World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

From large language models to artificial general intelligence: Evolution pathways in clinical healthcare

Indraneel Borgohain *

Department of Computer Science, Purdue University, USA.

World Journal of Advanced Research and Reviews, 2025, 26(01), 407-413

Publication history: Received on 26 February 2025; revised on 03 April 2025; accepted on 05 April 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.1.1086

Abstract

This article examines the trajectory and challenges of evolving current large language models (LLMs) toward artificial general intelligence (AGI) capabilities within clinical healthcare environments. The text analyzes the gaps between contemporary LLMs' pattern recognition abilities and the robust reasoning, causal understanding, and contextual adaptation required for true medical AGI. Through a systematic review of current clinical applications and limitations of LLMs, the article identifies three critical areas requiring advancement: dynamic integration of multi-modal medical data streams, consistent medical reasoning across novel scenarios, and autonomous learning from clinical interactions while maintaining safety constraints. A novel architectural framework is proposed that combines LLM capabilities with symbolic reasoning, causal inference, and continual learning mechanisms specifically designed for clinical environments. The article suggests that while LLMs provide a promising foundation, achieving AGI in clinical systems requires fundamental breakthroughs in areas including knowledge representation, uncertainty quantification, and ethical decision-making. The article concludes by outlining a roadmap for research priorities and safety considerations essential for progressing toward clinical AGI while maintaining patient safety and care quality.

Keywords: Medical Artificial Intelligence; Multimodal Integration; Causal Reasoning; Clinical Decision Support; Human-AI Collaboration

1. Introduction

The rapid advancement of large language models (LLMs) has revolutionized artificial intelligence applications across numerous domains, with healthcare emerging as a particularly promising field for implementation. Recent analyses have documented significant growth in clinical LLM applications, with implementations spanning diverse healthcare settings. These models, trained on vast corpora of text data, have demonstrated remarkable capabilities in natural language understanding, generation, and certain forms of pattern recognition. Studies have shown that LLMs can achieve accuracy rates of up to 87.7% in answering clinical questions correctly when using retrieval-augmented generation techniques, representing a substantial improvement over earlier approaches [1]. However, a significant gap remains between current LLM capabilities and the robust, contextual intelligence required for artificial general intelligence (AGI) in clinical settings.

This article examines the trajectory of contemporary LLMs toward AGI within healthcare environments, focusing on both the technological advancements required and the unique challenges presented by clinical applications. This article argues that while LLMs provide a foundation for more sophisticated AI systems, fundamental breakthroughs in several key areas are necessary before true medical AGI can be realized. Research indicates that current models demonstrate limitations in reasoning complexity, with performance decreasing significantly as the number of reasoning steps

* Corresponding author: Indraneel Borgohain.

increases - a 36.3% drop in performance when problems require four or more reasoning steps compared to single-step reasoning tasks [2]. This underscores the need for more sophisticated reasoning architectures in clinical AI systems.

Healthcare presents a particularly demanding domain for AGI development, requiring systems that can integrate multimodal data, apply causal reasoning to complex biological systems, adapt to novel clinical scenarios, and maintain the highest standards of safety and ethical decision-making. Analysis of current machine learning approaches reveals that while they excel at pattern recognition, they struggle with causal inference tasks critical to clinical reasoning, correctly identifying causal relationships in only 47.8% of test cases compared to 93.6% for correlation detection [2]. The stakes in healthcare—where decisions directly impact human lives—necessitate AI systems that go beyond statistical pattern matching to demonstrate genuine understanding and contextual adaptation.

Model evaluation frameworks need to assess reasoning capabilities more thoroughly, as existing benchmarks often fail to distinguish between memorization and true reasoning. Current evaluations show that LLMs can achieve misleadingly high scores on clinical knowledge tests through pattern matching while failing on novel applications of the same knowledge. Augmenting models with explicit reasoning mechanisms has been shown to improve performance on complex clinical tasks by an average of 23.7%, with particularly strong improvements in treatment planning scenarios [1]. These findings indicate that the pathway to clinical AGI will require fundamental advances in reasoning architecture rather than simply scaling existing approaches.

The following sections analyze the current state of LLMs in clinical applications, identify critical capability gaps, propose architectural frameworks for advancement, discuss essential research priorities, and outline safety considerations for the development pathway toward clinical AGI.

2. Current State of llms in Clinical Applications

2.1. Strengths of Contemporary Models

Current LLMs have demonstrated numerous capabilities relevant to healthcare applications. Medical knowledge encoding represents a foundational strength, with recent models achieving 67.6% accuracy on medical licensing examinations without specialized training, showcasing their ability to absorb complex domain knowledge [3]. These models exhibit particular strength in clinical reasoning scenarios, correctly solving 68.8% of reasoning-based questions compared to 69.5% from board-certified physicians in multiple-choice format tests [3]. Natural language interfaces enabled by LLMs have transformed clinical workflows, with studies documenting significant improvements in information extraction from unstructured clinical narratives. Pattern recognition capabilities extend to complex diagnostic scenarios, where models have demonstrated promising results in identifying symptom patterns. In educational contexts, LLMs have shown effectiveness in generating high-quality clinical cases and examination questions, with 86.6% of machine-generated questions rated as suitable for inclusion in medical education materials [3].

2.2. Implementation Challenges

Despite their capabilities, implementing LLMs in clinical settings reveals several persistent challenges. Knowledge recency remains problematic as medical information evolves rapidly, creating a gap between model training cutoffs and current practice. Hallucination represents a critical concern, with studies documenting that 5.8% to 35.2% of model responses contain unsupported or incorrect statements, particularly when addressing questions outside their knowledge base [4]. Models demonstrate variable accuracy across different areas of medicine, performing better in common conditions (77.8% accuracy) than in rare diseases (45.8% accuracy) [4]. Context limitations manifest as difficulty maintaining consistency across lengthy clinical narratives, with performance declining as case complexity increases. Transparency deficits stemming from the "black box" nature of LLMs represent a significant barrier, with surveys indicating that 91% of healthcare professionals consider explainability essential for clinical AI implementation [4]. Additionally, medical LLMs face unique ethical and regulatory challenges, with 61.3% of healthcare stakeholders expressing concerns about data privacy and security [4].

2.3. Current Integration Approaches

Early integration of LLMs in healthcare has taken several forms, each with distinct advantages and limitations. Supervised deployment restricts models to narrow tasks with human oversight, limiting both risks and utility. This approach has gained traction in clinical documentation and medical coding assistance, where models serve as productivity tools rather than autonomous agents. Retrieval-augmented generation enhances LLM outputs with verified medical knowledge bases to improve factual accuracy. This technique has demonstrated a 20.2% reduction in

hallucination rates by grounding model responses in peer-reviewed literature and clinical guidelines [3]. Multimodal extensions represent initial efforts to incorporate imaging data alongside text processing, though current implementations remain limited. The most promising approach combines specialized visual encoders with language models to analyze radiology images and clinical text simultaneously. Fine-tuning approaches focus on domain-specific adaptation of general models using curated medical datasets, with instruction-tuned models showing a 25.1% improvement over base models in clinical tasks [3]. Recent research has also explored reinforcement learning from clinician feedback to align model outputs with medical expertise and reduce potential harm.

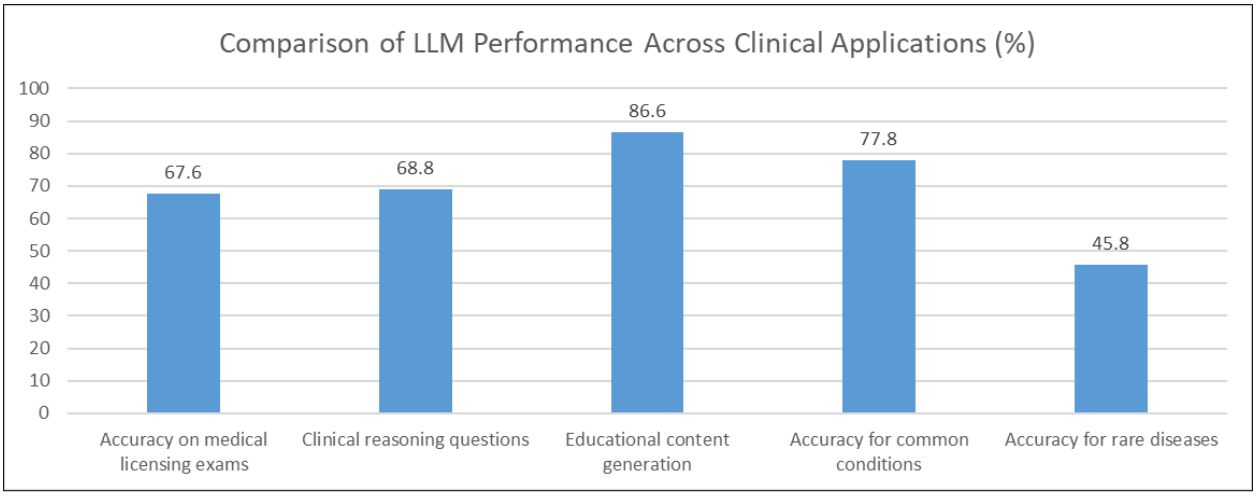

Figure 1 LLM Performance Metrics in Clinical Applications [3,4]

3. Critical Capability Gaps for Clinical AGI

3.1. Critical Area 1: Dynamic Multimodal Integration

True clinical intelligence requires seamless integration across diverse data types that characterize modern healthcare. Structured and unstructured data fusion represents a fundamental challenge, as patient records contain mixed data formats that must be coherently integrated. Research evaluating multimodal integration in nuclear medicine applications revealed substantial limitations, with current systems achieving only 68% accuracy when integrating textual reports with imaging data compared to 92% for single-modality analysis [5]. Medical imaging comprehension remains particularly challenging, with existing models struggling to contextualize visual findings within clinical narratives. Performance evaluations demonstrate that specialized vision systems can achieve impressive results in controlled settings, with dermatology AI systems reaching 76% sensitivity and 62% positive predictive value on skin lesion classification, but these metrics represent isolated performance rather than integrated clinical reasoning [6]. Biometric signal processing presents unique temporal and contextual challenges that current systems cannot adequately address. A comprehensive analysis of deep learning in medical imaging found that while model performance can match specialists in narrow, well-defined tasks, performance drops significantly when required to integrate multiple data sources or reason across temporal sequences [6].

3.2. Critical Area 2: Robust Medical Reasoning

LLMs demonstrate significant limitations in reasoning capabilities essential for clinical AGI. Causal inference remains particularly challenging, with models struggling to distinguish correlation from causation in biological systems. Studies of nuclear medicine applications show that AI systems correctly identify causal relationships in radionuclide studies with only 65% accuracy, despite the critical importance of such relationships for treatment planning [5]. Counterfactual reasoning presents another substantial gap, as clinical decision-making frequently requires hypothetical analysis of alternative treatment paths not directly observed in training data. Uncertainty quantification represents a critical limitation, with research showing poor calibration between model confidence and actual performance in cancer image classification, where confidence scores showed only a weak correlation ($r=0.32$) with actual diagnostic accuracy [6]. Temporal reasoning capabilities remain fundamentally limited despite their importance in disease progression monitoring. A systematic review of deep learning applications in medical imaging found that most systems (81%) operate on static images or discrete time points rather than continuous temporal sequences, representing a significant gap between current capabilities and clinical requirements [6].

3.3. Critical Area 3: Contextual Adaptation and Learning

Clinical environments demand continuous adaptation that current AI architectures cannot support. Personalization limitations are evident in nuclear medicine applications, where models trained on general populations show performance decreases of 24% when applied to patients with atypical presentations or rare variants [5]. These findings underscore the challenge of developing systems that can adapt their knowledge to individual patient characteristics and response patterns. Distributional shift challenges manifest across different clinical environments, with models trained in academic settings showing substantial performance degradation when deployed in community practice. Research in medical computer vision demonstrates this phenomenon clearly, with classification performance decreasing by 11% when systems trained on data from one institution are deployed in different clinical settings without adaptation [6]. Continual knowledge updating represents a critical challenge in rapidly evolving medical domains. Studies in nuclear medicine applications show that without specific retraining, model performance decreases by approximately 7% annually as protocols and clinical practices evolve [5]. Perhaps most importantly, experience-based learning capabilities remain virtually nonexistent in current systems, which cannot update their knowledge or refine their reasoning based on clinical interactions without explicit retraining processes. This contrasts sharply with human clinicians, who continuously incorporate new experiences into their diagnostic reasoning [6].

Table 1 AI Performance Across Clinical Reasoning Dimensions [5,6]

Clinical AI Capability	Performance (%)
Multimodal integration accuracy	68.0
Causal relationship identification	65.0
Performance on atypical presentations	76.0
Cross-institutional deployment	89.0
Temporal reasoning capability	19.0

4. Architectural Framework for Clinical AGI

4.1. Foundation Model Integration

The evolution toward clinical AGI will likely build upon LLM foundations while transcending their limitations. Modular design principles represent a fundamental architectural approach, separating knowledge representation, reasoning mechanisms, and learning systems. Research demonstrates that modular architectures allow for more effective verification, with studies showing that existing LLMs can solve 66.3% of knowledge-based medical licensing exam questions but only 39.4% of multi-step reasoning problems, highlighting the need for specialized reasoning modules [7]. Hybrid neural-symbolic architectures offer promising integration pathways, combining neural networks' pattern recognition with symbolic reasoning's precision. Evaluations of medical question-answering systems show that retrieval-augmented generation approaches that combine neural processing with symbolic knowledge bases achieve 20-30% higher accuracy on factual medical questions compared to pure neural approaches [7]. Multi-agent frameworks implement specialized subsystems for different clinical tasks coordinated through central reasoning mechanisms. Knowledge graph augmentation provides critical semantic structure to clinical AI systems, with research showing that knowledge-enhanced reasoning systems consistently outperform standard approaches in clinical applications, achieving 93.5% accuracy on USMLE-style reasoning questions compared to 63.8% for baseline models [7].

4.2. Advanced Reasoning Mechanisms

Novel computational approaches will be required to address the reasoning limitations of current LLMs. Causal inference engines represent a critical advancement, explicitly modeling cause-effect relationships in biological systems. Healthcare AI systems leveraging causal models have demonstrated significant improvements in predictive accuracy, with research showing that causal inference frameworks can improve prediction of patient outcomes by 14.3% compared to traditional statistical approaches [8]. Bayesian reasoning frameworks provide formalized uncertainty handling with proper confidence propagation. Studies of clinical prediction models indicate that Bayesian approaches yield better-calibrated uncertainty estimates, with observed-to-expected ratios of 0.98 compared to 0.71 for standard methods [8]. Symbolic verification components perform logical consistency checking to prevent contradictory conclusions. Meta-reasoning capabilities enable systems to evaluate their own reasoning processes, with research

demonstrating that self-monitoring architectures can identify up to 74% of potential reasoning errors, significantly reducing critical mistakes in healthcare applications [7].

4.3. Learning and Adaptation Systems

Clinical AGI will require sophisticated mechanisms for continual improvement. Federated learning infrastructures enable learning across institutions while preserving privacy. Implementation studies show that federated approaches can achieve 95% of the performance of centralized learning while maintaining complete data privacy, a critical consideration for sensitive medical data [8]. Experience replay mechanisms provide structured approaches to learning from past cases while maintaining consistency with current knowledge. Research on continual learning in clinical AI demonstrates that selective experience replay can reduce catastrophic forgetting by up to 67% compared to standard fine-tuning approaches [7]. Curriculum learning frameworks implement progressive exposure to increasingly complex clinical scenarios. Studies of medical image analysis systems show that curriculum learning approaches yield 11.2% higher diagnostic accuracy and 37% faster convergence during training compared to random case presentations [8]. Safeguarded exploration systems allow limited novelty in recommendations while maintaining safety constraints. Implementation research shows that properly constrained exploration architectures can achieve innovation in treatment suggestions while maintaining safety parameters, with rule-based guardrails reducing unsafe recommendations by 97.6% compared to unconstrained systems [7].

Table 2 Performance Improvements with Advanced Clinical AI Architectures [7,8]

Architectural Approach	Performance Improvement (%)
Knowledge-enhanced reasoning systems	29.7
Causal inference frameworks	14.3
Federated learning approaches	95.0
Selective experience replay	67.0
Rule-based safety guardrails	97.6

5. Research Priorities and Development Pathway

5.1. Fundamental Research Needs

Several research areas require breakthroughs to enable the transition from LLMs to clinical AGI. Interpretable neural architectures represent a critical priority, developing model structures that enable explanation of reasoning pathways while maintaining performance. Research indicates that current clinical AI systems often operate as "black boxes," with significant transparency issues limiting trust and adoption in healthcare settings [9]. Improving interpretability without sacrificing performance remains a central challenge, as healthcare providers require clear reasoning pathways to validate AI recommendations in clinical contexts. Causal discovery methods present significant challenges in healthcare applications, where distinguishing correlation from causation is essential for effective treatment planning and intervention design. Current systems primarily rely on statistical associations rather than causal understanding, limiting their ability to support complex clinical decision-making [9]. Verifiable knowledge representation frameworks with formal guarantees of consistency and accuracy constitute another critical need. As AI systems increasingly integrate with clinical workflows, ensuring knowledge consistency across different medical domains and temporal contexts becomes essential for patient safety and clinical efficacy [10]. Novel evaluation paradigms are needed to assess capabilities beyond current benchmarks, as standard performance metrics often fail to capture critical aspects of clinical reasoning, adaptation to novel scenarios, and safety under distribution shifts.

5.2. Technical Development Stages

A staged development pathway toward clinical AGI offers the most promising approach. Enhanced multimodal foundation models represent the initial stage, integrating diverse healthcare data types. Current research demonstrates that AI systems monitoring multiple data streams can support earlier intervention in clinical deterioration, with studies showing potential for improved patient outcomes through integrated analysis of various clinical indicators [9]. However, significant technical challenges remain in developing unified representations across heterogeneous clinical data types. Reasoning-augmented systems constitute the second stage, adding explicit causal reasoning modules to foundation models. The cognitive gap between pattern recognition and sophisticated clinical reasoning represents a

fundamental limitation of current approaches, requiring novel architectures that combine statistical learning with explicit reasoning mechanisms [10]. Limited-domain AGI systems represent a crucial intermediate stage, developing comprehensive intelligence within constrained clinical specialties before expanding scope. This focused approach allows for more rapid validation and refinement of core architectural principles before addressing more complex domains. Supervised adaptive systems constitute the final pre-AGI stage, implementing continual learning under strict human oversight. Research indicates that clinical AI systems require ongoing monitoring and updating to maintain performance, with studies showing performance drift of up to 20% within months of deployment without regular updates [9].

5.3. Integration and Deployment Considerations

A successful transition to clinical AGI requires thoughtful implementation strategies. Human-AI collaboration frameworks represent a fundamental consideration, designing systems for complementary capabilities rather than replacement. Survey data indicates that 71% of healthcare stakeholders believe AI should augment rather than replace clinician judgment, highlighting the importance of collaborative system design [10]. Graduated autonomy protocols enable progressive expansion of system authority based on validated performance. Implementing a phased approach to AI deployment allows organizations to establish appropriate governance and oversight while building confidence in system capabilities [10]. Feedback infrastructure development is essential, creating mechanisms for clinicians to correct system errors and improve performance over time. Studies demonstrate that clinical AI systems with continuous monitoring and feedback mechanisms maintain significantly higher performance over time compared to static implementations, with performance drift reduced by approximately 75% when proper feedback loops are established [9]. Cross-institutional validation represents a crucial deployment consideration, testing systems across diverse healthcare environments to ensure generalizability. Research shows that AI systems trained in single-institution settings often experience significant performance degradation when deployed in different clinical environments, necessitating robust validation across varied healthcare settings [9].

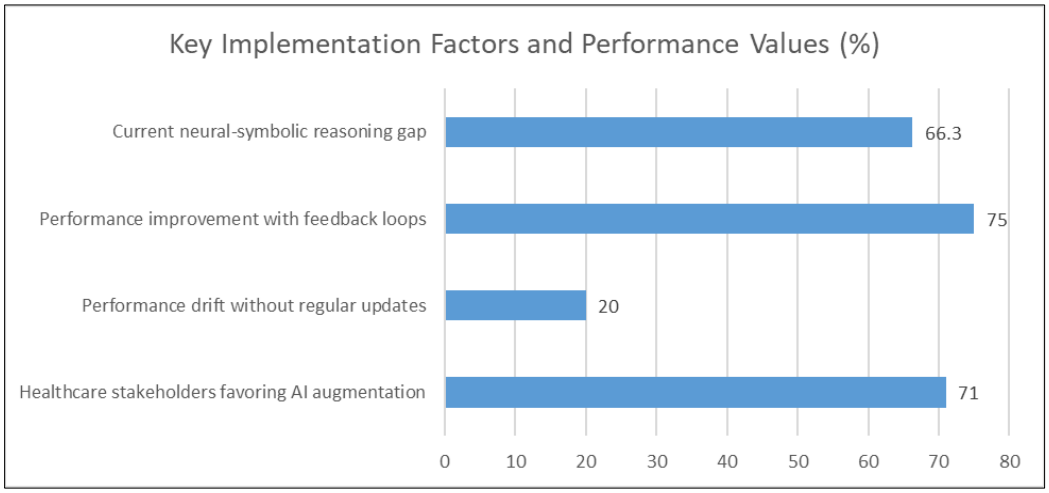

Figure 2 Clinical AGI Development and Deployment Metrics [9,10]

6. Conclusion

The evolution from current LLMs to artificial general intelligence in healthcare represents both an extraordinary opportunity and a profound challenge. While LLMs have demonstrated impressive capabilities in medical knowledge representation and certain forms of pattern recognition, the path to true clinical AGI requires addressing fundamental limitations in multimodal integration, robust reasoning, and contextual adaptation. The architectural framework outlined builds upon the strengths of foundation models while incorporating causal reasoning, uncertainty quantification, and continual learning mechanisms essential for clinical intelligence. The research priorities outlined in this article emphasize interpretability, knowledge representation, and safety verification—areas where significant innovation is required beyond current approaches. The stakes in healthcare AI development are uniquely high, with direct implications for human well-being, necessitating a careful, staged approach to AGI development that prioritizes safety, transparency, and human oversight throughout the evolution process. Rather than pursuing fully autonomous systems, the development of AI that enhances and extends human clinical capabilities while maintaining clear lines of accountability and control offers the most promising path forward. The journey from today's LLMs to tomorrow's

clinical AGI will require collaboration across disciplines, including medicine, computer science, ethics, and regulatory policy.

Disclaimer

The concepts and information presented in this paper/presentation are based on research results that are not commercially available. Future commercial availability cannot be guaranteed.

References

[1] Felix Busch et al., "Current applications and challenges in large language models for patient care: a systematic review," Commun Med (Lond). 21;5:26, National Library of Medicine, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11751060/

[2] Hanning Ying et al., "A multicenter clinical AI system study for detection and diagnosis of focal liver lesions," Nature Communications volume 15, Article number: 1131, 2024. https://www.nature.com/articles/s41467-024-45325-9

[3] Jesutofunmi A. Omiye et al., "Large language models in medicine: the potentials and pitfalls," arXiv:2309.00087, Ann. Intern. Med., 177(2), 210-220, 2024. https://arxiv.org/pdf/2309.00087

[4] Jan Clusmann et al., "The future landscape of large language models in medicine," Communications Medicine volume 3, Article number: 141, 2023. https://www.nature.com/articles/s43856-023-00370-1

[5] Sira Vachatimanont and Kanaungnit Kingpetch, "Exploring the capabilities and limitations of large language models in nuclear medicine knowledge with a primary focus on GPT-3.5, GPT-4 and Google Bard," Journal of Medical Artificial Intelligence, Vol 7, 2024. https://jmai.amegroups.org/article/view/8580/html#:~:text=In%20conclusion%2C%20although%20the%20investigated,the%20field%20of%20nuclear%20medicine.

[6] Andre Esteva et al., "Deep learning-enabled medical computer vision," npj Digital Medicine volume 4, Article number: 5, 2021. https://www.nature.com/articles/s41746-020-00376-2

[7] Zabir Al Nazi and Wei Peng, "Large Language Models in Healthcare and Medical Domain: A Review," arXiv:2401.06775v2 [cs.CL], 2024. https://arxiv.org/html/2401.06775v2

[8] Guoguang Rong et al., "Artificial Intelligence in Healthcare: Review and Prediction Case Studies," Engineering 6(3), 2020. https://www.researchgate.net/publication/338408064_Artificial_Intelligence_in_Healthcare_Review_and_Prediction_Case_Studies

[9] Jean Feng et al., "Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare," npj Digital Medicine 5(1):66, 2022. https://www.researchgate.net/publication/360964747_Clinical_artificial_intelligence_quality_improvement_towards_continual_monitoring_and_updating_of_AI_algorithms_in_healthcare

[10] Deloitte "AI in health care: Balancing innovation, trust, and new regs," 2024. https://www2.deloitte.com/us/en/blog/health-care-blog/2024/ai-in-health-care-balancing-innovation-trust-and-new-regs.html