JOURNAL OF ABDOMINAL WALL SURGERY | EUROPEAN HERNIA SOCIETY

MINI REVIEW
published: 26 March 2026
doi: 10.3389/jaws.2026.16349

Large Language Models in Surgery: Promise, Pitfalls, and Practical Use

*Danette T. Denham, Colin Y. Wang, Emil Maric, Lucy R. Hinton and B. Todd Heniford**

Division of Gastrointestinal and General Surgery, Department of Surgery, Endeavor Health, Evanston, IL, United States

Background: Large Language Models (LLMs) represent a transformative advancement in artificial intelligence (AI) with rapidly expanding applications in medicine. While AI-related medical publications increased 36-fold between 2000–2022, practical guidance for surgeons remains limited. This mini-review delineates pragmatic applications of LLMs in surgical practice while addressing key limitations, implementation considerations, and ethical considerations.

Methods: We reviewed contemporary LLM platforms and their integration into clinical workflows, patient communication, surgical research and academic writing, evaluating both benefits, constraints and risk mitigation relevant to practicing surgeons.

Findings: LLMs demonstrate significant utility across multiple domains. In clinical workflows, ambient documentation and chart summarization may reduce documentation burden and support rapid synthesis of complex patient data. For patient communication, these tools can simplify complex medical information, tailor or translate patient instructions to appropriate reading levels or languages, and generate empathetic responses to patient messages with improved efficiency. In research, LLMs assist with literature summarization, study design optimization, and risk of bias assessment in RCT, allowing surgeons to focus on higher-level scientific reasoning. Despite promising applications, several constraints demand attention. Effective prompting requires specific techniques including clear clinical objectives, explicit instructions, and iterative refinement. LLM outputs require verification to prevent "hallucinations" - fabricated or inaccurate information. Protected health information (PHI) must never be entered into public LLM platforms to maintain HIPAA compliance. Liability frameworks for AI-generated errors remain ambiguous, with unclear responsibility deferred amongst providers, institutions, and developers.

Conclusion: LLMs offer surgeons valuable tools for enhancing workflow efficiency and patient communication when deployed with appropriate oversight. Success requires understanding prompt engineering principles, maintaining rigorous fact-checking protocols, protecting patient privacy, and recognizing that human judgment remains irreplaceable in clinical decision-making.

Keywords: academic research, artificial intelligence (AI), large language models, patient outcomes, surgery

OPEN ACCESS

***Correspondence**
B. Todd Heniford,
todd.heniford@gmail.com

Received: *03 February 2026*
Revised: *04 March 2026*
Accepted: *16 March 2026*
Published: *26 March 2026*

Citation:
Denham DT, Wang CY, Maric E, Hinton LR and Heniford BT (2026) Large Language Models in Surgery: Promise, Pitfalls, and Practical Use. J. Abdom. Wall Surg. 5:16349. doi: 10.3389/jaws.2026.16349

INTRODUCTION

Artificial Intelligence (AI) represents one of the most disruptive technological advances of the modern era, with the potential to impact nearly every aspect of human life. Within the field of medicine, interest in AI has increased exponentially, with a 36-fold increase in AI-related medical publications between 2000–2022, growing from approximately 8,500 publications to over 307,000 [1]. Large language models (LLMs) are a subclassification of AI that use transformer-based neural networks trained on extensive datasets to learn semantic and statistical patterns of tokens, allowing interpretation and generation of human-like responses to prompts without task-specific training [2, 3].

The first widely recognized LLM was Generative Pre-trained Transformer (GPT)-3.5, released in November 2022 by OpenAI, followed by GPT-4 in March 2023 [4]. Subsequently, comparable models such as Bard [5] (now Google Gemini) and Bing Chat [6] (now Microsoft CoPilot) have entered the market (**Table 1**). Proposed clinical applications for these revolutionary LLMs span a diverse range of functions, including improving diagnostic capabilities, predicting outcomes, reducing clinical documentation burden, improving medical education, and filtering the expanding research literature. However, persistent uncertainties regarding the accuracy, reproducibility, and ethical governance of these systems impede their ubiquitous clinical adoption [7].

Despite expanding interest in medical LLMs, practical guidance tailored to surgical workflows remains limited. At the same time, physicians generally express positive attitudes toward adopting AI tools when they are feasible and clinically useful [8]. The aim of this review is to delineate clear and pragmatic uses for surgeons with currently available LLM platforms. The limitations of AI use within medicine will be evaluated, including the ethical considerations with LLM deployment in surgical practice.

LLM Integration With Electronic Health Record Systems

For many physicians, their initial clinical experience with LLMs will occur through their electronic health record (EHR) system. These tools fall into two categories: those developed natively by EHR vendors such as Epic or Oracle Health and third-party applications designed to integrate seamlessly with existing EHR platforms. The integration model offers distinct advantages, including smoother incorporation into established clinical workflows and institutional vetting of data security practices and regulatory compliance. However, this approach also introduces access barriers, as many tools require institutional licenses rather than individual subscriptions. Despite these constraints, LLM-integrated EHR tools are rapidly proliferating across healthcare settings and represent an important frontier in clinical AI adoption.

Peer-reviewed evidence for these commercial tools remains limited due to their proprietary nature and rapid development cycles. Much of the available information regarding features, performance metrics, and clinical benefits originates from vendor marketing materials, which should be evaluated with appropriate skepticism. The following discussion provides a sampling of currently available tools to illustrate the landscape of LLM integration in EHR systems and is not intended as an endorsement of any particular vendor or product. Healthcare organizations and individual clinicians should conduct thorough due diligence, including review of any available independent validation studies and consideration of institutional needs, before adopting these technologies.

Ambient Clinical Documentation

Perhaps the most widely adopted LLM application in EHRs is ambient clinical documentation, which uses AI to automatically generate clinical notes from recorded patient-clinician conversations. Third-party platforms such as Abridge [9], Nuance DAX Copilot [10], and DeepScribe [11] integrate with

TABLE 1 | Overview of relevant AI with parent company, price, and HIPAA compliance.

LLM name	Parent company	Cost	HIPAA compliance
Abridge [9]	None [9]	Must contact company for estimate [9]	Yes [70].
ChatGPT[a] [71]	OpenAI [71]	Pro: $20/month Plus: $200/month [72]	OpenAI for healthcare products are HIPAA compliant, not ChatGPT single subscriptions [73].
Claude[a] [74]	Anthropic [74]	Pro: $20/month Max:$100/month [75]	Healthcare option HIPAA-ready, not single subscriptions [76].
Copilot[a] [77] (formally bing chat)	Microsoft [77]	Personal: $9.99/month Premium: $19.99/month [78]	HIPAA compliance with copilot studio, a feature available to organizations [79].
Gemini[a] [80] (formally bard)	Alphabet inc [80]	Pro: $19.99/month Ultra [2]: $124.99/month [81]	Compliance with HIPAA accessible through google workspace or BAA [82].
Grok[a] [83]	xAI [83]	SuperGrok: $30/month SuperGrok heavy: $300/month [84]	Can support HIPAA under BAA [85].
OpenEvidence [86]	None [86]	Free for US healthcare professionals with NPI [86]	Yes [87].

BAA: Business Association Agreement; NPI: National provider identifier.
[a]*Free version available for public*

major EHR systems including Epic, Cerner/Oracle Health, and Athenahealth. A multicenter study by Olson et al. demonstrated that ambient AI scribes reduced clinician burnout from 51.9% to 38.8% within 30 days, while decreasing both note-related cognitive task load by 2.64 points on a 10-point scale and after-hours documentation time by 0.90 h [12]. A separate study by Moura et al. found that hybrid ambient clinical documentation combining generative AI with virtual scribes reduced "work outside of work" by 41.7% and improved financial productivity by 12.1% within 50 days [13]. DAX Copilot's has demonstrated ability to provide accurate, thorough inpatient note taking with succinct synthesis of information whilst limiting hallucinations and bias [14].

EHR vendors have also developed native ambient documentation capabilities. Epic's ambient solution [15] and Oracle Health's next-generation EHR [16] have incorporated ambient documentation features into their platforms. Additional entrants (Athelas AIR [17] and eClinicalWorks' Sunoh [18]) offer complete EHR platforms with built-in ambient AI functionality. Beyond real-time note generation, these tools often include autonomous medical coding capabilities that automatically suggest CPT and ICD-10 codes from the generated clinical notes, further streamlining administrative workflows and reducing administrative burden. A vendor-reported study facilitated by Microsoft showed a 3.4% increased level of service (LOS) when DAX Copilot for EPIC was used [10], but independent validation is limited.

Clinical Intelligence and Patient Data Aggregation

LLMs are increasingly being deployed within EHRs to synthesize complex patient data and generate actionable clinical insights. Note summarization features create context-specific summaries of patient charts tailored to different care settings, such as emergency department triage versus annual physicals [15]. These summaries include citations linking back to source information, allowing clinicians to quickly grasp a patient's status without manually reviewing multiple chart entries.

Athenahealth's Clinically Inferred Diagnosis feature uses multi-dimensional data analysis to suggest potential diagnoses and identify care gaps based on historical patient health data, medication lists, and past encounters [19]. Navina's clinical intelligence platform reconciles historical patient records with real-time data streams from multiple sources including labs, imaging, and clinical notes to support value-based care initiatives [20]. The platform identifies open care gaps and coding opportunities during patient encounters, integrating with ambient AI tools to align live patient dialogue with historical records.

PatientKeeper, now part of Commure, provides AI-driven 12-h patient summaries that highlight significant changes, treatments, and key patient data, along with a conversational AI chat feature that allows clinicians to quickly retrieve specific information without time-consuming manual searches [21]. Oracle Health's next-generation EHR incorporates Oracle Health Data Intelligence, which continuously integrates patient data from clinical, claims, social determinants, and pharmacy sources to deliver real-time insights for personalized care planning [16]. These tools aim to reduce the cognitive burden of information foraging while supporting more informed clinical decision-making at the point of care.

Revenue Cycle Management Automation

LLM's are transforming revenue cycle management by automating traditionally labor-intensive billing and coding tasks. Commure's Autonomous Coding platform integrates with Epic, Cerner/Oracle Health, MEDITECH, and over 30 other EHR systems, automatically generating CPT codes, ICD-10 diagnoses, and modifiers directly from clinical documentation [21]. Epic's "Penny" AI agent assists with billing code suggestions and generates appeal letters for denied insurance claims [15]. These tools address the growing complexity of medical coding.

By leveraging ambient documentation outputs, autonomous coding systems can achieve high accuracy rates while reducing manual review burden. The integration of these tools with upstream documentation platforms creates an end-to-end workflow from patient encounter to claim submission, potentially improving first-pass claim acceptance rates and reducing days in accounts receivable. However, implementation success depends on robust data flow from the EHR, documentation quality, and alignment between AI vendors and institutional coding practices.

Assisted Practice in the Clinical Setting

Current LLMs may assist clinicians in perioperative risk stratification; accurately predicting ICU admission, unplanned hospital admission, and mortality using real-world EHR data. The models performed well with categorical outcomes but were less accurate with continuous numerical predictions such as length of stay [22]. ChatGPT showed superior performance over competing LLMs in use of clinical risk assessments tools such as Charlson Comorbidity Index to estimate patient clinical risk [23]. ChatGPT can be used to provide recommendations for thromboembolic prophylaxis with reasonable accuracy, which can help to mitigate surgical risk whilst reducing cognitive load of the treating team [24].

PATIENT COMMUNICATION

Physician-patient communication is foundational component of medical care and is associated with patient satisfaction, outcomes, and medicolegal risk [1, 3, 25]. LLM's have the potential to streamline patient communication by drafting responses to patient questions, editing existing patient material, and translating medical information into understandable content [26]. As patient access to clinical notes, laboratory results, and imaging reports becomes more common, discordance between clinical language, and patient understanding can increase anxiety and confusion. Multiple studies have shown that LLMs can produce generally accurate summaries of radiology reports, whilst maintaining important diagnostic information. Similarly, LLMs can be used to convert discharge summaries to more patient-friendly formats [27, 28].

Health literacy is an important social determinant of health, with poor literacy associated with worse health outcomes [29, 30].

Written educational materials facilitate patient understanding and enable patients to revisit information beyond the clinical encounter; however they are often well above readability levels appropriate for the general population (8th grade or less) [31, 32]. ChatGPT-4 has been to shown to reliably "Rewrite the following at a 6th-grade reading level" resulting in patient material with improved readability [32]. Similarly, Abreu and colleagues evaluated patient-facing cancer information from 34 NCCN-affiliated institutions and found that ChatGPT-4 reliably improved readability from approximately a college-freshman level to roughly a ninth-grade level while preserving accuracy and overall quality [33]. These findings support the use of LLMs as an editing layer for expert-generated content to reduce health-literacy barriers without materially degrading informational fidelity.

Additionally, Spanish-speaking patients represents a substantial portion of the US population, and many prefer postoperative instructions in their native language [34, 35]. LLMs have been shown to more accurately translate discharge instructions in Spanish and Portuguese when compared to Google Translate [36]. When using utilizing LLM generative writing capabilities for translation, outputs should be reviewed for accuracy and to ensure they adhere to institutional policies.

Patient messaging can be a major time burden. One study noted physicians spend an average of 2.3 min responding to each patient message [37]. LLMs can draft responses that clinicians then review and edit, potentially saving time while maintaining quality. In comparative evaluations, LLM-generated responses having similar accuracy, empathy but significantly higher word count per question answered [38, 39]. Answers to patient questions can be increased with accurate LLM prompting, such as instructing the LLM to simulate an experienced orthopedic surgeon [40]. Used thoughtfully and with clinician oversight, this technology creates time-efficient, clear patient communication whilst concurrently improving patient understanding.

RESEARCH

The integration of LLMs into academic research workflows has grown rapidly. Recent analyses suggest a marked increase in AI assistance in manuscripts and preprints [41], and surveys of clinical researchers report leveraging LLMs for tasks perceived to improve efficiency (question formulation, literature review, data summarization, manuscript editing), but persistent concerns regarding accuracy, bias, and transparency remain [42].

Literature Review and Evidence Synthesis

Generative AI has been proposed as a support mechanism in evidence synthesis workflows, including strategy generation and screening assistance. Studies evaluating LLM performance for real-time systematic search tasks reveal mixed performance. While models like ChatGPT can generate structured search queries or assisted screening prompts, current evidence indicated these models frequently miss large proportions of relevant studies or may provide irrelevant outputs when used as standalone search agents [43]. Nonetheless, ChatGPT-4 and Gemini were able to draft literature search syntax for a systematic literature review [44]. The ability to generate search syntax streamlines the review process while preserving investigator authority over article selection.

AI in Study Design

Emerging research suggests that when guided by expert oversight, AI-generated frameworks can match conventional RCT design criteria, with the potential to improve representativeness and generalizability of trial populations [45]. ChatGPT and Claude demonstrate high rates of accuracy in structured tasks, such as assessing the Risk of Bias in RCTs, allowing for more efficient appraisal when paired with human verification of study validity [46]. ChatGPT can be assistive in generating queries for systematic review with high precision [47]. When prompted, ChatGPT can evaluate scientific claims and highlight unresolved research questions which can in turn act as an impetus for ongoing research [48].

Importantly, systematic reviews of generative AI tools in evidence synthesis show that major tasks such as literature searching and bias assessment still require human oversight, as models continue to produce false inclusions, omissions, and inconsistent interpretations without supervision.

AI Assistance in Academic Writing

LLMs influence how scientific writing itself is produced and evaluated. A cross-sectional study published in *JAMA Network Open* compared medical research abstracts written by surgical trainees, senior surgeons, and ChatGPT. ChatGPT was given previously written, unassociated abstracts by the senior author, three papers on the surgical topic and the current study data after processed by a statistician. Blinded, very experienced, surgical reviewers were unable to reliably distinguish AI-generated abstracts from those authored by humans, and ChatGPT outputs scored comparably to resident and senior surgeon abstracts, demonstrating that LLMs can produce high-quality academic text when provided structured prompts and appropriate oversight [49].

In addition, recent reviews report researchers utilizing AI to streamline manuscript writing tasks. An analysis of ChatGPT in medical research describes its use for drafting and editing assistance, generating structured outlines, citation and reference support, and table or figure creation [50]. While these tools may streamline writing and editing tasks, they do not replace the need for domain expertise, accurate citation practices, and rigorous verification of all claims.

LLM for Data Extraction and Audit

Extracting structured data from clinical notes remains a time consuming yet essential component of clinical research. LLMs have demonstrated the ability to convert unstructured EHR text into structured datasets for audit and research purposes, achieving reported accuracy rates of 90%–95% across a range of clinical tasks [51]. In Urology, ChatGPT extracted key variables from operative [52] and pathology [53] reports with high accuracy and generalizability. The ability of LLMs to rapidly process large volumes of text makes them particularly well suited to large-scale chart abstraction and have been used successfully to extract information from extensive medical records pertaining to

breast cancer [54]. Beyond data extraction, generative AI tools such as Google Bard have been leveraged to develop algorithms that analyze EHR-derived Excel datasets, significantly improving the efficiency of surgical quality assurance audits by reducing reliance on manual data review [55].

Transparency and Publication Guidelines

The academic community is actively developing ethical frameworks and reporting standards for use of LLMs in academic work. Existing guidance consistently emphasizes three core principles: 1) human authors must take responsibility and vet all AI outputs, 2) meaningful human intellectual contribution must be present, and 3) transparent acknowledgement of AI use should be included in publication submissions to support reproducibility, credibility, and research integrity [56]. Major medical publishers, including SAGE, Nature Portfolio, and Elsevier, generally permit limited use of LLMs for basic language editing tasks such as grammar, spelling, and punctuation, and such use typically does not require formal declaration. In contrast, these publishers prohibit the use of generative AI for peer review and editorial decision-making and caution against uploading full manuscripts or confidential review materials into external AI tools due to concerns regarding confidentiality, intellectual property, and data security.

Detection of LLM-generated text remains imperfect, and proposed methods include statistical approaches and experimental detection techniques; however, these strategies have limitations and are not uniformly adopted [57].

Guidelines for AI Use in Academic Research

As AI becomes more ubiquitous in academic research, authors must retain responsibility for ensuring that the tone, reasoning, and intellectual framing reflect their own understanding of the subject matter. LLMs can enhance efficiency and support idea development, but their use must not compromise academic authenticity or accuracy.

Mijatović and colleagues have proposed practical guidance for the responsible use of AI in scientific writing, emphasizing the need to verify references and factual claims against primary sources and to critically edit LLM-generated text to ensure alignment with the author's intent. These guidelines highlight that the research foundation and analysis should be the authors, with the LLM used as a support tool, not substitution for critical thinking. Individual institutional guidelines must be followed and it is recommended to discuss AI use with the collaborative research team and mentors to ensure safe use [58].

Future Applications

The near-term research applications of LLMs will likely center on reducing friction in study execution rather than replacing analytic judgment. These uses may include automated eligibility screening from structured EHR data, semi-automated chart abstraction with human verification, and generation of standardized case report forms and data dictionaries to improve multi-site alignment. LLMs may also assist with literature surveillance and adverse-event narrative classification, but these workflows require prospective validation and clear audit trails to ensure reproducibility.

DISCUSSION

As AI use becomes widespread in medicine, surgeons need to weigh potential technological, workflow benefit against the inherent risks of probabilistic language systems. This technology still requires close human oversight. It is still imperative to fact check LLM output and patient privacy must be protected. The limitations discussed highlight the key technical, ethical, and jurisdictional considerations fundamental to safe and effective LLM use in surgical practice.

Prompting (Prompt Engineering and Context Management)

The effectiveness of LLMs is highly dependent on the quality of the prompts provided to the generative AI. To provide an output, the LLM tokenizes individual prompt words or parts of words, which are then assigned a unique numerical value. These numerical tokens are fed into the model, and the quality of the output relies on the complexity of the artificial neural network. The program constructs its response based solely on pattern recognition and does not understand the meaning of the words [59]. Well-constructed prompts help align the AI system to generate clinically relevant and accurate outputs. Clear, specific, and context-appropriate prompts are essential. The following principles may help surgeons optimize their use of LLMs [60].

1. Define the objective. Clearly stating the intended goal of the LLM. Specify the clinical task, the intended audience, and the desired output format.
2. Provide explicit and specific instructions, including role assignment to align the LLM with appropriate clinical expertise (e.g., "assume the role of an experienced surgical attending").
3. Define contextually relevant parameters applicable to the clinical scenario.
4. Employ iterative refinement, using a structured feedback loop guided by clinical judgment.
5. Align outputs with evidence-based practice. Instruct the LLM to review literature such as reference guidelines or key studies. AI-generated outputs require independent verification to ensure accuracy and avoid hallucinated content [61].

Management of Protected Health Information

A principal advantage of this technology is the ability to recall prior use interactions, enhancing the accuracy and relevance of subsequent outputs [62]. This function enables surgeons to refine and optimize their prompts, facilitating the generation of more precise, clinically relevant responses. This beneficial capability requires caution in the clinical setting, particularly with respect to Protected Health Information (PHI). PHI entered into non-approved systems may be retained or reproduced in subsequent outputs depending on platform behavior and settings [63]. Surgeons should restrict generative AI use to scenarios that do not require entry of PHI, and should use only institutionally approved, HIPAA-aligned tools (e.g., EHR-integrated solutions) when patient data are involved.

Hallucinations

The accuracy of information is very important in clinical applications. LLM can produce inaccurate or fabricated information, known as 'hallucinations", resulting from an output being incoherent or misrepresentative of the source and are generally due to errors in coding and decoding. These discrepancies can arise in AI-generated responses as users lack visibility into the underlying source material from which specific outputs are derived [64]. Although the time-saving potential of LLMs is appealing in a time-constrained profession, particularly for automating repetitive tasks like discharge summaries, these applications still require close clinician oversight to ensure accuracy and mitigate harm. A study published in *JAMA* Internal Medicine reported that errors generated during LLM-assisted discharge summaries generally carried low potential for patient harm, but physician review remain necessary for safety and compliance [34].

Within surgical research, hallucinated citations are a well described failure mode. LLMs may fabricate references and inappropriately misattribute claims to nonexistent or unrelated publications [59, 65]. This risk is especially consequential in academic work and requires careful verification of all references and claims.

While LLMs are trained on large text corpora, the lack of transparency surrounding their training data necessitates caution. Clinicians should consider patient demographics carefully, as generalized outputs may not reflect the needs or characteristics of individual or underrepresented populations [66]. Implicit bias encompasses unconscious negative or positive stereotypes that unintentionally influence judgments, decisions, and actions; and can be present within the LLMs response as these are trained on datasets arising from internet and can reflect the common biases in culture [59].

Liability

The allocation of responsibility for errors and their potential impact on patient care needs to be considered closely. Due to their complexity and limited foreseeability, AI systems are not easily accommodated within traditional liability frameworks [67]. The delineation of responsibility among stakeholders in medical liability cases is complicated when AI systems are incorporated into clinical decision-making. As this technology increasingly mediates the relationship between clinician actions and patient outcomes, establishing causation, negligence, or harm becomes more challenging for patients [68]. Notably, public perceptions of liability differ from those of clinicians. Members of the public are significantly more likely than physicians to believe that physicians should be held responsible for errors occurring during AI-assisted care [69]. Without a clear framework for liability associated with AI in healthcare, surgeon hesitation is reasonable.

CONCLUSION

As AI adoption accelerates, prioritizing accuracy, privacy, and patient safety remains essential. Although LLMs can assist with multiple clinical and academic tasks, safe use currently requires consistent human oversight and verification. Surgeons should view LLMs as assistive tools rather than autonomous clinical agents, applying structured prompting, rigorous fact-checking, and privacy-conscious workflows while maintaining human judgment and accountability.

AUTHOR CONTRIBUTIONS

DD contributed to data collection, data curation, and initial manuscript drafting. CW and EM contributed to study design, data interpretation, and critical revision of the manuscript for important intellectual content. LH contributed to study conception, methodology, and manuscript drafting and revision. BH contributed to study conception, oversight of the project, and critical revision of the manuscript. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work. All authors contributed to the article and approved the submitted version.

FUNDING

The author(s) declared that financial support was not received for this work and/or its publication.

CONFLICT OF INTEREST

The authors(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

GENERATIVE AI STATEMENT

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

PUBLISHER'S NOTE

REFERENCES

1. Li H, Han Z, Wu H, Musaev ER, Lin Y, Li S, et al. Artificial Intelligence in Surgery: Evolution, Trends, and Future Directions. *Int J Surg Lond Engl* (2025) 111(2):2101–11. doi:10.1097/JS9.0000000000002159
2. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A Systematic Review of Large Language Model (LLM) Evaluations in Clinical Medicine. *BMC Med Inform Decis Mak* (2025) 25:117. doi:10.1186/s12911-025-02954-4
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large Language Models in Medicine. *Nat Med* (2023) 29(8):1930–40. doi:10.1038/s41591-023-02448-8
4. Varghese J, Chapiro J. ChatGPT: The Transformative Influence of Generative AI on Science and Healthcare. *J Hepatol* (2024) 80(6):977–80. doi:10.1016/j.jhep.2023.07.028
5. *An Important Next Step on Our AI Journey*. Mountain View, CA: Google. (2023). Available online at: https://blog.google/innovation-and-ai/technology/ai/bard-google-ai-search-updates/ (Accessed January 10, 2026).
6. Bing Chat | Microsoft Edge. (2026). Available online at: https://www.microsoft.com/en-us/edge/features/bing-chat (Accessed January 10, 2026).
7. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. *J Med Internet Res* (2023) 25:e51580. doi:10.2196/51580
8. Heinrichs H, Kies A, Nagel SK, Kiessling F. Physicians' Attitudes Toward Artificial Intelligence in Medicine: Mixed Methods Survey and Interview Study. *J Med Internet Res* (2025) 27:e74187. doi:10.2196/74187
9. Generative AI for Clinical Conversations | Abridge. (2026). Available online at: https://www.abridge.com (Accessed January 10, 2026).
10. *Microsoft Dragon Copilot | Microsoft for Healthcare*. (2026). Available online at: https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-copilot (Accessed January 10, 2026).
11. *DeepScribe AI Medical Scribe*. (2026). Available online at: https://www.deepscribe.ai/ (Accessed January 10, 2026).
12. Olson KD, Meeker D, Troup M, Barker TD, Nguyen VH, Manders JB, et al. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. *JAMA Netw Open* (2025) 8(10):e2534976. doi:10.1001/jamanetworkopen.2025.34976
13. Moura LM, Mishuris RG, Metlay JP, Habib M, Ting DY, Gallagher KL, et al. Hybrid Ambient Clinical Documentation and Physician Performance: Work Outside of Work, Documentation Delay, and Financial Productivity. *J Gen Intern Med* (2025). doi:10.1007/s11606-025-09979-5
14. Ghanem YK, Nation R, Sofield H, Rouhi AD, Gandhi SV, Sier R, et al. DAX Copilot: Ambient AI Scribe May Help Reduce Surgical Resident Clinical Documentation Burden. *Surg Endosc* (2026) 40(1):688–94. doi:10.1007/s00464-025-12404-x
15. Artificial Intelligence | Epic. (2026). Available online at: https://www.epic.com/software/ai/ (Accessed January 10, 2026).
16. *Oracle Health*. (2026). Available online at: https://www.oracle.com/health/ (Accessed January 10, 2026).
17. Ambient AI. *Athelas* (2026). Available online at: https://www.athelas.com/ambient-ai (Accessed January 10, 2026).
18. eClinicalWorks. Meet your AI-Powered EHR. *eClinicalWorks* (2026). Available online at: https://www.eclinicalworks.com/products-services/meet-your-ai-powered-ehr/(Accessed January 10, 2026).
19. Healthcare Products and Services for Ambulatory Care | Athenahealth. (2026). Available online at: https://www.athenahealth.com (Accessed January 10, 2026).
20. *The Clinician-First AI Copilot for Value-Based Success | Navina*. (2026). Available online at: https://www.navina.ai (Accessed January 10, 2026).
21. Commure - AI Solutions Co-Developed with Health Systems. (2026). Available online at: https://www.commure.com (Accessed January 10, 2026).
22. Chung P, Fong CT, Walters AM, Aghaepour N, Yetisgen M, O'Reilly-Shah VN. Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication. *JAMA Surg* (2024) 159(8):928–37. doi:10.1001/jamasurg.2024.1621
23. Kara K, Gunel T. Clinical Risk Computation by Large Language Models Using Validated Risk Scores. *J Med Syst* (2025) 49(1):121. doi:10.1007/s10916-025-02261-5
24. Duey AH, Nietsch KS, Zaidat B, Ren R, Ndjonko LCM, Shrestha N, et al. Thromboembolic Prophylaxis in Spine Surgery: An Analysis of ChatGPT Recommendations. *Spine J Off J North Am Spine Soc* (2023) 23(11):1684–91. doi:10.1016/j.spinee.2023.07.015
25. Levinson W, Hudak P, Tricco AC. A Systematic Review of surgeon-patient Communication: Strengths and Opportunities for Improvement. *Patient Educ Couns* (2013) 93(1):3–17. doi:10.1016/j.pec.2013.03.023
26. Aydin S, Karabacak M, Vlachos V, Margetis K. Large Language Models in Patient Education: A Scoping Review of Applications in Medicine. *Front Med* (2024) 11:1477898. doi:10.3389/fmed.2024.1477898
27. Shiraishi M, Tomioka Y, Miyakuni A, Moriwaki Y, Yang R, Oba J, et al. Generating Informed Consent Documents Related to Blepharoplasty Using Chatgpt. *Ophthal Plast Reconstr Surg* (2024) 40(3):316–20. doi:10.1097/IOP.0000000000002574
28. Chien A, Tang H, Jagessar B, Chang KW, Peng N, Nael K, et al. AI-Assisted Summarization of Radiologic Reports: Evaluating GPT3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical. *Am J Neuroradiol* (2024) 45(2):244–8. doi:10.3174/ajnr.A8102
29. Nutbeam D, Lloyd JE. Understanding and Responding to Health Literacy as a Social Determinant of Health. *Annu Rev Public Health* (2021) 42:159–73. doi:10.1146/annurev-publhealth-090419-102529
30. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low Health Literacy and Health Outcomes: An Updated Systematic Review. *Ann Intern Med* (2011) 155(2):97–107. doi:10.7326/0003-4819-155-2-201107190-00005
31. Hoek AE, Anker SCP, van Beeck EF, Burdorf A, Rood PPM, Haagsma JA. Patient Discharge Instructions in the Emergency Department and Their Effects on Comprehension and Recall of Discharge Instructions: A Systematic Review and Meta-analysis. *Ann Emerg Med* (2020) 75(3):435–44. doi:10.1016/j.annemergmed.2019.06.008
32. Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM. Enhancing Health Literacy: Evaluating the Readability of Patient Handouts Revised by Chatgpt's Large Language Model. *Otolaryngol–head Neck Surg Off J Am Acad Otolaryngol-head Neck Surg* (2024) 171(6):1751–7. doi:10.1002/ohn.927
33. Abreu AA, Murimwa GZ, Farah E, Stewart JW, Zhang L, Rodriguez J, et al. Enhancing Readability of Online Patient-Facing Content: The Role of AI Chatbots in Improving Cancer Information Accessibility. *J Natl Compr Cancer Netw JNCCN* (2024) 22(2 D):e237334. doi:10.6004/jnccn.2023.7334
34. Himmelstein J, Himmelstein DU, Woolhandler S, Bor DH, Gaffney A, Zallman L, et al. Health Care Spending and Use Among Hispanic Adults with and Without Limited English Proficiency, 1999-2018. *Health Aff Proj Hope* (2021) 40(7):1126–34. doi:10.1377/hlthaff.2020.02510
35. Jang M, Plocienniczak MJ, Mehrazarin K, Bala W, Wong K, Levi JR. Evaluating the Impact of Translated Written Discharge Instructions for Patients with Limited English Language Proficiency. *Int J Pediatr Otorhinolaryngol* (2018) 111:75–9. doi:10.1016/j.ijporl.2018.05.031
36. Brewster RCL, Gonzalez P, Khazanchi R, Butler A, Selcer R, Chu D, et al. Performance of Chatgpt and Google Translate for Pediatric Discharge Instruction Translation. *Pediatrics* (2024) 154(1):e2023065573. doi:10.1542/peds.2023-065573
37. Kummervold PE, Johnsen JAK. Physician Response Time when Communicating with Patients over the Internet. *J Med Internet Res* (2011) 13(4):e79. doi:10.2196/jmir.1583
38. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging Large Language Models for Generating Responses to Patient Messages—A Subjective Analysis. *J Am Med Inform Assoc JAMIA* (2024) 31(6):1367–79. doi:10.1093/jamia/ocae052
39. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* (2023) 183(6):589–96. doi:10.1001/jamainternmed.2023.1838

40. Chen YC, Lee SH, Sheu H, Lin SH, Hu CC, Fu SC, et al. Enhancing Responses from Large Language Models With Role-Playing Prompts: A Comparative Study on Answering Frequently Asked Questions About Total Knee Arthroplasty. *BMC Med Inform Decis Mak* (2025) 25:196. doi:10.1186/s12911-025-03024-5
41. Liang W, Zhang Y, Wu Z, Lepp H, Ji W, Zhao X, et al. Quantifying Large Language Model Usage in Scientific Papers. *Nat Hum Behav* (2025) 9:1–11. Published online. doi:10.1038/s41562-025-02273-8
42. Mishra T, Sutanto E, Rossanti R, Pant N, Ashraf A, Raut A, et al. Use of Large Language Models as Artificial Intelligence Tools in Academic Research and Publishing Among Global Clinical Researchers. *Sci Rep* (2024) 14(1):31672. doi:10.1038/s41598-024-81370-6
43. Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, et al. The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: Chatgpt and Microsoft Bing AI Performance Evaluation. *JMIR Med Inform* (2024) 12(1):e51187. doi:10.2196/51187
44. Boyle A, Huo B, Sylla P, Calabrese E, Kumar S, Slater BJ, et al. Large Language model-generated Clinical Practice Guideline for Appendicitis. *Surg Endosc* (2025) 39(6):3539–51. doi:10.1007/s00464-025-11723-3
45. Jin L, Ong JCL, Elangovan K, Ke Y, Pyle A, Ting DSW, et al. Large Language Models in Randomized Controlled Trials Design: Observational Study. *J Med Internet Res* (2025) 27:e67469. doi:10.2196/67469
46. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the Risk of Bias in Randomized Clinical Trials with Large Language Models. *JAMA Netw Open* (2024) 7(5):e2412687. doi:10.1001/jamanetworkopen.2024.12687
47. Wang S, Scells H, Koopman B, Zuccon G. Can Chatgpt Write a Good Boolean Query for Systematic Review Literature Search? *arXiv* (2026):03495. doi:10.48550/arXiv.2302.03495
48. Xie S, Zhao W, Deng G, He G, He N, Lu Z, et al. Utilizing ChatGPT as a Scientific Reasoning Engine to Differentiate Conflicting Evidence and Summarize Challenges in Controversial Clinical Questions. *J Am Med Inform Assoc JAMIA* (2024) 31(7):1551–60. doi:10.1093/jamia/ocae100
49. Holland AM, Lorenz WR, Cavanagh JC, Smart NJ, Ayuso SA, Scarola GT, et al. Comparison of Medical Research Abstracts Written by Surgical Trainees and Senior Surgeons or Generated by Large Language Models. *JAMA Netw Open* (2024) 7(8):e2425373. doi:10.1001/jamanetworkopen.2024.25373
50. Lee PY, Salim H, Abdullah A, Teo CH. Use of ChatGPT in Medical Research and Scientific Writing. *Malays Fam Physician Off J Acad Fam Physicians Malays* (2023) 18:58. doi:10.51866/cm0006
51. Gu B, Shao V, Liao Z, Carducci V, Brufau SR, Yang J, et al. Scalable Information Extraction from Free Text Electronic Health Records Using Large Language Models. *BMC Med Res Methodol* (2025) 25(1):23. doi:10.1186/s12874-025-02470-z
52. Hsueh JY, Nethala D, Singh S, Hyman JA, Gelikman DG, Linehan WM, et al. Exploring the Feasibility of GPT-4 as a Data Extraction Tool for Renal Surgery Operative Notes. *Urol Pract* (2024) 11(5):782–9. doi:10.1097/UPJ.0000000000000599
53. Kaufmann B, Busby D, Das CK, Tillu N, Menon M, Tewari AK, et al. Validation of a zero-shot Learning Natural Language Processing Tool to Facilitate Data Abstraction for Urologic Research. *Eur Urol Focus* (2024) 10(2):279–87. doi:10.1016/j.euf.2024.01.009
54. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing Prompts from Large Language Model for Extracting Clinical Information from Pathology and Ultrasound Reports in Breast Cancer. *Radiat Oncol J* (2023) 41(3):209–16. doi:10.3857/roj.2023.00633
55. McGowan M, Correia Martins F, Keen JL, Whitehead A, Davis E, Pathiraja P, et al. Can Natural Language Processing Be Effectively Applied for Audit Data Analysis in Gynaecological Oncology at a UK Cancer Centre? *Int J Med Inf* (2024) 182:105306. doi:10.1016/j.ijmedinf.2023.105306
56. Mann SP, Vazirani AA, Aboy M, Brian D, Earp BD, Minssen T, et al. Guidelines for Ethical Use and Acknowledgement of Large Language Models in Academic Writing. *Nat Mach Intell* (2024) 6, 1272–1274. doi:10.1038/s42256-024-00922-7
57. Rao VS, Kumar A, Lakkaraju H, Shah NB. Detecting LLM-Generated Peer Reviews. *PLoS One* (2025) 20(9):e0331871. doi:10.1371/journal.pone.0331871
58. Mijatović A, Žuljević MF, Ursić L, Marušić A. Responsible Use of Large Language Models in Manuscript Preparation. *Curr Protoc* (2026) 6(1):e70300. doi:10.1002/cpz1.70300
59. Laizure SC. Caution: Chatgpt Doesn'T Know what you Are Asking and Doesn'T Know what It Is Saying. *J Pediatr Pharmacol Ther JPPT* (2024) 29(5):558–60. doi:10.5863/1551-6776-29.5.558
60. Patil R, Heston T, Bhuse V. Prompt Engineering in Healthcare. *Electronics* (2024) 13:2961. doi:10.3390/electronics13152961
61. Liu J, Liu F, Wang C, Liu S. Prompt Engineering in Clinical Practice: Tutorial for Clinicians. *J Med Internet Res* (2025) 27:e72644. doi:10.2196/72644
62. Salvagno M, Taccone FS, Gerli AG. Artificial Intelligence Hallucinations. *Crit Care* (2023) 27(1):180. doi:10.1186/s13054-023-04473-y
63. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. *ACM Comput Surv* (2023) 55(12):248–38. doi:10.1145/3571730
64. Beutel G, Geerits E, Kielstein JT. Artificial Hallucination: GPT on LSD? *Crit Care Lond Engl* (2023) 27(1):148. doi:10.1186/s13054-023-04425-6
65. Hatem R, Simmons B, Thornton JE. A Call to Address AI "Hallucinations" and How Healthcare Professionals Can Mitigate Their Risks. *Cureus* (2023) 15(9):e44720. doi:10.7759/cureus.44720
66. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large Language Models Propagate Race-based Medicine. *NPJ Digit Med* (2023) 6:195. doi:10.1038/s41746-023-00939-z
67. Chan B. Applying a Common Enterprise Theory of Liability to Clinical AI Systems. *Am J L Med* (2021) 47(4):351–85. doi:10.1017/amj.2022.1
68. De Micco F, Grassi S, Tomassini L, Di Palma G, Ricchezze G, Scendoni R. Robotics and AI into Healthcare from the Perspective of European Regulation: Who Is Responsible for Medical Malpractice? *Front Med* (2024) 11:1428504. doi:10.3389/fmed.2024.1428504
69. Khullar D, Casalino LP, Qian Y, Lu Y, Chang E, Aneja S. Public Vs Physician Views of Liability for Artificial Intelligence in Health Care. *J Am Med Inform Assoc JAMIA* (2021) 28(7):1574–7. doi:10.1093/jamia/ocab055
70. Abridge HIPAA-Compliance. Abridge (2024). Available online at: https://support.abridge.com/hc/en-us/articles/30235280178195-Abridge-HIPAA-Compliance (Accessed January 22, 2026).
71. *Chatgpt | AI Chatbot to Discover, Learn and Create*. (2026). Available online at: https://chatgpt.com/overview/ (Accessed January 22, 2026).
72. Chatgpt Plans | Free, Plus, Pro, Business and Enterprise. (2026). Available online at: https://chatgpt.com/en-MX/pricing/ (Accessed January 22, 2026).
73. *Introducing Openai for Healthcare*. (2026). Available online at: https://openai.com/index/openai-for-healthcare/ (Accessed January 22, 2026).
74. *Introducing Claude*. (2026). Available online at: https://www.anthropic.com/news/introducing-claude (Accessed January 22, 2026).
75. Pricing | Claude. (2026). Available online at: https://claude.com/pricing (Accessed January 22, 2026).
76. Healthcare | Claude. (2026). Available online at: https://claude.com/solutions/healthcare? (Accessed January 22, 2026).
77. *Microsoft 365 Copilot | AI Productivity Tools for Work*. (2026). Available online at: https://www.microsoft.com/en-us/microsoft-365-copilot (Accessed January 22, 2026).
78. Copilot Pricing Plans for Individuals | Microsoft Copilot. (2026). Available online at: https://www.microsoft.com/en-us/microsoft-365-copilot/pricing/individuals (Accessed January 22, 2026).
79. iaanw. Review ISO. *SOC, and HIPAA Compliance - Microsoft Copilot Studio*. (2026). Available online at: https://learn.microsoft.com/en-us/microsoft-copilot-studio/admin-certification (Accessed January 22, 2026).
80. Learn About Gemini, the Everyday AI Assistant from Google. *Gemini*. (2026). Available online at: https://gemini.google/about/ (Accessed January 22, 2026).
81. *Google AI Pro and Ultra — Get Access to Gemini 3 Pro and More*. Gemini. (2026). Available online at: https://gemini.google/subscriptions/ (Accessed January 22, 2026).
82. Gemini for Google Workspace FAQ - Business/Enterprise - Google Workspace Admin Help. Mountain View, CA: Google LLC. (2026). Available online at: https://support.google.com/a/answer/14130944?hl=en&co=DASHER._Family%3DBusiness-Enterprise#zippy=%2Cis-gemini-hipaa-compliant (Accessed January 22, 2026).

83. Chat and Ask AI - AI Powered Chat Bot Assistant. (2026). Available online at: https://askaichat.app/onboarding/multi-model-tools/grok? (Accessed January 22, 2026).
84. Welcome | Xai. (2026). Available online at: https://x.ai/ (Accessed January 22, 2026).
85. Enterprise Faqs | xAI. (2026). Available online at: https://x.ai/legal/faq-enterprise (Accessed January 22, 2026).
86. OpenEvidence. OpenEvidence. (2026). Available online at: https://www.openevidence.com (Accessed January 22, 2026).
87. OpenEvidence. OpenEvidence Is now HIPAA Compliant: Health Care Professionals Using OpenEvidence Can Securely Upload PHI. OpenEvidence (2026). Available online at: https://www.openevidence.com/announcements/openevidence-is-now-hipaa-compliant (Accessed January 22, 2026).