

Rethinking The Integration of LLMs And Generative AI In Healthcare: Applications, Challenges and Future Directions

Akanksha Saxena¹, Fardeen Ahmad Khan², Sachin Kumar Saxena³

^{1,2}Assistant Professor, SRMS CET, Bareilly, Uttar Pradesh, India

³Senior Analyst Business Intelligence, Teleperformance Private Limited, Gurugram, India

Abstract—Competing AI models such as transformer-based large language models, diffusion models among others, combined with recent advances in deep learning architectures (GTM, GPT, CLM, etc.) have been responsible for recent breakthroughs in natural language processing, medical image analysis, predictive modeling and synthetic data generation. Proprietary and open-source models from leading providers such as GPT-5.5, Claude Opus, Gemini, DeepSeek and Nano Banana Pro have shown potential for data extraction, clinical documentation automation, diagnosis and decision support. The arrival of large language models (LLMs) and generative artificial intelligence (GenAI) has heralded a revolutionary era for healthcare, enabling un-precedented progress across clinical, research, and educational domains. This review systematically reviews the existing evidence on the implementation, performance, and impact of LLMs and GenAI in healthcare, including systematic reviews, comparative studies, and benchmark analyses. The clinical integration of these AI technologies has demonstrated potential to improve work-flow efficiency, data accuracy and patient communication within specialties like surgery, radiology, dentistry and medical education. GenAI allows for the creation of synthetic medical images, visualization of molecular mechanisms, and generation of educational resources, thereby speeding up research and dissemination of knowledge. However, challenges persist around the reliability, transparency and ethical deployment of LLMs and GenAI. The authors highlight issues of variability in model performance, risks of algorithmic bias, privacy concerns, and the need for expert oversight as important considerations. Open-source LLMs offer the potential for broader adoption due to scalability and affordability, but this necessitates robust governance and empirical validation. In this review, the authors highlight the opportunities and challenges that LLMs and GenAI bring to healthcare. It calls for interdisciplinary collaboration, integration of AI literacy in medical education, and establishment of

strong ethical frameworks. Further research and consensus guidelines are necessary to ensure these transformative technologies are implemented in a responsible, effective, and equitable fashion.

Index Terms—Large Language Models, Generative AI, healthcare automation, open-source LLMs, AI-driven diagnostics, ethical AI deployment, AI governance, AI literacy, responsible implementation.

I. INTRODUCTION

The advent and rapid evolution of large language models (LLMs) and generative artificial intelligence (GenAI) technologies in recent years have profoundly changed the field of artificial intelligence (AI) [1], with far-reaching implications for research, education, and clinical practice. The convergence of innovations [2] in deep neural network architectures including transformer-based and diffusion models has unleashed new levels of performance, scalability and versatility in tasks ranging from medical image analysis to natural language processing (NLP), and from predictive analytics to synthetic data generation. The emergence of models such as GPT-5.5, Claude Opus, Gemini, DeepSeek, and Nano Banana Pro has introduced a new era in AI system capabilities [3], facilitating substantial advancements in automation, decision support, and human-machine interaction within the biomedical domain. This is particularly true in the healthcare sector, where the use of LLMs and GenAI has accelerated and are driving transformative change in both clinical and non-clinical areas. These technologies have shown potential [4] for increasing the efficiency and accuracy of data extraction, streamlining clinical documentation, supporting diagnosis and facilitating complex

decision-making processes. Systematic reviews and comparative studies have demonstrated the diverse applications of LLMs in orthopaedic surgery, radiology, dentistry, and graduate medical education to extract clinical data, audit records, communicate with patients, and automate administrative workflows [5]. Moreover, GenAI models have been crucial in producing synthetic medical images, illustrating molecular mechanisms, and enhancing educational resources, thereby speeding up research and fostering the exchange of knowledge. However, despite these remarkable advancements, the application of LLMs and GenAI in healthcare presents significant methodological, ethical, and practical challenges. The quality and reliability of results generated by AI are still under evaluation, especially in high-stakes clinical settings. Research comparing models such as DeepSeek, GPT, Claude, [6] and Gemini has demonstrated differences in performance across different task types and clinical sub domains, particularly for complex or multimodal inputs. Some models are good at text-based reasoning or mathematical tasks, others at image-based or empathetic communication tasks. This shows the heterogeneity of current AI capabilities [7]. The risks of over-reliance, algorithmic bias, privacy breaches and lack of transparency make the responsible adoption of these technologies all the more difficult.

To unlock the full potential of LLMs and GenAI in healthcare, it is necessary to address these challenges through rigorous evaluation protocols, interdisciplinary collaboration, and development of comprehensive ethical and governance frameworks. Recent literature stresses the importance of integrating AI literacy into medical curricula, encourages continued research on the educational and clinical impact of GenAI and emphasizes the need for expert supervision especially in specialized or high-risk tasks. Their novelty and rapid evolution demand continuous empirical validation and the development of consensus guidelines to enable fair, effective and trustworthy AI-driven solutions. Moreover, as open-source LLMs such as DeepSeek-R1 are scalable and cost-effective, the

authors offer promising alternatives to proprietary models, expanding the availability of advanced AI tools to academia and clinical environments. Benchmark studies have shown that such models can achieve similar or better performance at a fraction of the cost, but need to be carefully managed with respect to privacy, censorship and technical constraints. Meanwhile, the development of multimodal foundation models [8] such as Gemini Nano Banana Pro has diversified the potential of AI in life sciences for advanced image generation, visualization, and data augmentation for research and clinical communication.

This review aims to provide a comprehensive synthesis of the current landscape, applications and impact of LLMs and GenAI in healthcare, drawing on recent systematic reviews, comparative studies and benchmark analyses. The review [9] critically appraises the strengths, limitations and future directions of these technologies to inform researchers, clinicians, educators and policymakers of the opportunities and challenges of the integration of AI into medical practice and research.

Special emphasis is given to the ethical, legal and social implications of AI [10] implementation with a focus on transparency, equity and patient safety. The review also emphasizes the transformative potential of GenAI in health promotion, risk prediction and digital health interventions, and its role in addressing workforce challenges such as provider burnout and staff shortages.

In summary, the fast pace of LLMs and GenAI represents an unprecedented opportunity and a complex challenge for the healthcare sector. As these technologies become more and more embedded in clinical and educational workflows [11], there is a need for a concerted effort to harness their benefits and mitigate risks and to ensure responsible and effective implementation. The journey [12] of AI in healthcare is ongoing, and there is a need for sustained research, inter-disciplinary dialogue and the development of strong frameworks to guide the ethical and equitable deployment of these transformative tools.

Table 1 Comparison Table for different Large Language Models Architectures

Model Name	Company	Date/Year	Processing	Architecture
GPT-4	OpenAI	15 Mar 2023	image and text	Transformer based
GPT 5.5	OpenAI	April 24, 2026	agentic model	Transformer based
Claude Haiku 4.5	Anthropic	October 15, 2025	coding, computer use, and agent tasks	hybridized transformer architecture

Claude: Opus 4.8	Anthropic	May 28,2026	dynamic workflows	Anthropic lineup
Claude: Sonnet 4.6	Anthropic	February 17, 2026	coding, agents, and professional work	decoder-only
Gemini 3.5	Google DeepMind	May 19, 2026	native multimodal processing (text, images, audio, video)	decoder-only Transformer-based Large Language Model
SeeDance 2.0	ByteDance	2026	native multimodal audio-video generation supporting Text, Image, Audio, Video	unified Multimodal Audio-Video Joint Generation Architecture
Nano Banana Pro	Google DeepMind	2025	Multimodal AI (text-to-image and image editing with reasoning)	a proprietary multimodal Transformer-based foundation model
Qwen3.6-35B-A3B	ElevenLabs	2022	multimodal speech AI	proprietary deep learning speech architecture based on neural sequence models and transformer-based speech generation

Table 1 presents a comparative overview of some of the most advanced large language models (LLMs) and multimodal foundation models developed by leading artificial intelligence organizations. The comparison highlights each model's developer, release year, primary processing capabilities, and underlying architectural approach.

This scoping review systematically [13] reviews the current landscape, applications and impact of natural language processing (NLP) tools in orthopedic surgery. The review synthesizes the results across 24 eligible studies and describes the main uses of NLP in extracting clinical data, auditing, diagnosis, communicating with patients and operational activities. The review highlights the potential of NLP to enhance efficiency and accuracy in orthopedic workflows, but also suggests that further research is needed to address methodological, ethical, and practical challenges in the integration of NLP.

II. RELATED WORK

The scoping review systematically evaluates current research on the application of natural language processing (NLP) tools in orthopedic surgery [14]. It highlights the potential of NLP tools to enhance efficiency, data extraction, and clinical processes, and emphasizes the need for more evidence and guidelines for their integration.

In this cross-sectional study, [15] the authors evaluated the accuracy of a large language model-based chat interface, “Versa Chat” (gpt-35-turbo), for extracting structured data elements from abdominal imaging reports of hepatocellular carcinoma (HCC)

compared with manual chart review. The LLM demonstrated high accuracy, especially for simple classification tasks, but performed poorly on tasks involving summation or comparison. These results confirm the feasibility of LLMs for efficient, semi-automated clinical data extraction, while also emphasizing current technical limitations and the need for further refinement prior to wide clinical adoption.

The study evaluated the performance of four large language models—ChatGPT-5, Gemini 2.5 Pro, DeepSeek-V3.2, and Claude Sonnet-4.5—in responding to common patient enquiries about restorative dentistry [16]. DeepSeek-V3.2 exhibited the most balanced and highest overall performance in terms of accuracy, empathy, and readability metrics. Claude Sonnet-4.5 and ChatGPT-5 demonstrated high accuracy but relatively more variability in empathy and readability metrics. All of the models produced text that was above the recommended readability level for patient education. The findings indicate that these models can aid in patient information processes, but clinical supervision remains essential, and additional research is necessary to enhance their effectiveness and accessibility in dental practice.

The authors systematically tested the performance of four large language models (LLMs), namely ChatGPT-4, DeepSeek-R1, Google Gemini, and Microsoft Copilot, in a pilot study in answering 240 multiple-choice questions (MCQs) in oral and maxillofacial radiology (OMFR), which included text-based and image-based questions [17]. The results showed that ChatGPT had the highest overall accuracy (90.4%) and performed well in both text-based (90.5%) and

image-based (90.0%) questions. DeepSeek achieved good performance on text questions (84.5%), but was unable to handle image questions due to the lack of multimodal capabilities. Copilot and Gemini exhibited moderate performance with marked variability depending on question type and subject. All LLMs achieved the educational committee's minimum accuracy threshold of 70%, but performance varied across some specific OMFR subdomains, with the lowest accuracy observed in questions on salivary gland diseases. The study highlights the supportive and complementary role of LLMs in OMFR education and clinical decision-making, but also underscores the need for expert oversight, especially in specialised or image-based tasks. Limitations include use of a single institution's set of questions, language limitations, and technical limitations specific to the model. Future work should consider larger datasets, more languages and further fine-tuning LLMs for domain-specific accuracy.

This comparative study assesses the reliability and effectiveness of DeepSeek-R1, an open-source large language model, against proprietary models, namely Claude-3.5-Sonnet and GPT-4o, for research and educational purposes [18]. Using benchmark datasets (MATH-500, HumanEval, MMLU, MedQA) and a prospective user study with medical researchers, DeepSeek-R1 demonstrated superior or comparable accuracy in mathematical, programming, and medical reasoning tasks with 96% cost savings and faster inference. However, there are some limitations, including a 4k-token output cap, real-time censorship, and data stored in China. The results demonstrate that DeepSeek-R1 is a cost-effective and scalable alternative for academic and clinical settings, but privacy and ethical considerations need to be carefully implemented.

This review article provides a comprehensive overview of generative artificial intelligence (Gen AI) applications and implications in healthcare, emphasizing its transformative potential in clinical and administrative domains [19]. Gen AI technologies enable personalized treatment planning, synthetic data generation, medical image analysis, nursing workflow optimization, risk prediction and pandemic preparedness. Gen AI adds more value to medical education, public relations, revenue cycle management, and healthcare marketing. The review points out the potentials of Gen AI to increase efficiency, reduce provider burnout and

improve patient outcomes, while emphasizing the need for strong ethical frameworks to address data privacy, transparency, and algorithmic bias to enable responsible and effective implementation in the healthcare context.

In this work [20], the authors provide a broad overview of recent advances of generative artificial intelligence (AI) in healthcare, in particular, diffusion models and transformer architectures. The paper classifies innovations in medical imaging (reconstruction, translation, generation, classification, segmentation), protein structure prediction, clinical documentation, diagnostic assistance, radiology interpretation, decision support, medical coding and drug design. Comparative analyses show that diffusion and transformer-based models outperform others in terms of performance, scalability, and utility in a range of clinical settings. It also points out major limitations including interpretability, computational costs, bias and privacy concerns, and suggests future research directions to enhance the clinical reliability, fairness, and ethical deployment of generative AI in healthcare.

This mini review critically discusses the integration of generative artificial intelligence (GenAI), specifically large language models (LLMs), into graduate medical education (GME) [21]. This paper reviews the current literature to identify five key opportunities for GenAI in GME: decreasing electronic health record (EHR) workload, improving clinical simulation, providing personalized education, assisting with research and analytics, and offering clinical decision support. It also covers in depth major risks, including inaccuracy, overreliance, challenges to authenticity and academic integrity, bias, and privacy concerns. The review discusses the promise of GenAI to revolutionize medical education but calls for a measured approach to its implementation, as its advantages and disadvantages are continuously evaluated. Recommendations include embedding AI literacy into curricula and ongoing research into GenAI's educational impact.

[22] This review article systematically discusses the potential of Generative Artificial Intelligence (Gen AI) to revolutionize health care, focusing on clinical and non-clinical domains. The paper highlights the progress in personalized treatment planning, medical image analysis, synthetic data generation, surgical care, risk prediction, and population health management through the use of Gen AI technologies such as

large language models, generative adversarial networks, and variational autoencoders. Other non-clinical applications discussed include improvements to medical education, nursing workflows, administrative automation, revenue cycle management and healthcare marketing. The review underscored the potential of Gen AI to enhance diagnostic accuracy, operational efficiency, and patient outcomes while tackling workforce challenges, including provider burnout and staffing shortages. The authors highlight the need for strong ethical guidelines and governance frameworks to facilitate responsible implementation, especially in the context of data privacy, algorithmic bias, and transparency. The limitations identified are the relative novelty of Gen AI in healthcare and the need for further empirical studies to validate emerging applications. Future directions highlight interdisciplinary collaboration and ongoing ethical oversight as key to harnessing the benefits of Gen AI in healthcare delivery and research.

This article reviews recent progress in artificial intelligence (AI) applications for early-stage human health promotion, with an emphasis on the emerging role of large language models, discriminative machine learning, and multimodal foundation models [23]. The discussion spans the current landscape including AI-assisted risk prediction, early disease screening, and health behavior interventions while critically assessing methodological strengths and limitations. Key challenges addressed are data representativeness, public risk interpretation and long-term impact evaluation. Future directions outlined in the article include digital twin technology, federated learning and the development of consensus guidelines to support equitable, effective and trustworthy AI-driven health promotion strategies.

Gemini Nano Banana Pro is a state-of-the-art multimodal generative AI model based on the architecture of Google DeepMind's Gemini 3 Pro, optimized for high-fidelity image generation and editing, specifically for life sciences applications [24]. The model creates professional quality visuals such as infographics, diagrams and realistic medical images from complex multimodal prompts using a 1-million-token context window, multilingual text rendering, advanced reasoning and integration with Google Search. Key contributions are the ability to maintain style and identity consistency, support of up to 14 reference images and responsible AI features such as

SynthIDwatermarking. Bench-marking shows state-of-the-art performance on reasoning and multimodal tasks. In life sciences, Nano Banana Pro allows fast generation of synthetic medical images for data augmentation, visualisation of molecular mechanisms and the automatic generation of educational materials, greatly accelerating research and communication workflows. Limitations include possible hallucinations and the need for expert validation of outputs; ethical and regulatory considerations are discussed, with responsible-use guidelines and traceability features offered. Ongoing developments such as the launch of Nano Banana 2 and incorporation into agentic AI workflows indicate a disruptive effect on scientific illustration, data visualisation and cross-disciplinary collaboration in the biotechnology and healthcare industries.

III. CONCLUSION AND FUTURE WORK

The rapid evolution and adoption of large language models (LLMs) and generative artificial intelligence (GenAI) have revolutionised the healthcare landscape, presenting novel avenues for innovation in the clinical, research, and educational realms. This review has summarised the current evidence from systematic reviews, comparative studies and benchmarking analyses, highlighting the extraordinary progress and ongoing challenges of these technologies.

LLMs and GenAI have demonstrated immense potential in automating data extraction, streamlining clinical documentation, supporting diagnostic decision-making, and enhancing communication between healthcare providers and patients. The ability to generate synthetic medical images, visualise molecular mechanisms and improve educational materials has accelerated research and helped spread knowledge. Benchmarking studies show that both proprietary and open-source models such as GPT-5.5, Claude Opus, Gemini, DeepSeek, and Nano Banana Pro have merits and limitations across clinical tasks from text-based reasoning to multimodal image analysis.

However, the adoption of LLMs and GenAI is not without its methodological, ethical, and practical concerns. Major challenges to the responsible deployment of these technologies include model performance variability, algorithmic bias, privacy risks and lack of transparency. The novelty and complexity of AI

models requires rigorous evaluation protocols, expert oversight, and continuous empirical validation. Embedding AI literacy in medical education and promoting interdisciplinary collaboration are essential for successful and safe integration.

Open-source LLMs like DeepSeek-R1 provide scalable, affordable options to proprietary models, opening access to advanced AI tools in academic and clinical settings. However, privacy, censorship and technical limitations should be carefully managed to preserve trust and protect sensitive data.

The development of multimodal foundation models such as Gemini Nano Banana Pro broadens the scope of AI applications to enable sophisticated data augmentation, visualisation and communication in life sciences and healthcare.

LLMs and GenAI are maturing quickly and health care continues to evolve—be it workflow, health promotion or workforce—because providers are feeling burnt out and suffering from staff shortages. The future of AI in healthcare will be built on the development of comprehensive ethical frameworks, consensus guidelines and governance models that emphasise transparency, fairness and patient safety. Harnessing the full potential of these transformative technologies while minimising risks requires ongoing research, empirical validation, and interdisciplinary dialogue.

To conclude, LLMs and GenAI are a powerful tool and a complex challenge for healthcare. Technological innovation will be needed to pursue their successful integration, but with a careful look at the ethical, legal and social considerations. Moving the field forward will require a cautious and collaborative approach to ensure these technologies contribute meaningfully to better patient outcomes, operational efficiency and equitable access to high-quality healthcare.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of interest: The authors declare that they have no conflicts of interest.

Research involving human participants and/or animals There is no direct human participation in the manuscript.

Informed consent Informed consent was obtained from all individual participants involved in the study

ACKNOWLEDGMENT

Authors thank God for providing wisdom and resources for human participation and informed content. I acknowledge my guidance in helping me while implement the prediction model. Under his umbrella, I was able to forward my research work with all ethical standards.

DATA AVAILABILITY STATEMENT

Not applicable.

REFERENCES

- [1] W. Abu-Ashour, S. Emil, and D. Poenaru, "Using artificial intelligence to label free-text operative and ultrasound reports for grading pediatric appendicitis," *J. Pediatr. Surg.*, vol. 59, pp. 783–790, 2024, doi: 10.1016/j.jpedsurg.2024.01.033.
- [2] C. Raffel, N. Shazeer, A. Roberts, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," 2020.
- [3] H. Zhou, F. Liu, J. Wu, et al., "A collaborative large language model for drug analysis," *Nat. Biomed. Eng.*, 2025, doi: 10.1038/s41551-025-01471-z.
- [4] S. Cangül, M. Taşyürek, T. Tunç, et al., "From algorithms to empathy: Can large language models effectively answer patients' questions in restorative dentistry?" *BMC Oral Health*, vol. 26, 2026, doi: 10.1186/s12903-026-08255-0.
- [5] S. S. Bhuyan, V. Sateesh, N. Mukul, et al., "Generative artificial intelligence use in healthcare: Opportunities for clinical excellence and administrative efficiency," *J. Med. Syst.*, vol. 49, 2025, doi: 10.1007/s10916-024-02136-1.
- [6] J. Zuo, H. Deng, H. Zhou, et al., "Is Nano Banana Pro a low-level vision all-rounder? A comprehensive evaluation on 14 tasks and 40 datasets," 2025.
- [7] T. Tung, S. M. N. Hasnaeen, and X. Zhao, "Ethical and practical challenges of generative AI in healthcare and proposed solutions: A survey," *Front. Digit. Health*, vol. 7, 2025.
- [8] A. Killekar, A. Shanbhag, R. J. Miller, et al., "Retrieval-augmented Claude Opus 4.7 and GPT-5.5 surpass human performance on the nuclear cardiology board preparation exam (and Claude drafts a paper about it)," 2026.

- [9] "Technical guide," 2026.
- [10] E. T. Jacques, E. Datuowei, E. Quaye, et al., "Authority signals in Claude AI health citations: A descriptive analysis using the authority signals framework," 2026, doi: 10.5281/zenodo.19616668.
- [11] A. W. Y. Chia, W. L. L. Teo, Y. L. Munro, and R. Dalan, "Evaluating the performance of a generative AI model in assessing qualitative health research articles adherence to objective reporting standards," *Sci. Rep.*, vol. 16, 2026, doi: 10.1038/s41598-025-29591-1.
- [12] A. Eisele-Metzger, J. L. Lieberum, M. Toews, et al., "Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2," *Res. Synth. Methods*, vol. 16, pp. 491–508, 2025, doi: 10.1017/rsm.2025.12.
- [13] Z. Al Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," *Informatics*, vol. 11, 2024.
- [14] F. Sasanelli, K. D. R. Le, S. B. P. Tay, et al., "Applications of natural language processing tools in orthopaedic surgery: A scoping review," *Appl. Sci.*, vol. 13, 2023.
- [15] J. Ge, M. Li, M. B. Delk, and J. C. Lai, "A comparison of large language model versus manual chart review for extraction of data elements from the electronic health record," 2023.
- [16] S. Cangül, M. Taşyürek, T. Tunç, et al., "From algorithms to empathy: Can large language models effectively answer patients' questions in restorative dentistry?" *BMC Oral Health*, vol. 26, 2026, doi: 10.1186/s12903-026-08255-0.
- [17] E. Haylaz, I. Gumussoy, F. Kalabalik, et al., "Evaluation of the performance of four different large language models (ChatGPT, DeepSeek, Copilot, and Gemini) in answering oral and maxillofacial radiology questions: Pilot study," *BMC Oral Health*, vol. 26, 2026, doi: 10.1186/s12903-026-07790-0.
- [18] R. Rakhshi, T. Khosravi, A. Rahimzadeh, et al., "DeepSeek-R1: Reliability for research and education—A comparative study with Claude-3.5-Sonnet and GPT-4o," 2025, doi: 10.29252/JCBR.9.3.1.
- [19] S. S. Bhuyan, V. Sateesh, N. Mukul, et al., "Generative artificial intelligence use in healthcare: Opportunities for clinical excellence and administrative efficiency," *J. Med. Syst.*, vol. 49, 2025, doi: 10.1007/s10916-024-02136-1.
- [20] Y. Shokrollahi, J. Colmenarez, W. Liu, et al., "Recent advances in generative AI for healthcare applications."
- [21] R. Janumpally, S. Nanua, A. Ngo, and K. Youens, "Generative artificial intelligence in graduate medical education," *Front. Med. (Lausanne)*, vol. 11, 2024.
- [22] S. S. Bhuyan, V. Sateesh, N. Mukul, et al., "Generative artificial intelligence use in healthcare: Opportunities for clinical excellence and administrative efficiency," *J. Med. Syst.*, vol. 49, 2025, doi: 10.1007/s10916-024-02136-1.
- [23] N. Zheng, X. Zhang, C. Li, et al., "Advances in artificial intelligence empowering early-stage human health: Current landscape and future directions," *The Innovation*, Art. no. 101222, 2025, doi: 10.1016/j.xinn.2025.101222.
- [24] Gemini, "Executive summary," 2026.