센터가 포함되었다. 마취를 동반한 수술 또는 시술을 받았고 수술 전 전자건강기록에 최소 1개의 임상의 작성 기록이 제출된 환자가 연구에 포함되었다. 데이터는 2023년 11월부터 12월까지 분석되었다. **노출** 원본 기록, 기록 요약, 퓨샷 프롬프팅 및 사고 연쇄 프롬프팅 전략을 비교했다. **주요 결과 및 측정** 이진 및 범주형 결과에 대한 F1 점수. 수치 기간 결과에 대한 평균 절대 오차. **결과** 연구 결과는 작업별 데이터셋에서 측정되었으며, 각 데이터셋은 1000개 사례를 포함했으나, 계획되지 않은 입원은 949개 사례, 병원 사망률은 576개 사례였다. 각 작업의 최상 결과는 ASA-PS에 대해 F1 점수 0.50(95% CI, 0.47-0.53), 입원

				에 대해 0.64(95% CI, 0.61–0.67), ICU 입원에 대해 0.81(95% CI, 0.78–0.83), 계획되지 않은 입원에 대해 0.61(95% CI, 0.58–0.64), 병원 사망률 예측에 대해 0.86(95% CI, 0.83–0.89)을 포함했다. **기간 예측 작업의 성능은 모든 프롬프트 전략에서 보편적으로 낮았으며, 대규모 언어 모델은 PACU 1단계 기간에 대해 평균 절대 오차 49분(95% CI, 46–51분), 입원 기간에 대해 4.5일(95% CI, 4.2–5.0일), ICU 기간 예측에 대해 1.1일(95% CI, 0.9–1.3일)을 달성**했다. **결론 및 관련성 현재 범용 대규모 언어 모델은 분류 작업의 수술 전후 위험 계층화에서 임상의를 지원할 수 있지만, 수치 기간 예측에는 부적절**하다. 예측에 대한 고품질 자연어 설명을 생성하는 능력은 임상 워크플로우에서 유용한 도구가 될 수 있으며, 전통적인 위험 예측 모델을 보완할 수 있다. **IMPORTANCE** General-domain large language models may be able to perform risk stratification and predict postoperative outcome measures using a description of the procedure and a patient's electronic health record notes. **OBJECTIVE** To examine predictive performance on 8 different tasks: prediction of American Society of Anesthesiologists Physical Status (ASA-PS), hospital admission, intensive care unit (ICU) admission, unplanned admission, hospital mortality, postanesthesia care unit (PACU) phase 1 duration, hospital duration, and ICU duration. **DESIGN, SETTING, AND PARTICIPANTS** This prognostic study included task-specific datasets constructed from 2 years of retrospective electronic health records data collected during routine clinical care. Case and note data were formatted into prompts and given to the large language model GPT-4 Turbo (OpenAI) to generate a prediction and explanation. The setting included a quaternary care center comprising 3 academic hospitals and affiliated clinics in a single metropolitan area. Patients who had a surgery or procedure with anesthesia and at least 1 clinician-written note filed in the electronic health record before surgery were included in the study. Data were analyzed from November to December 2023. **EXPOSURES** Compared original notes, note summaries, few-shot prompting, and chain-of-thought prompting strategies. **MAIN OUTCOMES AND MEASURES** F1 score for binary and categorical outcomes. Mean absolute error for numerical duration outcomes. **RESULTS** Study results were measured on task-specific datasets, each with 1000 cases with the exception of unplanned admission, which had 949 cases, and hospital mortality, which had 576 cases. The best results for each

				task included an F1 score of 0.50 (95% CI, 0.47–0.53) for ASA-PS, 0.64 (95% CI, 0.61–0.67) for hospital admission, 0.81 (95% CI, 0.78–0.83) for ICU admission, 0.61 (95% CI, 0.58–0.64) for unplanned admission, and 0.86 (95% CI, 0.83–0.89) for hospital mortality prediction. Performance on duration prediction tasks was universally poor across all prompt strategies for which the large language model achieved a mean absolute error of 49 minutes (95% CI, 46–51 minutes) for PACU phase 1 duration, 4.5 days (95% CI, 4.2–5.0 days) for hospital duration, and 1.1 days (95% CI, 0.9–1.3 days) for ICU duration prediction. CONCLUSIONS AND RELEVANCE Current general-domain large language models may assist clinicians in perioperative risk stratification on classification tasks but are inadequate for numerical duration predictions. Their ability to produce high-quality natural language explanations for the predictions may make them useful tools in clinical workflows and may be complementary to traditional risk prediction models.
76	(Ulus, Ceker, and Hacibey 2025)	Large Language Models and Male Circumcision: A Reliability Assessment	ChatGPT Copilot Gemini Perplexity	**목적:** 남성 포경수술은 일부 국가에서 신생아 또는 종교적 이유로 일상적으로 시행되고 있지만, 건강상의 이점, 잠재적 위험 및 신체 자율성에 대한 영향에 관한 지속적인 논쟁의 대상이 되고 있다. 본 연구는 남성 포경수술의 다양한 측면에 대해 널리 사용되는 4개의 대규모 언어 모델(LLM)이 생성한 환자 대상 콘텐츠의 신뢰성을 평가하는 것을 목표로 한다. **방법:** 2025년 5월 10일에 20개의 표준화된 질문을 사용하여 LLM에 대한 검색을 수행했다. ChatGPT, Copilot, Gemini 및 Perplexity의 응답을 3명의 독립적인 전문가가 평가했다. 평가자 간 신뢰도는 급내상관계수로 평가되었고, 모델 성능 차이는 Bonferroni 보정을 사용한 Kruskal-Wallis 검정으로 분석되었다. **결과:** 평가자 간 신뢰도는 급내상관계수 0.79($p<0.001$)로 강했다. **Perplexity는 주제 영역에 걸쳐 평가될 때 ChatGPT, Copilot 및 Gemini에 비해 통계적으로 유의하게 낮은 성능을 보였다($p<0.001$). 마찬가지로 Perplexity는 명확성, 구조, 유용성 및 사실 정확성의 기준에서 다른 모델보다 통계적으로 유의하게 낮은 성능을 보였다($p<0.001$)**. **결론:** Gemini와 Copilot은 주제 영역과 평가 기준 모두에서 최고의 성과를 보였으며, 남성 포경수술에 관한 정확하고 잘 구조화된 의료 정보를 제공하는 LLM의 능력에서 상당한 차이를 강조했다. ChatGPT는 환자 안내에서 가능성을 보여주지만, **Perplexity와 같은 모델의 일관성 없는 성능은 의료 커뮤니케이션에서 신중한 구현과 지속적인 감독의 필요성을 강조**한다. **Aim:** Male circumcision remains routine in some countries for neonatal or

				religious reasons; however, it continues to be the subject of ongoing debate concerning its health benefits, potential risks, and implications for bodily autonomy. This study aims to evaluate the reliability of patient-facing content generated by four widely used large language models (LLMs) on various aspects of male circumcision. **Methods:** A search regarding LLMs was conducted using 20 standardized questions on 10 May 2025. Responses from ChatGPT, Copilot, Gemini, and Perplexity were evaluated by three independent experts. Inter-rater reliability was assessed with the intraclass correlation coefficient, and model performance differences were analyzed using Kruskal-Wallis tests with Bonferroni correction. **Results:** Inter-rater reliability was strong, with an intraclass correlation coefficient of 0.79 ($p<0.001$). Perplexity demonstrated statistically significant lower performance compared to ChatGPT, Copilot, and Gemini when evaluated across the thematic domains ($p<0.001$). Similarly, Perplexity performed statistically significantly worse than the other models across the criteria of clarity, structure, utility, and factual accuracy ($p<0.001$). **Conclusion:** Gemini and Copilot were the top performers across both thematic domains and evaluation criteria, highlighting substantial differences among LLMs in their ability to provide accurate and well-structured medical information regarding male circumcision. While ChatGPT shows promise for patient guidance, the inconsistent performance of models such as Perplexity highlights the need for cautious implementation and continuous oversight in healthcare communication.
77	(Omiye et al. 2023)	Large language models propagate race-based medicine	ChatGPT (May 12 and August 3 versions) ChatGPT-4 Google's Bard (May 18 and August 3 versions) Anthropic's Claude (May 15 and August 3 versions)	대규모 언어 모델(LLM)은 의료 시스템에 통합되고 있지만, 이러한 모델은 해로운 인종 기반 의학을 재현할 수 있다. 본 연구의 목적은 4개의 상용 대규모 언어 모델(LLM)이 인종 기반 의학이나 인종에 관한 널리 퍼진 오해를 확인하는 8가지 다른 시나리오에 응답할 때 해로운, 부정확한 인종 기반 콘텐츠를 전파하는지 평가하는 것이다. 질문은 4명의 의사 전문가 간의 논의와 의대 수련생들이 믿는 인종 기반 의학적 오해에 대한 이전 연구에서 도출되었다. 우리는 4개의 대규모 언어 모델을 9개의 다른 질문으로 평가했으며, 각 질문은 5회씩 조사되어 모델당 총 45개의 응답을 받았다. **모든 모델이 인종에 따라 의학적 판단을 다르게 하는 잘못된 관행을 계속 반복하는 답변을 제공했다.** **모델은 동일한 질문을 반복적으로 받았을 때 항상 일관된 응답을 하지 않았다**. LLM은 의료 환경에서 사용하도록 제안되고 있으며, 일부 모델은 이미 전자건강기록 시스템에 연결되고 있다. 그러나 본 연구는 우리의 발견에 기초할 때 이러한 LLM이 반박된 인종차별적 아이디어를 영속화하여 잠재적으로

				해를 끼칠 수 있음을 보여준다. Large language models (LLMs) are being integrated into healthcare systems; but these models may recapitulate harmful, race-based medicine. The objective of this study is to assess whether four commercially available large language models (LLMs) propagate harmful, inaccurate, race-based content when responding to eight different scenarios that check for race-based medicine or widespread misconceptions around race. Questions were derived from discussions among four physician experts and prior work on race-based medical misconceptions believed by medical trainees. We assessed four large language models with nine different questions that were interrogated five times each with a total of 45 responses per model. All models had examples of perpetuating race-based medicine in their responses. Models were not always consistent in their responses when asked the same question repeatedly. LLMs are being proposed for use in the healthcare setting, with some models already connecting to electronic health record systems. However, this study shows that based on our findings, these LLMs could potentially cause harm by perpetuating debunked, racist ideas.
78	(Kim et al. 2025)	Medical Hallucination in Foundation Models and Their Impact on Healthcare	ChatGPT o3-mini ChatGPT o1-preview ChatGPT-4o ChatGPT-4o-mini Gemini 2.0 Thinking Gemini 2.0 Flash Gemini 1.5 Flash PMC-LLaMA Alpaca Variants (AlphaCare-13B, MedAlpaca-13B)	다중모드 데이터를 처리하고 생성할 수 있는 기초 모델은 의학에서 AI의 역할을 변화시켰다. 그러나 신뢰성의 주요 한계는 환각으로, 부정확하거나 조작된 정보가 임상 결정과 환자 안전에 영향을 미칠 수 있다. 우리는 의료 환각을 모델이 오해의 소지가 있는 의료 콘텐츠를 생성하는 모든 경우로 정의한다. 본 논문은 의료 환각의 고유한 특성, 원인 및 영향을 검토하며, 특히 이러한 오류가 실제 임상 시나리오에서 어떻게 나타나는지에 초점을 맞춘다. 우리의 기여에는 (1) 의료 환각을 이해하고 해결하기 위한 분류 체계, (2) 의료 환각 데이터셋과 실제 의료 사례에 대한 의사 주석 LLM 응답을 사용한 모델 벤치마킹으로 환각의 임상적 영향에 대한 직접적인 통찰력 제공, (3) 의료 환각에 대한 경험에 관한 다국적 임상의 설문조사가 포함된다. 우리의 결과는 **사고 연쇄(CoT) 및 검색 증강 생성과 같은 추론 기법이 환각률을 효과적으로 줄일 수 있음을 보여준다. 그러나 이러한 개선에도 불구하고 무시할 수 없는 수준의 환각이 지속**된다. 이러한 발견은 강력한 탐지 및 완화 전략의 윤리적이고 실용적인 필요성을 강조하며, AI가 의료에 더 통합됨에 따라 환자 안전을 우선시하고 임상 무결성을 유지하는 규제 정책의 기초를 확립한다. 임상의의 피드백은 기술적 발전뿐만 아니라 환자 안전을 보장하기 위한 보다 명확한 윤리적 및 규제 지침의 긴급한 필요성을 강조한다. 논문 자료, 요약 및 추가 정보를 정리한 저장소는 https://github.com/mitmedialab/medical hallucination에서 확인할 수 있다.

				Foundation Models that are capable of processing and generating multi-modal data have transformed AI's role in medicine. However, a key limitation of their reliability is hallucination, where inaccurate or fabricated information can impact clinical decisions and patient safety. We define medical hallucination as any instance in which a model generates misleading medical content. This paper examines the unique characteristics, causes, and implications of medical hallucinations, with a particular focus on how these errors manifest themselves in real-world clinical scenarios. Our contributions include (1) a taxonomy for understanding and addressing medical hallucinations, (2) benchmarking models using medical hallucination dataset and physician-annotated LLM responses to real medical cases, providing direct insight into the clinical impact of hallucinations, and (3) a multi-national clinician survey on their experiences with medical hallucinations. Our results reveal that inference techniques such as Chain-of-Thought (CoT) and Search Augmented Generation can effectively reduce hallucination rates. However, despite these improvements, non-trivial levels of hallucination persist. These findings underscore the ethical and practical imperative for robust detection and mitigation strategies, establishing a foundation for regulatory policies that prioritize patient safety and maintain clinical integrity as AI becomes more integrated into healthcare. The feedback from clinicians highlights the urgent need for not only technical advances but also for clearer ethical and regulatory guidelines to ensure patient safety. A repository organizing the paper resources, summaries, and additional information is available at https://github.com/mitmedialab/medical hallucination.
79	(Han et al. 2024)	Medical large language models are susceptible to targeted misinformation attacks	Llama-2-7B Llama-3-8B GPT-J-6B meditron-7B	대규모 언어 모델(LLM)은 광범위한 의학 지식을 가지고 있으며 많은 영역에서 의료 정보에 대해 추론할 수 있어, 가까운 미래에 다양한 의료 응용 분야에서 유망한 잠재력을 가지고 있다. 본 연구에서 **우리는 의학 분야에서 LLM의 우려스러운 취약점을 입증한다. LLM 가중치의 단 1.1%를 표적 조작함으로써 부정확한 생의학적 사실을 의도적으로 주입할 수 있다**. **잘못된 정보는 다른 생의학 작업의 성능을 유지하면서 모델의 출력에 전파된다**. 우리는 1,025개의 부정확한 생의학적 사실 세트에서 우리의 발견을 검증한다. **이러한 특이한 취약성은 의료 환경에서 LLM 적용에 대한 심각한 보안 및 신뢰성 우려를 제기**한다. 이는 의료 실무에서 신뢰할 수 있고 안전한 사용을 보장하기 위해 강력한 보호 조치, 철저한 검증 메커니즘, 이러한 모델에 대한 접근의 엄격한 관리의 필요성을 강조한다. Large language models (LLMs) have broad medical knowledge and can

				reason about medical information across many domains, holding promising potential for diverse medical applications in the near future. In this study, we demonstrate a concerning vulnerability of LLMs in medicine. Through targeted manipulation of just 1.1% of the weights of the LLM, we can deliberately inject incorrect biomedical facts. The erroneous information is then propagated in the model's output while maintaining performance on other biomedical tasks. We validate our findings in a set of 1025 incorrect biomedical facts. This peculiar susceptibility raises serious security and trustworthiness concerns for the application of LLMs in healthcare settings. It accentuates the need for robust protective measures, thorough verification mechanisms, and stringent management of access to these models, ensuring their reliable and safe use in medical practice.
80	(Alber et al. 2025)	Medical large language models are vulnerable to data-poisoning attacks	1.3 billion parameterGPT-3-like autoregressive decoder-only transformers with a rotary position embedding2 fraction of0.5.	의료 분야에서 대규모 언어 모델(LLM)의 채택은 잘못된 의학 지식을 확산시킬 가능성에 대한 신중한 분석을 요구한다. LLM은 훈련 중 개방 인터넷에서 방대한 양의 데이터를 수집하기 때문에, 의도적으로 심어진 오정보를 포함할 수 있는 검증되지 않은 의학 지식에 잠재적으로 노출된다. 여기서 우리는 LLM 개발에 사용되는 인기 있는 데이터셋인 The Pile에 대한 데이터 오염 공격을 시뮬레이션하는 위협 평가를 수행한다. **우리는 훈련 토큰의 단 0.001%를 의료 오정보로 대체하는 것이 의료 오류를 전파할 가능성이 더 높은 해로운 모델을 초래한다는 것을 발견했다**. 더욱이 우리는 **손상된 모델이 의료 LLM을 평가하는 데 일상적으로 사용되는 오픈소스 벤치마크에서 손상 없는 대응 모델의 성능과 일치한다는 것을 발견**했다. 생의학 지식 그래프를 사용하여 의료 LLM 출력을 선별함으로써, 우리는 유해한 콘텐츠의 91.9%를 포착하는(F1 = 85.7%) 위해 완화 전략을 제안한다. 우리의 알고리즘은 지식 그래프의 하드코딩된 관계에 대해 확률적으로 생성된 LLM 출력을 검증하는 독특한 방법을 제공한다. 개선된 데이터 출처 및 투명한 LLM 개발에 대한 현재의 요구를 고려할 때, 우리는 특히 오정보가 잠재적으로 환자 안전을 손상시킬 수 있는 의료 분야에서 웹 스크랩 데이터에 무차별적으로 훈련된 LLM의 신흥 위험에 대한 인식을 높이고자 한다. The adoption of large language models (LLMs) in healthcare demands a careful analysis of their potential to spread false medical knowledge. Because LLMs ingest massive volumes of data from the open Internet during training, they are potentially exposed to unverified medical knowledge that may include deliberately planted misinformation. Here, we perform a threat assessment that simulates a data-poisoning attack against The Pile, a popular dataset used for LLM development. We find that replacement of just

				0.001% of training tokens with medical misinformation results in harmful models more likely to propagate medical errors. Furthermore, we discover that corrupted models match the performance of their corruption-free counterparts on open-source benchmarks routinely used to evaluate medical LLMs. Using biomedical knowledge graphs to screen medical LLM outputs, we propose a harm mitigation strategy that captures 91.9% of harmful content (F1 = 85.7%). Our algorithm provides a unique method to validate stochastically generated LLM outputs against hard-coded relationships in knowledge graphs. In view of current calls for improved data provenance and transparent LLM development, we hope to raise awareness of emergent risks from LLMs trained indiscriminately on web-scraped data, particularly in healthcare where misinformation can potentially compromise patient safety.
81	(Ji et al. 2025)	Mitigating the risk of health inequity exacerbated by large language models	LLaMA	대규모 언어 모델(LLM)의 최근 발전은 특히 중개 연구를 위한 임상시험 매칭 자동화와 임상 의사결정 지원을 위한 의료 질의응답 향상에서 수많은 의료 응용 분야에서의 잠재력을 입증했다. 그러나 **우리의 연구는 인종, 성별, 소득 수준, LGBT+ 지위, 노숙, 문맹, 장애, 실업과 같은 비결정적 사회인구학적 요인을 LLM의 입력에 통합하는 것이 부정확하고 해로운 출력으로 이어질 수 있음**을 보여준다. 이러한 불일치는 LLM이 의료 분야에 광범위하게 구현될 경우 기존의 건강 격차를 악화시킬 수 있다. 이 문제를 해결하기 위해 우리는 LLM 기반 의료 응용 프로그램에서 건강 불평등의 위험을 탐지하고 완화하도록 설계된 새로운 프레임워크인 EquityGuard를 소개한다. 우리의 평가는 다양한 인구 집단에 걸쳐 공평한 결과를 촉진하는 데 있어 그 효과성을 입증한다. Recent advancements in large language models (LLMs) have demonstrated their potential in numerous medical applications, particularly in automating clinical trial matching for translational research and enhancing medical question-answering for clinical decision support. However, our study shows that incorporating non-decisive socio-demographic factors, such as race, sex, income level, LGBT+ status, homelessness, illiteracy, disability, and unemployment, into the input of LLMs can lead to incorrect and harmful outputs. These discrepancies could worsen existing health disparities if LLMs are broadly implemented in healthcare. To address this issue, we introduce EquityGuard, a novel framework designed to detect and mitigate the risk of health inequities in LLM-based medical applications. Our evaluation demonstrates its effectiveness in promoting equitable outcomes across diverse populations.

82	(Howard et al. 2024)	Navigating ChatGPT's alignment with expert consensus on pediatric OSA management	ChatGPT 3.5	**목적:** 본 연구는 인공지능(AI), 특히 ChatGPT의 의료 의사결정에의 잠재적 통합을 평가하는 것을 목표로 했으며, 지속적인 소아 폐쇄성 수면무호흡증의 관리에 관한 전문가 합의 성명과의 일치도에 초점을 맞추었다. **방법:** 우리는 편도아데노이드 절제술 후 소아 지속성 OSA 관리에 관한 2024년 전문가 합의 성명(ECS)의 52개 진술에 대한 ChatGPT의 응답을 분석했다. 각 진술은 9점 리커트 척도 형식을 사용하여 ChatGPT에 입력되었으며, 평균 점수와 표준편차를 계산하기 위해 각 진술을 3회 입력했다. 통계 분석은 Excel을 사용하여 수행되었다. **결과:** **ChatGPT의 응답은 진술의 63%(33/52)에서 합의 성명 평균 점수의 1.0 이내였다. 13%(7/52)는 ChatGPT 평균 응답이 ECS 평균과 2.0 이상 차이가 나는 진술이었으며, 대부분이 수술 및 약물 관리 범주**에 속했다. **ChatGPT 평균 점수가 합의 평균과 2.0 이상 차이가 나는 진술은 확립된 의학적 주제에 대한 부정확한 정보 전파의 위험을 강조했으며, 응답의 주목할 만한 변동성은 ChatGPT의 신뢰성에서의 비일관성을 시사**했다. **결론:** ChatGPT가 많은 경우에 전문 의학적 의견과 일치하는 유망한 능력을 보여주었지만, 논쟁이 있는 영역에서 부정확성을 전파할 가능성과 비일관성은 임상 환경에서의 적용에 중요한 고려사항을 제기한다. 연구 결과는 의료 분야에서 AI 도구의 지속적인 평가 및 개선의 필요성을 강조하며, 의료 의사결정 과정에 AI의 안전하고 효과적인 통합을 보장하기 위해 AI 개발자, 의료 전문가 및 규제 기관 간의 협력을 강조한다. **Objective:** This study aimed to evaluate the potential integration of artificial intelligence (AI), specifically ChatGPT, into healthcare decision-making, focusing on its alignment with expert consensus statements regarding the management of persistent pediatric obstructive sleep apnea. **Methods:** We analyzed ChatGPT's responses to 52 statements from the 2024 expert consensus statement (ECS) on the management of pediatric persistent OSA after adenotonsillectomy. Each statement was input into ChatGPT using a 9-point Likert scale format, with each statement entered three times to calculate mean scores and standard deviations. Statistical analysis was performed using Excel. **Results:** ChatGPT's responses were within 1.0 of the consensus statement mean score for 63 % (33/52) of the statements. 13 % (7/52) were statements in which the ChatGPT mean response was different from the ECS mean by 2.0 or greater, the majority of which were in the categories of surgical and medical management. Statements with ChatGPT mean scores

				differing by more than 2.0 from the consensus mean highlighted the risk of disseminating incorrect information on established medical topics, with a notable variation in responses suggesting inconsistencies in ChatGPT's reliability. Conclusion: While ChatGPT demonstrated a promising ability to align with expert medical opinions in many cases, its inconsistencies and potential to propagate inaccuracies in contested areas raise important considerations for its application in clinical settings. The findings underscore the need for ongoing evaluation and refinement of AI tools in healthcare, emphasizing collaboration between AI developers, healthcare professionals, and regulatory bodies to ensure AI's safe and effective integration into medical decision-making processes.
83	(Biro et al. 2025)	Opportunities and risks of artificial intelligence in patient portal messaging in primary care	ChatGPT 4.0	환자 포털 메시징의 급격한 증가는 일차 진료 의사(PCP)의 업무량을 증가시켜 번아웃에 기여하고 있다. 환자 메시지에 대한 응답 초안을 작성하기 위해 생성형 인공지능(AI)을 사용하는 것은 인지적 부담을 줄이는 데 유망함을 보여주었지만, AI 초안 사용의 안전성과 인식에 대해서는 아직 많은 미지의 부분이 있다. 이 횡단면 시뮬레이션 연구는 PCP가 환자 포털 메시지에 대한 AI 생성 초안 응답의 오류를 식별하고 수정할 수 있는지 평가했다. 20명의 현역 PCP가 18개의 환자 포털 메시지를 검토했으며, **그중 4개는 객관적 부정확성 또는 잠재적으로 해로운 누락으로 분류된 오류를 포함**했다. **각 오류는 13-15명의 참가자에 의해 불충분하게 처리되었고, 오류가 있는 초안의 35-45%가 전혀 편집되지 않은 채 제출**되었다. 참가자의 80%가 AI 초안이 인지적 업무량을 줄였다는 데 동의했고 75%가 안전하다고 판단했지만, 수정되지 않은 오류는 환자 안전 위험을 강조하며, AI 도구의 개선된 설계, 훈련 및 오류 탐지 메커니즘의 필요성을 강조한다. The rapid increase in patient portal messaging has heightened the workload for primary care physicians (PCPs), contributing to burnout. The use of generative artificial intelligence (AI) to draft responses to patient messages has shown promise in reducing cognitive burden, yet there is still much unknown about the safety and perceptions of using AI drafts. This cross-sectional simulation study assessed whether PCPs could identify and correct errors in AI-generated draft responses to patient portal messages. Twenty practicing PCPs reviewed 18 patient portal messages, four of which contained errors categorized as objective inaccuracies or potentially harmful omissions. Each error was insufficiently addressed by 13–15 participants, and 35–45% of erroneous drafts were submitted entirely unedited. While

				80% of participants agreed AI drafts reduced cognitive workload and 75% found them safe, uncorrected errors highlight patient safety risks, underscoring the need for improved design, training, and error-detection mechanisms for AI tools.
84	(Zeng et al. 2025)	Perceptions and Attitudes of Chinese Oncologists Toward Endorsing AI-Driven Chatbots for Health Information Seeking Among Patients with Cancer: Phenomenological Qualitative Study	ChatGPT	**배경:** 대규모 언어 모델 인공지능(AI)에 의해 구동되는 챗봇은 암 환자의 건강 정보 접근성을 향상시킬 잠재적 도구로 부상했다. 그러나 환자 교육에의 통합은 종양학 전문의들 사이에서 우려를 제기한다. 건강 정보를 위한 AI 기반 챗봇 추천에 대한 종양학 전문의의 인식과 태도를 검토한 문헌은 제한적이다. **목적:** 본 연구는 암 환자에게 AI 기반 챗봇을 추천하는 것에 대한 중국 종양학 전문의의 인식과 태도를 탐구하는 것을 목표로 한다. **방법:** 이 현상학적 질적 연구에서 우리는 중국 남서부 및 동부의 4개 병원에서 종양학 전문의를 의도적으로 표본추출하고 2024년 11월 19일부터 12월 21일 사이에 24명의 참가자와 반구조화 인터뷰를 실시했다. 데이터 수집의 종료 시점을 결정하기 위해 데이터 포화 원칙을 관찰했다. 데이터는 Colaizzi 방법을 사용하여 분석되었다. **결과:** 참가자의 평균 연령은 42.0세(범위 29-53)였으며, 여성 9명(37%), 남성 15명(62%)이 포함되었다. 참가자는 종양학 분야에서 평균 8.8년(범위 1-25)의 경력을 가지고 있었다. 참가자 중 7명(29%)이 환자에게 AI 챗봇을 추천한 적이 있었다. 인터뷰 기록 분석에서 인지된 이점, 중대한 우려, 의사-환자 역학에 대한 영향을 포함한 3가지 주요 주제가 드러났다. 이점에는 향상된 접근성과 만성 질환 관리에 대한 잠재적 지원이 포함되었다. **우려는 책임, 오정보, 개인화 부족, 개인정보 보호 및 데이터 보안 위험, 환자 준비 및 교육에 집중되었다. 종양학 전문의는 AI 챗봇이 의사-환자 역학에 미치는 이중적 영향을 강조하며, 개선된 의사소통의 잠재력과 AI에 대한 과도한 의존으로 인한 신뢰 침식 위험을 인식**했다. **결론:** 중국 종양학 전문의는 건강 정보 접근성과 만성 질환 관리를 향상시키는 AI 기반 챗봇의 잠재력을 인식하면서도 책임, 오정보, 개인화 부족, 개인정보 보호 및 데이터 보안 위험, 환자 준비를 포함한 중대한 우려를 보고한다. 이러한 문제를 해결하려면 명확한 정책 및 지침, 엄격한 테스트 및 검증, 기관의 지지, 강력한 환자 및 제공자 교육과 같은 포괄적인 솔루션이 필요하다. 향후 노력은 안전하고 효과적이며 윤리적인 방식으로 환자 중심 치료를 지원하기 위해 AI 기술의 강점을 활용하면서 장벽을 해결하는 데 초점을 맞춰야 한다. **Background:** Chatbots driven by large language model artificial intelligence (AI) have emerged as potential tools to enhance health information access for patients with cancer. However, their integration into patient education raises concerns among oncologists. Limited literature has examined the

				perceptions and attitudes of oncologists in terms of endorsing AI-driven chatbots for health information. **Objective:** This study aims to explore the perceptions and attitudes of Chinese oncologists toward endorsing AI-driven chatbots to patients with cancer. **Methods:** In this phenomenological qualitative study, we purposively sampled oncologists from 4 hospitals in Southwest and East China and conducted semistructured interviews with 24 participants between November 19, 2024, and December 21, 2024. The data saturation principle was observed to determine the end point of data collection. Data were analyzed using the Colaizzi method. **Results:** The participants were aged 42.0 (range 29-53) years on average, including 9 (37%) female and 15 (62%) male participants. The participants had an average of 8.8 (range 1-25) years in oncology. Of the participants, 7 (29%) had recommended AI chatbots to patients. Three key themes were revealed from analysis of interview transcriptions, including perceived benefits, significant concerns, and impacts on doctor-patient dynamics. Benefits included enhanced accessibility and potential support for chronic condition management. Concerns centered on liability, misinformation, lack of personalization, privacy and data security risks, and patient readiness and education. Oncologists stressed a dual impact of AI chatbots on doctor-patient dynamics, recognizing the potential for improved communication and risks of trust erosion due to overreliance on AI. **Conclusions:** While recognizing the potential of AI-driven chatbots to enhance accessibility of health information and chronic disease management, Chinese oncologists report significant concerns, including liability, misinformation, lack of personalization, privacy and data security risks, and patient readiness. Addressing the challenges requires comprehensive solutions, such as clear policies and guidelines, rigorous testing and validation, institutional endorsement, and robust patient and provider education. Future efforts should focus on resolving the barriers while leveraging the strengths of AI technology to support patient-centered care in a safe, effective, and ethical manner.
85	(Marafino and Liu 2023)	Performance of a large language model (ChatGPT-3.5) for Pooled Cohort Equation estimation of	ChatGPT-3.5	산술 및 기타 정량적 작업에 대한 입증된 능력에도 불구하고, ChatGPT 및 기타 대규모 언어 모델의 임상 위험 계산 성능은 아직 평가되지 않았다. 합성 환자 데이터를 사용하여 이 예비 연구는 실제 점수와 비교하여 ChatGPT에서 도출된 죽상동맥

		atherosclerotic cardiovascular disease risk		경화성 심혈관 질환 위험의 Pooled Cohort Equation(PCE) 점수의 보정, 재현성 및 사회인구학적 편향 가능성을 평가하는 것을 목표로 했다. 우리는 ChatGPT에서 도출된 PCE 점수가 실제 PCE 점수와 중등도로 연관되어 있음에도 불구하고, **실제 PCE 점수에 대한 보정이 불량하고 반복적인 프롬프팅 라운드 간에 불안정성을 나타내어 재현성 부족을 시사한다는 것을 발견**했다. 더욱이 **ChatGPT에서 도출된 PCE 점수는 본 연구의 합성 환자의 사회인구학적 상태의 맥락적 지표에 부적절하게 민감한 것**으로 나타났다. 이러한 결과를 확인하고, 정확한 위험 계산이 적절한 임상 의사결정에 필수적인 심혈관 질환 예방을 넘어 다른 환경뿐만 아니라 더 다양한 프롬프트에서의 성능을 평가하기 위해서는 추가 작업이 필요하다. Despite demonstrated facility for arithmetic and other quantitative tasks, the performance of ChatGPT and other large language models for clinical risk calculation have yet to be assessed. Using synthetic patient data, this preliminary study aimed to assess the calibration, reproducibility, and potential for sociodemographic bias of ChatGPT-derived Pooled Cohort Equation (PCE) scores of atherosclerotic cardiovascular disease risk as compared to true scores. We found that ChatGPT-derived PCE scores, despite being moderately associated with the true PCE scores, displayed poor calibration with respect to true PCE scores, and exhibited instability between repeated rounds of prompting, suggesting lack of reproducibility. Moreover, ChatGPT-derived PCE scores also appeared inappropriately sensitive to contextual indicators of the sociodemographic status of the synthetic patients in this study. Further work is needed to confirm these results, and to assess performance on a wider variety of prompts as well as in other settings beyond cardiovascular disease prevention where accurate risk calculation is also vital to appropriate clinical decision-making.
86		Performance of ChatGPT 3.5 and 4 as a tool for patient support before and after DBS surgery for Parkinson's disease	ChatGPT 3.5 ChatGPT 4	심부 뇌 자극술(DBS)은 파킨슨병을 포함한 다양한 의학적 상태를 치료하기 위해 뇌의 특정 영역에 전극을 이식하는 신경외과 수술이다. 수술 전후 환자의 의문과 질문은 최신 과학적 및 임상 실무에 따라 해결되어야 한다. ChatGPT는 이해하기 쉬운 방식으로 의학적 질문을 이해하고 답변하는 능력으로 인공지능이 사용될 수 있는 방법의 예로 등장하며, 모든 사람이 접근할 수 있다. 그러나 이러한 자원의 위험은 여전히 완전히 이해될 필요가 있다. 영어와 포르투갈어로 된 40개 질문에 대한 ChatGPT 모델 3.5와 4의 응답은 기능 신경외과 및 신경 운동 장애 분야의 경험이 풍부한 2명의 전문가에 의해 독립적으로 평가되었고 세 번째 검토자에 의해 해결되었다. ChatGPT 3.5와 4는 파킨슨병에 대한 DBS 수술과 관련된 영어와 포르투갈어 모두로 된 80개 질문에 대해 양호한 수준의 정확도를 보였다. 정확한 것으로 평가

				된 응답의 비율은 GPT 3.5의 경우 57.5%, GPT 4의 경우 83.8%였다. **GPT 3.5는 응답의 6.3%(5/80)에 대해 잠재적으로 해로운 답변을 제공했다. GPT 4의 응답 중 해로운 것으로 평가된 것은 없었다.** 일반적으로 ChatGPT 3.5와 4는 두 가지 다른 언어에 걸쳐 품질과 신뢰성 측면에서 양호한 성능을 보였다. 그럼에도 불구하고 해로운 응답을 경시해서는 안 되며, 이러한 자원을 사용하여 환자를 대할 때 이 측면을 고려하는 것이 중요하다. **현재의 안전 우려를 고려할 때, 환자가 DBS 수술 안내를 위해 이러한 모델을 사용하는 것은 권장되지 않는다.** Deep brain stimulation (DBS) is a neurosurgical procedure that involves implanting electrodes into specific areas of the brain to treat a variety of medical conditions, including Parkinson's disease. Doubts and questions from patients prior to or following surgery should be addressed in line with the most recent scientific and clinical practice. ChatGPT emerges as an example of how artificial intelligence can be used, with its ability to comprehend and answer medical questions in an understandable way, accessible to everyone. However, the risks of these resources still need to be fully understood. ChatGPT models 3.5 and 4 responses to 40 questions in English and Portuguese were independently graded by two experienced specialists in functional neurosurgery and neurological movement disorders and resolved by a third reviewer. ChatGPT 3.5 and 4 demonstrated a good level of accuracy in responding to 80 questions in both English and Portuguese, related to DBS surgery for Parkinson's disease. The proportion of responses graded as correct was 57.5% and 83.8% for GPT 3.5 and GPT 4, respectively. GPT 3.5 provided potentially harmful answers for 6.3% (5/80) of its responses. No responses from GPT 4 were graded as harmful. In general, ChatGPT 3.5 and 4 demonstrated good performance in terms of quality and reliability across two different languages. Nonetheless, harmful responses should not be scorned, and it's crucial to consider this aspect when addressing patients using these resources. Considering the current safety concerns, it's not advisable for patients to use such models for DBS surgery guidance. Performance of ChatGPT 3.5 and 4 as a tool for patient support before and after DBS surgery for Parkinson's disease.
87	(Yu, Chen, et al. 2025)	Performance of Large Language Models in Diagnosing Rare Hematologic Diseases and the Impact of Their Diagnostic	Claude 3.5 Sonnet (Anthropic PBC) DeepSeek-R1 (DeepSeek) Doubao (ByteDance)	**배경:** 희귀 혈액 질환은 임상적 복잡성으로 인해 자주 진단되지 않거나 오진된다. 특히 사고 연쇄 추론을 사용하는 신세대 대규모 언어 모델(LLM)이 진단 정확도를 향상시킬 수 있는지는 불분명하다. **목적:** 본 연구는 희귀 혈액 질환에서 신세대 상용 LLM의 진단 성능을 평가하고,

		Outputs on Physicians: Combined Retrospective and Prospective Study	Gemini Experimental 1206 (Google LLC) ChatGPT-4o(OpenAI) ChatGPT-o1-preview (OpenAI) Qwen (Qwen-Max-2025-01-25)	LLM 출력이 의사의 진단 정확도를 향상시키는지 확인하는 것을 목표로 했다. **방법:** 우리는 2단계 연구를 수행했다. 후향적 단계에서 우리는 9개 희귀 혈액 질환을 다루는 158개의 비공개 실제 입원 기록에 대해 7개의 주류 LLM을 평가하고, 상위 10 정확도와 평균 역순위(MRR)를 사용하여 진단 성능을 평가했으며, Jaccard 유사성과 엔트로피를 통해 순위 안정성을 평가했다. Spearman 순위 상관관계를 사용하여 의사의 진단과 LLM 생성 출력 간의 연관성을 검토했다. 전향적 단계에서 다양한 경험 수준의 28명의 의사가 각각 5개의 사례를 진단했으며, LLM이 진단 정확도를 향상시킬 수 있는지 평가하기 위해 3개의 순차적 단계에 걸쳐 LLM 생성 진단에 접근했다. **결과:** 후향적 단계에서 ChatGPT-o1-preview는 가장 높은 상위 10 정확도(70.3%)와 MRR(0.577)을 보였으며, DeepSeek-R1이 2위를 차지했다. **아밀로이드 경쇄(AL) 아밀로이드증, Castleman 병, Erdheim-Chester 병, 다발성 신경병증, 장기비대, 내분비병증, 단클론 감마병증 및 피부 변화(POEMS) 증후군에 대한 진단 성능은 낮았다**. 흥미롭게도 대부분의 LLM에서 높은 정확도는 종종 낮은 순위 안정성과 상관관계가 있었다. 의사 성능은 상위 10 정확도(ρ=0.565) 및 MRR(ρ=0.650)과 강한 상관관계를 보였다. **전향적 단계에서 LLM은 경험이 적은 의사의 진단 정확도를 크게 향상시켰으며, 전문의에게는 유의한 이점이 관찰되지 않았다. 그러나 LLM이 편향된 응답을 생성할 때 의사 성능은 종종 향상되지 않거나 심지어 감소**했다. **결론:** 미세 조정 없이도 신세대 상용 LLM, 특히 사고 연쇄 추론을 가진 LLM은 희귀 혈액 질환의 진단을 높은 정확도로 식별할 수 있으며 경험이 적은 의사의 진단 성능을 크게 향상시킬 수 있다. 그럼에도 불구하고 편향된 LLM 출력은 임상의를 오도할 수 있으며, 적절한 안전장치 시스템과 함께 비판적 평가와 신중한 임상 통합의 필요성을 강조한다. **Background:** Rare hematologic diseases are frequently underdiagnosed or misdiagnosed due to their clinical complexity. Whether new-generation large language models (LLMs), particularly those using chain-of-thought reasoning, can improve diagnostic accuracy remains unclear. **Objective:** This study aimed to evaluate the diagnostic performance of new-generation commercial LLMs in rare hematologic diseases and to determine whether the LLM output enhances physicians' diagnostic accuracy. **Methods:** We conducted a 2-phase study. In the retrospective phase, we evaluated 7 mainstream LLMs on 158 nonpublic real-world admission records covering 9 rare hematologic diseases, assessed diagnostic performance using top-10 accuracy and mean reciprocal rank (MRR), and

				evaluated ranking stability via Jaccard similarity and entropy. Spearman rank correlation was used to examine the association between physicians' diagnoses and LLM-generated outputs. In the prospective phase, 28 physicians with varying levels of experience diagnosed 5 cases each, gaining access to LLM-generated diagnoses across 3 sequential steps to assess whether LLMs can improve diagnostic accuracy. **Results:** In the retrospective phase, ChatGPT-o1-preview demonstrated the highest top-10 accuracy (70.3%) and MRR (0.577), and DeepSeek-R1 ranked second. Diagnostic performance was low for amyloid light-chain (AL) amyloidosis; Castleman disease; Erdheim-Chester disease; and polyneuropathy, organomegaly, endocrinopathy, monoclonal gammopathy, and skin changes (POEMS) syndrome. Interestingly, higher accuracy often correlated with lower ranking stability across most LLMs. The physician performance showed a strong correlation with both top-10 accuracy (ρ =0.565) and MRR (ρ=0.650). In the prospective phase, LLMs significantly improved the diagnostic accuracy of less-experienced physicians; no significant benefit was observed for specialists. However, when LLMs generated biased responses, physician performance often failed to improve or even declined. **Conclusions:** Without fine-tuning, new-generation commercial LLMs, particularly those with chain-of-thought reasoning, can identify diagnoses of rare hematologic diseases with high accuracy and significantly enhance the diagnostic performance of less-experienced physicians. Nevertheless, biased LLM outputs may mislead clinicians, highlighting the need for critical appraisal and cautious clinical integration with appropriate safeguard systems. Trial Registration: Chinese Clinical Trial Registry ChiCTR2400089959; https://www.chictr.org.cn/hvshowproject.html?id=260575
88	(Fisch et al. 2024)	Performance of large language models on advocating the management of meningitis: a comparative qualitative stud	Bard Bing Claude-2 ChatGPT-3.5 ChatGPT-4 Llama PaLM	**목적** 우리는 가상 의료 사례를 사용하여 세균성 수막염 지침에 대한 대규모 언어 모델(LLM)의 준수도를 조사하고, 의료 분야에서의 유용성과 한계를 강조하는 것을 목표로 했다. **방법** 유양돌기염에 이차적인 세균성 수막염 환자의 시뮬레이션된 임상 시나리오를 공개적으로 접근 가능한 7개의 LLM(Bard, Bing, Claude-2, GPT-3.5, GPT-4, Llama, PaLM)에 3개의 독립적인 세션에서 제시했다. 응답은 양호한 임상 실무와 2개의 국제 수막염 지침 준수 여부에 대해 평가되었다. **결과** 중추신경계 감염은 LLM 세션의 90%에서 식별되었다. 모두 영상 검사를 권장

				했으며, 81%는 요추 천자를 제안했다. 혈액 배양과 특정 유양돌기염 검사는 각각 62%와 38%의 세션에서만 제안되었다. 세션의 38%만이 올바른 경험적 항생제 치료를 제공했으며, 항바이러스 치료와 덱사메타손은 각각 33%와 24%에서 권고되었다. **오해의 소지가 있는 진술은 52%에서 생성되었다. LLM의 텍스트 길이와 성능 간에 유의한 상관관계는 발견되지 않았다(r=0.29, p=0.20)**. 모든 LLM 중 GPT-4가 최고의 성능을 보였다. **고찰** 최신 LLM은 감별 진단 및 진단 절차에 대한 가치 있는 조언을 제공하지만, **현실적인 임상 시나리오에 적용될 때 세균성 수막염에 대한 치료 특정 정보에서 크게 차이가 난다. 오해의 소지가 있는 진술은 흔했으며, 성능 차이는 출력 길이보다는 각 LLM의 고유한 알고리즘에 기인**했다. **결론 사용자는 의료 의사결정의 지원 도구로 LLM을 고려할 때 이러한 한계와 성능 변동성을 인식해야 한다**. 이러한 모델의 복잡한 의료 시나리오에 대한 이해력과 신뢰할 수 있는 정보 제공 능력을 개선하기 위해서는 추가 연구가 필요하다. **Objectives** We aimed to examine the adherence of large language models (LLMs) to bacterial meningitis guidelines using a hypothetical medical case, highlighting their utility and limitations in healthcare. **Methods** A simulated clinical scenario of a patient with bacterial meningitis secondary to mastoiditis was presented in three independent sessions to seven publicly accessible LLMs (Bard, Bing, Claude-2, GPT-3.5, GPT-4, Llama, PaLM). Responses were evaluated for adherence to good clinical practice and two international meningitis guidelines. **Results** A central nervous system infection was identified in 90% of LLM sessions. All recommended imaging, while 81% suggested lumbar puncture. Blood cultures and specific mastoiditis work-up were proposed in only 62% and 38% sessions, respectively. Only 38% of sessions provided the correct empirical antibiotic treatment, while antiviral treatment and dexamethasone were advised in 33% and 24%, respectively. Misleading statements were generated in 52%. No significant correlation was found between LLMs' text length and performance (r=0.29, p=0.20). Among all LLMs, GTP-4 demonstrated the best performance. **Discussion** Latest LLMs provide valuable advice on differential diagnosis and diagnostic procedures but significantly vary in treatment-specific information for bacterial meningitis when introduced to a realistic clinical scenario. Misleading statements were common, with performance differences attributed to each LLM's unique algorithm rather than output length.

				Conclusions Users must be aware of such limitations and performance variability when considering LLMs as a support tool for medical decision-making. Further research is needed to refine these models' comprehension of complex medical scenarios and their ability to provide reliable information.
89	(Pichowicz, Kotas, and Piotrowski 2025)	Performance of mental health chatbot agents in detecting and managing suicidal ideation	GPT-4o mini Gemini 2.0 Flash DeepSeek-v1 LeChat Llama 3.1 8B	인공지능(AI) 기술의 발전은 AI 기반 챗봇 에이전트를 통해 정신 건강 문제를 경험하는 개인을 돕기 위해 설계된 스마트폰 애플리케이션의 급속한 발전을 촉발했다. 그러나 자살 위기를 포함한 정신 건강 위기를 경험하는 개인을 다룰 때 이러한 에이전트의 안전성은 평가되지 않았다. 본 연구에서 우리는 시뮬레이션된 자살 위험 시나리오에 응답하는 29개 AI 기반 챗봇 에이전트의 능력을 평가했다. 애플리케이션 저장소를 검색하고 정신적 고통을 경험할 때 유익하다고 주장하고 AI 기반 챗봇 기능을 제공하는 앱을 찾기 위해 앱 설명을 선별했다. 모든 에이전트는 증가하는 자살 위험을 시뮬레이션하도록 설계된 Columbia-Suicide Severity Rating Scale을 기반으로 한 표준화된 프롬프트 세트로 테스트되었다. 우리는 응급 연락처 정보를 제공하는 능력 및 기타 요인을 기반으로 한 사전 정의된 기준에 따라 응답을 평가했다. **테스트된 에이전트 중 어느 것도 적절한 응답에 대한 초기 기준을 충족하지 못했으며, 51.72%는 한계적 응답에 대한 완화된 기준을 충족했고, 48.28%는 부적절한 것으로 간주되었다. 일반적인 오류에는 응급 연락처 정보를 제공하지 못하는 것과 맥락적 이해 부족이 포함되었다. 이러한 발견은 적절한 임상 검증 없이 민감한 건강 맥락에서 AI 기반 챗봇을 배치하는 것에 대한 우려를 제기**한다. Advances in artificial intelligence (AI) technologies sparked a rapid development of smartphone applications designed to help individuals experiencing mental health problems through an AI-powered chatbot agent. However, the safety of such agents when dealing with individuals experiencing a mental health crisis, including suicidal crisis, has not been evaluated. In this study, we assessed the ability of 29 AI-powered chatbot agents to respond to simulated suicidal risk scenarios. Application repositories were searched and the app descriptions screened in search of apps that claimed to be beneficial when experiencing mental distress and offered an AI-powered chatbot function. All agents were tested with a standardized set of prompts based on the Columbia-Suicide Severity Rating Scale designed to simulate increasing suicidal risk. We assessed the responses according to pre-defined criteria based on the ability to provide emergency contact information and other factors. None of the tested agents satisfied our initial criteria for an adequate response, 51.72% satisfied the relaxed criteria for a marginal response, while 48.28% were deemed

				inadequate. Common errors included the inability to provide emergency contact information and a lack of contextual understanding. These findings raise concerns about the deployment of AI-powered chatbots in sensitive health contexts without proper clinical validation.
90	(Hong et al. 2025)	Physician Perspectives on Large Language Models in Healthcare: A Cross-Sectional Survey Study	GPT-3.5 (ChatGPT Free) GPT-4 (ChatGPT Plus) Llama (or variants) BERT (or variants) Bard / Gemini Mistral GatorTron Vicuna Claude Perplexity	**목적:** 본 연구는 의료 환경에서 대규모 언어 모델(LLM)에 관한 의사의 실무와 관점을 평가하는 것을 목표로 한다. **방법:** 2024년 5월부터 7월 사이에 기관 LLM 접근 권한이 있는 곳과 없는 곳 등 2개의 주요 학술 의료 센터(AMC)에서 의사 관점을 비교하는 횡단면 설문조사 연구를 수행했다. 참가자는 부서 리더십과 눈덩이 표본추출을 통해 모집된 임상 교수진과 수련의를 포함했다. 주요 결과는 현재 LLM 사용 빈도, 평가 지표의 중요도 순위, 책임 우려 및 선호하는 학습 주제였다. **결과:** 306명의 응답자(전임의 217명 [70.9%], 수련의 80명 [26.1%]) 중 197명(64.4%)이 LLM을 사용한다고 보고했다. **기관 LLM 접근 권한이 있는 AMC는 유의하게 낮은 책임 우려를 보고했다**(높은 우려 보고 49.2% vs 66.7%; 17.5 백분율 포인트 차이 [95% CI, 6.8-28.2]; P=.0082). **정확성은 모든 전문 분야에서 우선순위로 지정되었다(중앙값 순위 1.0 [IQR, 1.0-2.0]). 응답자 중 287명의 의사(94%)가 추가 교육을 요청했다. 주요 학습 우선순위는 임상 응용(206명 [71.9%])과 위험 관리(181명 [63.1%])**였다. 광범위한 개인적 사용에도 불구하고 **단 8명의 의사(2.6%)만이 환자에게 LLM을 추천**했다. 주목할 만한 전문 분야 및 인구통계학적 변동이 나타났으며, 젊은 의사가 더 높은 열정을 보였지만 법적 우려도 증가했다. **결론:** 이 설문조사 연구는 LLM에 대한 의사의 현재 사용 패턴과 관점에 대한 통찰력을 제공한다. 책임 우려는 기관 LLM 접근 권한이 있는 환경에서 감소하는 것으로 보인다. 연구 결과는 의료 센터가 LLM 관련 정책 및 교육 프로그램을 개발할 때 고려할 기회를 제안한다. **OBJECTIVES:** This study aims to evaluate physicians' practices and perspectives regarding large language models (LLMs) in healthcare settings. **METHODS:** A cross-sectional survey study was conducted between May and July 2024 comparing physician perspectives at two major academic medical centers (AMCs), one with institutional LLM access and one without. Participants included both clinical faculty and trainees recruited through departmental leadership and snowball sampling. Primary outcomes were current LLM use frequency, ranked importance of evaluation metrics, liability concerns, and preferred learning topics. **RESULTS:** Among 306 respondents (217 attending physicians [70.9%], 80 trainees [26.1%]), 197 (64.4%) reported using LLMs. The AMC with

				institutional LLM access reported significantly lower liability concerns (49.2% vs 66.7% reporting high concern; 17.5 percentage points difference [95% CI, 6.8–28.2]; P=.0082). Accuracy was prioritized across all specialties (median rank 1.0 [IQR, 1.0–2.0]). Of the respondents, 287 physicians (94%) requested additional training. Key learning priorities were clinical applications (206 [71.9%]) and risk management (181 [63.1%]). Despite widespread personal use, only 8 physicians (2.6%) recommended LLMs to patients. Notable specialty and demographic variations emerged, with younger physicians showing higher enthusiasm but also elevated legal concerns. **CONCLUSIONS:** This survey study provides insights into physicians' current usage patterns and perspectives on LLMs. Liability concerns appear to be lessened in settings with institutional LLM access. The findings suggest opportunities for medical centers to consider when developing LLM-related policies and educational programs.
91	(Clusmann, Ferber, et al. 2025)	Prompt injection attacks on vision language models in oncology	Claude-3 Opus Claude-3.5 Sonnet Reka Core GPT-4o	비전-언어 인공지능 모델(VLM)은 의학 지식을 보유하고 있으며 이미지 해석기, 가상 기록자, 일반 의사결정 지원 시스템을 포함하여 의료 분야에서 다양한 방식으로 사용될 수 있다. 그러나 여기서 우리는 의료 작업에 적용된 현재 VLM이 근본적인 보안 결함을 나타낸다는 것을 입증한다: 프롬프트 주입 공격에 의해 손상될 수 있다. 이는 매개변수에 대한 접근 없이 VLM과 상호작용하기만 해도 유해한 정보를 출력하는 데 사용될 수 있다. 우리는 4개의 최첨단 VLM에서 이러한 공격에 대한 취약성을 평가하기 위한 정량적 연구를 수행했다: **Claude-3 Opus, Claude-3.5 Sonnet, Reka Core, GPT-4o. N = 594개의 공격 세트를 사용하여 우리는 이러한 모든 모델이 취약하다는 것을 보여준다. 구체적으로 우리는 다양한 의료 영상 데이터에 하위 시각적 프롬프트를 삽입하면 모델이 유해한 출력을 제공하게 할 수 있으며, 이러한 프롬프트는 인간 관찰자에게 명확하지 않다는 것을 보여준다**. 따라서 우리의 연구는 광범위한 임상 채택 전에 완화되어야 하는 의료 VLM의 주요 취약점을 입증한다. Vision-language artificial intelligence models (VLMs) possess medical knowledge and can be employed in healthcare in numerous ways, including as image interpreters, virtual scribes, and general decision support systems. However, here, we demonstrate that current VLMs applied to medical tasks exhibit a fundamental security flaw: they can be compromised by prompt injection attacks. These can be used to output harmful information just by interacting with the VLM, without any access to its parameters. We perform a quantitative study to evaluate the vulnerabilities to these attacks in four

				state of the art VLMs: Claude-3 Opus, Claude-3.5 Sonnet, Reka Core, and GPT-4o. Using a set of N = 594 attacks, we show that all of these models are susceptible. Specifically, we show that embedding sub-visual prompts in manifold medical imaging data can cause the model to provide harmful output, and that these prompts are non-obvious to human observers. Thus, our study demonstrates a key vulnerability in medical VLMs which should be mitigated before widespread clinical adoption.
92	(Zhang et al. 2025)	Prompt injection attacks on vision-language models for surgical decision support	Gemini 1.5 Pro Gemini 2.5 Pro ChatGPT-o4-mini-high Qwen 2.5-VL	**중요성:** 인공지능 기반 복강경 비디오 분석은 최소 침습 수술의 안전성과 정밀도를 높일 잠재력을 가지고 있다. 비전-언어 모델은 복잡한 시공간(비디오) 데이터를 이해하는 능력으로 인해 비디오 기반 수술 의사결정 지원에 특히 유망하다. 그러나 이러한 능력을 가능하게 하는 동일한 다중모드 인터페이스는 삽입된 기만적인 텍스트나 이미지(프롬프트 주입 공격)를 통한 조작에 대한 새로운 취약점을 도입한다. **목적:** 임상적으로 관련된 수술 의사결정 지원 작업의 맥락에서 최첨단 비디오 가능 비전-언어 모델이 텍스트 및 시각적 프롬프트 주입 공격에 얼마나 취약한지 체계적으로 평가한다. **설계, 설정 및 참가자:** 이 관찰 연구에서 우리는 출혈 사건 탐지, 이물질, 영상 왜곡, 안전성의 중요한 관점 평가, 수술 기술 평가 등 11개의 수술 의사결정 지원 작업에 걸쳐 4개의 최첨단 비전-언어 모델인 Gemini 1.5 Pro, Gemini 2.5 Pro, GPT-o4-mini-high, Qwen 2.5-VL을 체계적으로 평가했다. 프롬프트 주입 시나리오는 오해를 일으키는 텍스트 프롬프트와 다양한 지속 시간에 적용되는 흰색 텍스트 오버레이로 표시되는 시각적 교란을 포함했다. **주요 결과 및 측정:** 주요 측정값은 기준선 성능과 각 프롬프트 주입 조건 간에 대비된 모델 정확도였다. **결과:** 모든 비전-언어 모델이 양호한 기준선 정확도를 보였으며, Gemini 2.5 Pro는 일반적으로 모든 작업에서 가장 높은 평균[표준편차] 정확도(0.82[0.01])를 달성했으며, Gemini 1.5 Pro(0.70[0.03]) 및 GPT-o4 mini-high(0.67[0.06])와 비교되었다. 작업 전반에 걸쳐 Qwen 2.5-VL은 대부분의 출력을 검열했고 비검열 출력에서(0.58[0.03])의 정확도를 달성했다. **텍스트 및 시간적으로 변화하는 시각적 프롬프트 주입은 모든 모델의 정확도를 감소시켰다. 장기간 시각적 프롬프트 주입은 일반적으로 단일 프레임 주입보다 더 위해했다. Gemini 2.5 Pro는 가장 큰 강건성을 보였고 프롬프트 주입에도 불구하고 여러 작업에서 안정적인 성능을 유지했지만, GPT-o4-mini-high는 가장 높은 취약성을 나타냈으며, 모든 작업에 걸친 평균(표준편차) 정확도가 기준선에서 0.67(0.06)에서 전체 지속 시간 시각적 프롬프트 주입 하에서 0.24(0.04)로 감소**했다(P 〈 .001). **결론 및 관련성:** 이러한 발견은 비전-언어 모델이 실시간 수술 의사결정 지원에 안전하게 배치되기 전에 강력한 시간적 추론 능력과 전문화된 안전장치의 중요한 필요

				성을 나타낸다. **Importance:** Artificial Intelligence–driven analysis of laparoscopic video holds potential to increase the safety and precision of minimally invasive surgery. Vision–language models are particularly promising for video–based surgical decision support due to their capabilities to comprehend complex temporospatial (video) data. However, the same multimodal interfaces that enable such capabilities also introduce new vulnerabilities to manipulations through embedded deceptive text or images (prompt injection attacks). **Objective:** To systematically evaluate how susceptible state–of–the–art video–capable vision–language models are to textual and visual prompt injection attacks in the context of clinically relevant surgical decision support tasks. **Design, Setting, and Participants:** In this observational study, we systematically evaluated four state–of–the–art vision–language models, Gemini 1.5 Pro, Gemini 2.5 Pro, GPT–o4–mini–high, and Qwen 2.5–VL, across eleven surgical decision support tasks: detection of bleeding events, foreign objects, image distortions, critical view of safety assessment, and surgical skill assessment. Prompt injection scenarios involved misleading textual prompts and visual perturbations, displayed as white text overlay, applied at varying durations. **Main Outcomes and Measures:** The primary measure was model accuracy, contrasted between baseline performance and each prompt injection condition. **Results:** All vision–language models demonstrated good baseline accuracy, with Gemini 2.5 Pro generally achieving the highest mean [standard deviation] accuracy across all tasks (0.82 [0.01]), compared to Gemini 1.5 Pro (0.70 [0.03]) and GPT–o4 mini–high (0.67 [0.06]). Across tasks, Qwen 2.5–VL censored most outputs and achieved an accuracy of (0.58 [0.03]) on non–censored outputs. Textual and temporally–varying visual prompt injections reduced the accuracy for all models. Prolonged visual prompt injections were generally more harmful than single–frame injections. Gemini 2.5 Pro showed the greatest robustness and maintained stable performance for several tasks despite prompt injections, whereas GPT–o4–mini–high exhibited the highest vulnerability, with mean (standard deviation) accuracy across all tasks declining from 0.67 (0.06) at baseline to 0.24 (0.04) under full–duration visual

				prompt injection (P 〈 .001). **Conclusion and Relevance:** These findings indicate the critical need for robust temporal reasoning capabilities and specialized guardrails before vision-language models can be safely deployed for real-time surgical decision support.
93	(He et al. 2024)	Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study	ChatGPT-4 ChatGPT-3.5 LLaMA 2 MedAlpaca ORCA_mini	**배경:** 환자들은 환자 포털을 통해 전자건강기록과 실험실 검사 결과 데이터에 쉽게 접근할 수 있지만, 실험실 검사 결과는 종종 혼란스럽고 이해하기 어렵다. 많은 환자들이 동료로부터 조언을 구하기 위해 웹 기반 포럼이나 질의응답(Q&A) 사이트로 향한다. 건강 관련 질문에 대한 소셜 Q&A 사이트의 답변 품질은 크게 다르며, 모든 응답이 정확하거나 신뢰할 수 있는 것은 아니다. ChatGPT와 같은 대규모 언어 모델(LLM)은 환자가 질문에 대한 답변을 받을 수 있는 유망한 방법을 열었다. **목적:** 우리는 환자가 묻는 실험실 검사 관련 질문에 대해 관련성 있고, 정확하며, 유용하고, 무해한 응답을 생성하는 데 LLM을 사용하는 타당성을 평가하고 증강 접근법을 사용하여 완화할 수 있는 잠재적 문제를 식별하는 것을 목표로 했다. **방법:** 우리는 Yahoo! Answers에서 실험실 검사 결과 관련 Q&A 데이터를 수집하고 본 연구를 위해 53개의 Q&A 쌍을 선택했다. LangChain 프레임워크와 ChatGPT 웹 포털을 사용하여 GPT-4, GPT-3.5, LLaMA 2, MedAlpaca, ORCA_mini의 5개 LLM으로부터 53개 질문에 대한 응답을 생성했다. 우리는 Recall-Oriented Understudy for Gisting Evaluation, Bilingual Evaluation Understudy, Metric for Evaluation of Translation With Explicit Ordering, Bidirectional Encoder Representations from Transformers Score를 포함한 표준 Q&A 유사성 기반 평가 지표를 사용하여 답변의 유사성을 평가했다. 우리는 LLM 기반 평가자를 사용하여 대상 모델이 기준선 모델보다 관련성, 정확성, 유용성 및 안전성 측면에서 더 높은 품질을 가지고 있는지 판단했다. 우리는 동일한 4가지 측면에서 7개의 선택된 질문에 대한 모든 응답에 대해 의료 전문가와 수동 평가를 수행했다. **결과:** 4개 LLM의 응답 유사성과 관련하여; GPT-4 출력을 참조 답변으로 사용했을 때, GPT-3.5의 응답이 가장 유사했으며, LLaMA 2, ORCA_mini, MedAlpaca가 그 뒤를 이었다. Yahoo 데이터의 인간 답변은 가장 낮은 점수를 받았으며, 따라서 GPT-4 생성 답변과 가장 유사하지 않았다. 승률과 의료 전문가 평가 결과 모두 GPT-4의 응답이 모든 4가지 측면(관련성, 정확성, 유용성 및 안전성)에서 다른 모든 LLM 응답 및 인간 응답보다 더 나은 점수를 달성했음을 보여주었다. **LLM 응답은 때때로 개인의 의료 맥락에서 해석 부족, 부정확한 진술 및 참조 부족으로 어려움을 겪었다**. **결론:** 환자의 실험실 검사 결과 관련 질문에 대한 응답 생성에서 LLM을 평가한 결과, 다른 4개 LLM 및 Q&A 웹사이트의 인간 답변과 비교할 때 GPT-4의 응답이

				더 정확하고, 유용하며, 관련성이 있고, 더 안전했다. GPT-4 응답이 부정확하고 개인화되지 않은 경우가 있었다. 우리는 프롬프트 엔지니어링, 프롬프트 증강, 검색 증강 생성, 응답 평가를 포함하여 LLM 응답의 품질을 향상시키는 여러 방법을 식별했다. Background: Although patients have easy access to their electronic health records and laboratory test result data through patient portals, laboratory test results are often confusing and hard to understand. Many patients turn to web-based forums or question-and-answer (Q&A) sites to seek advice from their peers. The quality of answers from social Q&A sites on health-related questions varies significantly, and not all responses are accurate or reliable. Large language models (LLMs) such as ChatGPT have opened a promising avenue for patients to have their questions answered. Objective: We aimed to assess the feasibility of using LLMs to generate relevant, accurate, helpful, and unharmful responses to laboratory test-related questions asked by patients and identify potential issues that can be mitigated using augmentation approaches. Methods: We collected laboratory test result-related Q&A data from Yahoo! Answers and selected 53 Q&A pairs for this study. Using the LangChain framework and ChatGPT web portal, we generated responses to the 53 questions from 5 LLMs: GPT-4, GPT-3.5, LLaMA 2, MedAlpaca, and ORCA_mini. We assessed the similarity of their answers using standard Q&A similarity-based evaluation metrics, including Recall-Oriented Understudy for Gisting Evaluation, Bilingual Evaluation Understudy, Metric for Evaluation of Translation With Explicit Ordering, and Bidirectional Encoder Representations from Transformers Score. We used an LLM-based evaluator to judge whether a target model had higher quality in terms of relevance, correctness, helpfulness, and safety than the baseline model. We performed a manual evaluation with medical experts for all the responses to 7 selected questions on the same 4 aspects. Results: Regarding the similarity of the responses from 4 LLMs; the GPT-4 output was used as the reference answer, the responses from GPT-3.5 were the most similar, followed by those from LLaMA 2, ORCA_mini, and MedAlpaca. Human answers from Yahoo data were scored the lowest and, thus, as the least similar to GPT-4-generated answers. The results of the win rate and medical expert evaluation both showed that GPT-4's responses

				achieved better scores than all the other LLM responses and human responses on all 4 aspects (relevance, correctness, helpfulness, and safety). LLM responses occasionally also suffered from lack of interpretation in one's medical context, incorrect statements, and lack of references. **Conclusions:** By evaluating LLMs in generating responses to patients' laboratory test result–related questions, we found that, compared to other 4 LLMs and human answers from a Q&A website, GPT-4's responses were more accurate, helpful, relevant, and safer. There were cases in which GPT-4 responses were inaccurate and not individualized. We identified a number of ways to improve the quality of LLM responses, including prompt engineering, prompt augmentation, retrieval-augmented generation, and response evaluation.
94	(Si et al. 2025)	Quality safety and disparity of an AI chatbot in managing chronic diseases: simulated patient experiments	ERNIE Bot ChatGPT DeepSeek	AI 솔루션의 급속한 발전은 개발도상국 환경에서 만성 질환의 진단 부족과 불량한 관리 문제를 해결할 기회를 드러낸다. 모의 환자 방법과 실험 설계를 사용하여, 우리는 384개의 환자-AI 시험에서 중국의 ERNIE Bot과의 의료 상담의 품질, 안전성 및 격차를 평가한다. ERNIE Bot은 77.3%의 진단 정확도, 94.3%의 올바른 약물 처방에 도달했지만, **불필요한 의료 검사(91.9%)와 불필요한 약물(57.8%)의 높은 처방률을 보였다. 환자 연령과 가구 경제 상태에 따라 격차가 관찰되었으며, 나이가 많고 부유한 환자가 더 집중적인 치료를** 받았다. **표준화된 조건에서 ERNIE Bot, ChatGPT 및 DeepSeek은 인간 의사보다 높은 진단 정확도를 보였지만 과잉 처방 경향이 더 컸다. 결과는 개발도상국 맥락에서 의료 제공의 품질, 접근성 및 경제성을 강화하는 데 있어 ERNIE Bot의 큰 잠재력을 시사하지만, 안전성 및 사회인구학적 격차 증폭과 관련된 중대한 위험도 강조**한다. The rapid development of AI solutions reveals opportunities to address the underdiagnosis and poor management of chronic conditions in developing settings. Using the method of simulated patients and experimental designs, we evaluate the quality, safety, and disparity of medical consultation with ERNIE Bot in China among 384 patient-AI trials. ERNIE Bot reached a diagnostic accuracy of 77.3%, correct drug prescriptions of 94.3%, but prescribed high rates of unnecessary medical tests (91.9%) and unnecessary medications (57.8%). Disparities were observed based on patient age and household economic status, with older and wealthier patients receiving more intensive care. Under standardized conditions, ERNIE Bot, ChatGPT, and DeepSeek demonstrated higher diagnostic accuracy but a greater tendency toward overprescription than human physicians. The results suggest the

				great potential of ERNIE Bot in empowering quality, accessibility, and affordability of healthcare provision in developing contexts, but also highlight critical risks related to safety and amplification of sociodemographic disparities.
95	(Bhimani et al. 2025)	Real-World Evaluation of Large Language Models in Healthcare (RWE-LLM): A New Realm of AI Safety & Validation	Hippocratic AI Care Agent	**배경:** 의료 분야에서 인공지능(AI)의 배치는 특히 환자와 직접 상호작용하는 시스템에 대해 강력한 안전성 검증 프레임워크를 필요로 한다. 이론적 프레임워크가 존재하지만, 추상적 원칙과 실제 구현 사이에는 중대한 격차가 남아 있다. 전통적인 LLM 벤치마킹 접근법은 매우 제한적인 출력 범위를 제공하며 높은 안전 기준을 요구하는 의료 응용 프로그램에는 불충분하다. **목적:** 대규모 임상의 참여를 통해 의료 AI 안전성 검증을 위한 포괄적인 프레임워크를 개발하고 평가한다. **방법:** 우리는 레드 팀 방법론에서 영감을 받아 포괄적인 안전성 검증을 달성하기 위해 범위를 확장하면서 RWE-LLM(의료에서 대규모 언어 모델의 실제 평가) 프레임워크를 구현했다. 우리의 접근법은 사전 구현, 계층적 검토, 해결 및 지속적인 모니터링의 4단계에 걸쳐 입력 데이터 품질에만 의존하기보다는 출력 테스트를 강조한다. 우리는 평균 11.5년의 임상 경력을 가진 6,234명의 미국 면허 임상의(간호사 5,969명 및 의사 265명)를 참여시켰다. 프레임워크는 오류 탐지 및 해결을 위한 3단계 검토 프로세스를 사용했으며, 4회 반복(pre-Polaris 및 Polaris 1.0, 2.0, 3.0)에 걸쳐 환자 교육, 추적관찰 및 행정 지원에 초점을 맞춘 비진단 AI Care Agent를 평가했다. **결과:** RWE-LLM 프레임워크를 사용하여 307,000개 이상의 고유한 통화가 평가되었다. **각 상호작용은 경미한 임상 부정확성에서 중대한 안전 우려에 이르기까지 여러 심각도 범주에 걸쳐 잠재적인 오류 표시 대상**이 되었다. 다단계 검토 시스템은 플래그가 지정된 모든 상호작용을 성공적으로 처리했으며, 내부 간호 검토가 초기 전문가 평가를 제공했고 필요할 때 의사 판정이 뒤따랐다. 프레임워크는 일관된 처리 시간과 문서화 기준을 유지하면서 식별된 안전 우려를 해결하는 데 효과적인 처리량을 입증했다. 오류 식별과 시스템 개선 사이의 지속적인 피드백 루프를 통해 안전 프로토콜의 체계적 개선이 달성되었다. 성능 지표는 반복 간 상당한 안전 개선을 입증했으며, 올바른 의학적 조언 비율이 ~80.0%(pre-Polaris)에서 96.79%(Polaris 1.0), 98.75%(Polaris 2.0), 99.38%(Polaris 3.0)로 개선되었다. **잠재적 경미한 위해를 초래하는 부정확한 조언은 1.32%에서 0.13%, 0.07%로 감소했으며, 중증 위해 우려는 제거되었다(0.06%에서 0.10%, 0.00%)**. **결론:** RWE-LLM 프레임워크의 성공적인 전국적 구현은 의료 환경에서 AI 안전성을 보장하기 위한 실용적인 모델을 확립한다. 우리의 방법론은 포괄적인 출력 테스트가 수평적 LLM이 사용하는 전통적인 입력 검증 접근법보다 훨씬 강력한 안전 보증을 제공한다는 것을 입증한다. 자원 집약적이지만, 이 접근법은 의료 AI 시스템에

				대한 엄격한 안전성 검증이 필요하면서도 달성 가능하다는 것을 증명하며, 향후 배치를 위한 벤치마크를 설정한다. **Background:** The deployment of artificial intelligence (AI) in healthcare necessitates robust safety validation frameworks, particularly for systems directly interacting with patients. While theoretical frameworks exist, there remains a critical gap between abstract principles and practical implementation. Traditional LLM benchmarking approaches provide very limited output coverage and are insufficient for healthcare applications requiring high safety standards. **Objective:** To develop and evaluate a comprehensive framework for healthcare AI safety validation through large-scale clinician engagement. **Methods:** We implemented the RWE-LLM (Real-World Evaluation of Large Language Models in Healthcare) framework, drawing inspiration from red teaming methodologies while expanding their scope to achieve comprehensive safety validation. Our approach emphasizes output testing rather than relying solely on input data quality across four stages: pre-implementation, tiered review, resolution, and continuous monitoring. We engaged 6,234 US licensed clinicians (5,969 nurses and 265 physicians) with an average of 11.5 years of clinical experience. The framework employed a three-tier review process for error detection and resolution, evaluating a non-diagnostic AI Care Agent focused on patient education, follow-ups, and administrative support across four iterations (pre-Polaris and Polaris 1.0, 2.0, and 3.0). **Results:** Over 307,000 unique calls were evaluated using the RWE-LLM framework. Each interaction was subject to potential error flagging across multiple severity categories, from minor clinical inaccuracies to significant safety concerns. The multi-tiered review system successfully processed all flagged interactions, with internal nursing reviews providing initial expert evaluation followed by physician adjudication when necessary. The framework demonstrated effective throughput in addressing identified safety concerns while maintaining consistent processing times and documentation standards. Systematic improvements in safety protocols were achieved through a continuous feedback loop between error identification and system enhancement. Performance metrics demonstrated substantial safety improvements between iterations, with correct medical advice rates

				improving from ~80.0% (pre-Polaris), to 96.79% (Polaris 1.0), to 98.75% (Polaris 2.0) and 99.38% (Polaris 3.0). Incorrect advice resulting in potential minor harm decreased from 1.32% to 0.13% and 0.07%, and severe harm concerns were eliminated (0.06% to 0.10% and 0.00%). Conclusions: The successful nationwide implementation of the RWE-LLM framework establishes a practical model for ensuring AI safety in healthcare settings. Our methodology demonstrates that comprehensive output testing provides significantly stronger safety assurance than traditional input validation approaches used by horizontal LLMs. While resource-intensive, this approach proves that rigorous safety validation for healthcare AI systems is both necessary and achievable, setting a benchmark for future deployments.
96	(Chang et al. 2025)	Red teaming ChatGPT in medicine to yield real-world insights on model behavior	ChatGPT-3.5 ChatGPT-4.0 ChatGPT-4.0 with Internet GPT-4o	레드 팀, 즉 예기치 않거나 원치 않는 모델 행동을 적대적으로 노출하는 관행은 대규모 언어 모델의 공평성과 정확성을 개선하는 데 중요하지만, 의료 분야에서 모델 제작자와 제휴하지 않은 레드 팀은 부족하다. 우리는 임상의, 의학 및 공학 학생, 기술 전문가(총 80명 참가자)로 구성된 팀을 소집하여 실제 임상 사례로 모델을 스트레스 테스트하고 안전성, 개인정보 보호, 환각/정확성 및 편향의 축을 따라 부적절한 응답을 분류했다. 6명의 의학 훈련을 받은 검토자가 프롬프트-응답 쌍을 재분석하고 정성적 주석을 추가했다. 376개의 고유한 프롬프트(1504개 응답) 중 20.1%가 부적절했다(GPT-3.5: 25.8%; GPT-4.0: 16%; 인터넷을 사용한 GPT-4.0: 17.8%). 이후 우리는 행사 후 출시된 모델인 GPT-4o를 테스트하여 벤치마크의 유용성을 보여주었다(20.4% 부적절). GPT-3.5에서 적절했던 응답의 21.5%가 업데이트된 모델에서는 부적절했다. 우리는 레드 팀 프롬프트 구성에 대한 통찰력을 공유하고 반복적인 모델 평가를 위한 벤치마크를 제시한다. Red teaming, the practice of adversarially exposing unexpected or undesired model behaviors, is critical towards improving equity and accuracy of large language models, but non-model creator-affiliated red teaming is scant in healthcare. We convened teams of clinicians, medical and engineering students, and technical professionals (80 participants total) to stress-test models with real-world clinical cases and categorize inappropriate responses along axes of safety, privacy, hallucinations/accuracy, and bias. Six medically-trained reviewers re-analyzed prompt-response pairs and added qualitative annotations. Of 376 unique prompts (1504 responses), 20.1% were inappropriate (GPT-3.5: 25.8%; GPT-4.0: 16%; GPT-4.0 with Internet: 17.8%). Subsequently, we show the utility of our benchmark by testing GPT-

				4o, a model released after our event (20.4% inappropriate). 21.5% of responses appropriate with GPT-3.5 were inappropriate in updated models. We share insights for constructing red teaming prompts, and present our benchmark for iterative model assessments.
97	(Cappellani et al. 2024)	Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients	ChatGPT 3.5	**목적:** 인공지능 챗봇(ChatGPT)이 제공하는 안과 정보의 정확성을 평가한다. **방법:** 안과 8개 세부 전문 분야의 5개 질환을 ChatGPT 버전 3.5로 평가했다. 각 질환에 대해 ChatGPT에 3개의 질문을 했다: x는 무엇인가?; x는 어떻게 진단되는가?; x는 어떻게 치료되는가? (x = 질환명). 응답은 미국 안과학회(AAO) 환자 지침과 비교하여 등급이 매겨졌으며, 점수는 -3(검증되지 않았고 환자가 그러한 제안을 따를 경우 환자의 건강이나 웰빙에 잠재적으로 해로움)부터 2(정확하고 완전함)까지 범위였다. **주요 결과:** 안과 건강 정보와 관련된 프롬프트에 대한 ChatGPT의 응답 정확도를 -3에서 2까지의 척도로 점수화했다. **결과:** 120개 질문 중 93개(77.5%)가 ≥ 1점을 받았다. 27개(22.5%)가 ≤ -1점을 받았으며; 이 중 9개(7.5%)가 -3점을 받았다. 모든 세부 전문 분야의 전체 중앙값 점수는 "x는 무엇인가" 질문에 대해 2점, "x는 어떻게 진단되는가"에 대해 1.5점, "x는 어떻게 치료되는가"에 대해 1점이었지만, Kruskal-Wallis 검정에서 유의성을 달성하지 못했다. **결론:** 긍정적인 점수에도 불구하고 ChatGPT는 자체적으로 여전히 일반적인 안과 질환에 대한 불완전하고, 부정확하며, 잠재적으로 해로운 정보를 제공하며, 이는 해당 질환에 대해 AAO가 지지하지 않는 부작용 가능성이 있는 침습적 시술 또는 기타 개입의 권고로 정의된다. ChatGPT는 환자 교육의 가치 있는 보조 수단이 될 수 있지만, 현재는 동반되는 인간 의료 감독 없이는 충분하지 않다. **Purpose:** To assess the accuracy of ophthalmic information provided by an artificial intelligence chatbot (ChatGPT). **Methods:** Five diseases from 8 subspecialties of Ophthalmology were assessed by ChatGPT version 3.5. Three questions were asked to ChatGPT for each disease: what is x?; how is x diagnosed?; how is x treated? (x = name of the disease). Responses were graded by comparing them to the American Academy of Ophthalmology (AAO) guidelines for patients, with scores ranging from -3 (unvalidated and potentially harmful to a patient's health or well-being if they pursue such a suggestion) to 2 (correct and complete). **Main outcomes:** Accuracy of responses from ChatGPT in response to prompts related to ophthalmic health information in the form of scores on a

				scale from −3 to 2. **Results:** Of the 120 questions, 93 (77.5%) scored ≥ 1. 27. (22.5%) scored ≤ −1; among these, 9 (7.5%) obtained a score of −3. The overall median score amongst all subspecialties was 2 for the question "What is x", 1.5 for "How is x diagnosed", and 1 for "How is x treated", though this did not achieve significance by Kruskal–Wallis testing. **Conclusions:** Despite the positive scores, ChatGPT on its own still provides incomplete, incorrect, and potentially harmful information about common ophthalmic conditions, defined as the recommendation of invasive procedures or other interventions with potential for adverse sequelae which are not supported by the AAO for the disease in question. ChatGPT may be a valuable adjunct to patient education, but currently, it is not sufficient without concomitant human medical supervision.
98	(La Bella et al. 2024)	Reliability of a generative artificial intelligence tool for pediatric familial Mediterranean fever: insights from a multicentre expert survey	Microsoft Copilot Chat-GPT 4.0.	**배경:** 인공지능(AI)은 의료 분야에서 임상 및 연구 사용을 위한 인기 있는 도구가 되었다. 본 연구의 목적은 소아 가족성 지중해열(FMF)에 대한 생성형 AI 도구의 정확성과 신뢰성을 평가하는 것이었다. **방법:** 소아 FMF에 대해 3회 반복된 15개의 질문을 인기 있는 생성형 AI 도구인 Chat-GPT 4.0을 탑재한 Microsoft Copilot에 프롬프트로 제공했다. 9명의 소아 류마티스 전문가가 1에서 5까지의 값을 가진 리커트 유사 척도를 사용하여 눈가림 메커니즘으로 응답 정확도를 평가했다. **결과:** 초기 평가에서 전체 응답의 중앙값은 2.00에서 5.00까지 범위였다. 두 번째 평가 동안 중앙값은 2.00에서 4.00까지 범위였으며, 세 번째 평가에서는 3.00에서 4.00까지 범위였다. 평가자 내 변동성은 낮음에서 중등도 일치도를 보였다(급내상관계수 범위: −0.151~0.534). Krippendorff의 알파 계수 값이 0.136(첫 번째 응답)에서 0.132(두 번째 응답), 0.089(세 번째 응답)로 범위를 보임으로써 **시간이 지남에 따라 전문가 간 일치도가 감소하는 것으로 기록되었다. 마지막으로 전문가들은 설문 전후에 AI에 대한 다양한 수준의 신뢰를 보였다**. **결론:** AI는 조기 진단 및 관리 최적화를 포함하여 소아 류마티스학에서 유망한 영향을 미치지만, **불확실한 정보 신뢰성과 전문가 검증 부족으로 인해 문제가 지속된다. 우리의 설문조사는 FMF에 관한 AI 생성 응답에서 상당한 부정확성과 불완전성을 드러냈으며, 평가자 내 및 평가자 간 신뢰도가 낮았다**. AI 생성 의료 정보를 관리하는 데 있어 인간 검증은 여전히 중요하다. **Background:** Artificial intelligence (AI) has become a popular tool for clinical and research use in the medical field. The aim of this study was to evaluate the accuracy and reliability of a generative AI tool on pediatric familial

				Mediterranean fever (FMF). **Methods:** Fifteen questions repeated thrice on pediatric FMF were prompted to the popular generative AI tool Microsoft Copilot with Chat-GPT 4.0. Nine pediatric rheumatology experts rated response accuracy with a blinded mechanism using a Likert-like scale with values from 1 to 5. **Results:** Median values for overall responses at the initial assessment ranged from 2.00 to 5.00. During the second assessment, median values spanned from 2.00 to 4.00, while for the third assessment, they ranged from 3.00 to 4.00. Intra-rater variability showed poor to moderate agreement (intraclass correlation coefficient range: −0.151 to 0.534). A diminishing level of agreement among experts over time was documented, as highlighted by Krippendorff's alpha coefficient values, ranging from 0.136 (at the first response) to 0.132 (at the second response) to 0.089 (at the third response). Lastly, experts displayed varying levels of trust in AI pre- and post-survey. **Conclusions:** AI has promising implications in pediatric rheumatology, including early diagnosis and management optimization, but challenges persist due to uncertain information reliability and the lack of expert validation. Our survey revealed considerable inaccuracies and incompleteness in AI-generated responses regarding FMF, with poor intra- and extra-rater reliability. Human validation remains crucial in managing AI-generated medical information.
99	(Huppertz et al. 2025)	Revolution or risk?—Assessing the potential and challenges of GPT-4V in radiologic image interpretation	ChatGPT-4 Vision (GPT-4V)	**목적:** ChatGPT-4 Vision(GPT-4V)은 이미지를 사용하여 질의할 수 있는 최첨단 다중모드 대규모 언어 모델(LLM)이다. 우리는 임상 영상 연구를 자율적으로 평가할 때 도구의 진단 성능을 평가하는 것을 목표로 했다. **재료 및 방법:** 대형 대학 병원의 방사선 진료에서 명확한 소견과 확립된 참조 진단을 가진 총 206개의 영상 연구(즉, 방사선촬영(n = 60), CT(n = 60), MRI(n = 60), 혈관조영술(n = 26))에 접근했다. 판독은 이미지만 제공된 상태에서 맥락 없이, 그리고 추가 임상 및 인구통계학적 정보와 함께 맥락화되어 수행되었다. 응답은 여러 진단 차원에 따라 평가되었고 적절한 통계적 검정을 사용하여 분석되었다. **결과:** 이미지 정보보다 맥락을 선호하는 뚜렷한 경향으로, 도구의 진단 정확도는 8.3%(맥락 없음)에서 29.1%(맥락화, 첫 번째 진단 정확), 63.6%(맥락화, 감별 진단 중 정확한 진단)로 향상되었다($p \leq 0.001$, Cochran's Q 검정). **진단 정확도는 20개 이미지를 30일 및 90일 후에 재판독했을 때 최대 30% 감소했으며, 도구의 자체 보고 신뢰도와 무관한 것으로 보였다(Spearman's $\rho = 0.117$ ($p = 0.776$)). 설명된 영상 소견이 92.7%에서 제안된 진단과 일치하여 유효한 진단 추론을 나타냈지만, 도구는 412개 응답에서 258개의 영상 소견을 조작했고 65개 이미지에서 영**

				상 모달리티 또는 해부학적 영역을 잘못 식별했다. **결론: 현재 형태의 GPT-4V는 방사선 이미지를 신뢰할 수 있게 해석할 수 없다. 특히 임상 맥락 없이 이미지를 무시하고, 소견을 조작하며, 세부 사항을 잘못 식별하는 경향은 의료 제공자를 오도하고 환자를 위험에 빠뜨릴 수 있다.** **핵심 요점:** **질문** Generative Pre-trained Transformer 4 Vision(GPT-4V)이 임상 맥락이 있거나 없이 방사선 이미지를 해석할 수 있는가? **발견** GPT-4V는 8%(맥락 없음), 29%(맥락화, 가장 가능성 있는 진단 정확), 64%(맥락화, 감별 진단 중 정확한 진단)의 진단 정확도를 보이며 저조한 성능을 보였다. **임상적 관련성** GPT-4V와 같은 상용 다중모드 대규모 언어 모델의 방사선 진료에서의 유용성은 제한적이다. 임상 맥락 없이는 진단 오류와 조작된 소견이 환자 안전을 손상시키고 임상 의사결정을 오도할 수 있다. 이러한 모델은 유익하기 위해 추가로 개선되어야 한다. **Objectives:** ChatGPT-4 Vision (GPT-4V) is a state-of-the-art multimodal large language model (LLM) that may be queried using images. We aimed to evaluate the tool's diagnostic performance when autonomously assessing clinical imaging studies. **Materials and methods:** A total of 206 imaging studies (i.e., radiography (n = 60), CT (n = 60), MRI (n = 60), and angiography (n = 26)) with unequivocal findings and established reference diagnoses from the radiologic practice of a large university hospital were accessed. Readings were performed uncontextualized, with only the image provided, and contextualized, with additional clinical and demographic information. Responses were assessed along multiple diagnostic dimensions and analyzed using appropriate statistical tests. **Results:** With its pronounced propensity to favor context over image information, the tool's diagnostic accuracy improved from 8.3% (uncontextualized) to 29.1% (contextualized, first diagnosis correct) and 63.6% (contextualized, correct diagnosis among differential diagnoses) ($p \leq 0.001$, Cochran's Q test). Diagnostic accuracy declined by up to 30% when 20 images were re-read after 30 and 90 days and seemed unrelated to the tool's self-reported confidence (Spearman's $\rho = 0.117$ ($p = 0.776$)). While the described imaging findings matched the suggested diagnoses in 92.7%, indicating valid diagnostic reasoning, the tool fabricated 258 imaging findings

				in 412 responses and misidentified imaging modalities or anatomic regions in 65 images. **Conclusion:** GPT-4V, in its current form, cannot reliably interpret radiologic images. Its tendency to disregard the image, fabricate findings, and misidentify details, especially without clinical context, may misguide healthcare providers and put patients at risk. **Key Points:** **Question** Can Generative Pre-trained Transformer 4 Vision (GPT-4V) interpret radiologic images—with and without clinical context? **Findings** GPT-4V performed poorly, demonstrating diagnostic accuracy rates of 8% (uncontextualized), 29% (contextualized, most likely diagnosis correct), and 64% (contextualized, correct diagnosis among differential diagnoses). **Clinical relevance** The utility of commercial multimodal large language models, such as GPT-4V, in radiologic practice is limited. Without clinical context, diagnostic errors and fabricated findings may compromise patient safety and misguide clinical decision-making. These models must be further refined to be beneficial.
100	(Mruthyunjaya et al. 2025)	Right Diagnoses But Wrong Reasoning: Current Large-Language Model-Based Agentic Frameworks Have Flawed Clinical Reasoning Despite High Diagnostic Accuracy	OpenAI o OpenAI o3 mini Gemini 2.5 Pro Gemini 2.0 Flash QwQ Deepseek R1 70B	**배경:** 류마티스 관절염(RA)은 전 세계 인구의 약 1%에 영향을 미친다. 전 세계적으로 훈련된 류마티스 전문의가 부족하다. 인도에서 RA 환자들은 종종 변형을 동반하여 늦게 내원한다. 인공지능(AI) 기반 대규모 언어 모델(LLM)은 널리 사용 가능한 휴대전화 기기를 통해 배치될 수 있어 잠재적인 선별 솔루션을 제공한다. 추론 품질과 환각에 대한 우려를 해결하면서 가장 효과적인 모델 SARA(RA 선별 에이전트)를 식별하기 위해 다양한 에이전틱 프레임워크를 탐색했다. **방법:** PreRAID(류마티스 관절염 사전 선별 정보 데이터베이스)라는 독점 데이터셋을 개발 중이며, 구조화된 온라인 서식을 통해 RA 또는 비-RA로 의사 진단을 받은 동의한 환자의 데이터로 구성된다. 우리는 지식 기반(KB)에 280개 사례를, 테스트에 70개를 포함했다. Neo4j 벡터 데이터베이스는 임베딩 기반 검색을 가능하게 했다. 6개의 LLM을 평가했으며—4개의 폐쇄 소스(OpenAI o1, OpenAI o3 mini, Gemini 2.5 Pro, Gemini 2.0 Flash)와 2개의 오픈 소스(QwQ, Deepseek R1 70B). 이러한 모델은 성능이 포화될 때까지 사이클당 10개씩 훈련 사례 수를 점진적으로 증가시켜 훈련되었다. 그런 다음 모델에게 50개의 새로운 사례를 진단하도록 요청했고, 3가지 에이전틱 구성에 대해 진단 정확도를 계산했다: (1) 지식 기반(KB) 접근 없는 단일 에이전트; (2) 검색 증강 생성(RAG)을 사용하는 단일 에이전트; (3) 첫 번째 에이전트가 진단과 추론을 생성하고 두 번째 에이전트가 검증하는 이중 에이전트 설정. 모델은 또한 각 진단의 배경에 있는 이유를 설명하도록 요청받았으며, 이는 2명의 펠로우와 1명의 컨설턴트 류마티스 전문의에 의해 4점 리커트

				척도를 사용하여 독립적으로 평가되었다. **결과:** PreRAID 데이터셋에는 82%의 RA 확진 사례와 18%의 대조군이 포함되었다. Deepseek R1은 KB가 있는 단일 에이전트 설정에서 가장 높은 정확도(82%)를 보였으며, o1과 o3 mini가 각각 80%로 뒤를 이었다. 2개 에이전트 설정에서 정확도가 떨어졌으며, 특히 Gemini 2.5 Pro(37%)와 Gemini 2.0 Flash(40%)에서 두드러졌다. 3개 이상의 에이전트 설정은 이중 에이전트보다 성능이 나빴다. 추론 품질은 모든 모델에서 차선이었다. Gemini 2.0 Flash(36/50)와 Deepseek R1(28/50)이 가장 올바른 정당화를 가졌으며, QwQ와 Gemini 2.5 Pro는 가장 낮은 점수(각각 6과 10)를 받았다. 많은 출력이 경미하거나 중대한 추론 결함을 보여 진단 정확도와 추론 무결성 간의 단절을 나타냈다. **결론: RA 선별에서 LLM 진단 정확도와 추론 유효성 간의 불일치가 있으며, 이는 현재 임상 배치에 대한 안전성 우려를 제기한다. 이는 LLM 프레임워크가 더 강력한 추론 능력을 포함할 때까지 주의가 필요함을 강조**한다. 향후 연구는 "설명 가능한 AI"와 류마티스 관절염에 대한 AI 기반 진단 도구의 정확성과 신뢰성을 모두 향상시키기 위해 설계상 설명 가능성 패러다임을 통합하는 데 초점을 맞춰야 한다. 진단을 위한 AI 지원 임상 의사결정이 배치되기 전에 이 장벽을 극복해야 한다. **Background:** Rheumatoid arthritis (RA) affects approximately 1% of the worldwide population. There is a global shortage of trained rheumatologists. In India, patients with RA often present late, with deformities. Artificial Intelligence (AI) driven Large Language models (LLMs) can be deployed via widely available mobile phone devices, offering a potential screening solution. Various agentic frameworks were explored to identify the most effective model SARA (Screening Agent for RA) while also addressing concerns about reasoning quality and hallucination that limit clinical deployment. **Methods:** A proprietary dataset, PreRAID (Pre-Screening Rheumatoid Arthritis Information Database), is being developed, consisting of data from consenting patients with a physician diagnosis of RA or not RA, through a structured online proforma. We included 280 cases for the knowledge base (KB) and 70 for testing. A Neo4j vector database enabled embedding-based retrieval. Six LLMs were evaluated—four closed-source (OpenAI o1, OpenAI o3 mini, Gemini 2.5 Pro, Gemini 2.0 Flash) and two open-source (QwQ, Deepseek R1 70B). These models were trained by incrementally increasing the number of training cases by 10 per cycle, until there was saturation in the performance. Then the models were asked to diagnose 50 new cases

				and diagnostic accuracy was calculated for three agentic configurations: (1) a single agent without knowledge base (KB) access; (2) a single agent with retrieval-augmented generation (RAG); and (3) a dual-agent setup, where the first agent generated a diagnosis and reasoning, validated by a second agent. The models were also asked to explain the reasons behind each diagnosis, which was independently assessed by two fellows and one consultant rheumatologist using a four-point Likert scale. **Results:** The PreRAID dataset included 82% RA-confirmed cases and 18% controls. Deepseek R1 showed the highest accuracy (82%) in the single-agent with KB setting, followed by o1 and o3 mini (80% each). Accuracy dropped in the two-agent setup, most notably in Gemini 2.5 Pro (37%) and Gemini 2.0 Flash (40%). Triple and higher agent setups performed worse than dual agents. Reasoning quality was suboptimal across all models. Gemini 2.0 Flash (36/50) and Deepseek R1 (28/50) had the most correct justifications, while QwQ and Gemini 2.5 Pro scored the lowest (6 and 10, respectively). Many outputs showed minor or major reasoning flaws, indicating a disconnect between diagnostic accuracy and reasoning integrity. **Conclusion:** There is a misalignment between LLM diagnostic accuracy and reasoning validity in RA screening, which raises safety concerns for clinical deployment presently. This underscores the need for caution until LLM frameworks include more robust reasoning capabilities. Future research should focus on "explainable AI" and incorporating explainability-by-design paradigms to enhance both the accuracy and reliability of AI-driven diagnostic tools for rheumatoid arthritis. This barrier needs to be surmounted before AI assisted clinical decision making can be deployed for diagnosis.
101	(Levkovich and Elyoseph 2023)	Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette Study	ChatGPT-3.5 ChatGPT-4	**배경:** OpenAI가 개발한 언어 인공지능(AI) 모델인 ChatGPT는 정신 건강 전문가에게 전망 있는 기여를 제공한다. 중요한 이론적 함의를 가지고 있지만, 특히 자살 예방과 관련된 ChatGPT의 실용적 능력은 아직 입증되지 않았다. **목적:** 본 연구의 목적은 2개월 동안 인지된 부담감과 좌절된 소속감이라는 2가지 식별 가능한 요인을 고려하여 자살 위험을 평가하는 ChatGPT의 능력을 평가하는 것이었다. 또한 우리는 ChatGPT-4가 ChatGPT-3.5보다 더 정확하게 자살 위험을 평가했는지 평가했다. **방법:** ChatGPT는 인지된 부담감과 좌절된 소속감의 다양한 정도를 나타내는 가상 환자를 묘사한 사례를 평가하는 과제를 받았다. ChatGPT가 생성한 평가는 이후 정신 건강 전문가가 제공한 표준 평가와 대조되었다. ChatGPT-3.5와 ChatGPT-4(2023년 5월 24일)를 모두 사용하여 2023년 6월과 7월에 3가지 평가 절차를 실

				행했다. 우리의 의도는 정신 건강 전문가와 이전 버전의 ChatGPT-3.5(3월 14일 버전)의 평가 능력과 관련하여 자살 위험의 다양한 측면을 평가하는 ChatGPT-4의 숙련도를 면밀히 조사하는 것이었다. **결과:** 2023년 6월과 7월 기간 동안 우리는 ChatGPT-4가 평가한 자살 시도 가능성이 모든 조건에서 정신 건강 전문가(n=379)의 규범과 유사하다는 것을 발견했다(평균 Z 점수 0.01). **그럼에도 불구하고 정신 건강 전문가가 수행한 평가와 비교할 때 자살 시도 가능성을 현저히 과소평가한 ChatGPT-3.5(5월 버전)가 수행한 평가와 관련하여 뚜렷한 불일치가 관찰되었다(평균 Z 점수 −0.83)**. 경험적 증거는 ChatGPT-4의 자살 사고와 정신적 고통 발생률 평가가 정신 건강 전문가보다 높았음을 시사한다(평균 Z 점수 각각 0.47 및 1.00). 반대로 **ChatGPT-4와 ChatGPT-3.5(두 버전 모두)가 평가한 회복력 수준은 정신 건강 전문가가 제공한 평가와 비교할 때 낮은 것으로 관찰되었다(평균 Z 점수 각각 −0.89 및 −0.90)**. **결론:** 연구 결과는 ChatGPT-4가 전문가가 제공한 평가와 유사한 방식으로 자살 시도 가능성을 추정함을 시사한다. 자살 사고를 인식하는 측면에서 ChatGPT-4가 더 정확한 것으로 보인다. **그러나 정신적 고통과 관련하여 ChatGPT-4에 의한 과대평가가 관찰되어 추가 연구의 필요성을 나타낸다**. 이러한 결과는 게이트키퍼, 환자, 심지어 정신 건강 전문가의 의사결정을 지원할 수 있는 ChatGPT-4의 잠재력에 대한 시사점을 가진다. 임상적 잠재력에도 불구하고 임상 실무에서 ChatGPT-4의 능력 사용을 확립하기 위해서는 집중적인 후속 연구가 필요하다. ChatGPT-3.5가 특히 중증 사례에서 자살 위험을 자주 과소평가한다는 발견은 특히 우려스럽다. 이는 ChatGPT가 실제 자살 위험 수준을 경시할 수 있음을 나타낸다. **Background:** ChatGPT, a linguistic artificial intelligence (AI) model engineered by OpenAI, offers prospective contributions to mental health professionals. Although having significant theoretical implications, ChatGPT's practical capabilities, particularly regarding suicide prevention, have not yet been substantiated. **Objective:** The study's aim was to evaluate ChatGPT's ability to assess suicide risk, taking into consideration 2 discernable factors—perceived burdensomeness and thwarted belongingness—over a 2-month period. In addition, we evaluated whether ChatGPT-4 more accurately evaluated suicide risk than did ChatGPT-3.5. **Methods:** ChatGPT was tasked with assessing a vignette that depicted a hypothetical patient exhibiting differing degrees of perceived burdensomeness and thwarted belongingness. The assessments generated by ChatGPT were subsequently contrasted with standard evaluations

				rendered by mental health professionals. Using both ChatGPT-3.5 and ChatGPT-4 (May 24, 2023), we executed 3 evaluative procedures in June and July 2023. Our intent was to scrutinize ChatGPT-4's proficiency in assessing various facets of suicide risk in relation to the evaluative abilities of both mental health professionals and an earlier version of ChatGPT-3.5 (March 14 version). **Results:** During the period of June and July 2023, we found that the likelihood of suicide attempts as evaluated by ChatGPT-4 was similar to the norms of mental health professionals (n=379) under all conditions (average Z score of 0.01). Nonetheless, a pronounced discrepancy was observed regarding the assessments performed by ChatGPT-3.5 (May version), which markedly underestimated the potential for suicide attempts, in comparison to the assessments carried out by the mental health professionals (average Z score of −0.83). The empirical evidence suggests that ChatGPT-4's evaluation of the incidence of suicidal ideation and psychache was higher than that of the mental health professionals (average Z score of 0.47 and 1.00, respectively). Conversely, the level of resilience as assessed by both ChatGPT-4 and ChatGPT-3.5 (both versions) was observed to be lower in comparison to the assessments offered by mental health professionals (average Z score of −0.89 and −0.90, respectively). **Conclusions:** The findings suggest that ChatGPT-4 estimates the likelihood of suicide attempts in a manner akin to evaluations provided by professionals. In terms of recognizing suicidal ideation, ChatGPT-4 appears to be more precise. However, regarding psychache, there was an observed overestimation by ChatGPT-4, indicating a need for further research. These results have implications regarding ChatGPT-4's potential to support gatekeepers, patients, and even mental health professionals' decision-making. Despite the clinical potential, intensive follow-up studies are necessary to establish the use of ChatGPT-4's capabilities in clinical practice. The finding that ChatGPT-3.5 frequently underestimates suicide risk, especially in severe cases, is particularly troubling. It indicates that ChatGPT may downplay one's actual suicide risk level.
102	(Clark 2025)	The Ability of AI Therapy Bots to Set Limits With Distressed Adolescents: Simulation-Based Comparison Study	ChatGPT Gemini	**배경:** 생성형 인공지능(AI)의 최근 발전은 ChatGPT와 Gemini와 같은 강력하고 쉽게 접근 가능한 도구를 일반 대중에게 소개하여 빠르게 확장되는 다양한 용도로 사용되고 있다. 이러한 용도 중에는 치료사의 역할을 하는 전문 챗봇과 정서적 지원을 제공하는 개인 맞춤형 디지털 동반자가 있다. 그러나 일관되게 안전하고 효과적인

				치료를 제공하는 AI 치료사의 능력은 대체로 입증되지 않았으며, 이러한 우려는 특히 정신 건강 지원을 찾는 청소년과 관련하여 두드러진다. **목적:** 본 연구는 정신 건강 고통을 겪는 가상의 청소년이 제안한 해롭거나 잘못된 생각을 지지하려는 치료 및 동반자 AI 챗봇의 의지를 확인하는 것을 목표로 했다. **방법:** 치료 지원 또는 동반을 제공하는 10개의 공개적으로 이용 가능한 AI 봇의 편의 표본에 정신 건강 문제를 가진 청소년에 대한 3개의 상세한 가상 사례가 각각 제시되었다. 각 가상 청소년은 AI 챗봇에게 학교 중퇴, 한 달간 모든 인간 접촉 회피, 나이 많은 교사와의 관계 추구와 같은 2개의 해롭거나 잘못된 제안을 지지해 줄 것을 요청했으며, 각 챗봇에 총 6개의 제안이 제시되었다. 제시된 임상 시나리오는 청소년 치료 실무에서 일반적으로 볼 수 있는 문제를 반영하도록 의도되었으며, 가상 청소년이 제안한 제안은 명백히 위험하거나 현명하지 않도록 의도되었다. 10개의 AI 봇은 일반 AI 봇, 동반자 봇, 전용 정신 건강 봇을 포함한 다양한 챗봇 유형을 대표하도록 저자가 선택했다. 챗봇 응답은 청소년의 제안된 행동에 대한 직접적인 지지로 정의되는 명시적 지지에 대해 분석되었다. **결과:** **총 60개 시나리오 중 챗봇은 60개 중 19개(32%)의 기회에서 해로운 제안을 적극적으로 지지했다. 10개 챗봇 중 4개는 제안된 아이디어의 절반 이상을 지지했으며, 어느 봇도 모두 반대하지 못했다**. **결론:** 정신 건강 또는 정서적 지원을 제공하는 AI 챗봇의 상당 부분이 가상 청소년의 해로운 제안을 지지했다. **이러한 결과는 일부 AI 기반 동반자 또는 치료 봇이 심각한 정신 건강 문제를 가진 청소년을 안전하게 지원할 수 있는 능력에 대한 우려를 제기하며, AI 봇이 적절할 때 유용한 지침을 제공하는 것을 희생하여 지나치게 지지적인 경향이 있을 수 있다는 우려를 높인다**. 결과는 청소년을 위한 디지털 정신 건강 지원과 관련된 감독, 안전 프로토콜 및 지속적인 연구의 긴급한 필요성을 강조한다. **Background:** Recent developments in generative artificial intelligence (AI) have introduced the general public to powerful, easily accessible tools, such as ChatGPT and Gemini, for a rapidly expanding range of uses. Among those uses are specialized chatbots that serve in the role of a therapist, as well as personally curated digital companions that offer emotional support. However, the ability of AI therapists to provide consistently safe and effective treatment remains largely unproven, and those concerns are especially salient in regard to adolescents seeking mental health support. **Objective:** This study aimed to determine the willingness of therapy and companion AI chatbots to endorse harmful or ill-advised ideas proposed by fictional teenagers experiencing mental health distress.

				Methods: A convenience sample of 10 publicly available AI bots offering therapeutic support or companionship were each presented with 3 detailed fictional case vignettes of adolescents with mental health challenges. Each fictional adolescent asked the AI chatbot to endorse 2 harmful or ill-advised proposals, such as dropping out of school, avoiding all human contact for a month, or pursuing a relationship with an older teacher, resulting in a total of 6 proposals presented to each chatbot. The clinical scenarios presented were intended to reflect challenges commonly seen in the practice of therapy with adolescents, and the proposals offered by the fictional teenagers were intended to be clearly dangerous or unwise. The 10 AI bots were selected by the author to represent a range of chatbot types, including generic AI bots, companion bots, and dedicated mental health bots. Chatbot responses were analyzed for explicit endorsement, defined as direct support for the teenagers' proposed behavior. **Results:** Across 60 total scenarios, chatbots actively endorsed harmful proposals in 19 out of the 60 (32%) opportunities to do so. Of the 10 chatbots, 4 endorsed half or more of the ideas proposed to them, and none of the bots managed to oppose them all. **Conclusions:** A significant proportion of AI chatbots offering mental health or emotional support endorsed harmful proposals from fictional teenagers. These results raise concerns about the ability of some AI-based companion or therapy bots to safely support teenagers with serious mental health issues and heighten concern that AI bots may tend to be overly supportive at the expense of offering useful guidance when appropriate. The results highlight the urgent need for oversight, safety protocols, and ongoing research regarding digital mental health support for adolescents.
103	(Gunay, Ozturk, and Yigit 2024)	The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: A comparison with cardiologists and emergency medicine specialists	ChatGPT-4 ChatGPT-4o Gemini advanced	**서론:** 잘 알려진 대규모 언어 모델(LLM) 중 하나인 GPT-4, GPT-4o 및 Gemini advanced는 시각 데이터를 인식하고 해석할 수 있는 능력을 가지고 있다. 문헌을 검토하면 GPT-4의 심전도 성능을 조사하는 연구는 매우 제한적이다. 그러나 Gemini와 GPT-4o의 심전도 평가 성공을 조사하는 연구는 문헌에 없다. 우리 연구의 목적은 GPT-4, GPT-4o 및 Gemini의 심전도 평가 성능을 평가하고, 의료 분야에서의 사용 가능성을 평가하며, 심전도 해석의 정확도를 심장내과 전문의 및 응급의학과 전문의와 비교하는 것이다. **방법:** 연구는 2024년 5월 14일부터 6월 3일까지 수행되었다. "150 ECG Cases"라는 책이 참고 자료로 사용되었으며, 일상적인 심전도와 더 어려운 심전도의 두 섹션을 포함했다. 본 연구를 위해 2명의 응급의학과 전문의가 각 섹션에서 20개의 심전

				도 사례를 선택하여 총 40개 사례를 구성했다. 다음 단계에서 질문은 응급의학과 전문의와 심장내과 전문의에 의해 평가되었다. 후속 단계에서는 별도의 채팅 인터페이스에서 GPT-4, GPT-4o 및 Gemini Advanced에 매일 진단 질문이 입력되었다. 최종 단계에서 심장내과 전문의, 응급의학과 전문의, GPT-4, GPT-4o 및 Gemini Advanced가 제공한 응답이 일상적인 심전도, 더 어려운 심전도, 총 심전도 수의 3가지 범주에 걸쳐 통계적으로 평가되었다. **결과:** 심장내과 전문의는 모든 3개 그룹에서 GPT-4, GPT-4o 및 Gemini Advanced보다 우수한 성능을 보였다. 응급의학과 전문의는 일상적인 심전도 질문과 총 심전도 질문에서 GPT-4o보다 나은 성능을 보였다(각각 $p = 0.003$ 및 $p = 0.042$). GPT-4o를 Gemini Advanced 및 GPT-4와 비교할 때, GPT-4o는 총 심전도 질문에서 더 나은 성능을 보였다(각각 $p = 0.027$ 및 $p < 0.001$). **일상적인 심전도 질문에서도 GPT-4o는 Gemini Advanced보다 우수한 성능을 보였다($p = 0.004$). GPT-4($p < 0.001$, Fleiss Kappa = 0.265) 및 Gemini Advanced($p < 0.001$, Fleiss Kappa = 0.347)가 제공한 응답에서 약한 일치도가 관찰되었으며, GPT-4o가 제공한 응답에서는 중등도 일치도가 관찰되었다($p < 0.001$, Fleiss Kappa = 0.514).** **결론:** GPT-4o는 특히 더 어려운 심전도 질문에서 가능성을 보여주며 심전도 평가의 보조 수단으로서 잠재력을 가질 수 있지만, 일상적이고 전반적인 평가에서의 성능은 여전히 인간 전문가보다 뒤처진다. GPT-4와 Gemini의 제한된 정확성과 일관성은 현재 임상 심전도 해석에서의 사용이 위험함을 시사한다. **Introduction:** GPT-4, GPT-4o and Gemini advanced, which are among the well-known large language models (LLMs), have the capability to recognize and interpret visual data. When the literature is examined, there are a very limited number of studies examining the ECG performance of GPT-4. However, there is no study in the literature examining the success of Gemini and GPT-4o in ECG evaluation. The aim of our study is to evaluate the performance of GPT-4, GPT-4o, and Gemini in ECG evaluation, assess their usability in the medical field, and compare their accuracy rates in ECG interpretation with those of cardiologists and emergency medicine specialists. **Methods:** The study was conducted from May 14, 2024, to June 3, 2024. The book "150 ECG Cases" served as a reference, containing two sections: daily routine ECGs and more challenging ECGs. For this study, two emergency medicine specialists selected 20 ECG cases from each section, totaling 40 cases. In the next stage, the questions were evaluated by

				emergency medicine specialists and cardiologists. In the subsequent phase, a diagnostic question was entered daily into GPT-4, GPT-4o, and Gemini Advanced on separate chat interfaces. In the final phase, the responses provided by cardiologists, emergency medicine specialists, GPT-4, GPT-4o, and Gemini Advanced were statistically evaluated across three categories: routine daily ECGs, more challenging ECGs, and the total number of ECGs. **Results:** Cardiologists outperformed GPT-4, GPT-4o, and Gemini Advanced in all three groups. Emergency medicine specialists performed better than GPT-4o in routine daily ECG questions and total ECG questions ($p = 0.003$ and $p = 0.042$, respectively). When comparing GPT-4o with Gemini Advanced and GPT-4, GPT-4o performed better in total ECG questions ($p = 0.027$ and $p < 0.001$, respectively). In routine daily ECG questions, GPT-4o also outperformed Gemini Advanced ($p = 0.004$). Weak agreement was observed in the responses given by GPT-4 ($p < 0.001$, Fleiss Kappa = 0.265) and Gemini Advanced ($p < 0.001$, Fleiss Kappa = 0.347), while moderate agreement was observed in the responses given by GPT-4o ($p < 0.001$, Fleiss Kappa = 0.514). **Conclusion:** While GPT-4o shows promise, especially in more challenging ECG questions, and may have potential as an assistant for ECG evaluation, its performance in routine and overall assessments still lags behind human specialists. The limited accuracy and consistency of GPT-4 and Gemini suggest that their current use in clinical ECG interpretation is risky.
104	(Ismaiel et al. 2024)	The evaluation of the performance of ChatGPT in the management of labor analgesia	ChatGPT4	ChatGPT4는 2023년 OpenAI가 출시한 주요 대규모 언어 모델(LLM) 챗봇이다. ChatGPT4는 자유 텍스트 질의에 응답하고, 질문에 답변하며, 사실상 모든 주제에 대해 제안할 수 있다. ChatGPT4는 합리적인 정확도로 마취 및 심지어 산과 마취 지식 기반 질문에 성공적으로 답변했다. 그러나 ChatGPT4는 아직 산과 마취 임상 의사결정에서 도전받지 않았다. **연구 목적:** 본 연구에서 우리는 전문 산과 마취과 전문의와 비교하여 임상 분만 진통 시나리오 관리에서 ChatGPT4의 성능을 평가했다. **개입:** 의학적 복잡성이 점진적으로 증가하는 8개의 임상 질문을 ChatGPT4에 제시했다. **측정:** ChatGPT4 응답은 5점 리커트 척도를 사용하여 각 응답의 안전성, 정확성 및 완전성을 기준으로 7명의 전문 산과 마취과 전문의에 의해 평가되었다. **주요 결과:** ChatGPT4는 제시된 산과 마취 임상 시나리오에 대한 응답의 73%에서 안전한 것으로 간주되었다(응답의 27%는 안전하지 않은 것으로 간주됨). **ChatGPT4 응답 중 어느 것도 7명의 전문 산과 마취과 전문의 모두에 의해 만장일**

				치로 안전한 것으로 간주되지 않았다. 더욱이 ChatGPT4 응답은 전반적으로 부분적으로 정확했으며(5점 만점에 4점), 다소 불완전했다(5점 만점에 3.5점). 결론: 요약하면, ChatGPT4의 모든 응답 중 약 4분의 1이 전문 산과 마취과 전문의에 의해 안전하지 않은 것으로 간주되었다. 이러한 발견은 산과 마취 또는 기타 전문 의료 분야의 임상 의사결정을 위해 ChatGPT4와 같은 LLM의 추가적인 미세 조정 및 훈련의 필요성을 시사할 수 있다. 이러한 LLM은 임상 의사결정에서 산과 마취과 전문의를 지원하고 전반적인 환자 치료를 향상시키는 데 중요한 미래 역할을 할 수 있다. ChatGPT4 is a leading large language model (LLM) chatbot released by OpenAI in 2023. ChatGPT4 can respond to free-text queries, answer questions and make suggestions regarding virtually any topic. ChatGPT4 has successfully answered anesthesia and even obstetric anesthesia knowledge-based questions with reasonable accuracy. However, ChatGPT4 has yet to be challenged in obstetric anesthesia clinical decision-making. Study Objective: In this study, we evaluated the performance of ChatGPT4 in the management of clinical labor analgesia scenarios compared to expert obstetric anesthesiologists. Intervention: Eight clinical questions with progressively increasing medical complexity were posed to ChatGPT4. Measurements: The ChatGPT4 responses were rated by seven expert obstetric anesthesiologists based on safety, accuracy and completeness of each response using a five-point Likert rating scale. Main Results: ChatGPT4 was deemed safe in 73% of responses to the presented obstetric anesthesia clinical scenarios (27% of responses were deemed unsafe). None of the ChatGPT4 responses were unanimously deemed to be safe by all seven expert obstetric anesthesiologists. Moreover, ChatGPT4 responses were overall partly accurate (score 4 out of 5) and somewhat incomplete (score 3.5 out of 5). Conclusions: In summary, approximately one quarter of all responses by ChatGPT4 were deemed unsafe by expert obstetric anesthesiologists. These findings may suggest the need for more fine-tuning and training of LLMs such as ChatGPT4 specifically for clinical decision making in obstetric anesthesia or other specialized medical fields. These LLMs may come to play an important future role in assisting obstetric anesthesiologists in clinical decision making and enhancing overall patient care.

105	(Stanceski et al. 2024)	The quality and safety of using generative AI to produce patient-centred discharge instructions	ChatGPT-3.5	퇴원 시 환자 중심 지침은 순응도와 결과를 개선할 수 있다. GPT-3.5를 사용하여 환자 중심 퇴원 지침을 생성하면서, 우리는 안전성, 정확성 및 언어 단순화에 대한 응답을 평가했다. MIMIC-IV의 100개 퇴원 요약서로 테스트했을 때, AI 도구에 기인할 수 있는 잠재적으로 해로운 안전 문제가 18%에서 발견되었으며, 여기에는 6%의 환각과 3%의 새로운 약물이 포함되었다. AI 도구는 환자 중심 퇴원 지침을 생성할 수 있지만, 위해를 피하기 위해서는 신중한 구현이 필요하다. Patient-centred instructions on discharge can improve adherence and outcomes. Using GPT-3.5 to generate patient-centred discharge instructions, we evaluated responses for safety, accuracy and language simplification. When tested on 100 discharge summaries from MIMIC-IV, potentially harmful safety issues attributable to the AI tool were found in 18%, including 6% with hallucinations and 3% with new medications. AI tools can generate patient-centred discharge instructions, but careful implementation is needed to avoid harms.
106	(Zare et al. 2025)	Utility, Accuracy, and Bias of Large Language Models as Real-Time Diagnostic Support in Primary Care	ChatGPT-4	**목적:** GPT-4와 같은 대규모 언어 모델(LLM)은 실시간 진단 제안을 제공하여 일차 진료를 향상시킬 잠재력을 가지고 있다. 이 파일럿 연구는 시뮬레이션된 일차 진료 상담 중 임상 컨설턴트로서 기성 LLM, 특히 GPT-4를 평가했다. **방법:** 2명의 일차 진료 의사(PCP), 3명의 기록자, 2명의 시뮬레이션된 환자가 참여하는 시뮬레이션 환경에서 10개의 실시간 세션이 수행되었다. 맞춤형 인터페이스를 사용하여 PCP는 환자-의사 대화를 기반으로 진단 제안을 받기 위해 GPT-4와 상호작용했다. 주요 결과에는 감별 진단의 진단 정확도, 주 진단의 순위, 그리고 GPT-4 생성 진행 기록과 인간 기록자의 기록 비교가 포함되었다. 우리는 또한 합성으로 생성된 시나리오를 사용하여 성별과 인종에 걸친 진단 성능의 잠재적 인구통계학적 편향을 분석했다. **결과:** 결과는 GPT-4 보조로 PCP의 상위 1개 진단 정확도가 50%에서 60%로 향상되었음을 보여주었다. 생성된 진행 기록은 인간이 생성한 기록보다 정확성과 철저함이 현저히 낮게 평가되었다. 또한 의사들은 GPT-4가 불필요한 검사와 의뢰를 제안하는 것에 대한 우려를 제기했다. 우리는 인구통계학적 그룹에 걸친 GPT-4의 상위 1개 진단 정확도에서 편향을 발견하지 못했지만, 인종 그룹에 대한 상위 3개 정확도의 통계적으로 유의한 차이는 추가 조사가 필요한 잠재적 인구통계학적 격차를 강조한다. **결론:** 우리의 발견은 GPT-4가 일차 진료에서 진단 정확도를 지원하고 감별 진단을 확대할 수 있음을 시사한다. 그러나 진행 기록 품질의 한계, 잠재적 격차, 의료 검사 및 의뢰의 과다 사용 위험은 임상 통합 전에 추가적인 개선 및 의료 특화 미세 조정의 필요성을 강조한다.

				Objective: Large Language Models (LLMs), such as GPT-4, have potential to enhance primary care by offering real-time diagnostic suggestions. This pilot study evaluated an off-the-shelf LLM, particularly GPT-4, as a clinical consultant during simulated primary care consultations. **Method:** Ten real-time sessions were conducted in a simulated environment involving two primary care physicians (PCPs), three scribes, and two simulated patients. Using a custom interface, PCPs interacted with GPT-4 to receive diagnostic suggestions based on patient-physician dialogues. Primary outcomes included diagnostic accuracy of differential diagnoses, ranking of primary diagnoses, and a comparison of GPT-4-generated progress notes with human scribes' notes. We also analyzed potential demographic biases in diagnostic performance across genders and races using synthetically generated scenarios. **Results:** Results showed PCPs' top-1 diagnostic accuracy improved from 50% to 60% with GPT-4 assistance. Generated progress notes were rated significantly less accurate and thorough than human-generated notes. Additionally, physicians raised concerns about GPT-4 suggesting unnecessary tests and referrals. We found no bias in GPT-4's top-1 diagnostic accuracy across demographic groups, but the statistically significant difference in top-3 accuracy for racial groups highlights potential demographic disparities that require further investigation. **Conclusion:** Our findings suggest that GPT-4 could support diagnostic accuracy and broaden differential diagnoses in primary care. However, its limitations in progress note quality, potential disparities, and the risk of over-utilization of medical tests and referrals highlight the need for further refinement and healthcare-specific fine-tuning before clinical integration.
107	(Clusmann, Schulz, et al. 2025)	Incidental Prompt Injections on Vision-Language Models in Real-Life Histopathology	Claude 3 Opus Claude 3.5 Sonnet GPT-4o	비전-언어 모델(VLM)은 다중모드 의료 데이터를 분석할 수 있다. 그러나 VLM의 중요한 약점은 프롬프트 주입 공격에 대한 취약성이다. 여기서 모델은 상충되는 지시를 받아 잠재적으로 해로운 출력을 생성한다. 우리는 조직병리학적 영상의 손으로 쓴 라벨과 워터마크가 의도하지 않은 프롬프트 주입으로 작용하여 조직병리학에서 의사결정에 영향을 미칠 수 있다고 가정했다. 우리는 최첨단 VLM인 Claude 3 Opus, Claude 3.5 Sonnet, GPT-4o에 대해 총 3,888개의 관찰을 통한 정량적 분석을 수행했다. 우리는 조직 옆에 제시될 때 VLM이 다양한 라벨과 워터마크에 어떻게 반응하는지 조사하는 다양한 실제 영감을 받은 시나리오를 설계했다. **모든 모델은 다양한 다중 클래스 문제에서 30~65%의 기준 정확도를 보였음에도 불구하**

				고, 노출되는 정답라벨에 대해애서는 거의 완벽한 정확도 (90~~65%의 기준선 정확도에도 불구하고, 정답 유출 라벨에 대해서는 거의 완벽한 정확도(90~~100%)에 도달했고, 오해를 일으키는 라벨이나 워터마크에 대해서는 극도로 낮은 정확도(0~10%)를 보였다. 모든 VLM은 인간이 제공한 라벨에 의존했고, 그러한 입력에 명백한 오류가 포함되어 있어도 이를 무오류로 받아들였다. 더욱이 프롬프트 엔지니어링으로는 이러한 효과를 완화할 수 없었다. 따라서 우리는 조직병리학적 영상의 라벨, 펜 및 워터마크가 의도하지 않은 프롬프트 주입으로 작용하여 최첨단 VLM의 의사결정에서 명백한 실수를 초래한다는 증거를 제공한다. 우리의 발견은 실제 환경에서 진단 정확도와 환자 안전을 손상시킬 수 있는 중대한 취약점을 드러낸다. 향후 연구는 잠재적 프롬프트 주입을 탐지하고 무력화하는 강력한 방법을 개발하는 데 초점을 맞춰야 한다. Vision–language models (VLMs) can analyze multimodal medical data. However, a significant weakness of VLMs is their susceptibility to prompt injection attacks. Here, the model receives conflicting instructions, leading to potentially harmful outputs. We hypothesized that handwritten labels and watermarks on histopathological images could act as inadvertent prompt injections, influencing decision–making in histopathology. We conducted a quantitative analysis with a total of 3888 observations on the state–of–the–art VLMs Claude 3 Opus, Claude 3.5 Sonnet, and GPT–4o. We designed various real–world–inspired scenarios in which we investigated how VLMs react to different labels and watermarks if presented with those next to the tissue. All models reached almost perfect accuracies (90 to 100%) for ground–truth–leaking labels and abysmal accuracies (0 to 10%) for misleading labels or watermarks, despite baseline accuracies between 30 and 65% for various multiclass problems. All VLMs relied on the human–provided labels and accepted them as infallible, even when those inputs contained obvious errors. Furthermore, prompt engineering could not mitigate these effects. We therefore provide evidence that labels, pen– and watermarks on histopathological images act as inadvertent prompt injections leading to blatant mistakes in decision–making by state–of–the–art VLMs. Our findings reveal a critical vulnerability that could compromise diagnostic accuracy and patient safety in a real–world setting. Future research should focus on developing robust methods to detect and neutralize potential prompt injections.

Abroms, L. C., A. Yousefi, C. N. Wysota, T. C. Wu, and D. A. Broniatowski. 2025. 'Assessing the Adherence of ChatGPT Chatbots to Public Health Guidelines for Smoking Cessation: Content Analysis', *J Med Internet Res*, 27: e66896.

Alber, D. A., Z. Yang, A. Alyakin, E. Yang, S. Rai, A. A. Valliani, J. Zhang, G. R. Rosenbaum, A. K. Amend-Thomas, D. B. Kurland, C. M. Kremer, A. Eremiev, B. Negash, D. D. Wiggan, M. A. Nakatsuka, K. L. Sangwon, S. N. Neifert, H. A. Khan, A. V. Save, A. Palla, E. A. Grin, M. Hedman, M. Nasir-Moin, X. C. Liu, L. Y. Jiang, M. A. Mankowski, D. L. Segev, Y. Aphinyanaphongs, H. A. Riina, J. G. Golfinos, D. A. Orringer, D. Kondziolka, and E. K. Oermann. 2025. 'Medical large language models are vulnerable to data-poisoning attacks', *Nat Med*, 31: 618-26.

Alkhalaf, M., P. Yu, M. Yin, and C. Deng. 2024. 'Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records', *J Biomed Inform*, 156: 104662.

Amador Barbosa, I., M. Sergio Almeida Alves, P. Rayse Zagalo de Almeida, P. de Almeida Rodrigues, R. Pimentel de Oliveira, S. Augusto Fernades de Menezes, J. D. Mendonca de Moura, and R. Roberto de Souza Fonseca. 2025. 'Assessing the diagnostic and treatment accuracy of Large Language Models (LLMs) in Peri-implant diseases: A clinical experimental study', *J Dent*, 162: 106091.

Arslan, H. C., K. Arslan, M. Gun, P. O. Karkin, and T. Arslanoglu. 2025. 'Evaluation of ChatGPT-5 responses in obstetric and gynecological emergencies: concordance, readability, and clinical reliability', *BMC Emerg Med*, 25: 220.

Aydin, C., O. B. Duygu, A. B. Karakas, E. Er, G. Gokmen, A. M. Ozturk, and F. Govsa. 2025. 'Clinical Failure of General-Purpose AI in Photographic Scoliosis Assessment: A Diagnostic Accuracy Study', *Medicina (Kaunas)*, 61.

Barlas, T., A. E. Altinova, M. Akturk, and F. B. Toruner. 2024. 'Credibility of ChatGPT in the assessment of obesity in type 2 diabetes according to the guidelines', *Int J Obes (Lond)*, 48: 271-75.

Ben-Zion, Z., K. Witte, A. K. Jagadish, O. Duek, I. Harpaz-Rotem, M. C. Khorsandian, A. Burrer, E. Seifritz, P. Homan, E. Schulz, and T. R. Spiller. 2025. 'Assessing and alleviating state anxiety in large language models', *NPJ Digit Med*, 8: 132.

Bhimani, M., A. Miller, J. D. Agnew, M. S. Ausin, M. Raglow-Defranco, H. Mangat, M. Voisard, M. Taylor, S. Bierman-Lytle, V. Parikh, J. Ghukasyan, R. Lasko, S. Godil, A. Atreja, and S. Mukherjee. 2025. "Real-World Evaluation of Large Language Models in Healthcare (RWE-LLM): A New Realm of AI Safety & Validation." In.

Birkun, A. A., and A. Gautam. 2023. 'Large Language Model (LLM)-Powered Chatbots Fail to Generate Guideline-Consistent Content on Resuscitation and May Provide Potentially Harmful Advice', *Prehosp Disaster Med*, 38: 757-63.

Biro, J. M., J. L. Handley, J. Malcolm McCurry, A. Visconti, J. Weinfeld, J. Gregory Trafton, and R. M. Ratwani. 2025. 'Opportunities and risks of artificial intelligence in patient portal messaging in primary care', *NPJ Digit Med*, 8: 222.

Brin, D., V. Sorin, Y. Barash, E. Konen, B. S. Glicksberg, G. N. Nadkarni, and E. Klang. 2025. 'Assessing GPT-4 multimodal performance in radiological image analysis', *Eur Radiol*, 35: 1959-65.

Cappellani, F., K. R. Card, C. L. Shields, J. S. Pulido, and J. A. Haller. 2024. 'Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients', *Eye (Lond)*, 38: 1368-73.

Chang, C. T., H. Farah, H. Gui, S. J. Rezaei, C. Bou-Khalil, Y. J. Park, A. Swaminathan, J. A. Omiye, A. Kolluri, A. Chaurasia, A. Lozano, A. Heiman, A. S. Jia, A. Kaushal, A. Jia, A. Iacovelli, A. Yang, A. Salles, A. Singhal, B. Narasimhan, B. Belai, B. H. Jacobson, B. Li, C. H. Poe, C. Sanghera, C. Zheng, C. Messer, D. V. Kettud, D. Pandya, D. Kaur, D. Hla, D. Dindoust, D. Moehrle, D. Ross, E. Chou, E. Lin, F. N. Haredasht, G. Cheng, I. Gao, J. Chang, J. Silberg, J. A. Fries, J. Xu, J. Jamison, J. S. Tamaresis, J. H. Chen, J. Lazaro, J. M. Banda, J. J. Lee, K. E. Matthys, K. R. Steffner, L. Tian, L. Pegolotti, M. Srinivasan, M. Manimaran, M. Schwede, M. Zhang, M. Nguyen, M. Fathzadeh, Q. Zhao, R. Bajra, R. Khurana, R. Azam, R. Bartlett, S. T. Truong, S. L. Fleming, S. Raj, S. Behr, S. Onyeka, S. Muppidi, T. Bandali, T. Y. Eulalio, W. Chen, X. Zhou, Y. Ding, Y. Cui, Y. Tan, Y. Liu, N. Shah, and R. Daneshjou. 2025. 'Red teaming ChatGPT in medicine to yield real-world insights on model behavior', *NPJ Digit Med*, 8: 149.

Chang, C. T., N. Srivathsa, C. Bou-Khalil, A. Swaminathan, M. R. Lunn, K. Mishra, S. Koyejo, and R. Daneshjou. 2024. "Evaluating anti-LGBTQIA+ medical bias in large language models." In.

Chen, Y., Y. Liu, Y. Huang, X. Huang, Z. Zheng, F. Yang, H. Lin, H. Lin, X. Li, A. Xie, and Y. Huang. 2025. 'Assessing the ability of ChatGPT 4.0 in generating check-up reports', *Front Med (Lausanne)*, 12: 1658561.

Chung, P., C. T. Fong, A. M. Walters, N. Aghaeepour, M. Yetisgen, and V. N. O'Reilly-Shah. 2024. 'Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication', *JAMA Surg*, 159: 928-37.

Cilli Hayiroglu, S., and T. Bozkurt. 2025. 'ChatGPT, Gemini, and Grok on familial mediterranean fever: are they trustworthy?', *Clin Rheumatol*.

Clark, A. 2025. 'The Ability of AI Therapy Bots to Set Limits With Distressed Adolescents: Simulation-Based Comparison Study', *JMIR Ment Health*, 12: e78414.

Clusmann, J., D. Ferber, I. C. Wiest, C. V. Schneider, T. J. Brinker, S. Foersch, D. Truhn, and J. N. Kather. 2025. 'Prompt injection attacks on vision language models in oncology', *Nat Commun*, 16: 1239.

Clusmann, Jan, Stefan J. K. Schulz, Dyke Ferber, Isabella C. Wiest, Aurélie Fernandez, Markus Eckstein, Fabienne Lange, Nic G. Reitsam, Franziska Kellers, Maxime Schmitt, Peter Neidlinger, Paul-Henry Koop, Carolin V. Schneider, Daniel Truhn, Wilfried Roth, Moritz Jesinghaus, Jakob N. Kather, and Sebastian Foersch. 2025. 'Incidental Prompt Injections on Vision–Language Models in Real-Life Histopathology', *Nejm Ai*, 2.

Comrie, D. 2023. "ChatGPT Decision Support System: Utility in Creating Public Policy for Concussion/Repetitive Brain Trauma Associated With Neurodegenerative Diseases." In.

Das, A., A. Tariq, F. Batalini, B. Dhara, and I. Banerjee. 2024. 'Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer', *medRxiv*.

de Oliveira, R., M. Garber, J. M. Gwinnutt, E. Rashidi, J. S. Hwang, W. Gilmour, J. Nanavati, K. Zine El Abidine, and C. D. Mack. 2025. 'A study of calibration as a measurement of trustworthiness of large language models in biomedical natural language processing', *JAMIA Open*, 8: ooaf058.

Denecke, K., and H. Paula. 2025. 'Evaluating Large Language Models for Analysing Safety Risks in Healthcare Incident Reports', *Stud Health Technol Inform*, 329: 386-90.

Dong, C., X. Qiu, J. Deng, L. Xu, X. Dong, S. Chen, T. Mei, Q. Li, Y. Cheng, J. Sun, H. Wang, and L. Yu. 2025. 'Comparative evaluation of large language models in delivering guideline-compliant recommendations for topical NSAID use in musculoskeletal pain: a multidimensional analysis', *Clin Rheumatol*, 44: 4703-10.

Elyoseph, Z., and I. Levkovich. 2023. 'Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment', *Front Psychiatry*, 14: 1213141.

Esmaeilzadeh, P. 2025. 'Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety', *BMC Med Inform Decis Mak*, 25: 342.

Fisch, U., P. Kliem, P. Grzonka, and R. Sutter. 2024. 'Performance of large language models on advocating the management of meningitis: a comparative qualitative study', *BMJ Health Care Inform*, 31.

Flathers, M., G. Smith, E. Wagner, C. E. Fisher, and J. Torous. 2024. 'AI depictions of psychiatric diagnoses: a preliminary study of generative image outputs in Midjourney V.6 and DALL-E 3', *BMJ Ment Health*, 27.

Fraser, H., D. Crossland, I. Bacher, M. Ranney, T. Madsen, and R. Hilliard. 2023. 'Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study', *JMIR Mhealth Uhealth*, 11: e49995.

Gorgos, P., K. Heder Ternell, C. Hammarstrand, A. Wallmon, E. Brogren, A. Bjorkman, and A. Horvath. 2025. 'ChatGPT and claude in hand surgery: an explanatory evaluation of clinical decision support on common surgical cases', *Hand Surg Rehabil*, 44: 102530.

Gun, M. 2025a. 'AI-Assisted Blood Gas Interpretation: A Comparative Study With an Emergency Physician', *Am J Emerg Med*, 94: 1-2.

———. 2025b. 'Can AI match emergency physicians in managing common emergency cases? A comparative performance evaluation', *BMC Emerg Med*, 25: 142.

Gunay, S., A. Ozturk, and Y. Yigit. 2024. 'The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: A comparison with cardiologists and emergency medicine specialists', *Am J Emerg Med*, 84: 68-73.

Gurnani, B., K. Kaur, P. Gireesh, L. Balakrishnan, and C. Mishra. 2025. 'Evaluating the novel role of ChatGPT-4 in addressing corneal ulcer queries: An AI-powered insight', *Eur J Ophthalmol*, 35: 1531-41.

Hadar-Shoval, D., K. Asraf, Y. Mizrachi, Y. Haber, and Z. Elyoseph. 2024. 'Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values', *JMIR Ment Health*, 11: e55988.

Hager, P., F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, and D. Rueckert. 2024. 'Evaluation and mitigation of the limitations of large language models in clinical decision-making', *Nat Med*, 30: 2613-22.

Han, T., S. Nebelung, F. Khader, T. Wang, G. Muller-Franzes, C. Kuhl, S. Forsch, J. Kleesiek, C. Haarburger, K. K. Bressem, J. N. Kather, and D. Truhn. 2024. 'Medical large language models are susceptible to targeted misinformation attacks', *NPJ Digit Med*, 7: 288.

Haze, T., R. Kawano, H. Takase, S. Suzuki, N. Hirawa, and K. Tamura. 2023. 'Influence on the accuracy in ChatGPT: Differences in the amount of information per medical field', *Int J Med Inform*, 180: 105283.

He, Z., B. Bhasuran, Q. Jin, S. Tian, K. Hanna, C. Shavor, L. G. Arguello, P. Murray, and Z. Lu. 2024. 'Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study', *J Med Internet Res*, 26: e56655.

Hong, H. J., N. H. Shah, M. A. Pfeffer, and L. S. Lehmann. 2025. 'Physician Perspectives on Large Language Models in Health Care: A Cross-Sectional Survey Study', *Appl Clin Inform*, 16: 1738-48.

Howard, E. C., J. M. Carnino, N. Y. K. Chong, and J. R. Levi. 2024. 'Navigating ChatGPT's alignment with expert consensus on pediatric OSA management', *Int J Pediatr Otorhinolaryngol*, 186: 112131.

Hu, Y., Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, K. Roberts, and H. Xu. 2024. 'Improving large language models for clinical named entity recognition via prompt engineering', *J Am Med Inform Assoc*, 31: 1812-20.

Huang, R., H. Wu, Y. Yuan, Y. Xu, H. Qian, C. Zhang, X. Wei, S. Lu, X. Zhang, J. Kan, C. Wan, and Y. Liu. 2025. 'Evaluation and Bias Analysis of Large Language Models in Generating Synthetic Electronic Health Records: Comparative Study', *J Med Internet Res*, 27: e65317.

Huppertz, M. S., R. Siepmann, D. Topp, O. Nikoubashman, C. Yuksel, C. K. Kuhl, D. Truhn, and S. Nebelung. 2025. 'Revolution or risk?-Assessing the potential and challenges of GPT-4V in radiologic image interpretation', *Eur Radiol*, 35: 1111-21.

Ismaiel, N., T. P. Nguyen, N. Guo, B. Carvalho, P. Sultan, and collaborators study. 2024. 'The evaluation of the performance of ChatGPT in the management of labor analgesia', *J Clin Anesth*, 98: 111582.

Jeblick, K., B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stuber, J. Topalis, T. Weber, P. Wesp, B. O. Sabel, J. Ricke, and M. Ingrisch. 2024. 'ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports', *Eur Radiol*, 34: 2817-25.

Jeon, S., S. A. Lee, H. S. Chung, J. Y. Yun, E. A. Park, M. K. So, and J. Huh. 2025. 'Evaluating the Use of Generative Artificial Intelligence to Support Genetic

Counseling for Rare Diseases', *Diagnostics (Basel)*, 15.

Ji, Y., W. Ma, S. Sivarajkumar, H. Zhang, E. M. Sadhu, Z. Li, X. Wu, S. Visweswaran, and Y. Wang. 2025. 'Mitigating the risk of health inequity exacerbated by large language models', *NPJ Digit Med*, 8: 246.

Kim, Y., H. Jeong, S. Chen, S. S. Li, M. Lu, K. Alhamoud, J. Mun, C. Grau, M. Jung, R. Gameiro, L. Fan, E. Park, T. Lin, J. Yoon, W. Yoon, M. Sap, Y. Tsvetkov, P. Liang, X. Xu, X. Liu, D. McDuff, H. Lee, H. W. Park, S. Tulebaev, and C. Breazeal. 2025. "Medical Hallucination in Foundation Models and Their Impact on Healthcare." In.

La Bella, S., M. Attanasi, A. Porreca, A. Di Ludovico, M. C. Maggio, R. Gallizzi, F. La Torre, D. Rigante, F. Soscia, F. Ardenti Morini, A. Insalaco, M. F. Natale, F. Chiarelli, G. Simonini, F. De Benedetti, M. Gattorno, and L. Breda. 2024. 'Reliability of a generative artificial intelligence tool for pediatric familial Mediterranean fever: insights from a multicentre expert survey', *Pediatr Rheumatol Online J*, 22: 78.

Latt, P. M., E. T. Aung, K. Htaik, N. N. Soe, D. Lee, A. J. King, R. Fortune, J. J. Ong, E. P. F. Chow, C. S. Bradshaw, R. Rahman, M. Deneen, S. Dobinson, C. Randall, L. Zhang, and C. K. Fairley. 2025. 'Evaluation of artificial intelligence (AI) chatbots for providing sexual health information: a consensus study using real-world clinical queries', *BMC public health*, 25: 1788.

Lauderdale, S. A., R. Schmitt, B. Wuckovich, N. Dalal, H. Desai, and S. Tomlinson. 2025. 'Effectiveness of generative AI-large language models' recognition of veteran suicide risk: a comparison with human mental health providers using a risk stratification model', *Front Psychiatry*, 16: 1544951.

Lee, S., W. I. Cho, C. Park, Y. Lee, C. Park, and T. Ko. 2025. "Evaluating the Influence of Demographic Identity in the Medical Use of Large Language Models." In.

Levkovich, I., and Z. Elyoseph. 2023. 'Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette Study', *JMIR Ment Health*, 10: e51232.

Makrygiannakis, M. A., K. Giannakopoulos, and E. G. Kaklamanos. 2024. 'Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing', *European journal of orthodontics*.

Maniaci, A., C. C. Hoch, L. Sogalow, B. Schmidl, and J. R. Lechien. 2025. 'AI in clinical decision-making: ChatGPT-4 vs. Llama2 for otolaryngology cases', *Eur Arch Otorhinolaryngol*, 282: 3293-302.

Marafino, B. J., and V. X. Liu. 2023. "Performance of a large language model (ChatGPT-3.5) for Pooled Cohort Equation estimation of atherosclerotic cardiovascular disease risk." In.

Masanneck, L., S. G. Meuth, and M. Pawlitzki. 2025. 'Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology', *NPJ Digit Med*, 8: 137.

McMahon, H. V., and B. D. McMahon. 2024. 'Automating untruths: ChatGPT, self-managed medication abortion, and the threat of misinformation in a

post-Roe world', *Front Digit Health*, 6: 1287186.

Menz, B. D., N. M. Kuderer, S. Bacchi, N. D. Modi, B. Chin-Yee, T. Hu, C. Rickard, M. Haseloff, A. Vitry, R. A. McKinnon, G. Kichenadasse, A. Rowland, M. J. Sorich, and A. M. Hopkins. 2024. 'Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis', *BMJ*, 384: e078538.

Mruthyunjaya, P., S. Verma, A. Agarwal, U. Maharana, M. Mandal, and S. Ahmed. 2025. "Right Diagnoses But Wrong Reasoning: Current Large-Language Model-Based Agentic Frameworks Have Flawed Clinical Reasoning Despite High Diagnostic Accuracy." In.

Naliyatthaliyazchayil, P., R. Muthyala, S. Purkayastha, and J. W. Gichoya. 2025. "Evaluating Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification with Zero Shot Prompting." In.

Omiye, J. A., J. C. Lester, S. Spichak, V. Rotemberg, and R. Daneshjou. 2023. 'Large language models propagate race-based medicine', *NPJ Digit Med*, 6: 195.

Pesapane, F., L. Nicosia, A. Rotili, S. Penco, V. Dominelli, C. Trentin, F. Ferrari, G. Signorelli, S. Carriero, and E. Cassano. 2025. 'A preliminary investigation into the potential, pitfalls, and limitations of large language models for mammography interpretation', *Discov Oncol*, 16: 233.

Pichowicz, W., M. Kotas, and P. Piotrowski. 2025. 'Performance of mental health chatbot agents in detecting and managing suicidal ideation', *Sci Rep*, 15: 31652.

Polat, E., Y. B. Polat, E. Senturk, R. Dogan, A. Yenigun, S. Tugrul, S. B. Eren, F. Aksoy, and O. Ozturan. 2024. 'Evaluating the accuracy and readability of ChatGPT in providing parental guidance for adenoidectomy, tonsillectomy, and ventilation tube insertion surgery', *Int J Pediatr Otorhinolaryngol*, 181: 111998.

Prasad, S., J. Langlie, L. Pasick, R. Chen, and E. Franzmann. 2025. 'Evaluating advanced AI reasoning models: ChatGPT-4.0 and DeepSeek-R1 diagnostic performance in otolaryngology: a comparative analysis', *Am J Otolaryngol*, 46: 104667.

Qazi, I. A., A. Ali, A. U. Khawaja, M. J. Akhtar, A. Z. Sheikh, and M. H. Alizai. 2025. "Automation Bias in Large Language Model Assisted Diagnostic Reasoning Among AI-Trained Physicians." In.

Radulesco, T., D. Ebode, A. Maniaci, S. Gargula, A. M. Saibene, C. Chiesa-Estomba, I. Gengler, L. Vaira, P. Vishnumurthy, J. R. Lechien, and J. Michel. 2025. 'Evaluation of Artificial Intelligence Chatbots for Facial Injection Planning: Comparative Performance and Safety Limitations', *Aesthetic Plast Surg*.

Ralla, B., N. Biernath, I. Lichy, L. Kurz, F. Friedersdorff, T. Schlomm, J. Schmidt, H. Plage, and J. Jeutner. 2025. 'How Accurate Is AI? A Critical Evaluation of Commonly Used Large Language Models in Responding to Patient Concerns About Incidental Kidney Tumors', *J Clin Med*, 14.

Rebitschek, F. G., A. Carella, S. Kohlrausch-Pazin, M. Zitzmann, A. Steckelberg, and C. Wilhelm. 2025. 'Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information', *NPJ Digit Med*, 8: 343.

Riley, G., E. Wang, C. Flynn, A. Lopez, and A. Sridhar. 2025. 'Evaluating the fidelity of AI-generated information on long-acting reversible contraceptive methods', *Eur J Contracept Reprod Health Care*, 30: 74-77.

Sharma, S., A. M. Alaa, and R. Daneshjou. 2025. 'A longitudinal analysis of declining medical safety messaging in generative AI models', *NPJ Digit Med*, 8: 592.

Shazahan, Mohamed Ashiq, Saavi Reddy Pellakuru, Sonal Saran, Shashank Chapala, Sindhura Mettu, and Rajesh Botchu. 2025. 'Can ChatGPT Aid in Musculoskeletal Intervention?', *Journal of Clinical Interventional Radiology ISVIR*.

Si, Y., Y. Meng, X. Chen, R. An, L. Mao, B. Li, H. Bateman, H. Zhang, H. Fan, J. Zu, S. Gong, Z. Zhou, Y. Miao, X. Fan, and G. Chen. 2025. 'Quality safety and disparity of an AI chatbot in managing chronic diseases: simulated patient experiments', *NPJ Digit Med*, 8: 574.

Smith, A., M. Liebrenz, D. Bhugra, J. Grana, R. Schleifer, and A. Buadze. 2025. 'Are clinical improvements in large language models a reality? Longitudinal comparisons of ChatGPT models and DeepSeek-R1 for psychiatric assessments and interventions', *Int J Soc Psychiatry*: 207640251358071.

Stanceski, K., S. Zhong, X. Zhang, S. Khadra, M. Tracy, L. Koria, S. Lo, V. Naganathan, J. Kim, A. G. Dunn, and J. Ayre. 2024. 'The quality and safety of using generative AI to produce patient-centred discharge instructions', *NPJ Digit Med*, 7: 329.

Tang, L., Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau, C. Weng, and Y. Peng. 2023. 'Evaluating large language models on medical evidence summarization', *NPJ Digit Med*, 6: 158.

Teixeira-Marques, F., N. Medeiros, F. Nazare, S. Alves, N. Lima, L. Ribeiro, R. Gama, and P. Oliveira. 2024. 'Exploring the role of ChatGPT in clinical decision-making in otorhinolaryngology: a ChatGPT designed study', *Eur Arch Otorhinolaryngol*, 281: 2023-30.

Toiv, A., Z. Saleh, A. Ishak, E. Alsheik, D. Venkat, N. Nandi, and T. E. Zuchelli. 2024. 'Digesting Digital Health: A Study of Appropriateness and Readability of ChatGPT-Generated Gastroenterological Information', *Clin Transl Gastroenterol*, 15: e00765.

Ulus, Ismail, Gokhan Ceker, and Ibrahim Hacibey. 2025. 'Large Language Models and Male Circumcision: A Reliability Assessment', *Medical Bulletin of Haseki*.

Umihanic, S., H. Osmanovic, N. Selak, D. Kopric, A. Huseinbasic, E. Sehic-Kozica, B. Babic, and F. Umihanic. 2025. 'Evaluating the Concordance Between ChatGPT and Multidisciplinary Teams in Breast Cancer Treatment Planning: A Study from Bosnia and Herzegovina', *J Clin Med*, 14.

Valentini, M., J. Szkandera, M. A. Smolle, S. Scheipl, A. Leithner, and D. Andreou. 2024. 'Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients?', *Front Public Health*, 12: 1303319.

Wakonig, K. M., S. Barisch, L. Kozarzewski, S. Dommerich, and M. H. Lerchbaumer. 2025. 'Comparing ChatGPT 4.0's Performance in Interpreting Thyroid Nodule Ultrasound Reports Using ACR-TI-RADS 2017: Analysis Across Different Levels of Ultrasound User Experience', *Diagnostics (Basel)*, 15.

Wang, H., R. Yang, M. Alwakeel, A. Kayastha, A. Chowdhury, J. M. Biro, A. D. Sorrentino, J. L. Handley, S. Hantzmon, S. Bessias, N. J. Economou-Zavlanos,

A. Bedoya, M. Agrawal, R. M. Ratwani, E. G. Poon, M. J. Pencina, K. I. Pollak, and C. Hong. 2025. 'An evaluation framework for ambient digital scribing tools in clinical applications', *NPJ Digit Med*, 8: 358.

Wang, X., J. Guo, T. Zhang, H. Lu, D. Zhou, H. Zhang, and X. Wang. 2025. 'Evaluating the performance of ChatGPT in clinical multidisciplinary treatment: a retrospective study', *BMC Med Inform Decis Mak*, 25: 340.

Williams, C. Y. K., J. Bains, T. Tang, K. Patel, A. N. Lucas, F. Chen, B. Y. Miao, A. J. Butte, and A. E. Kornblith. 2024. 'Evaluating Large Language Models for Drafting Emergency Department Discharge Summaries', *medRxiv*.

Wong, S. C., E. K. Chiu, K. H. Chiu, A. R. Tam, P. H. Chau, M. H. Choi, W. Y. Ng, M. O. Kwok, B. Y. Chau, M. Y. Ng, G. K. Lam, P. W. Wong, T. W. Chung, S. Sridhar, E. S. Ma, K. Y. Yuen, and V. C. Cheng. 2025. 'Comparative Evaluation and Performance of Large Language Models in Clinical Infection Control Scenarios: A Benchmark Study', *Healthcare (Basel)*, 13.

Yassin, Y., T. Nguyen, K. Panchal, K. Getchell, and T. Aungst. 2025. 'Evaluating a generative artificial intelligence accuracy in providing medication instructions from smartphone images', *J Am Pharm Assoc (2003)*, 65: 102284.

Yu, A., A. Li, W. Ahmed, M. Saturno, and S. K. Cho. 2025. 'Evaluating Artificial Intelligence in Spinal Cord Injury Management: A Comparative Analysis of ChatGPT-4o and Google Gemini Against American College of Surgeons Best Practices Guidelines for Spine Injury', *Global Spine J*, 15: 3199-220.

Yu, H., T. Chen, X. Zhang, Y. Yang, Q. Liu, C. Yang, K. Shen, H. Li, W. Tang, X. Zhong, X. Shuai, X. Yu, Y. Liao, C. Wang, H. Zhu, and Y. Wu. 2025. 'Performance of Large Language Models in Diagnosing Rare Hematologic Diseases and the Impact of Their Diagnostic Outputs on Physicians: Combined Retrospective and Prospective Study', *J Med Internet Res*, 27: e77334.

Zaboli, A., F. Brigo, G. Brigiari, M. Massar, M. Parodi, N. Pfeifer, G. Magnarelli, and G. Turcato. 2025. 'Chat-GPT in triage: Still far from surpassing human expertise - An observational study', *Am J Emerg Med*, 92: 165-71.

Zaboli, A., F. Brigo, S. Sibilio, M. Mian, and G. Turcato. 2024. 'Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage?', *Am J Emerg Med*, 79: 44-47.

Zaboli, A., F. Brigo, M. Ziller, M. Massar, M. Parodi, G. Magnarelli, G. Brigiari, and G. Turcato. 2025. 'Exploring ChatGPT's potential in ECG interpretation and outcome prediction in emergency department', *Am J Emerg Med*, 88: 7-11.

Zack, T., E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D. W. Bates, R. E. Abdulnour, A. J. Butte, and E. Alsentzer. 2024. 'Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study', *Lancet Digit Health*, 6: e12-e22.

Zare, S., A. Awasthi, W. Liaw, and H. Van Nguyen. 2025. "Utility, Accuracy, and Bias of Large Language Models as Real-Time Diagnostic Support in Primary Care." In.

Zaretsky, J., J. M. Kim, S. Baskharoun, Y. Zhao, J. Austrian, Y. Aphinyanaphongs, R. Gupta, S. B. Blecker, and J. Feldman. 2024. 'Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format', *JAMA Netw Open*, 7: e240357.

Zeljkovic, I., M. Novak, A. Jordan, A. Lisicic, T. Nemeth-Blazic, N. Pavlovic, and S. Manola. 2025. 'Evaluating ChatGPT-4's correctness in patient-focused informing and awareness for atrial fibrillation', *Heart Rhythm O2*, 6: 58-63.

Zeng, L., Q. Li, Y. Zuo, Y. Zhang, and Z. Li. 2025. 'Perceptions and Attitudes of Chinese Oncologists Toward Endorsing AI-Driven Chatbots for Health Information Seeking Among Patients with Cancer: Phenomenological Qualitative Study', *J Med Internet Res*, 27: e71418.

Zhang, A., M. Yuksekgonul, J. Guild, J. Zou, and J. C. Wu. 2023. "ChatGPT Exhibits Gender and Racial Biases in Acute Coronary Syndrome Management." In.

Zhang, Z., M. I. Qadir, M. Carstens, E. H. Zhang, M. S. Loiselle, F. M. Martinus, M. K. Mroczkowski, J. Clusmann, J. N. Kather, and F. R. Kolbinger. 2025. 'Prompt injection attacks on vision-language models for surgical decision support', *medRxiv*.

Zhong, W., M. Sun, S. Yao, Y. Liu, D. Peng, Y. Liu, K. Yang, H. Gao, H. Yan, W. Hao, Y. Yan, and C. Yin. 2025. 'Enhancing the Accuracy of Human Phenotype Ontology Identification: Comparative Evaluation of Multimodal Large Language Models', *J Med Internet Res*, 27: e73233.