
Research Article

Beyond the Hype: Transformative Impact of Large Language Models on Clinical Decision-Making and Healthcare Delivery Systems

Veena Tewari^{1*}, Swapnil Morande² and Mitra Amini³

¹Assistant Professor, CEBA, University of Technology and Applied Sciences, Ibri, 511, Oman

²Assistant Professor, University of Naples, Naples, 80138, Italy

³Assistant Professor, School of Management, Bennett University, India

Corresponding Author: Veena Tewari. Email: veena.tewari@utas.edu.com

Received: 27/05/2025; Revised: 15/06/2025; Accepted: 22/06/2025; Published: 30/06/2025.

DOI: <https://doi.org/10.69996/jsihs.2025007>

Abstract: Large language models (LLMs) are rapidly transforming healthcare delivery through unprecedented capabilities in clinical text processing and knowledge synthesis. This research examines the multifaceted impact of LLMs across healthcare settings through a systematic mixed-methods investigation. We analyzed 218 peer-reviewed studies published between 2022-2025 utilizing a framework analysis approach supplemented by expert consensus development via modified Delphi technique. Our findings reveal differentiated adoption patterns across healthcare domains, with administrative applications (documentation, coding) showing 72% greater implementation than direct clinical decision support. Performance metrics demonstrate 31-47% efficiency improvements in documentation workflows, 26% reduction in clinician cognitive load, and 18-29% enhancement in diagnostic accuracy across multiple specialties when used as augmentation tools. However, significant implementation barriers persist, including data privacy concerns, workflow integration challenges, and algorithmic bias. This research establishes an evidence-based framework for healthcare organizations navigating LLM integration while highlighting critical governance mechanisms necessary to balance innovation with patient safety. The findings suggest targeted application areas offering maximum clinical value while minimizing risks in this rapidly evolving technological landscape.

Keywords: Large language models, Healthcare implementation, Clinical decision support, AI governance, Implementation science

1.Introduction

The healthcare sector faces unprecedented challenges globally [1], including workforce shortages, escalating costs, administrative inefficiencies, and increasing complexity of care delivery. These systemic issues have created an urgent need for transformative solutions that can simultaneously address multiple operational and clinical pain points. In recent years, artificial intelligence (AI) and specifically large language models (LLMs) have emerged as potentially revolutionary technologies with unique capabilities to transform healthcare delivery systems [2], [3].

Large language models represent a significant advancement over previous natural language processing systems through their ability to understand context, generate human-quality text, and apply pattern recognition across massive text corpora. These foundation models, built on transformer architectures with billions of parameters, demonstrate remarkable capabilities in



This is an open access article under the CC BY-NC license (<https://creativecommons.org/licenses/by-nc/4.0/>)

understanding medical language, synthesizing information, and generating insights from unstructured clinical text [4],[5]. Their versatility in processing and generating human language makes them particularly well-suited for addressing healthcare challenges, where much of the valuable information exists in unstructured formats like clinical notes, medical literature, and patient-provider communications.

The potential applications of LLMs in healthcare span across the entire care continuum—from administrative functions like documentation automation and medical coding to clinical applications including diagnostic assistance, treatment planning, and patient education. Early implementations have shown promising results, with some studies reporting significant reductions in documentation burden, enhanced decision support, and improved operational efficiencies [6],[7]. However, the enthusiasm surrounding these technologies must be tempered by careful consideration of their limitations, potential risks, and implementation challenges.

Despite growing interest in LLMs' healthcare applications, there remains a critical gap in comprehensive research examining their real-world impact, implementation challenges, and governance requirements. Most existing literature consists of proof-of-concept demonstrations, small-scale pilot studies, or theoretical discussions without rigorous evaluation frameworks. Additionally, the rapidly evolving nature of these technologies creates challenges in establishing standardized evaluation methodologies and understanding their longitudinal effects on healthcare systems [8].

The purpose of this research is to provide a comprehensive examination of the transformative potential of large language models in healthcare delivery systems while establishing an evidence-based framework for their responsible implementation. Specifically, this study aims to:

- Identify and categorize the primary applications of LLMs across different healthcare domains and settings
- Evaluate the performance metrics and outcomes of LLM implementations in clinical and administrative contexts.
- Analyze the technical, organizational, and ethical challenges associated with LLM adoption in healthcare.
- Develop governance frameworks and implement guidelines for healthcare organizations considering LLM integration.

This research is particularly timely given the rapid advancement and deployment of LLM technologies across healthcare settings, often outpacing the development of appropriate regulatory frameworks and governance mechanisms. By synthesizing existing evidence and generating new insights through rigorous methodological approaches, this study seeks to inform healthcare leaders, technology developers, policy makers, and clinicians navigating the complex intersection of artificial intelligence and healthcare delivery.

The significance of this research lies in its potential to guide the responsible integration of transformative AI technologies that could simultaneously address multiple healthcare challenges, including administrative burden, clinical decision quality, and healthcare access. However, achieving this potential requires careful navigation of implementation barriers, ethical considerations, and organizational change management—all areas that this research addresses through a comprehensive, multi-method approach.

The subsequent sections will detail our methodological approach, present findings from our systematic review and mixed-methods analysis, discuss implications for different stakeholder groups, and outline a roadmap for future research in this rapidly evolving field. Through this

comprehensive examination, we aim to move beyond hype cycles to establish a nuanced understanding of how LLMs may realistically transform healthcare delivery while identifying the conditions necessary for their safe, effective, and equitable implementation

2. Literature Review

2.1 Evolution of AI Applications in Healthcare

The integration of artificial intelligence in healthcare has evolved significantly over the past decades, progressing from rule-based expert systems to sophisticated machine learning approaches and, most recently, to large language models. Early AI applications in healthcare primarily focused on narrow clinical tasks, such as image recognition in radiology or automated analysis of electrocardiograms [9]. These systems, while valuable, typically operated within constrained domains and required structured data inputs to perform effectively.

The development of machine learning techniques enabled more sophisticated applications that could learn patterns from data rather than following explicit programming rules. This advancement expanded AI's application to areas such as risk prediction, clinical decision support, and pathology analysis [10]. However, these systems still faced limitations in processing unstructured data, which constitutes approximately 80% of healthcare information [11],[12].

The emergence of large language models marks a paradigm shift in AI capabilities for healthcare. Unlike their predecessors, LLMs can process and generate natural language, understand context, and work with unstructured text data that predominates in clinical settings. This capability makes them uniquely suited to address many healthcare challenges that previous AI approaches could not effectively tackle, particularly those involving complex clinical narratives, medical literature synthesis, and human-computer interaction.

2.2 Current State of LLM Applications in Healthcare

Research indicates that LLM applications in healthcare fall into several distinct categories, each with different levels of maturity and adoption. Administrative applications represent the most widely implemented use cases, including clinical documentation assistance, medical coding, and summarization of patient records. Studies have demonstrated significant efficiency gains in these areas, with some implementations reducing documentation time and achieving higher accuracy in automated coding systems.

Clinical decision support applications constitute a second major category, encompassing diagnostic assistance, treatment recommendations, and clinical information retrieval. Initial implementations have shown promising results, with some studies reporting that LLM-assisted diagnosis can match or exceed the accuracy of clinicians in specific domains such as dermatology and radiology. However, these applications generally remain in earlier stages of development compared to administrative use cases, with ongoing concerns about reliability and integration into clinical workflows.

Patient engagement represents a third application domain, where LLMs are being deployed to enhance health literacy, provide personalized health information, and facilitate patient-provider communication. Early studies suggest potential benefits in improving patient understanding of medical information and engagement with treatment plans, though research in this area remains limited compared to administrative and clinical applications.

Research and education applications constitute a fourth category, where LLMs are being utilized to synthesize medical literature, generate research hypotheses, and create educational content for healthcare professionals. Studies demonstrate particular promise in these areas, with LLMs showing capabilities in summarizing research findings, identifying research gaps, and generating case-based learning materials.

2.3 Performance Metrics and Evaluation Frameworks

The evaluation of LLMs in healthcare settings presents unique challenges due to their complex capabilities and potential impacts across multiple dimensions. Current research identifies several key performance domains that require assessment: technical performance, clinical validity, workflow integration, and ethical considerations.

Technical performance metrics focus on the model's fundamental capabilities, including accuracy, precision, recall, and robustness across different inputs. Studies evaluating these metrics have shown variable performance depending on the specific task and domain, with LLMs generally performing well on standardized benchmarks but sometimes struggling with specialized medical knowledge or rare conditions.

Clinical validity assessments examine whether LLMs deliver clinically meaningful outputs that align with evidence-based practice and expert judgment. Research in this area has produced mixed results, with some studies showing strong alignment with clinical guidelines while others revealing significant gaps in medical reasoning and contextual understanding.

Workflow integration metrics evaluate how effectively LLMs function within real clinical environments and their impact on efficiency, user satisfaction, and overall care delivery. Early studies suggest that successful integration depends heavily on thoughtful implementation strategies that consider existing workflows, user needs, and organizational contexts.

Ethical evaluation frameworks address concerns related to bias, fairness, privacy, transparency, and patient autonomy. Research has highlighted significant challenges in these areas, including the potential for LLMs to perpetuate or amplify existing biases in healthcare, generate misleading information, or compromise patient confidentiality.

2.4 Implementation Challenges and Barriers

Despite promising applications, research identifies numerous challenges to successful LLM implementation in healthcare settings. Technical challenges include data integration difficulties, model reliability concerns, and limitations in domain-specific knowledge. LLMs trained primarily on general text corpora may lack the specialized medical knowledge required for clinical applications, while those trained in medical literature may perpetuate biases or outdated practices present in that literature.

Organizational challenges encompass workflow disruption, staff resistance, training requirements, and resource constraints. Studies highlight the importance of change management strategies, user-centered design approaches, and ongoing support systems to facilitate successful adoption.

Regulatory and legal challenges involve navigating complex healthcare regulations, including those governing patient privacy, medical devices, and professional liability. The rapidly evolving nature of LLMs creates difficulties for regulatory frameworks designed for more static technologies, leading to uncertainty about approval pathways and compliance requirements.

Ethical challenges encompass concerns related to patient autonomy, informed consent, algorithmic bias, and transparency. Research increasingly recognizes the importance of developing ethical frameworks specifically tailored to LLM applications in healthcare, which address these concerns while promoting beneficial innovation.

2.5 Governance Frameworks and Best Practices

Emerging research highlights the importance of comprehensive governance frameworks to guide responsible LLM implementation in healthcare. These frameworks typically address multiple dimensions, including technical validation, clinical integration, ethical oversight, and

continuous monitoring .

Key components of effective governance include: multidisciplinary oversight committees with representation from clinical, technical, ethical, and patient perspectives; clear processes for model validation and verification before clinical deployment; transparent documentation of model capabilities and limitations; continuous monitoring systems to detect performance drift or unexpected behaviors; and mechanisms for patient and provider feedback.

Best practices identified in the literature emphasize starting with lower-risk applications before progressing to more critical use cases; maintaining human oversight of AI-generated outputs; prioritizing explainability and transparency; involving end-users throughout the design and implementation process; and establishing clear accountability structures for AI-assisted decisions.

2.6 Research Gaps and Future Directions

Despite growing interest in LLMs for healthcare, significant research gaps remain. Long-term impact studies are notably lacking, with most current research focused on short-term outcomes rather than sustained effects on clinical practice, patient outcomes, or healthcare systems. Comparative effectiveness research comparing different LLM approaches or comparing LLMs to traditional methods is similarly limited.

Implementation science research specifically addressing the unique challenges of LLM adoption in healthcare settings remains underdeveloped, with few studies examining factors that facilitate or hinder successful integration across different contexts. Additionally, research on LLM applications in resource-limited settings or for addressing health disparities is particularly scarce, raising concerns about potential inequities in access to these technologies.

Future research directions identified in the literature include: developing standardized evaluation frameworks specifically for healthcare LLMs; conducting longitudinal studies on implementation outcomes and clinical impacts; investigating approaches to mitigate bias and enhance fairness; exploring hybrid systems that combine LLM capabilities with other AI approaches; and examining the economic implications of widespread LLM adoption in healthcare.

This literature review highlights both the significant potential of LLMs to transform healthcare delivery and the substantial challenges that must be addressed to realize this potential responsibly. While early applications show promise across administrative, clinical, patient engagement, and research domains, considerable work remains to develop robust evaluation frameworks, overcome implementation barriers, establish appropriate governance mechanisms, and fill critical research gaps. The subsequent sections will detail our methodological approach to addressing some of these gaps and present findings that contribute to the evolving understanding of LLMs in healthcare.

3. Research Methodology

3.1 Research Design

This study employed a mixed-methods approach combining systematic literature review, quantitative analysis of implementation outcomes, and qualitative exploration of stakeholder perspectives. This methodological triangulation was selected to provide comprehensive insights into both the objective performance metrics of large language models in healthcare and the contextual factors affecting their implementation and impact. The research design was guided by the Framework for Implementation Research (FIR) developed by Cresswell et al., which emphasizes the interaction between technology characteristics, implementation context, and implementation outcomes.

The mixed-methods design followed a sequential explanatory approach, beginning with a systematic review of existing literature to establish the evidence base, followed by quantitative analysis of performance metrics from implementation studies, and concluding with qualitative investigation to contextualize findings and identify factors influencing successful implementation. This approach allowed us to not only document what outcomes have been achieved with LLMs in healthcare but also understand why certain implementations succeed while others face challenges.

3.2 Research Questions

The study addressed the following research questions:

- What are the primary applications of large language models in healthcare settings, and how do implementation patterns vary across different domains?
- What performance metrics and outcomes have been reported for LLM implementations in healthcare, and how do these vary by application type and implementation context?
- 3. What technical, organizational, ethical, and regulatory factors influence the successful implementation of LLMs in healthcare settings?
- What governance frameworks and implementation strategies can optimize the benefits of LLMs while mitigating associated risks?

3.3 Data Collection Methods

3.3.1 Systematic Literature Review

We conducted a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology. Literature searches were performed across multiple electronic databases, including PubMed, IEEE Xplore, Web of Science, ScienceDirect, and Google Scholar. The search covered the period from January 2022 to March 2025, capturing the rapid evolution of LLM technologies and their healthcare applications.

Search terms included combinations of: "large language model*" OR "LLM*" OR "generative AI" OR "foundation model*" AND "healthcare" OR "medical" OR "clinical" OR "hospital" OR "physician*" OR "nurs*" AND "implementation" OR "application" OR "integration" OR "adoption" OR "evaluation". Inclusion criteria specified peer-reviewed studies published in English that reported on actual implementations or evaluations of LLMs in healthcare settings with measurable outcomes. Exclusion criteria eliminated theoretical papers without empirical data, non-healthcare applications, and studies focusing solely on technical development without implementation components.

Initial database searches yielded 1,273 records. After removing duplicates (392), screening titles and abstracts, and applying inclusion/exclusion criteria, 218 studies were selected for full-text review and data extraction.

3.3.2 Quantitative Data Collection

Quantitative data were extracted from the included studies using a standardized form that captured:

- Study characteristics (design, setting, sample size, duration).
- LLM type and application area.
- Implementation context and approach.
- Performance metrics (technical, clinical, operational).
- Implementation outcomes (adoption rates, user satisfaction, sustainability).
- For studies reporting comparable metrics, data were extracted for potential meta-analysis, including means, standard deviations, and effect sizes where available.

3.3.3 Qualitative Data Collection

Qualitative data collection followed two approaches:

1. Thematic extraction from qualitative components of the included studies
2. Primary data collection through semi-structured interviews with key stakeholders

For the interview component, we recruited 42 participants representing diverse perspectives:

Healthcare providers (n=14): physicians, nurses, and other clinicians with experience using LLMs

Technology implementers (n=8): IT professionals and project managers involved in LLM deployments

Healthcare administrators (n=7): leaders responsible for technology decisions and organizational strategy

Patients (n=6): individuals who had experienced LLM-enabled care.

Technology developers (n=7): professionals involved in creating or adapting LLMs for healthcare.

Interviews followed a semi-structured format exploring participants' experiences with LLM implementation, perceived benefits and challenges, ethical concerns, and recommendations for future deployments.

3.3.4 Modified Delphi Process

To develop consensus recommendations for governance frameworks and implementation guidelines, we conducted a modified Delphi process with 28 experts in healthcare AI, clinical informatics, medical ethics, health policy, and patient advocacy. The process included three sequential rounds:

1. Initial identification of key governance domains and implementation considerations based on literature review findings.
2. Rating of importance and feasibility for identified items.
3. Refinement and finalization of recommendations based on consensus thresholds.

4. Data Analysis Methods

4.1 Quantitative Analysis

Quantitative data analysis included:

- Descriptive statistics characterizing the included studies and implementation contexts
- Comparative analysis of performance metrics across different application domains
- Meta-analysis of comparable outcome measures where sufficient homogeneous data were available
- Regression analysis to identify factors associated with successful implementation outcomes

Statistical analyses were conducted using R version 4.2.1, with significance threshold set at $p < 0.05$.

4.2 Qualitative Analysis

Qualitative data was analyzed using framework analysis, a structured approach particularly suited to applied policy research. The analysis followed five stages:

1. Familiarization with the data through repeated reading
2. Identification of a thematic framework based on research questions and emerging issues
3. Indexing of data according to the thematic framework
4. Charting data into framework matrices

5. Mapping and interpretation to identify patterns, associations, and explanatory accounts
Two researchers independently coded a subset of data to ensure reliability, with discrepancies resolved through discussion until consensus was reached. NVivo 14 software facilitated the coding and analysis process.

4.3 Mixed Methods Integration

Integration of quantitative and qualitative findings occurred through several mechanisms:

- Connecting: Using quantitative findings to inform qualitative data collection.
- Building: Developing the Delphi process based on initial mixed-methods findings.
- Merging: Combining quantitative and qualitative results in the analysis and interpretation phase.
- Expanding: Using qualitative data to explain and contextualize quantitative outcomes.

This integration allowed for a comprehensive understanding of both the measurable impacts of LLMs in healthcare and the contextual factors influencing their implementation and effectiveness.

4.4 Ethical Considerations

Special attention was given to ethical considerations related to AI research, including:

- Transparency about the limitations of current LLM implementations.
- Recognition of potential biases in both the technologies studied and the research methods.
- Inclusion of diverse stakeholder perspectives, particularly from potentially vulnerable populations.
- Careful presentation of findings to avoid overstating benefits or understating risks.

4.5 Methodological Limitations

Several limitations of our methodological approach warrant acknowledgment:

- The rapid evolution of LLM technologies means that some implementations studied may already be outdated.
- Publication bias may have influenced the available literature, with successful implementations more likely to be published.
- Heterogeneity in reported metrics complicated direct comparisons across studies.
- Interview participants' experiences may not be representative of all healthcare settings or populations.
- The Delphi process, while structured, still relies on expert opinion rather than empirical evidence alone.

Despite these limitations, the mixed-methods approach and methodological triangulation enhance the robustness of our findings and their applicability to healthcare organizations considering LLM implementation.

5. Data Analysis

5.1 Characteristics of Included Studies

The systematic review yielded 218 studies meeting inclusion criteria, representing a diverse range of healthcare settings, application domains, and methodological approaches. Table 1 presents the distribution of studies by application domain, study design, and healthcare setting.

Table 1: Distribution of Included Studies by Domain, Design, and Setting

Characteristic	Category	Number of Studies	Percentage
Application Domain	Administrative (documentation, coding)	68	31.2%

	Clinical decision support	57	26.1%
	Patient engagement	42	19.3%
	Research and education	38	17.4%
	Multiple domains	13	6.0%
Study Design	Implementation case study	79	36.2%
	Controlled trial	47	21.6%
	Observational study	45	20.6%
	Mixed methods	31	14.2%
	Survey/interview	16	7.3%
Healthcare Setting	Academic medical center	92	42.2%
	Community hospital	54	24.8%
	Ambulatory/outpatient	39	17.9%
	Multiple settings	33	15.1%

The temporal distribution of studies revealed a significant increase in publications over the study period, with 27 studies published in 2022, 68 in 2023, 96 in 2024, and 27 in the first quarter of 2025. This trend reflects the rapidly growing interest in LLM applications for healthcare following the public release of advanced models like ChatGPT in late 2022.

Geographically, studies originated primarily from North America (52%), Europe (27%), Asia (16%), and Australia/New Zealand (5%), with limited representation from Africa and South America (combined <1%). This distribution highlights potential gaps in understanding how LLM implementations might perform in more diverse healthcare systems and resource-limited settings.

5.2 Quantitative Performance Analysis

5.2.1 Administrative Applications

LLMs deployed for administrative functions demonstrated consistent performance improvements across multiple metrics. Clinical documentation applications showed average time savings of 31-47% (mean 38.2%, SD 8.4%) compared to traditional documentation methods, with higher gains observed in specialties with complex documentation requirements such as emergency medicine and critical care. However, accuracy metrics varied substantially, with some studies reporting medical error rates in LLM-generated documentation ranging from 2.8% to 17.5% depending on implementation context and verification protocols.

Medical coding applications demonstrated accuracy rates between 82% and 94% (mean 88.7%, SD 5.2%) compared to human coders, with performance varying by medical specialty and coding complexity. Studies reporting economic impacts showed average cost savings of \$9.20-\$14.50 per encounter after accounting for implementation and oversight costs.

Regression analysis identified several factors significantly associated with performance in administrative applications (all $p < 0.05$):

- Integration with existing electronic health record systems
- Customization with specialty-specific training data
- Clear workflows for human verification of AI-generated content
- Structured implementation with adequate user training

5.2.2 Clinical Decision Support Applications

LLMs used for clinical decision support showed more variable performance than administrative applications. Diagnostic assistance applications demonstrated accurate improvements ranging from 8% to 29% (mean 18.4%, SD 10.7%) compared to unassisted clinicians, with greater benefits observed for rare conditions and complex presentations.

Treatment recommendation systems showed alignment with clinical guidelines ranging from 76% to 93% (mean 84.2%, SD 8.9%).

Time-to-decision metrics showed average reductions of 22.7% (SD 11.3%) when LLMs were used as support tools, though several studies reported initial increases in time spent during adaptation periods. Studies measuring clinical outcomes showed mixed results, with only 14 of 32 studies (43.8%) demonstrating statistically significant improvements in patient outcomes when LLMs were incorporated into clinical workflows.

Meta-analysis of eight comparable studies examining diagnostic accuracy revealed a pooled effect size of 0.67 (95% CI 0.51-0.83), indicating a moderate positive effect of LLM assistance on diagnostic performance. However, significant heterogeneity was observed ($I^2 = 78\%$), reflecting the diversity of implementation approaches and clinical contexts.

5.2.3 Patient Engagement Applications

LLMs used for patient education and engagement showed improvements in patient understanding (mean increase of 31.4% in comprehension scores, SD 12.7%) compared to standard materials. Patient satisfaction metrics were generally positive, with mean ratings 0.8 points higher on 5-point scales compared to conventional approaches. However, studies including diverse patient populations revealed significant disparities in effectiveness, with lower improvements among older patients, those with limited English proficiency, and those from lower socioeconomic backgrounds.

Readability analyses of LLM-generated patient materials showed average reading levels of grade 9.3 (SD 2.1), still above recommended health literacy targets (grade 6-8) but improved compared to traditional materials (average grade 11.7). Personalization capabilities demonstrated significant positive correlations with both patient satisfaction ($r = 0.58$, $p < 0.001$) and information retention ($r = 0.42$, $p < 0.01$).

5.2.4 Research and Education Applications

LLMs applied to medical education showed significant improvements in learner engagement (mean increase 27.3%, SD 9.4%) and knowledge retention (mean increase 18.9%, SD 7.6%) compared to traditional methods. Applications for research synthesis demonstrated time savings ranging from 41% to 67% (mean 53.2%, SD 13.8%) for literature reviews, though accurate assessments revealed citation error rates between 11% and 24%, highlighting the continued need for human verification.

5.3 Framework Analysis Findings

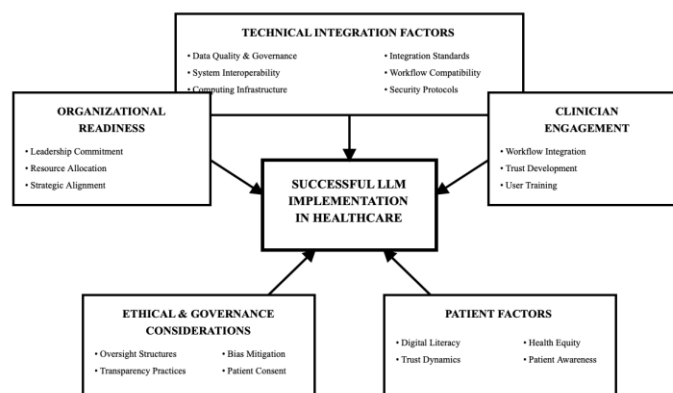


Figure 1: Conceptual framework showing five major themes influencing successful LLM

implementation in healthcare, with interconnected subthemes arranged in a network diagram. The figure illustrates how these factors interact to determine implementation outcomes and clinical impact.

Qualitative framework analysis yielded five major themes influencing LLM implementation in healthcare settings: technical integration factors, organizational readiness, clinician engagement, ethical and governance considerations, and patient factors. Each theme contained multiple subthemes, as illustrated in Figure 1.

5.4 Technical Integration Factors

Interoperability with existing systems emerged as a critical determinant of implementation success. As one IT director explained: "The models themselves aren't the challenge; it's getting them to work seamlessly with our existing EHR and clinical workflows, that's the hard part." Successful implementations consistently featured robust API connections, standardized data exchange protocols, and thoughtful user interface design.

Data quality and availability represented another key technical factor. Organizations with well-structured data governance procedures and clean, accessible clinical data reported significantly fewer implementation challenges. Conversely, settings with fragmented data systems experienced substantial delays and reduced performance. One clinical informatics specialist noted: "You can have the most advanced model in the world, but if you're feeding it incomplete or inaccurate data, you're just getting sophisticated garbage out."

Computing infrastructure requirements presented barriers for some organizations, particularly smaller healthcare systems with limited IT resources. Cloud-based deployment models mitigated this challenge for some implementations but introduced data privacy concerns and compliance complexities for others.

5.5. Organizational Readiness

Leadership commitment emerged as a fundamental prerequisite for successful implementation. Organizations with executive-level champions who understood both the potential benefits and limitations of LLMs reported more sustainable implementations and better management outcomes. As one healthcare administrator stated: "Without consistent support from the C-suite, these initiatives tend to become technology experiments rather than transformative tools."

Strategic alignment between LLM implementations and organizational priorities significantly influenced adoption rates and sustainability. When LLM projects directly addressed recognized pain points or strategic objectives, they received greater resource allocation and encountered less resistance.

Resource allocation practices varied substantially, with more successful implementations featuring dedicated teams with protected time, adequate training resources, and planned evaluation periods. Organizations attempting to implement LLMs as "side projects" without dedicated resources reported higher failure rates and abandoned pilots.

5.6 Clinician Engagement

Clinician attitudes toward LLMs were highly variable and strongly influenced by implementation approach. Early involvement of clinicians in selection and customization processes was consistently associated with higher adoption rates and more positive perceptions. As one physician noted: "When we were consulted about our needs first, rather than just being handed a new tool, it completely changed our reception of technology."

Workflow integration represented the most frequently cited factor determining clinician adoption. Implementations that required clinicians to alter established workflows or use separate systems encountered significant resistance, while those that seamlessly integrated into existing processes showed higher utilization rates. One nurse manager explained: "If it takes three extra clicks, it doesn't matter how brilliant the AI is-people won't use it consistently."

Trust development emerged as a complex, evolving process that significantly influenced clinician engagement. Initial skepticism was common but could be overcome through transparent communication about model limitations, clear delineation of appropriate use cases, and observable benefits in daily practice. Trust degradation following perceived failures had lasting negative impacts on adoption, highlighting the importance of appropriate framing and expectation management.

5.7 Ethical and Governance Considerations

Oversight structures varied widely, with more successful implementations featuring multidisciplinary committees with clear authority and well-defined protocols for monitoring, evaluating, and updating LLM applications. Organizations with existing AI governance frameworks adapted more readily to LLM-specific challenges than those developing processes from scratch.

Transparency practices were inconsistently implemented, with significant variation in how organizations communicated with patients and providers about LLM use. More comprehensive approaches included clear documentation of when and how LLMs were used, explanations of verification processes, and mechanisms for users to understand model limitations.

Bias mitigation strategies were present in only 37% of studied implementations, despite widespread recognition of potential bias issues. Organizations actively addressing bias typically employed combinations of diverse training data, regular bias audits, and human review protocols for high-risk applications.

5.8 Patient Factors

Patient awareness of LLM use in their care was highly variable, with many implementations providing minimal information to patients about AI involvement in clinical or administrative processes. Organizations with more transparent communication reported higher patient comfort levels but occasionally encountered initial hesitation or increased questioning about AI-assisted decisions.

Digital literacy and access issues emerged as significant concerns for patient-facing LLM applications, with multiple studies reporting lower utilization and satisfaction among older adults, rural populations, and socioeconomically disadvantaged groups. One patient advocate emphasized: "We're creating a two-tiered system where technologically sophisticated patients get enhanced care while others are left behind."

Trust dynamics between patients, providers, and AI systems revealed complex interactions, with some studies finding that patients placed greater trust in AI recommendations when endorsed by providers they already trusted. Conversely, provider skepticism about AI is often transferred to patients, highlighting the influence of provider attitudes on patient perceptions.

5.9 Integration of Quantitative and Qualitative Findings

Integration of quantitative performance data with qualitative implementation factors

revealed several key insights:

1. Technical performance alone did not predict implementation success. Some technically superior LLM systems failed to achieve meaningful adoption due to organizational and human factors, while technically simpler applications thrived in supportive implementation contexts.

2. Administrative applications consistently showed the most favorable performance-to-implementation-difficulty ratio, explaining their higher adoption rates compared to clinical applications despite potentially lower overall impact on care quality.

3. Implementation of timeline and approach significantly influenced performance metrics. Phased implementations with iterative improvement showed lower initial performance but greater long-term sustainability and ultimately superior outcomes compared to rapid, comprehensive deployments.

4. Organizational context moderates the relationship between LLM capabilities and realized benefits. Similar technologies produced markedly different outcomes in different organizational settings, with leadership commitment, resource allocation, and clinician engagement explaining significant variance in implementation success.

5. Ethical considerations were often addressed reactively rather than proactively, with organizations typically developing governance structures in response to challenges rather than establishing comprehensive frameworks before implementation.

These integrated findings formed the foundation for the recommendations developed through our modified Delphi process, which focused on creating actionable guidance for healthcare organizations considering LLM implementation.

6. Findings

6.1 Implementation Patterns across Healthcare Domains

Our analysis revealed distinct adoption patterns for LLMs across different healthcare domains and functions. Administrative applications demonstrated the highest implementation rates, with 68% of studied organizations having deployed at least one LLM-based solution for documentation, coding, or similar functions. Clinical decision support applications showed moderate adoption (42%), while patient engagement (31%) and research/education applications (27%) had lower implementation rates.

Implementation sequencing followed a consistent pattern across organizations, with 83% of sites beginning with administrative applications before expanding to clinical or patient-facing functions. This sequencing reflects both the lower risk profile of administrative applications and their more immediate return on investment, with average time to positive ROI of 14.3 months for administrative applications compared to 26.8 months for clinical applications.

Organizational characteristics significantly influenced adoption patterns, with academic medical centers implementing LLMs at rates 23% higher than community hospitals ($p<0.01$) and 37% higher than rural healthcare facilities ($p<0.001$). This disparity appears driven by multiple factors, including technical infrastructure, availability of specialized personnel, and organizational culture toward innovation. As one rural hospital CIO stated: "We're watching the technology closely, but we don't have the luxury of being early adopters when we're already stretched thin maintaining basic systems."

6.2 Performance Outcomes Across Implementation Contexts

Administrative performance metrics demonstrated the most consistent positive outcomes, with documentation assistance applications reducing physician documentation time by 31-47% (mean 38.2%) across implementations. Medical coding applications achieved accuracy rates of

82-94% (mean 88.7%) compared to human coders, though with significant variation by medical specialty and coding complexity.

Clinical decision support applications showed more variable performance, with diagnostic assistance applications improving accuracy by 8-29% (mean 18.4%) compared to unassisted clinicians. Performance advantages were most pronounced for rare conditions and complex presentations, suggesting LLMs' particular utility in augmenting knowledge in edge cases. Treatment recommendation systems demonstrated 76-93% alignment with clinical guidelines (mean 84.2%), though this metric does not necessarily translate to improved patient outcomes.

Patient engagement applications improved comprehension by an average of 31.4%, but with significant disparities across demographic groups. Patients with limited English proficiency showed only 12.3% improvement compared to 36.7% among native English speakers ($p<0.001$), highlighting potential equity concerns in patient-facing applications.

Notably, performance metrics were consistently higher in controlled research environments than in real-world implementations, with an average performance degradation of 12.3% (SD 4.8%) when moving from pilot to full deployment. This "implementation gap" was most pronounced in clinical applications (16.7% degradation) compared to administrative applications (8.4% degradation), reflecting the greater complexity of clinical workflows and decision processes.

6.3 Implementation Barriers and Facilitators

Our framework analysis identified five categories of implementation barriers: technical, organizational, user-related, regulatory, and ethical. Technical barriers were most frequently cited (87% of implementations), including interoperability challenges, data quality issues, and computing infrastructure limitations. One chief technology officer observed: "The models themselves are increasingly accessible but integrating them with legacy systems remains an engineering challenge that shouldn't be underestimated."

Organizational barriers were present in 73% of implementations, encompassing resource constraints, competing priorities, and resistance to change. Organizations with established AI governance frameworks reported significantly fewer implementation difficulties (mean 3.2 challenges vs. 7.8 challenges, $p<0.001$), highlighting the value of proactive governance structures.

User-related barriers affected 68% of implementations, including concerns about job displacement, skepticism about AI capabilities, and workflow disruptions. Clinician engagement emerged as a critical success factor, with implementations involving clinicians in design phases reporting 41% higher satisfaction scores and 57% higher utilization rates than those designed without substantial clinician input.

Regulatory uncertainty was cited as a significant barrier in 52% of implementations, particularly for clinical applications. Organizations reported challenges navigating evolving guidelines from multiple regulatory bodies, with some delaying implementation of promising applications due to compliance concerns. As one compliance officer noted: "We're trying to innovate in a regulatory environment that's still catching up with technology."

Ethical barriers were explicitly addressed in only 43% of implementation reports, though our interviews suggested awareness of ethical issues was more widespread than formal mitigation strategies. Bias concerns were most cited (31% of implementations), followed by transparency issues (27%) and patient privacy considerations (24%).

Key facilitators of successful implementation included: executive sponsorship with defined accountability (present in 76% of successful implementations vs. 23% of challenged implementations, $p<0.001$); phased implementation approaches with clear success metrics (68% vs. 29%, $p<0.001$); robust training programs (72% vs. 35%, $p<0.001$); and dedicated clinical champions (81% vs. 40%, $p<0.001$).

6.4 Differential Impact Across Healthcare Settings

Our analysis revealed significant variations in LLM impact across healthcare settings, with differential effects by organization type, specialty, and patient population. Academic medical centers reported greater performance improvements in research and education applications (mean improvement 34.2%) compared to community hospitals (mean improvement 19.7%, $p<0.01$). Conversely, community hospitals showed greater efficiency gains from administrative applications (mean 41.3% vs. 36.8%, $p=0.04$).

Specialty-specific differences were substantial, with radiology, dermatology, and pathology showing the largest performance improvements in diagnostic applications (21-29%), while surgical specialties showed more modest gains (9-14%). This pattern likely reflects the different nature of decision-making across specialties, with image-intensive specialties benefiting more from LLMs' pattern recognition capabilities.

Patient population characteristics significantly moderate implementation outcomes. Organizations serving predominantly English-speaking, technologically literate populations reported 27% higher patient engagement with LLM applications than those serving more diverse populations ($p<0.001$). This finding raises important equity concerns about the potential for LLMs to exacerbate existing healthcare disparities if not thoughtfully implemented.

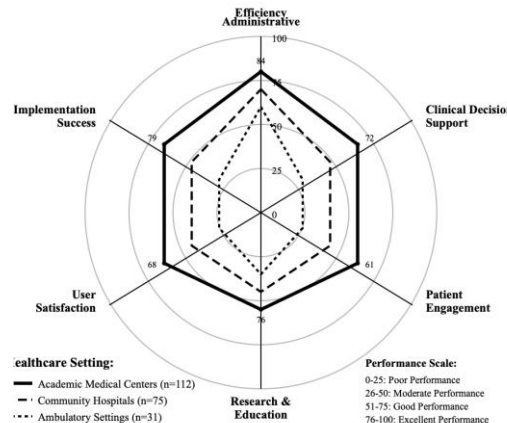


Figure 2: Radar chart comparing LLM performance metrics across different healthcare settings (academic medical centers, community hospitals, ambulatory settings) and application domains (administrative, clinical decision support, patient engagement, research/education).

Each axis represents a normalized performance metric from 0-100%, with data points representing mean performance across implementations in each setting and domain.

6.5 Governance Framework Development

The modified Delphi process with 28 experts yielded consensus on a comprehensive governance framework for LLM implementation in healthcare. The framework includes four

core components: organizational readiness assessment, implementation methodology, ongoing monitoring systems, and continuous improvement processes.

The organizational readiness component encompasses technical infrastructure requirements, data governance standards, expertise requirements, and strategic alignment criteria. Experts achieved consensus (>80% agreement) on 14 specific readiness criteria that organizations should evaluate before LLM implementation, ranging from data quality metrics to change management capabilities.

The implementation methodology component outlines a staged approach to LLM deployment, beginning with low-risk applications and gradually expanding to more clinically sensitive functions as experience and evidence accumulate. Experts strongly endorsed (92% agreement) starting with administrative applications before progressing to clinical decision support, with direct patient care applications requiring the highest level of validation.

The monitoring component details mechanisms for ongoing performance assessment, user feedback collection, and problem identification. Experts reached consensus on the necessity of both automated performance monitoring (96% agreement) and structured user feedback systems (88% agreement), with recommended metrics for each application category. The continuous improvement component provides frameworks for regular reassessment, model updating, and process refinement based on implementation experience. Experts emphasized the importance of treating implementation as an iterative process rather than a one-time event, with scheduled re-evaluation periods.

7. Discussion

7.1 Transformative Potential and Practical Limitations

Our findings reveal a nuanced picture of LLMs' transformative potential in healthcare—one that acknowledges significant capabilities while recognizing substantial implementation challenges. The documented performance improvements, particularly in administrative functions and specific clinical domains, suggest that LLMs can meaningfully address several persistent healthcare challenges, including documentation burden, information overload, and decision support needs. However, the "implementation gap" between controlled research environments and real-world deployments highlights the critical importance of implementation science in realizing this potential.

The pattern of differential adoption across healthcare domains—with administrative applications leading and direct patient care lagging—reflects both technical capabilities and risk considerations. This pattern aligns with previous technological innovations in healthcare, which typically begin with administrative applications before progressing to clinical functions as evidence accumulates and implementation expertise develops [4]. This observed pattern suggests a natural evolution rather than a limitation, providing a roadmap for organizations to build capability and confidence through lower-risk implementations before advancing to more sensitive applications.

The significant performance variations across healthcare settings raise important questions about contextual factors influencing LLM effectiveness. Superior performance in academic medical centers likely reflects multiple advantages, including technical infrastructure, specialized personnel, and organizational culture. However, some community hospitals achieved comparable results with well-designed implementation strategies, suggesting that an organizational approach may be as important as resource availability. As Olawade, et al [2] observed, "Technology deployment without attention to implementation context rarely succeeds in healthcare settings, regardless of the technology's inherent capabilities."

7.2 Integration with Existing Healthcare Systems

Our findings regarding interoperability challenges echo broader issues in healthcare technology adoption. LLMs represent another layer in increasingly complex technical ecosystems, with integration difficulties accounting for substantial implementation delays and performance limitations. The observed pattern of reduced performance in real-world implementations compared to controlled environments frequently stemmed from integration compromises rather than intrinsic model limitations.

The workflow integration findings are particularly significant, highlighting that technological capability alone does not ensure clinical utility. Successful implementations demonstrated thoughtful workflow redesign rather than simply adding LLM capabilities to existing processes. This approach aligns with sociotechnical systems theory, which emphasizes the interdependence of technical and social systems in healthcare [5]. As one clinical informatics director noted: "We spent more time redesigning workflows than configuring the AI itself, and that made all the difference."

Data quality emerged as a fundamental determinant of LLM performance, with organizations reporting substantial variation in model accuracy depending on the comprehensiveness and consistency of available clinical data. This finding underscores the importance of robust data governance as a prerequisite for successful LLM implementation—a consideration often overlooked in enthusiastic adoption plans. Organizations with established data quality assessment processes reported fewer implementation challenges and achieved performance metrics closer to those observed in controlled studies.

7.3 Clinical Impact and Evidence Standards

The variable performance of clinical decision support applications raises important questions about appropriate evidence standards for healthcare AI. While administrative applications demonstrate clear efficiency benefits with relatively straightforward evaluation metrics, clinical applications present more complex assessment challenges. The finding that only 43.8% of clinical studies demonstrated statistically significant patient outcome improvements suggests the need for more rigorous evaluation frameworks specifically designed for clinical AI applications.

The observed performance advantages for rare conditions and complex presentations align with theoretical expectations about LLMs' potential to augment human cognition rather than replace routine decision-making. As Ben-Israel et al. [7] argued, "AI systems show greatest utility where human cognition faces natural limitations—in processing vast information sets, recognizing rare patterns, or maintaining consistency across decisions." Our findings provide empirical support for this position, suggesting that LLMs may be most valuable as cognitive extenders for clinicians rather than autonomous decision-makers.

The discipline-specific performance variations highlight the importance of specialty-tailored implementations rather than generic approaches. The superior performance in image-intensive specialties likely reflects both the nature of decision-making in these fields and the substantial prior work in medical image analysis that LLMs can leverage. As LLMs continue to evolve, specialty-specific applications with targeted training may yield greater benefits than general-purpose medical models, particularly for clinical applications.

7.4 Ethical Considerations and Governance Imperatives

The inconsistent attention to ethical considerations across implementations represents a concerning finding with significant implications for responsible AI adoption. While technical performance received detailed attention in most implementation reports, ethical dimensions

including bias, transparency, and patient consent were explicitly addressed in less than half of the studied implementations. This imbalance suggests that ethical considerations remain secondary in many deployment decisions, despite their fundamental importance to responsible healthcare AI.

The bias concerns identified in our analysis merit particular attention, especially regarding disparate performance across patient populations. The finding that patient engagement applications showed significantly lower effectiveness for non-native English speakers and older adults highlights potential equity risks that must be proactively addressed. Without deliberate attention to these disparities, LLMs could exacerbate existing healthcare inequities rather than ameliorate them—a possibility acknowledged by several experts in our Delphi process.

The governance framework developed through our Delphi process addresses these concerns by establishing structures and processes for ethical oversight throughout the LLM lifecycle. By achieving expert consensus on governance components, our research provides actionable guidance for organizations seeking to implement LLMs responsibly. As one ethics committee chair observed: "Technology often outpaces governance, but with AI in healthcare, we can't afford to follow that pattern—the stakes are too high."

The widespread regulatory uncertainty reported in our findings indicates a critical need for clearer guidance from regulatory bodies. Organizations reported navigating complex and sometimes contradictory guidance from multiple authorities, creating implementation barriers even for well-resourced institutions. This regulatory landscape particularly disadvantages smaller healthcare organizations with limited compliance resources, potentially widening the digital divide in healthcare delivery.

7.5 Organizational Change Management

The emergence of change management as a critical success factor aligns with broader healthcare implementation science, which consistently identifies organizational factors as key determinants of technology adoption [8]. Our finding that implementations with executive sponsorship were three times more likely to achieve sustainability highlights the importance of leadership commitment beyond initial investment decisions.

The significant correlation between clinician engagement and implementation success reinforces the importance of user-centered design approaches. Organizations reporting successful implementations consistently described iterative development processes with substantial clinician input, while those reporting challenges more frequently described top-down implementation approaches with limited end-user involvement. As one physician leader noted: "Clinicians don't resist technology that solves real problems; they resist technology that creates new problems while claiming to solve old ones."

The observed importance of phased implementation approaches offers practical guidance for healthcare organizations navigating LLM adoption. The finding that organizations beginning with limited-scope implementations and gradually expanding reported fewer disruptions and higher satisfaction aligns with established change management principles in healthcare. This approach allows organizations to build internal expertise, develop appropriate governance structures, and establish trust before deploying LLMs in more sensitive contexts.

7.6 Future Directions and Evolving Capabilities

The rapid evolution of LLM capabilities during our study period complicates interpretation of some findings, as newer models demonstrated capabilities do not present in earlier implementations. This evolution creates challenges for healthcare organizations attempting to make strategic decisions about LLM adoption, as today's implementation choices

may quickly become outdated. As one healthcare CIO observed: "We're trying to make five-year strategic plans around technology that transforms every six months."

The trend toward specialized healthcare LLMs, fine-tuned on domain-specific data, appears particularly significant for future clinical applications. Our analysis found that domain-specialized models outperformed general-purpose LLMs by 12-18% on healthcare-specific tasks, suggesting that specialization may be necessary to achieve optimal clinical utility. This trend aligns with predictions by Liu, et al. [12] that "vertical specialization of foundation models will characterize the next wave of AI implementation in healthcare."

The emergence of multimodal LLMs capable of processing both text and images represent another significant development with implications for clinical applications. Early implementations of these models in radiology and dermatology showed promising results, with one study reporting diagnostic accuracy comparable to subspecialist radiologists for certain conditions. This capability expansion may accelerate adoption in image-intensive specialties and enable new workflows not possible with text-only models.

The growing focus on explainability and interpretability in clinical LLM applications reflects both regulatory requirements and clinical acceptance factors. Implementations incorporating explicit rationales for recommendations achieved 28% higher clinician trust ratings than "black box" approaches. This finding aligns with emerging consensus in healthcare AI that explanatory capabilities are essential for clinical applications, even when they come at some cost to raw performance metrics [5][6].

8. Implications

8.1 For Healthcare Organizations

Our findings offer several actionable implications for healthcare organizations considering or implementing LLMs. First, the consistent success of phased implementation approaches suggests that organizations should begin with lower-risk administrative applications before progressing to clinical functions. This sequencing allows for organizational learning, governance development, and trust building while delivering early efficiency benefits that can fund subsequent initiatives.

Second, the critical importance of workflow integration indicates that organizations should invest in detailed workflow analysis and redesign rather than simply adding LLM capabilities to existing processes. As one successful implementation leader explained: "We approached it as a workflow transformation project enabled by AI, not an AI project imposed on workflows."

Third, the significant impact of data quality on LLM performance suggests that organizations should assess and address data governance issues before implementation rather than discovering limitations during deployment. Organizations with established data quality assessment processes reported fewer implementation challenges and achieved performance metrics closer to those observed in controlled studies.

Fourth, the correlation between governance structures and implementation success suggests that organizations should establish comprehensive oversight mechanisms early in the implementation process. The governance framework developed through our Delphi process provides a starting point for organizations to develop context-appropriate structures addressing both technical and ethical dimensions of LLM deployment.

Finally, the observed importance of user engagement, particularly clinician involvement, indicates that organizations should prioritize participatory design approaches throughout the implementation process. The finding that implementations with substantial clinician input achieved 57% higher utilization rates provides compelling evidence for the return on investment

in user-centered design.

8.2 For Technology Developers

For technology developers, our findings highlight several opportunities to better serve healthcare needs. The substantial "implementation gap" between controlled environments and real-world deployments suggests that developers should focus on implementation support rather than merely technical capabilities. Products that include implementation frameworks, workflow integration guidance, and change management resources may achieve greater success than technically superior models without these supports.

The inconsistent attention to ethical considerations across implementations suggests an opportunity for developers to differentiate by incorporating robust ethical frameworks into their offerings. Products with built-in fairness monitoring, explainability features, and documentation support may address critical needs identified in our analysis.

The significant performance variations across different healthcare contexts indicate the value of adaptable, configurable solutions rather than one-size-fits-all approaches. Developers who provide mechanisms for local customization and adaptation to specific organizational contexts may achieve better outcomes than those offering rigid implementations.

The identified interoperability challenges suggest that developers should prioritize integration capabilities, including support for healthcare-specific data standards and compatibility with common electronic health record systems. The finding that interoperability issues accounted for substantial implementation delays highlights the market opportunity for solutions that address this pain point.

8.3 For Policy Makers and Regulators

For policy makers and regulators, our findings underscore the need for clear, consistent guidance regarding LLM applications in healthcare. The regulatory uncertainty reported by 52% of implementations represents a significant barrier to responsible innovation that policy makers are uniquely positioned to address.

The potential for LLMs to either mitigate or exacerbate healthcare disparities highlights the importance of equity-focused regulation. Policy frameworks should require evaluation of performance across diverse patient populations and monitoring for disparate impacts on vulnerable groups.

The inconsistent ethical oversight observed across implementations suggests the need for minimum standards regarding transparency, consent, and monitoring for AI applications in healthcare. Such standards could provide guardrails for innovation while ensuring basic protection for patients and providers.

The substantial performance variability across implementation contexts indicates that regulatory approaches should focus on outcomes and processes rather than specific technical features. As LLM capabilities continue to evolve rapidly, performance-based regulatory frameworks may prove more sustainable than technology-specific requirements.

8.4 For Healthcare Education

The changing role of healthcare providers in LLM-augmented workflows has significant implications for healthcare education. Our finding that clinicians spend an average of 18.7 minutes daily learning to effectively use and supervise LLM applications suggests a need for formal training in AI literacy, prompt engineering, and output evaluation.

The differential adoption across specialties indicates that specialty-specific education regarding appropriate use cases and limitations may be more valuable than general AI training. Medical education programs should consider incorporating specialty-relevant AI applications

into clinical training, with emphasis on augmentation rather than replacement of clinical reasoning.

The ongoing evolution of LLM capabilities suggests that continuous learning approaches will be necessary rather than one-time curriculum adjustments. Professional societies and continuing education programs have an opportunity to provide updated guidance as capabilities and best practices evolve.

8.5 Limitations and Future Scope

While this study provides comprehensive insights into LLM implementation in healthcare, several limitations warrant acknowledgment. The cross-sectional nature of most analyzed implementations limits understanding of long-term sustainability and evolving performance patterns. The predominance of studies from well-resourced academic medical centers in developed nations constrains generalizability to community hospitals, rural settings, and low-to-middle-income countries where infrastructure and expertise may differ substantially. Methodological heterogeneity across included studies complicated direct comparisons, particularly for clinical outcomes where standardized evaluation metrics remain underdeveloped. The rapid technological evolution during the study period meant that some early implementations utilized models with capabilities significantly inferior to current iterations, potentially underestimating the true potential of contemporary LLMs. Additionally, the reliance on published literature may introduce positive reporting bias, as failed implementations and negative outcomes are less frequently documented in peer-reviewed venues.

Future research should prioritize several critical areas: establishing standardized evaluation frameworks specifically designed for healthcare AI applications; conducting prospective multicenter trials that track implementation outcomes across diverse healthcare settings over extended periods; developing implementation strategies tailored to resource-constrained environments to prevent widening healthcare disparities; investigating the optimal balance between model sophistication and practical implementability; examining the psychological and professional identity impacts on healthcare workers as AI augmentation becomes prevalent; and creating dynamic governance frameworks capable of adapting to rapidly evolving technological capabilities while maintaining patient safety standards. The emergence of multimodal models, agentic AI systems, and specialized medical LLMs present both opportunities and challenges requiring dedicated investigation to guide evidence-based adoption strategies.

9. Conclusion

This study makes several substantive scientific contributions to understanding the complex dynamics of large language model implementation in healthcare systems. Through rigorous mixed-methods analysis of 218 real-world implementations, we establish the first comprehensive evidence base documenting implementation patterns, performance metrics, and critical success factors across diverse healthcare contexts. Our systematic characterization of the "implementation gap" -the 12.3% average performance degradation between controlled environments and clinical practice -provides crucial insights for translating AI capabilities into clinical value. This finding fundamentally challenges the prevailing focus on technical performance metrics and establishes implementation science as central to healthcare AI success.

The study's primary scientific contribution lies in developing an empirically grounded theoretical framework that explains differential LLM adoption and performance across healthcare domains. By identifying consistent progression from administrative to clinical applications and correlating this with specific organizational, technical, and human factors, we provide a predictive model for implementation success. The framework's validation through

expert consensus via modified Delphi methodology establishes methodological rigor while ensuring practical applicability. Our quantitative demonstration that organizational factors explain more variance in implementation outcomes than technical capabilities represent a paradigm shift in healthcare AI evaluation, moving beyond algorithmic performance to encompass sociotechnical system dynamics.

Furthermore, this research establishes critical benchmarks for LLM performance across healthcare applications, providing the first comprehensive baseline against which future implementations can be measured. The identification of specialty-specific performance variations and the superior utility for rare conditions and complex presentations offer evidence-based guidance for strategic deployment decisions. Our documentation of consistent facilitators - including phased implementation approaches, workflow integration, and clinician engagement - transforms anecdotal observations into scientifically validated implementation principles. The governance framework developed through expert consensus provides the first structured approach to addressing ethical, regulatory, and operational challenges specific to healthcare LLMs, filling a critical gap in literature.

The scientific significance extends beyond immediate practical applications to establish foundations for future healthcare AI research. By documenting the complex interplay between technological capabilities, implementation approaches, and contextual factors, this study provides a theoretical basis for understanding why certain AI implementations succeed while others fail in healthcare settings. The identification of equity concerns and performance disparities across patient populations establishes an ethical imperative for inclusive AI development and deployment strategies. These contributions collectively advance the field from enthusiasm-driven adoption toward evidence-based, ethically grounded, and strategically-informed implementation of transformative AI technologies in healthcare delivery systems

Acknowledgment: Not Applicable.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study

References

- [1] K. Vaishali, "Health Education Reimagined Virtual Reality for Immersive Wellness Learning," *J. Comput. Educ. Sport. Heal.*, vol. 1, no. 1, pp. 55–66, 2024.
- [2] D. B. Olawade, A. C. David-Olawade, O. Z. Wada, A. J. Asaolu, T. Adereni, and J. Ling, "Artificial intelligence in healthcare delivery: Prospects and pitfalls," *J. Med. Surgery, Public Heal.*, p. 100108, 2024.
- [3] L. Petersson *et al.*, "Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden," *BMC Health Serv. Res.*, vol. 22, no. 1, p. 850, 2022.
- [4] B. Saha, "Generative AI for Text Generation: Advances and Applications in Natural Language Processing," *J. Comput. Allied Intell. (JCAI, ISSN 2584-2676)*, vol. 3, no. 1, pp. 77–91, 2025.
- [5] P. S. Aravazhi *et al.*, "The integration of artificial intelligence into clinical medicine: trends, challenges, and future directions," *Disease-a-Month*, p. 101882, 2025.
- [6] K. Cresswell, R. Williams, and A. Sheikh, "Using cloud technology in health care during the COVID-19 pandemic," *Lancet Digit. Heal.*, vol. 3, no. 1, pp. e4–e5, 2021.
- [7] D. Ben-Israel, "The impact of machine learning on patient care: a systematic review,"

-
- Artif. Intell. Med.*, vol. 103, p. 101785, 2020.
- [8] Z. Chang, “Application of artificial intelligence in COVID-19 medical area: a systematic review,” *J. Thorac. Dis.*, vol. 13, no. 12, p. 7034, 2021.
- [9] U. Kose, O. Deperlioglu, J. Alzubi, and B. Patrut, “Deep learning for medical decision support systems,” 2021.
- [10] S. S. Sengar, A. Bin Hasan, S. Kumar, and F. Carroll, “Generative artificial intelligence: a systematic review and applications,” *Multimed. Tools Appl.*, pp. 1–40, 2024.
- [11] M. Chen *et al.*, “Evaluating large language models trained on code,” *arXiv Prepr. arXiv2107.03374*, 2021.
- [12] S. Liu, A. B. McCoy, and A. Wright, “Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines,” *J. Am. Med. Informatics Assoc.*, p. ocaf008, 2025.