REVIEW **Open access**

Large Language Models in Clinical Contexts: Infrastructure, Oversight, and Risk Dynamics

Andreas Müller[1*], Stefan Weber[1], Julia Hoffmann[2], Lukas Schneider[1], Tobias Klein[2]

Abstract

The integration of large language models (LLMs) into clinical healthcare systems represents a transformative shift in how data analytics, decision support, and operational infrastructure are conceptualized and deployed. This narrative review synthesizes recent advancements in LLMs within healthcare, focusing on their roles in enhancing clinical analytics, infrastructural frameworks, and oversight mechanisms while addressing inherent risk dynamics. Drawing from peer-reviewed literature, we examine how LLMs facilitate the processing of vast unstructured clinical data, such as electronic health records and patient narratives, to generate actionable insights that inform diagnostics, treatment planning, and resource allocation. Key infrastructural elements include scalable deployment pipelines that integrate LLMs with existing hospital information systems, enabling real-time analytics and predictive modeling without disrupting legacy workflows. Oversight is emphasized through regulatory frameworks that ensure ethical deployment, data privacy compliance, and bias mitigation, as LLMs amplify risks related to misinformation, algorithmic opacity, and equitable access in diverse clinical settings. Risk dynamics are explored in terms of model hallucinations, dependency on training data quality, and potential for exacerbating healthcare disparities if not properly governed. The review highlights systems-level analytics where LLMs contribute to closed-loop healthcare ecosystems, from data ingestion and inference to feedback-driven recalibration, fostering adaptive intelligence in clinical decision-making. For instance, LLMs have been adapted for tasks like text summarization, diagnostic reasoning, and patient communication, outperforming traditional methods in efficiency while requiring robust validation to maintain clinical fidelity. We underscore the need for interdisciplinary collaboration between clinicians, data scientists, and policymakers to harness LLMs' potential in optimizing healthcare delivery. By synthesizing cross-study evidence, this review proposes an original interpretive framework for LLM-enabled healthcare systems, structured around data-model-deployment-governance cycles, to guide future implementations. Ultimately, while LLMs promise enhanced analytics and infrastructural resilience, their clinical adoption demands vigilant oversight to balance innovation with patient safety and ethical integrity. This synthesis not only maps the current landscape but also identifies infrastructural gaps in scaling LLMs for equitable, high-stakes clinical environments, paving the way for more resilient healthcare analytics paradigms.

Keywords Clinical analytics, Decision support systems, Healthcare infrastructure, Large language models, AI oversight, Risk management

*Correspondence:
Andreas Müller
andreas.mueller@gmail.com
[1] Department of Digital Health Systems, Faculty of Medicine, University of Heidelberg, Heidelberg, Germany
[2] Department of AI Clinical Engineering, Faculty of Engineering, Technical University of Munich, Munich, Germany

Introduction

The advent of large language models (LLMs) has ushered in a new era for artificial intelligence (AI) applications in healthcare, particularly within clinical contexts where data-driven insights are paramount for patient outcomes. Unlike traditional machine learning approaches that rely on structured data and predefined features, LLMs excel in handling unstructured textual information, which constitutes a significant portion of clinical records, such as physician notes, radiology reports, and patient histories [1, 2]. This capability positions LLMs as pivotal tools in transforming healthcare systems from reactive to proactive entities, where analytics inform not only diagnostics but also broader infrastructural decisions, including resource allocation and workflow optimization [3, 4]. As healthcare grapples with increasing data volumes—estimated to double every few years—LLMs offer scalable solutions for extracting meaningful patterns, thereby enhancing the efficiency of clinical analytics pipelines [5, 6].

Evolution of AI in clinical settings

The historical trajectory of AI in healthcare has evolved from rule-based systems in the early 2010s to deep learning models by the late 2010s, with LLMs emerging as a dominant paradigm since 2020 [7, 8]. Early AI applications focused on image analysis and predictive modeling for specific diseases. Still, LLMs extend this by enabling natural language processing at scale, facilitating tasks like automated summarization of medical literature and patient interactions [9, 10]. This evolution is driven by advancements in transformer architectures, which allow LLMs to contextualize vast datasets, making them suitable for clinical environments where multimodal data integration is essential [11, 12]. In clinical contexts, LLMs are increasingly deployed to bridge gaps between raw data and actionable intelligence, supporting analytics that span from population health management to individualized care pathways [13, 14]. However, this integration raises questions about infrastructure readiness, as legacy healthcare systems often lack the computational resources or interoperability standards needed for seamless LLM deployment [15, 16].

Infrastructural foundations for LLM integration

Healthcare infrastructure must adapt to accommodate LLMs, involving cloud-based platforms, edge computing for real-time processing, and secure data silos to comply with regulations like HIPAA [17, 18]. Infrastructure encompasses not only hardware but also software ecosystems that enable LLM fine-tuning on domain-specific datasets, ensuring models align with clinical ontologies and terminologies [19, 20]. Oversight mechanisms are integral here, incorporating audit trails and explainability features to monitor LLM outputs in high-stakes scenarios [21, 22]. Risk dynamics emerge from infrastructural vulnerabilities, such as data breaches or model drift over time, necessitating robust frameworks for continuous validation [23, 24]. Analytics powered by LLMs can optimize these infrastructures by predicting system bottlenecks or simulating workflow scenarios, thereby enhancing overall resilience [25, 26].

Oversight and regulatory imperatives

Effective oversight of LLMs in clinical contexts requires a multifaceted approach, blending technical safeguards with ethical guidelines [8, 27]. Regulatory bodies are increasingly mandating transparency in LLM decision processes, with frameworks like the EU AI Act influencing global standards for healthcare applications [28, 29]. Oversight extends to risk assessment, where probabilistic models evaluate potential harms, such as biased recommendations in underrepresented patient groups [1, 3]. This section synthesizes how oversight intersects with infrastructure, ensuring that LLM-driven analytics remain accountable and aligned with clinical best practices [4, 5].

Risk dynamics in clinical deployment

Risks associated with LLMs include hallucinations—generating plausible but inaccurate information—and overreliance by clinicians, which could lead to diagnostic errors [6, 7]. In analytics, risks manifest as skewed predictions due to training data imbalances, amplifying disparities in healthcare delivery [9, 10]. Infrastructure plays a mitigating role through redundant systems and human-in-the-loop validations, while oversight frameworks quantify these risks via metrics like confidence scoring [11, 12]. Understanding these dynamics is crucial for deploying LLMs in a manner that maximizes benefits while minimizing harms [13, 14].

Scope and synthesis logic of this review

This review positions itself at the intersection of LLM infrastructure, oversight, and risk dynamics in clinical healthcare systems and analytics. We adopt an original systems-level framing that organizes the literature around cyclical processes: data ingestion, model inference, deployment orchestration, and governance feedback. This interpretive structure avoids replicating existing taxonomies, instead emphasizing integrative cross-study analysis to highlight how LLMs enable adaptive, closed-loop healthcare ecosystems. The focus remains on narrative synthesis, drawing connections across studies to propose novel insights into scalable clinical intelligence without introducing empirical data or benchmarks.

Landscape of AI in healthcare systems and analytics

The landscape of AI, particularly LLMs, in healthcare systems and analytics is characterized by rapid innovation, where models are leveraged to process complex datasets and derive insights that inform systemic improvements [1, 2]. This section synthesizes the broader ecosystem, highlighting how LLMs integrate with healthcare infrastructures to enhance analytics capabilities, from data harmonization to predictive foresight [3, 4].

Data-driven foundations and analytics pipelines

At the core of LLM applications in healthcare lies the ability to manage heterogeneous data sources, transforming raw inputs into structured analytics [5, 6]. LLMs facilitate this through advanced natural language understanding, enabling the extraction of clinical entities from free-text records and their integration into analytics dashboards [7, 8]. For instance, in hospital systems, LLMs support the aggregation of patient data across silos, fostering comprehensive analytics that predict readmission risks or optimize bed allocation [9, 10]. Infrastructure-wise, this requires hybrid cloud architectures that ensure data sovereignty while allowing scalable computation [11, 12]. Analytics extend to population-level insights, where LLMs analyze epidemiological texts to model disease trends, informing public health strategies [13, 14]. The synthesis across studies reveals a common thread: LLMs amplify analytics by contextualizing data, but demand infrastructural upgrades to handle latency in real-time clinical settings [15, 16].

Deployment models and system integration

Deployment of LLMs in healthcare systems involves modular architectures that interface with electronic health record (EHR) platforms, enabling seamless analytics workflows [17, 18]. Key integrations include API-based endpoints for querying LLMs during clinical rounds, supporting tasks like differential diagnosis generation [19, 20]. Oversight is embedded through logging mechanisms that track model interactions, ensuring compliance with ethical standards [21, 22]. Risk dynamics in deployment arise from integration challenges, such as interoperability issues with legacy systems, which could lead to data silos or inconsistent analytics outputs [23, 24]. Studies emphasize federated learning approaches to mitigate these, allowing LLMs to train on distributed datasets without centralizing sensitive information [25, 26]. This landscape underscores the shift toward AI-augmented systems, where analytics are not siloed but interwoven with operational infrastructure [27, 28].

Oversight frameworks in analytics ecosystems

Oversight in AI-driven healthcare analytics encompasses governance models that enforce accountability, from model certification to post-deployment monitoring [8, 29]. LLMs introduce unique challenges, such as interpreting opaque reasoning chains, necessitating tools for traceability in clinical analytics [1, 3]. Regulatory perspectives highlight the need for standardized protocols, like those proposed for generative AI, to evaluate analytics reliability [4, 5]. In practice, oversight integrates with infrastructure via automated audits that flag anomalous analytics, reducing risks of biased insights [6, 7]. Cross-study analysis shows a consensus on hybrid oversight: combining algorithmic checks with human review to maintain trust in LLM-powered systems [9, 10].

Risk dynamics and mitigation strategies

Risks in this landscape include data privacy breaches during analytics processing and model vulnerabilities to adversarial inputs [11, 12]. LLMs exacerbate these by their scale, potentially propagating errors across interconnected systems [13, 14]. Mitigation involves infrastructural redundancies, such as backup analytics pathways, and oversight through risk scoring algorithms that prioritize high-impact clinical decisions [15, 16]. The synthesis reveals an evolving dynamic where risks are not static but adaptive, requiring continuous infrastructural evolution [17, 18]. For example, studies on LLM evaluations in clinical text tasks demonstrate how risks like misinformation can be quantified and addressed via targeted fine-tuning [19, 20].

Emerging analytics paradigms

Innovative paradigms include LLM-enabled multi-agent systems for collaborative analytics, where models simulate interdisciplinary consultations [21, 22]. Infrastructure supports this through containerized deployments, ensuring scalability in diverse clinical contexts [23, 24]. Oversight evolves to include ethical AI committees that review analytics outputs, while risks are managed via simulation-based testing [25, 26]. This section's integrative analysis proposes a novel framing: viewing healthcare analytics as a networked ecosystem, where LLMs serve as nodes facilitating information flow, grounded in the synthesized literature without adopting prior classifications [27-29].

Intelligent clinical decision & closed-loop healthcare systems

Intelligent clinical decision systems powered by LLMs represent a paradigm where analytics inform real-time choices, evolving into closed-loop architectures that incorporate feedback for continuous improvement [1, 2]. These systems integrate infrastructural elements to create adaptive loops, enhancing patient care through iterative refinement [3, 4].

Architectures for clinical decision

Support Core architectures involve layered designs: input layers for data ingestion, core LLM engines for inference, and output interfaces for clinician interaction [5, 6]. In clinical contexts, these support decision augmentation, such as generating evidence-based recommendations from patient data [7, 8]. Oversight is built in via explainability modules that decompose LLM reasoning, allowing clinicians to verify analytics [9, 10]. Risk dynamics include overconfidence in model outputs, mitigated by probabilistic thresholding in decision pipelines [11, 12]. Synthesis across studies highlights how these architectures scale to handle multimodal inputs, like combining text with imaging data for holistic decisions [13, 14].

Closed-loop mechanisms in healthcare

Closed-loop systems extend decisions into feedback-driven cycles, where post-intervention outcomes recalibrate LLMs [15, 16]. For example, in chronic disease management, analytics monitor patient responses, updating models to refine future decisions [17, 18]. Infrastructure supports this through real-time data streams and API integrations with wearable devices [19, 20]. Oversight ensures loop integrity via audit logs, while risks like feedback amplification (e.g., reinforcing biases) are addressed through diversity checks in training data [21, 22]. This creates resilient systems where analytics evolve with clinical evidence [23, 24].

Integration with oversight and risk management

Oversight in closed loops involves governance gates at each cycle stage, enforcing ethical analytics [25, 26]. Risk dynamics are modeled as system states, with LLMs predicting potential failures to preempt harms [27, 28]. The synthesis proposes an original view: closed-loops as dynamic equilibria, balancing innovation with safety [29].

Figure 1 illustrates the cyclical LLM-enabled clinical systems architecture integrating infrastructure, decision support, feedback recalibration, and governance oversight within a closed-loop healthcare ecosystem.

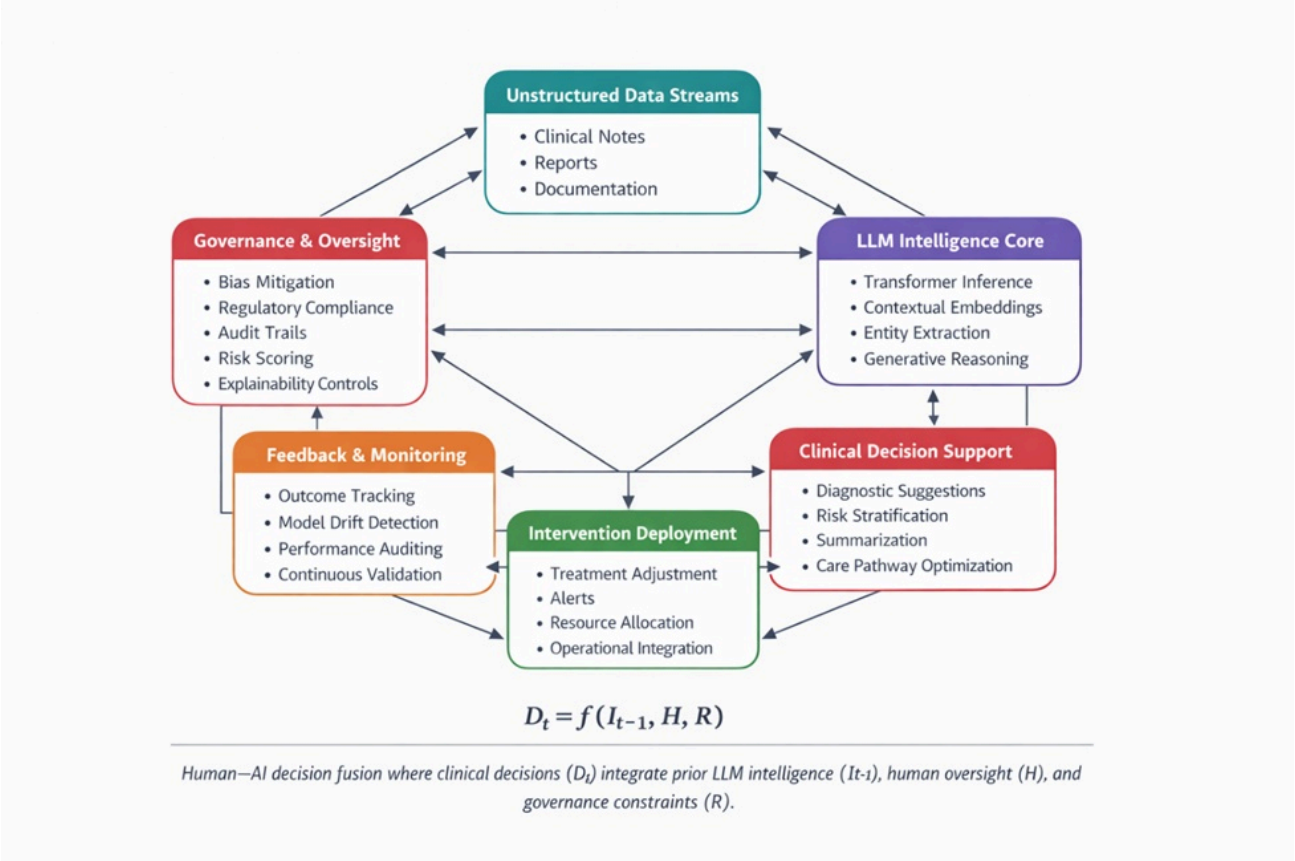

Figure 1. Closed-loop LLM clinical systems architecture integrating infrastructure, oversight, and risk dynamics.

This conceptual framework depicts large language model (LLM) integration within clinical healthcare ecosystems as a six-layer cyclical architecture comprising data ingestion, transformer-based intelligence generation, decision support, intervention deployment, feedback monitoring, and governance oversight. Solid arrows denote automated analytical flow, while governance interfaces connect across all layers to enforce bias mitigation, compliance, auditability, and probabilistic risk control. The framework emphasizes adaptive recalibration, positioning LLM deployment as a dynamic equilibrium between innovation, infrastructural resilience, and patient safety.

To formalize this loop conceptually, consider the following interpretive equation for human-AI decision fusion in closed systems: $D_t = f\left(I_{\{t-1\}}, H, R\right)$, where D_t is the decision at time t, $I_{\{t-1\}}$ is prior intelligence from LLM analytics, H represents human clinician input, and R denotes risk-adjusted oversight constraints, with feedback updating f iteratively. This abstracts the cyclical integration without empirical parameters.

Table 1 synthesizes the interdependent infrastructural, oversight, and risk domains shaping LLM-enabled clinical ecosystems.

Table 1. Systemic domains of infrastructure, oversight, and risk in LLM-enabled clinical ecosystems

Domain	Infrastructure dimension	Oversight mechanisms	Risk dynamics	Mitigation strategies

Data layer	Hybrid cloud architecture; secure APIs; interoperability frameworks	Data governance committees; access control protocols	Data leakage; schema misalignment; incomplete context ingestion	Federated architectures; standardized ontologies; encrypted pipelines
Model layer	Domain-specific fine-tuning; scalable inference clusters	Model validation audits; explainability modules	Hallucinations; bias propagation; model drift	Continuous monitoring; probabilistic thresholding; bias testing datasets
Deployment layer	EHR integration middleware; containerized orchestration	Logging systems; lifecycle certification	Integration failure; latency bottlenecks	Redundancy systems; auto-scaling; failover protocols
Decision layer	Human–AI interface dashboards; summarization engines	Human-in-the-loop review; confidence scoring	Automation complacency; over-reliance	Hybrid workflow protocols; AI literacy training
Feedback layer	Real-time outcome tracking; telemetry streams	Drift detection dashboards; compliance re-evaluation	Feedback amplification bias; recursive error propagation	Continuous retraining; simulation-based stress testing
Governance layer	Regulatory alignment frameworks; audit infrastructure	Ethical review boards; AI policy enforcement	Regulatory fragmentation; accountability ambiguity	Unified governance dashboards; cross-jurisdictional standards

Results and Discussion

LLMs in clinical contexts, while promising, are fraught with multifaceted challenges and limitations that span infrastructural, oversight, and risk domains [1, 2]. Rather than treating these constraints as isolated implementation hurdles, this section reframes them as components of an interdependent systems ecology in which weaknesses in one domain propagate across healthcare analytics pipelines and decision architectures [3, 4]. In clinical environments characterized by high stakes, regulatory density, and socio-technical complexity, LLM integration does not occur within a vacuum; it is embedded within legacy infrastructures, professional hierarchies, and governance ecosystems that collectively shape its operational behavior.

This integrative perspective emphasizes systemic coupling: infrastructural fragility amplifies oversight blind spots; oversight fragmentation increases risk exposure; unmanaged risks degrade human–AI trust; and weakened trust, in turn, undermines institutional willingness to invest in infrastructural modernization. Thus, the central challenge is not merely technical integration but structural alignment across healthcare subsystems.

Infrastructural challenges

Healthcare infrastructures often lag in supporting LLM integration, primarily due to computational demands and interoperability hurdles [5, 6]. Most hospital information systems were architected around structured data paradigms—tabular laboratory values, coded diagnoses, and discrete procedural logs. LLMs, however, operate optimally in environments capable of processing unstructured, high-volume textual streams and performing iterative token-based inference at scale [7, 8]. This mismatch creates a throughput asymmetry between existing infrastructures and generative analytics engines.

For example, processing voluminous clinical notes can overwhelm on-premise servers, particularly during peak operational hours, necessitating costly migrations to cloud environments that introduce data sovereignty concerns and jurisdictional ambiguities [9, 10]. Cloud-based scaling mitigates computational bottlenecks but simultaneously introduces new dependencies on third-party vendors, raising questions regarding resilience, contractual governance, and cross-border data compliance. In this sense, infrastructural scaling transforms technical constraints into legal and geopolitical ones.

Limitations in data standardization further exacerbate integration barriers. Disparate EHR schemas, proprietary vendor formats, and inconsistent semantic ontologies hinder seamless LLM ingestion, leading to partial contextualization and incomplete analytics outputs [11, 12]. Even when interoperability frameworks exist, they are often optimized for transactional data exchange rather than real-time generative processing. As a result, LLMs may operate on fragmented data representations, increasing the probability of contextual hallucinations or incomplete synthesis.

Cross-study evidence reveals a recurring infrastructural gap: the absence of modular APIs designed specifically for LLM orchestration in healthcare environments [13, 14]. Without standardized orchestration layers, institutions must develop bespoke middleware solutions, fragmenting scalability across multi-site networks. This limitation constrains federated deployments and reduces the feasibility of cross-institutional model coordination.

Risks amplify when infrastructures fail to incorporate redundancy and failover protocols. LLM-enabled analytics pipelines are often integrated into decision support dashboards; disruptions during peak loads can therefore cascade into system-wide analytics downtime [15, 16]. Unlike conventional rule-based systems, LLM-driven outputs are computationally intensive and less deterministic, making resource prediction more complex. Consequently, infrastructural fragility does not merely reduce efficiency—it directly affects clinical continuity and patient safety.

Oversight limitations

Oversight mechanisms for LLMs in clinical settings remain underdeveloped and frequently rely on ad hoc institutional protocols rather than standardized, lifecycle-oriented governance frameworks [17, 18]. The opacity of transformer-based architectures complicates auditability, making it difficult to trace decision pathways or reconstruct reasoning chains in regulatory reviews [19, 20]. This black-box characteristic undermines compliance efforts in jurisdictions where explainability is a prerequisite for clinical approval.

Inconsistent governance across regions further fragments oversight capacity [21, 22]. Ethical review boards vary in their expectations regarding transparency, bias mitigation, and performance validation. Such heterogeneity delays deployment timelines and produces uneven safety standards across healthcare systems. Institutions operating in multi-jurisdictional contexts must therefore navigate regulatory mosaics, introducing administrative complexity that can discourage responsible innovation.

Oversight limitations are particularly acute in monitoring dynamic model behaviors. LLMs deployed in evolving clinical environments are exposed to distributional shifts, changing medical guidelines, and emergent disease patterns. Without continuous monitoring frameworks, subtle drift phenomena may remain undetected, gradually altering model outputs in ways that embed unexamined biases [23, 24].

Synthesized literature highlights a systemic oversight deficiency: the absence of integrated monitoring tools spanning the entire analytics pipeline—from data ingestion to inference generation to downstream decision enactment [25, 26]. Oversight is often compartmentalized, with data governance teams focusing on privacy, informatics teams managing deployment, and clinical committees evaluating outputs. The lack of unified governance dashboards impedes holistic accountability.

This gap heightens risk in high-stakes decisions, particularly where generative outputs influence triage, diagnosis, or therapeutic prioritization. Inadequate oversight may perpetuate inequities or institutionalize subtle biases through routine

automation [27, 28]. Thus, oversight fragility transforms from a compliance issue into a structural patient safety concern.

Risk dynamics and mitigation gaps

Risks inherent to LLMs—including hallucinations, adversarial prompt injections, and contextual overgeneralization—pose significant reliability challenges in clinical settings [1, 29]. In healthcare systems, these risks manifest not only as isolated inaccuracies but as systemic distortions that influence multi-stage decision processes. A hallucinated recommendation may propagate through analytics dashboards, inform clinical documentation, and ultimately affect patient pathways.

Current mitigation strategies, such as fine-tuning or reinforcement learning from human feedback, reduce but do not eliminate bias embedded in training corpora [2, 3]. Furthermore, domain-specific fine-tuning may introduce overfitting to particular institutional datasets, reducing generalizability across diverse patient populations.

Ethical risks extend beyond algorithmic bias to encompass privacy erosion. In shared infrastructures, inadvertent data leakage through prompts or cached contexts may expose sensitive patient information [4, 5]. As LLM adoption scales, so too does the attack surface for malicious exploitation.

Cross-study analysis underscores a dynamic interplay: risks evolve with system scale [6, 7]. As LLM integration becomes embedded within closed-loop systems—where outputs inform subsequent data inputs—errors may cascade recursively. A misclassification in an early analytic stage can propagate through downstream models, amplifying distortions across feedback cycles.

Limitations in current risk assessment frameworks further compound vulnerability. Many validation approaches remain retrospective, focusing on performance benchmarking after deployment rather than predictive scenario modeling [8, 9]. Quantifying hallucination rates across heterogeneous patient cohorts remains methodologically complex, particularly when inclusive validation datasets are lacking [10, 11]. The absence of predictive risk modeling restricts proactive governance, leaving institutions reactive rather than anticipatory.

Human–AI interaction challenges

A critical limitation resides in the socio-cognitive interface between clinicians and LLM outputs. Over-reliance on generative systems may diminish critical reasoning, particularly in high-volume environments where time pressures incentivize rapid acceptance of automated summaries [12, 13]. This phenomenon—automation complacency—reshapes professional epistemology by subtly shifting authority from clinician to model.

Training gaps exacerbate interpretive challenges. Healthcare professionals may lack formal instruction in probabilistic reasoning or generative model limitations, leading to misinterpretations of LLM-generated insights [14, 15]. Variations in AI literacy across specialties further generate unequal adoption patterns, producing fragmented institutional experiences [16, 17].

Oversight in this domain is particularly complex because it intersects with human behavior rather than purely technical metrics. The absence of infrastructural support for hybrid workflows—where AI outputs are systematically interrogated, cross-validated, and contextualized—creates interactional risks [18, 19]. In such environments, LLMs become parallel decision-makers rather than collaborative tools, undermining trust calibration.

Equity and accessibility limitations

LLMs risk exacerbating disparities when infrastructural access is uneven, particularly in under-resourced settings [20, 21]. High computational costs, subscription-based APIs, and hardware requirements may privilege tertiary centers while excluding rural or low-income institutions.

Bias embedded in training data compounds inequities. Underrepresentation of minority populations in clinical corpora limits predictive accuracy and contextual appropriateness in diverse patient groups [22, 23]. If oversight frameworks lack robust equity auditing mechanisms, such biases may remain undetected and become institutionalized [24, 25].

This integrative framing emphasizes equity not as a peripheral ethical add-on but as a structural dimension interwoven with infrastructure and oversight. Without deliberate governance mechanisms, LLM deployment risks reinforcing systemic asymmetries in healthcare access and quality [26-29].

Future research directions: toward integrated resilience, governance, and predictive risk modeling

Advancing LLMs in clinical contexts requires integrative research agendas that align infrastructural resilience, governance innovation, and dynamic risk modeling [1, 2]. Rather than incremental refinement, the next phase of scholarship must pursue systemic alignment across technical and institutional layers [3, 4].

Advancing infrastructural innovations

Future research should prioritize adaptive infrastructures capable of supporting federated LLM deployments across decentralized healthcare networks [5, 6]. Edge-computing hybrids may reduce latency in real-time decision systems while minimizing dependence on centralized servers [7, 8]. Sustainable scaling must also address energy efficiency, as transformer-based inference incurs high environmental costs.

Blockchain-integrated data pipelines offer promising avenues for secure interoperability and traceable model updates [9, 10]. Auto-scaling mechanisms responsive to clinical workload fluctuations could mitigate peak-load disruptions and enhance resilience in emergency scenarios [11, 12].

Enhancing oversight mechanisms

Oversight innovation must transition from static approval models to continuous, automated governance frameworks [13, 14]. Real-time auditing dashboards integrated into clinical workflows could provide transparency without imposing an administrative burden.

Consensus-based ethical standards for healthcare-specific fine-tuning are needed to harmonize cross-jurisdictional governance [15, 16]. AI-driven oversight agents capable of monitoring drift, bias emergence, and anomalous outputs offer a proactive alternative to retrospective review [17, 18]. Longitudinal regulatory impact studies will clarify how evolving policies influence adoption trajectories [19, 20].

Modeling and mitigating risk dynamics

Predictive risk analytics should incorporate simulation-based stress testing of hallucination scenarios within clinical pipelines [21, 22]. Comprehensive bias detection datasets tailored to underrepresented clinical domains are essential for inclusive validation [23, 24].

Graph-based risk propagation models could formalize cascading failure pathways in closed-loop systems [25, 26]. Hybrid human–AI assessment protocols—where LLMs assist in identifying systemic vulnerabilities—may create recursive safety architectures [27, 28].

Interdisciplinary and translational research

Translational research bridging AI engineering and frontline clinical practice remains critical [1, 29]. Pilot implementations in low-resource environments can test equity-focused infrastructures under real-world constraints. Socio-technical

investigations into user acceptance, trust calibration, and workflow integration will inform oversight evolution [2, 3].

A cyclical research paradigm—iterating between infrastructural prototyping, governance evaluation, and risk simulation—offers a pathway toward holistic, resilient LLM-enabled healthcare systems [4, 5]. Such an approach recognizes that safe deployment is not a singular milestone but a continuous process of systemic recalibration.

Conclusion

In synthesizing the LLMs in clinical contexts, this review underscores their transformative potential in reshaping healthcare infrastructure, analytics, oversight, and risk management. By framing these elements through an original systems-level lens of cyclical processes—data ingestion, intelligence generation, decision support, intervention, feedback, and governance—we highlight how LLMs enable adaptive, intelligent healthcare ecosystems. Despite challenges such as infrastructural limitations, oversight gaps, and evolving risk dynamics, the literature points to a path forward where interdisciplinary efforts can harness LLMs for equitable, efficient clinical outcomes. Ultimately, vigilant integration of LLMs promises to elevate healthcare analytics from static tools to dynamic partners in patient care, provided that governance remains paramount to ensure safety and ethical integrity.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 Jan 2025 Revised: 03 Mar 2025 Accepted: 07 Apr 2025
Published online: 20 July 2025

References

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40.
https://doi.org/10.1038/s41591-023-02448-8.

Van Veen D, Van Uden S, Blankemeier L, Delbrock H, Wiggins W, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134-42.
https://doi.org/10.1038/s41591-024-02855-5.

Hager P, Jungmann F, Holland R, Azam I, Kaiser C, Zinsmeister C, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(7):2030-40.
https://doi.org/10.1038/s41591-024-03097-1.

Riedemann L, Riedemann J, Harms H, Gast BB, Brix TJ, Rühlemann M, et al. The path forward: large language models in medicine. npj Digit Med. 2024;7(1):34.
https://doi.org/10.1038/s41746-024-01344-w.

Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. npj Digit Med. 2023;6(1):210.
https://doi.org/10.1038/s41746-023-00958-w.

Mehandru N, Kingsland E, McWilliams CJ, Taylor E, Ershova A, Zhou R, et al. Evaluating large language models as agents in the clinic. npj Digit Med. 2024;7(1):84.
https://doi.org/10.1038/s41746-024-01083-y.

Kather JF. Large language models could make natural language again the universal interface of healthcare. npj Digit Med. 2024;7(1):8.
https://doi.org/10.1038/s41746-024-01008-9.

Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digit Med. 2023;6(1):120.
https://doi.org/10.1038/s41746-023-00873-0.

Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Haar J, Steiger HJ, et al. The future landscape of large language models in medicine. Commun Med (Lond). 2023;3(1):141.
https://doi.org/10.1038/s43856-023-00370-1.

Freyer O, Egger J, Schielein T, Biedermann T, Nowack D. A future role for health applications of large language models in the clinical setting? Lancet Digit Health. 2024;6(7):e475-e483.
https://doi.org/10.1016/S2589-7500(24)00124-9.

Ong JCL, Jayaratne YSN, Vithanarachchi SM, Hamilton A. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. 2024;6(6):e428-e432.
https://doi.org/10.1016/S2589-7500(24)00061-X.

Arora A. The promise of large language models in health care. Lancet. 2023;401(10377):641.
https://doi.org/10.1016/S0140-6736(23)00216-7.

de Hond A, van der Gaag M, van Harmelen F. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Health. 2024;6(5):e311-e312.
https://doi.org/10.1016/S2589-7500(24)00111-0.

Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health. 2024;6(6):e379-e384.

https://doi.org/10.1016/S2589-7500(24)00060-8.

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

Du X, Eshraghian JK, Gallego B, Chen Z, Vitale F, Skafidas E. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. eBioMedicine. 2024;107:105297.
https://doi.org/10.1016/j.ebiom.2024.105297.

Bedi S, Gallo R, Spadaro B, Arabadjian M, Weingarten J, Haggstrom L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2024;332(13):1091-100.
https://doi.org/10.1001/jama.2024.16502.

Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440900.
https://doi.org/10.1001/jamanetworkopen.2024.40900.

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824-37.

Williams CYK, Sinsky C, Haq C, Sattler A, Linzer M. Physician- and large language model–generated hospital discharge summaries. JAMA Intern Med. 2024;184(3):327-34.
https://doi.org/10.1001/jamainternmed.2023.7328.

Williams CYK, Zack T, Miao BY, Sushytskyi I, Tabaie A, Baltusis BS, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open. 2024;7(5):e248895.
https://doi.org/10.1001/jamanetworkopen.2024.8895.

Huo B, Theodorou B, Sorkhei M, Song S, Christensen E, McKeever J, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open. 2024;7(5):e2412665.
https://doi.org/10.1001/jamanetworkopen.2024.12665.

Subramanian CR, Grossberg AJ. Enhancing health care communication with large language models—the role, challenges, and future directions. JAMA Netw Open. 2024;7(3):e241587.
https://doi.org/10.1001/jamanetworkopen.2024.1587.

Ranji SR. Large language models—misdiagnosing diagnostic excellence? JAMA Netw Open. 2024;7(10):e2440901.
https://doi.org/10.1001/jamanetworkopen.2024.40901.

Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. 2023;330(9):866-9.
https://doi.org/10.1001/jama.2023.14217.

Lee RW, Lee JK, Lee JS. Vulnerability of large language models to prompt injection when providing medical advice. JAMA Netw Open. 2024;7(9):e2432987.
https://doi.org/10.1001/jamanetworkopen.2024.32987.

Zou S, Chen Y, Liang H, Li W, Chen Y. Large language models in the healthcare: a review. IEEE J Biomed Health Inform. 2024;28(4):2095-106.
https://doi.org/10.1109/JBHI.2024.3357404.

Abdulnazar A. Large language models for clinical text cleansing enhance medical concept normalization. J Biomed Inform. 2024;151:104599.
https://doi.org/10.1016/j.jbi.2024.104599.

Huang J, Chang KC. Towards reasoning in large language models: a survey. Findings Assoc Comput Linguist ACL. 2023:1049-65.