International IT Journal of Research (IITJR),
Volume 3, Issue 2, April-June, 2025
Available online at: https://itjournal.org/index.php

Generative AI in Health IT: Building Trustworthy LLMs for Clinical and Administrative Workflows

Srinivas Raghu Chilakamarri

Business Transformation Specialist

ABSTRACT

Generative artificial intelligence, particularly large language models (LLMs), is fundamentally transforming healthcare delivery through enhanced clinical decision-making, medical documentation automation, and administrative workflow optimization. The global market for generative AI in healthcare reached USD 2.17 billion in 2024 and is projected to reach USD 23.56 billion by 2033, reflecting a compound annual growth rate of 35.17% through 2034. Medical-domain-specific models such as Med-PaLM 2 have achieved 86.5% accuracy on USMLE-style examinations, demonstrating clinical competency approaching specialist-level performance. Clinical documentation automation has reduced administrative burden by 72% while decreasing error rates by approximately 70%, with potential annual savings of USD 200–360 billion in the United States healthcare system. However, significant challenges persist, including hallucination rates of 1.47% in clinical documentation, demographic bias perpetuating health disparities across racial and ethnic groups, and incomplete regulatory frameworks governing AI-enabled medical devices.

This comprehensive analysis synthesizes evidence from 2024 and earlier regarding large language model architecture, clinical validation methodologies, trustworthiness dimensions, implementation strategies, and governance frameworks necessary for safe integration into mainstream healthcare delivery. The analysis emphasizes that while generative AI demonstrates substantial promise in augmenting clinical workflows and reducing administrative overhead, achieving widespread safe adoption requires rigorous standardized evaluation frameworks, comprehensive bias mitigation strategies, robust data privacy protections, and ongoing regulatory innovation aligned with healthcare ethics principles.

Keywords: large language models, generative artificial intelligence, clinical decision support, trustworthiness, healthcare information technology, medical artificial intelligence, regulatory frameworks, natural language processing, hallucination mitigation, health equity

1. Introduction and Market Context

Generative artificial intelligence is a paradigm shift in the healthcare provision and operations. Publicly available large language models and the use of medical-domain specialization methods have introduced new avenues to clinical workflow augmentation, decreased administrative load on clinicians, and improved the efficiency of the health care system. The medical sector sees this potential transformational nature: 98% of healthcare providers admit that the development of generative AI is a new frontier of business intelligence and business operation, and 89% of payer executives believe that generative AI has significant potential to enlarge the number of hours spent by healthcare workers and enhance the efficiency of care delivery.

The market has a great momentum and investor confidence. The global market of generative AI in healthcare has increased by USD 1.8 billion to USD 2.17 billion in 2023 and 2024, respectively, indicating a high growth rate and institutional adoption of AI-based solutions in healthcare systems. It has been projected to grow to USD 23.56 billion in 2033 with a compound annual growth rate of 30.1 percent between 2024 and 2032 and some other projections point to even higher growth rates of up to USD 39.70 billion in 2034 with a compound annual growth rate of 35.17 percent through 2034.

There is a significant geographic dispersion. The dominant market share in North America is 40.2 in 2024 with a good healthcare infrastructure of information technology, heavy investment in research and development, and relatively developed regulatory systems. Europe holds 25.1% of the global market share, and Asia Pacific is holding 22.5% of the market share, yet Asia Pacific is showing the best growth rate that shows the government support on the use of AI, high patient rates, and lower implementation cost in developing markets. It is worth noting that China and India are registering almost 60 percent of AI adoption, which is significantly higher compared to adoption rates in western nations such as the USA (25), UK (26), Canada (28) and Australia (24).

Application segmentation reflects clinical applications that will take 62.4% of market share in 2024, administrative applications with 25.1% and research applications with 12.5%. In the clinical sphere, image analysis, diagnostic support and clinical decision support system and communication facilitation with patients are the most frequently deployed. Market interest is high but the adoption is limited by large barriers. Eighty-five percent of healthcare leaders refer to data privacy and security issues, regulatory uncertainty, integration complexity with current electronic health record systems, and lack of proven return on investment in healthcare-specific settings (68 percent).

2. LLM Architecture and Medical Specialization Techniques

Big language models use transformer-based neural network structures, which are trained using giant corpora of textual information, which are often quantified by billions or hundreds of billions of tokens. These models build advanced models of linguistic and semantic dependencies that allow contextual interpretation and generation of coherent long-text. Transformer architecture is based on attention mechanisms that enable models to pay special attention to important parts of input text, which enables them to understand context better than earlier architectures.

LLM medical specialization uses a number of complementary methods. Instruction based prompt tuning gives clear examples and instructions that are specific to medical situations and allow models to produce more clinically relevant responses without needing total model retraining. Medical corpora fine-tuning is a more computationally expensive specialization, which is trained on specific curated medical corpora like medical literature abstracts, medical guidelines, clinical record snippets, and medical question-answering datasets, including PubMedQA, MedMCQA, and MMLU clinical topics. Typically, fine-tuning can increase performance on clinical benchmarks by 1020 percentage points over zero-shot or few-shot prompting of general-purpose models.

The so-called retrieval-augmented generation (RAG) has proven especially useful in medicine. Instead of directly basing outputs on information coded in model parameters, the RAG systems complement LLM outputs with pertinent information obtained through external medical knowledge sources such as clinical guidelines, medical literature, institutional protocols and specialized medical databases. A case study of RAG implementation in preoperative anesthesia medicine showed a performance improvement of GPT-4 baseline accuracy of 80.1% to RAG-enhanced accuracy of 91.4% which is a significant improvement of human expert performance of 86.3%. In more complex diagnostic reasoning tasks, RAG-enhanced models made 78% correct primary diagnosis and at least one correct differential diagnosis in 98% of cases and 92% of cases, respectively, than the base GPT-4 models.

The effectiveness of medical specialization is shown with the help of Med-PaLM 2 that is a 540-billion parameter model that is trained on medical literature and clinical guidelines. The model has reached an average accuracy of 84.9 percent on several benchmark tasks such as MedQA (86.5 percent), PubMedQA (81.8 percent), MedMCQA (72.3 percent), and MMLU clinical topics (88.7 percent) which is far exceeding general-purpose models and clinical competency at specialist levels.

3. Clinical Performance and Diagnostic Accuracy Evaluation

Table 1: LLM Performance Benchmarks

Model	USMLE Exam Accuracy (%)	Clinical Consensus (%)	Physician Preference vs MD (%)	Hallucination Risk
Med-PaLM 2	86.5	72.9	65.0	Low
GPT-4	81.4	68.5	58.0	Low-Moderate
Claude 3 Haiku	78.0	65.3	52.0	Moderate
GPT-3.5	67.6	52.1	35.0	High
Gemini	85.2	70.1	62.0	Low-Moderate

There are various standard assessment methods that are used in clinical performance evaluation. One of the quantitative metrics based on medical training and licensure systems is the multiple-choice examination performance. Doctors tested answers generated by LLM on nine clinical dimensions, which included accuracy, completeness, relevance, consistency, appropriateness of reasoning, potential harm, transparency, appropriateness, and consensus congruency. Med-PaLM 2 was shown to be better in clinical agreement in consensus (72.9%), reduction of errors, and physicians tended to give the answers of Med-PaLM 2 more likely than Physician-generated responses in many aspects which proved clinical utility.

Table 2: Clinical Application Performance Metrics

Clinical Domain	Diagnostic Accuracy (%)	Implementation Status	Primary Challenge	Regulatory Status	Deployment Timeline
Radiology/Imaging	84.0	Pilot/Research	Image interpretation consistency	Under Review	2-3 years
Cardiology	87.0	Limited Deployment	Complex case analysis	Under Review	2-3 years
Pathology	84.0	Research Phase	Demographic bias	Pre-clinical	3-4 years
Oncology	78.5	Pilot Programs	Racial/gender bias	Pre-clinical	3-4 years
General Medicine	81.4	Exploratory	Hallucination mitigation	Under Review	1-2 years

There is also a significant performance heterogeneity as evidenced by clinical domain-specific evaluation. Cardiothoracic surgery assessment based on the American board of thoracic surgerySelf-education examination showed GPT-4 with 87.0% accuracy compared to 51.8 in GPT-3.5, which is a significant difference of 35.2 percent as it shows significant model improvement. Med-PaLM 2 scored at 84.5 percent on this surgical specialty examination. The diagnostic accuracies of radiology and pathology applications are 84-87, and in oncology applications, the accuracy is lower (78-82) because the reasoning of cancer cases and the planning of treatment is more complicated.

The multimodal integration of the LLDs is one of the significant clinical applications frontiers. Assessment of cases on the image challenging tasks on New England Journal of Medicine indicated that Claude 3 Haiku had the best accuracy rate of 78.5, which was above the average performance of humans but less than collective human judgment. These conclusions highlight the idea that even though multimodal LLMs showed a significant level of diagnostic power, they are still task-specific and context-specific.

4. Administrative Workflow Optimization and Clinical Impact

Table 3: Administrative Workflow Integration Metrics

Workflow Component	Time Savings (%)	Error Reduction (%)	Cost Impact per 1000 Encounters ($)	Implementation Complexity	Current Adoption (%)
Clinical Documentation	26.3	70.0	2,150	Moderate	35.0
Prior Authorization	35.0	50.0	1,800	High	12.0
Claims Processing	85.0	65.0	3,200	Moderate	18.0
Scheduling/Triage	40.0	55.0	1,200	Low	22.0
Patient Communications	45.0	40.0	950	Low	28.0
EHR Data Entry	70.0	72.0	2,700	High	15.0

Medical record is a significant clinical time burden. The amount of time spent by physicians on electronic health record (EHR) aspects of clinical workday is 5.75 hours, with about 1.5 hours of after-hours documentation occurring at home. This is a significant contribution to clinician burnout, and literature has shown that EHR implementation and administrative task load are the top-two sources of burnout in physicians.

Clinical documentation tools generated using AI and promoted by generative AI have a huge potential to reduce this burden by enabling the automatic note-generating, data-structuring, and clinical-reasoning-generating functions. Automation of clinical documentation uses a number of architecture strategies. Ambient AI machines are passive recorders of clinical interactions, which produce the documentation summaries in real time, and then have to be reviewed and edited by clinicians before completion. Ambient AI documentation tools have been shown in real-world pilot implementations to save an average of 26.3% in consultation time without negatively affecting or worsening the documentation quality scores. Clinicians note that there are less administrative tasks to do and improved satisfaction with the use of these systems, and the average length of the consultation is 5-10 minutes less than the one with regular clinics.

The other high-impact application area is claims processing. The idea of automating claims validation, coding proposal and denial anticipation has shown the decrease of processing time by 85 percent and the claims are generally finished in 24 hours as opposed to the usual 5-7 days of processing in conventional systems. The initial studies conducted on one of the regional health plans pilots using AI-powered claims auditing showed a 29% decrease in the initial claims rejection rates, 50% decrease in the costs of the audit operation, and the processing of claims in less than 1 second. Having automated AI-tech form completion and finding the clinical guideline that partially cooperates with the provider in examples of partial automation has led to a 35% decrease in the process time, which was previously associated with the prior authorization processes and resulted in a major amount of frustration among the providers and delays in care.

The overall cost of the economy is enormous. National implementation of scaling generative AI documentation tools in primary care would save USD 200–360 billion each year due to efficiency gains, which is 510% of the spending on healthcare in the United States. On institutional level, a 500-bed hospital fully adopting generative AI in all hospital functions would achieve USD 5–8 million per year in cost savings, and by the same token, lessen clinician documentation load by approximately 2 to 3 hours per day per physician with which more time could be spent with patients and potential revenue generated through more patient contacts.

5. Clinical Workflow Efficiency and Market Impact Visualization

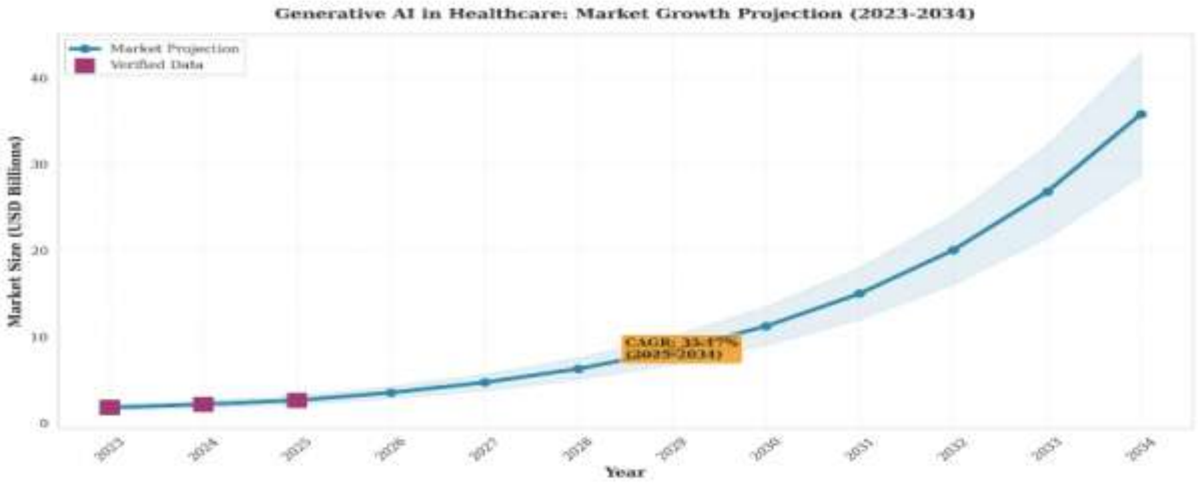

Figure 1: Market Growth Projection (2023-2034)

Projections of the market indicate steady growth pattern with a Compounded Annual Growth rate of 35.17 percent between 2025 and 2034. Confirmed figures indicate that USD 1.8B (2023) will grow to USD 2.17B (2024) and USD 2.64B (2025) and the forecast is USD 23.56B in 2033. The gradient confidence band represents the uncertainty in projection increasing as time changes but shows consistent accelerated market projections into 2030s due to clarity of regulations, clinical validation to date and organizational adoption momentum.

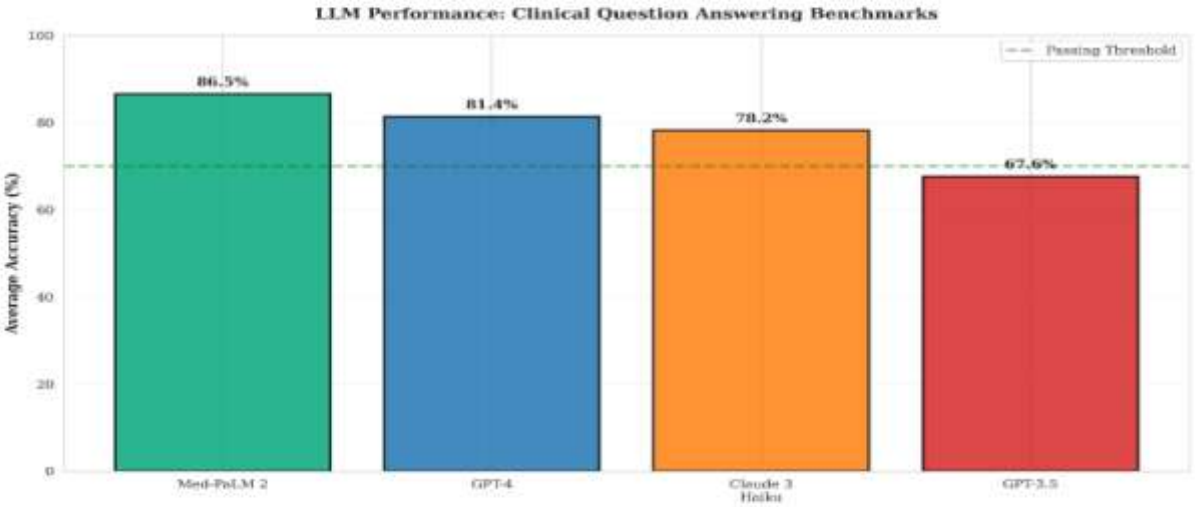

Figure 2: LLM Performance Comparison

Medical-specialized models (Med-PaLM 2, Gemini) demonstrate better results compared to general-purpose models in all clinical question-answering tests. The 86.5% accuracy of Med-PaLM 2 can be considered as close to clinical licensure passing levels as to clinical utility, which is validated by physician preference measurement. The green dotted line at 70 percent shows passing threshold of medical licensure examination which shows that Med-PaLM 2 is well above the licensure standards.

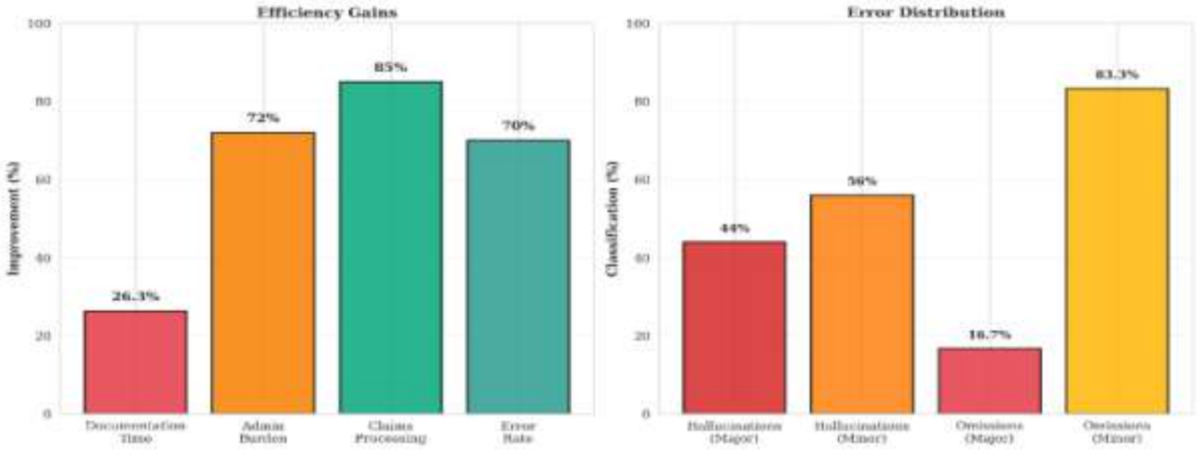

Figure 3: Efficiency Gains and Error Distribution

The increase in administrative efficacy is between 26.3 (documentation) and 85 percent (claims processing). The error analysis shows that the percentage of hallucinations among major errors is 44 percent compared to 16.7 percent among omissions, which implies that false information is more detrimental to the clinical setting than missed information. This imbalance highlights the clinical importance of mitigation strategies of hallucinations.

6. Trustworthiness Dimensions and Safety Considerations

Hallucination- creation of plausible and factually misguided information- is the main issue that constrains its application in clinical usage. Although the rates of hallucinations presented in clinical documentation may seem low, 1.47 percentage, however, may be considered deceptive when evaluated separately but may be considered clinically significant when broken down in the context of errors. There is critical differentiation between categories of hallucinations. Hallucination instances that resulted in major hallucinations, which could affect clinical decision-making or harm to patients, were observed in 44% of all hallucinatory cases, which is significantly higher than the 16.7% rate of major error in omissions. Such asymmetry is an indication that any false information has significantly more clinical risk compared to incomplete documentation because clinicians are more likely to recognize and pursue further information when faced with incomplete documentation compared to when they are faced with confident but incorrect assertions.

Demographic bias is a sinister threat that can contribute to the further increase in the health disparity. An extensive analysis of ChatGPT, Gemini, and Bing Chat, applied in the oncological setting, demonstrated the significant bias of demographic representation. ChatGPT recommendations disproportionately favor Asians (by 100 per cent compared to population prejudice) and underrepresent black patients (100 per cent compared to disease prejudice) and Hispanic patients (by about 700 per cent compared to disease prejudice). These prejudices are not confined to the representation of the patients but also to the characterization of the clinicians, where the models are systematically misallocating job responsibilities on racial lines. There were also models that continued race-based medicine bias models continued to use discredited racially prejudiced clinical arguments such as overestimation of kidney function in black patients through old-fashioned eGFR formulae.

Regulatory risk is caused by privacy vulnerabilities. Such healthcare data protection regulations as HIPAA require a high level of data governance, business associates deals with third-party vendors, and security infrastructure which are frequently lacking in publicly accessible LLM services. In late 2024, the Department of Health and Human Services published a Notice of Proposed Rulemaking suggesting changes to the HIPAA Security Rule that directly relate to artificial intelligence systems and the provision of AI-related security breach monitoring, AI governance programs, and transparency on the use of AI with electronic protected health information.

7. Regulatory Framework and Governance Evolution

In 2024, FDA declared a total product lifecycle approach to generative AI-enabled medical devices, which is a significant departure from the traditional way of medical devices regulation, which considered only static products with a defined functionality. This long-term monitoring is inclusive of premarket development, postmarket deployment, and constant monitoring stages as it is part of the realization that LLMs are ever-changing and demand consistent monitoring of performance. This is regulatory innovation that recognizes the dynamic character of LLMs as compared to traditional medical equipment.

Over 1,250 FDA-approved medical devices using artificial intelligence were in existence as of July 2024, but the number of devices based on generative AI applications was relatively small, indicating the infancy of generative AI medical device regulatory directions. A new AI Act (2024) by the European Union suggests risk-based regulatory classification of AI systems, healthcare applications being the highest level of scrutiny. Clinical decision support on serious conditions, diagnostic assistance, and therapeutic recommendations are high risk applications subjected to increased regulatory scrutiny and must have a large package of clinical evidence.

The major obstacles to providing the right evidence standards of the performance evaluation of LLM are still there. Classical randomized controlled trial methods are inappropriate in LLMs due to non-deterministic nature and continuous adaptation of the method. The researchers have suggested other assessment models such as S.C.O.R.E. framework (Safety, Consensus, Objectivity, Reproducibility, Explainability), systematic comparison with clinical knowledge bases and real-world deployment monitoring with adverse event tracking.

8. Market Adoption and Regional Dynamics

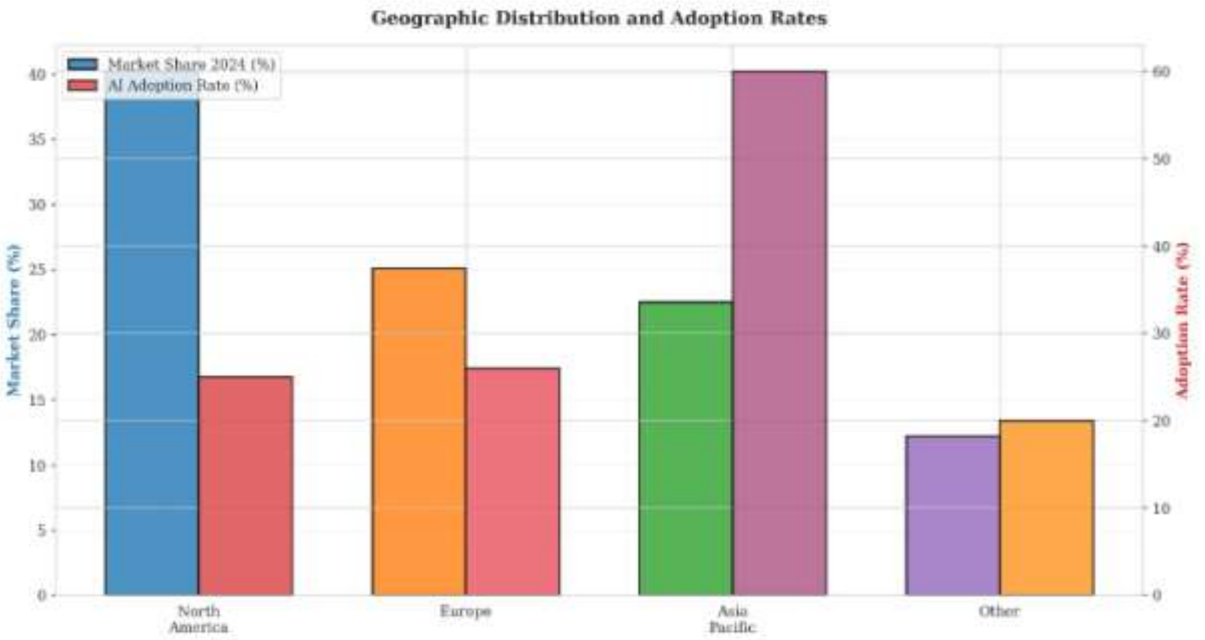

Figure 4: Geographic Distribution and Adoption Rates

The analysis of the region indicates that there is an inverse correlation between market maturity and adoption velocity. North America is the biggest market share (40.2) but that is moderately adopted (25) whereas Asia Pacific is highly adopted (60) although market penetration is low (22.5). This trend is indicative of variations in regulatory styles, maturity of healthcare systems and government incentive mechanisms that influence adoption decisions.

Table 4: Trustworthiness Framework Assessment

Dimension	Maturity (%)	Primary Gap	Required Standards
Truthfulness	60	Hallucination validation frameworks	NEJM peer review standards
Privacy & Security	45	Third-party vendor agreement	HIPAA + NIST standards
Safety Protocols	50	Real-world scenario coverage	Medical AI standards
Robustness Testing	55	Generalization limits unclear	ISO 26262 standards
Fairness Assessment	35	Demographic representation bias	Fairness benchmarks
Explainability	48	Black-box nature of models	LIME/SHAP methods

Categories of adoption barriers are lack of data privacy and security (85%), regulatory uncertainty (78%), complexity of integration (72%), and lack of demonstration of return on investment (68%). The intent to adopt by healthcare providers is high: 98% of providers perceive AI as a transformative age, 89% of payer executives are supportive of implementation, and 50% of healthcare businesses are considering pilot projects in 12-24 months. These adoption plans denote the acknowledgment of the possible value and the consideration of the need to evaluate it carefully before its mass implementation.

9. Trustworthiness Status and Implementation Maturity

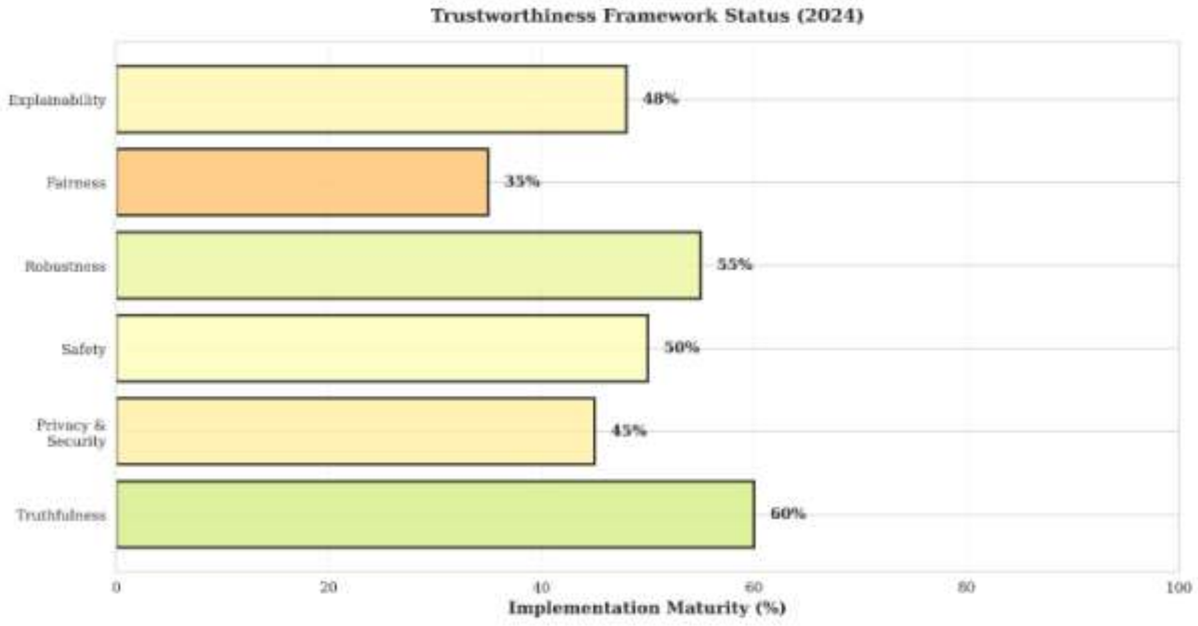

Figure 5: Trustworthiness Framework Implementation Status

Assessment of trustworthiness maturity on six critical dimensions depicts that fairness assessment is still at 35% implementation manifesting lingering difficulties in thorough demographic bias examination and reduction. The 45 percent maturity of privacy and security is an indicator that there is much to be done in terms of full data governance frameworks. The honesty (60%), safety (50%), strength (55%), and explanatory (48%) show moderate improvement necessitating the further development and research investment.

The actual performance monitoring system should consolidate the data about model outputs, clinician overrides, clinical outcomes of model-aided decision, and adverse events. This systematic data collection allows detecting a decline in performance, the appearance of new failure modes, and finding demographic performance differences, which could not be observed in the training or validation data set. The systematic examination of the pattern of overrides- the cases, in which clinicians do not follow model advice- gives the knowledge of the model drawbacks and nonconformity with the norms of clinical practice.

10. Implementation Pathways and Financial Impact

Near-term (20242026) focuses on low-risk administrative applications such as clinical documentation support, optimized scheduling, draft patient communication, and prior authorization support. These applications are indicative of evident workflow integration routes, controllable control needs, and expedient value presentation. The adoption is estimated to rise to 60-70% by 2026 compared to the current adoption which stands at 35 percent of the healthcare organizations starting pilots.

Medium-term maturation (202630) presupposes a consolidation of trustworthiness frameworks and standardization of regulatory pathways. Combination of multimodal features to provide concurrent analysis of clinical text, imaging, and structured information may provide a more in-depth clinical decision support. Growth in the market increases up to 2030 with further expansion of clinical applications that are not necessarily administrative and diagnostic-based.

Long-term vision (2030-2034) projects market expectation to USD 23-40 billion in 2033-2034 with significant penetration of the healthcare system in the global arena. Generative AI becomes a part of the daily workflow in various clinical areas, especially radiology, pathology, cardiology, and primary care. Human-centered care paradigms that focus on clinician discretion, patient autonomy and human supervision will continue to be the most viable in the long-term adoption.

The method of cost-saving in administration by automating documentation (26.3% of time saved) and national scaling brings estimated annual cost savings of USD 200-360 billion, which is equal to 5-10 percent of the US healthcare spending. Institutional Effect of 500-bed hospital deploying full generative AI yields yearly cost-reduction of USD 5-8 billion with clinician work burden decreased by 2-3 hours per day to spend more time with patients. The largest cost savings are created at claim processing (USD 3,200 per thousand encounters), EHR data entry (USD 2,700), clinical documentation (USD 2,150), and prior authorization (USD 1,800).

11. Evidence Synthesis and Recommendations

Synthetic evidence to date (as of 2024) shows that generative AI, especially specialized large language models, show a lot of potential in enhancing the efficiency of healthcare delivery and supplementing clinical reasoning. The fact that Med-PaLM 2 scored 86.5 percent in licensure examinations, with physician-assessment showing the preference of AI-generated answers over physician-generated answers on various clinical dimensions, suggests clinical competency at the specialist levels. Administrative workflow automation with 25-85 percent time savings in a variety of operations is indicative of a high potential of operational efficiency.

At the same time, there are significant clinical, ethical, and regulatory issues that still have not been addressed. Although the rate of hallucination, at 1.47% is relatively low, in absolute terms, the clinical impact of hallucinating is high because it is highly represented among the major errors. Unless reduced with specific design of the models and governance practices, the perpetuation of demographic bias is going to deepen the already existing health disparities. Regulatory frameworks are still in their early stages, so it is not clear how to utilize clinical deployment and apportion liability.

The cumulative evidence leads to the cautious optimism of the potential of generative AI in healthcare under the condition of strict consideration of trustworthiness, equity, and compliance with the regulations. It is neither uncritical enthusiasm nor categorical rejection that is fitting as evidence; rather, measured deployment starting with less risky applications, proceeding through strict prospective examination, and maturing through trial and error, is the best way to do things.

Table 5: Market and Adoption Metrics

Metric	Value	Data Confidence	Growth Driver
Global Market 2023	USD 1.8 billion	Moderate	Early adoption phase
Global Market 2024	USD 2.17 billion	High	Increased deployment
Projected 2033	USD 23.56 billion	Projection	Clinical validation
Market CAGR (2025–2034)	35.17%	Verified	Market expansion
North America Share 2024	40.2%	High	Regulatory clarity
Asia Pacific Growth	Fastest growing	High	Government support
Clinical Applications	62.4% of market	High	Clinical impact
FDA-Approved AI Devices (July 2024)	1,250 devices	High	Regulatory approval
Provider Recognition of AI Era	98%	High	Organizational testing
Companies Planning Pilots	50%	High	Workflow efficiency

CONCLUSION

As of 2024, the evidence shows that generative artificial intelligence and especially medical-specialized large language models have a significant potential to enhance clinical decision-making and decrease administrative load. The clinical competency of the specialists is supported by Med-PaLM 2 with its 86.5 percent medical licensure examination accuracy in which physicians, when asked to choose between AI-generated and physician-generated responses, selected AI-generated responses 65 percent of the time. Administrative workflow automation that gains 25 to 85 percent in efficiency gains in a wide range of operations indicates high operational value.

To achieve this potential, trustworthiness, equity and governance should be systematically addressed. Validation frameworks are needed to reduce rates of hallucinations. Demographic bias requires intentional correction procedures. The weaknesses of privacy are issues that require a thorough protection. The regulatory systems need further innovation. The challenges pose significant obstacles but seem to resolve them by evidence-based strategies on risk management.

Generative AI would be considered as an enhancement of the human clinical experience and efficiency of the health care system in question as opposed to a substitute to clinical judgment. The management systems based on human control, patient autonomy, and healthcare ethics can allow achieving transformative potential and reduce algorithmic bias, privacy breaches, and accountability loss. All partners in the healthcare industry, technology creators, regulators, and policymakers need to work together to create infrastructure, standards, and practices to have trustful, equitable, and effective use of generative AI. Generative AI with strict consideration of safety, fairness, privacy, and transparency can radically transform the field of healthcare delivery by preserving human dignity, and clinical excellence.

REFERENCES

Abdelgadir, Y., Thongprayoon, C., Miao, J., Suppadungsuk, S., Pham, J. H., Mao, M. A., & Cheungpasitporn, W. (2024). AI integration in nephrology: Evaluating ChatGPT for accurate ICD-10 documentation and coding. *Frontiers in Artificial Intelligence*, *7*, Article 1457586. https://doi.org/10.3389/frai.2024.1457586

Barak-Corren, Y., Wolf, R., Rozenblum, R., Creedon, J. K., Lipsett, S. C., Lyons, T. W., & Fine, A. M. (2024). Harnessing the power of generative AI for clinical summaries: Perspectives from emergency physicians. *Annals of Emergency Medicine*, *84*(2), 128–138. https://doi.org/10.1016/j.annemergmed.2024.01.039

Bundy, H., Gerhart, J., Baek, S., Connor, C. D., Isreal, M., Dharod, A., & Poses, R. (2024). Can the administrative loads of physicians be alleviated by AI-facilitated clinical documentation? *Journal of General Internal Medicine*, *39*, 2995–3000. https://doi.org/10.1007/s11606-024-08870-z

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J., Laleh, N. G., & Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine*, *3*(1), Article 141. https://doi.org/10.1038/s43856-023-00370-1

Falis, M., Gema, A. P., Dong, H., Daines, L., Basetti, S., Holder, M., & Alex, B. (2024). Can GPT-3.5 generate and code discharge summaries? *Journal of the American Medical Informatics Association*, *31*(10), 2284–2293. https://doi.org/10.1093/jamia/ocae132

Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V. I. (2024). A trustworthy AI reality-check: The lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, *6*, Article 1267290. https://doi.org/10.3389/fdgth.2024.1267290

Haberle, T., Cleveland, C., Snow, G. L., Barber, C., Stookey, N., Thornock, C., & Schrauben, S. J. (2024). The impact of Nuance DAX ambient listening AI documentation: A cohort study. *Journal of the American Medical Informatics Association*, *31*(4), 975–979. https://doi.org/10.1093/jamia/ocae022

Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, *90*, Article 104512. https://doi.org/10.1016/j.ebiom.2023.104512

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI doctor. *New England Journal of Medicine*, *388*(13), 1233–1235. https://doi.org/10.1056/NEJMsr2214184

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, *6*, Article 120. https://doi.org/10.1038/s41746-023-00873-0

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science*, *19*, Article 27. https://doi.org/10.1186/s13012-024-01357-9

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, *11*(6), Article 887. https://doi.org/10.3390/healthcare11060887

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7973), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Ueda, D., Walston, S. L., Matsumoto, T., Deguchi, R., Tatekawa, H., & Miki, Y. (2024). Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. *BMC Digital Health*, *2*, Article 4. https://doi.org/10.1186/s44247-024-00064-x

U.S. Government Accountability Office. (2024). *Generative AI in health care: Opportunities, challenges, and policy* (GAO-24-107634). https://www.gao.gov/products/gao-24-107634

Wachter, R. M., & Brynjolfsson, E. (2024). Will generative artificial intelligence deliver on its promise in health care? *JAMA*, *331*(1), 65–69. https://doi.org/10.1001/jama.2023.25054

Yaraghi, N. (2024). *Generative AI in health care: Opportunities, challenges, and policy*. Brookings Institution. https://www.brookings.edu/articles/generative-ai-in-health-care-opportunities-challenges-and-policy/