REVIEW **Open access**

Generative AI in Clinical Workflows: Documentation Utility, Failure Modes, and Oversight Mechanisms

Elena Petrova[1*], Ivan Georgiev[1]

Abstract

The integration of generative artificial intelligence (AI) into clinical workflows represents a transformative shift in healthcare systems and analytics, promising enhanced efficiency in documentation tasks while introducing novel challenges in reliability and governance. This narrative review synthesizes recent literature on the utility of generative AI models, such as large language models (LLMs), in automating clinical documentation, including patient notes, discharge summaries, and diagnostic reports, which traditionally consume significant clinician time. Studies highlight how these tools can streamline data ingestion from electronic health records (EHRs), generating coherent narratives that align with clinical standards, thereby reducing administrative burdens and allowing more focus on patient care. For instance, generative AI has demonstrated proficiency in summarizing complex medical dialogues and classifying clinical notes, often outperforming traditional methods in speed and accuracy, as evidenced by evaluations in German healthcare settings and emergency departments. However, the utility is tempered by inherent failure modes, including hallucinations—where models produce factually incorrect information—and biases amplified from training data, which can propagate errors in clinical decision-making. Oversight mechanisms are critical to mitigate these risks, encompassing human-in-the-loop verification, regulatory frameworks like the EU AI Act, and ethical guidelines for deployment in high-stakes environments. From a systems-level perspective, generative AI enables closed-loop analytics in healthcare infrastructure, where data flows from ingestion to inference, informing interventions and feeding back for model recalibration. This review examines how LLMs facilitate intelligent clinical decision support, such as in patient care document verification using EHRs and prompt engineering for medical education. Yet, failures such as catastrophic errors in multimodal AI applications underscore the need for robust oversight, including transparency in model training and post-deployment monitoring. Comparative analyses reveal that while generative AI excels in low-risk documentation tasks, its application in critical sectors demands interdisciplinary expertise to address trust deficits and ensure equitable outcomes. The review integrates cross-study insights, proposing an original framework for AI-enabled healthcare loops that emphasizes governance at each stage to balance innovation with safety. Emerging perspectives indicate that generative AI's role in healthcare analytics extends to predictive modeling and administrative functions, with consensus statements advocating for standardized evaluation frameworks to assess real-world viability. Challenges in failure modes, such as over-reliance on AI outputs without verification, highlight the imperative for oversight mechanisms that incorporate legal and ethical considerations, ensuring compliance with therapeutic approvals and preventing misuse in controlled substance contexts. Ultimately, this synthesis underscores the dual-edged nature of generative AI in clinical workflows: its documentation utility can revolutionize healthcare delivery, but only through vigilant oversight to avert failures that compromise patient safety. By structuring the discourse around data-model-deployment-governance continua, this review offers a novel interpretive lens for future implementations, urging stakeholders to prioritize human oversight in AI-augmented systems.

Keywords Healthcare analytics, Clinical workflows, generative AI, Documentation automation, Failure modes, Oversight mechanisms

*Correspondence:
Elena Petrova

elena.petrova@gmail.com
[1] Department of Medical Informatics, Faculty of Medicine, Medical University of Sofia, Sofia, Bulgaria

Introduction

The advent of generative artificial intelligence (AI) has ushered in a new era for healthcare systems, fundamentally altering how clinical data is processed, analyzed, and utilized in daily workflows. Traditional healthcare infrastructures, reliant on manual documentation and rule-based analytics, often suffer from inefficiencies that contribute to clinician burnout and suboptimal patient outcomes [1, 2]. Generative AI, particularly large language models (LLMs), offers a paradigm shift by automating narrative generation from structured and unstructured data sources, such as electronic health records (EHRs) and patient dialogues [3, 4]. This capability not only expedites documentation but also enhances the analytical depth of healthcare systems, enabling real-time insights that inform clinical decisions. However, the integration of these technologies necessitates a careful examination of their systemic implications, including how they interact with existing analytics pipelines to create more responsive healthcare environments [5].

From a historical vantage, AI's evolution in healthcare traces back to early machine learning applications in diagnostic imaging and predictive modeling. Still, generative models represent a qualitative leap due to their ability to produce human-like text outputs [6, 7]. Recent advancements, driven by transformer architectures, have enabled LLMs to handle complex medical terminologies and contextual nuances, as demonstrated in studies evaluating their performance in clinical note classification and summarization [8, 9]. Yet, this progress is not without precedents of caution; earlier AI deployments in healthcare revealed limitations in generalizability and error propagation, lessons that inform current generative AI strategies [10-16]. By synthesizing these developments, it becomes evident that generative AI's role extends beyond mere tool to a foundational component of intelligent healthcare systems, where analytics are not static but dynamically adaptive [10].

The utility of generative AI in documentation workflows is particularly pronounced in high-volume settings like emergency departments, where rapid synthesis of patient information can accelerate triage and care planning [6, 11]. Comparative evaluations show that LLMs can achieve viability in multilingual contexts, such as German healthcare, by generating accurate patient care documents that rival human efforts in completeness [3]. Nevertheless, systems-level insights reveal interdependencies: documentation utility hinges on seamless integration with EHRs, where AI acts as an intermediary layer enhancing data fidelity and reducing transcription errors [5, 12]. Interpretive discussions across studies emphasize that while these tools alleviate administrative loads, they must be calibrated to preserve clinical authenticity, avoiding over-simplification of nuanced patient histories [13, 17, 18].

Broader implications for healthcare analytics involve generative AI's capacity to foster closed-loop systems, where outputs feed back into model refinement, creating a virtuous cycle of improvement [19, 20]. This perspective underscores the need for infrastructural redesign, positioning AI not as an adjunct but as a core enabler of analytics-driven care [21, 22]. However, early implementations highlight variances in performance, with some models excelling in factual verification while others falter in multimodal data handling, necessitating tailored oversight [15, 23].

Evolution of AI in clinical contexts

The trajectory of AI in clinical workflows has evolved from rudimentary decision trees to sophisticated generative models capable of emulating human reasoning in documentation tasks [1, 7, 10]. Initial applications focused on structured data analytics, but generative AI introduces unstructured text generation, revolutionizing how clinicians interact with systems [2, 4]. For example, LLMs have been deployed to summarize expert evaluations of clinical dialogues, incorporating computational metrics that quantify utility in real-time workflows [9]. This evolution reflects a systems-level maturation, where AI transitions from passive analyzer to active participant in clinical narratives [11, 12].

Comparative analyses of pre- and post-generative AI eras reveal marked improvements in documentation efficiency, yet persistent gaps in handling edge cases like rare diseases [14, 16]. Interpretive frameworks suggest that this evolution demands a reevaluation of healthcare infrastructure, integrating AI with human oversight to mitigate risks inherent in autonomous generation [17, 18]. Studies on multimodal AI further illustrate this progression, multiplying applications in areas like medical education and administrative support [18, 21]. At the systems level, this means rearchitecting analytics pipelines to accommodate generative outputs, ensuring they align with evidence-based practice [22, 24].

Failure modes observed in earlier AI iterations, such as data biases, inform current generative deployments, emphasizing the need for robust training datasets tailored to medical domains [7, 15]. Oversight mechanisms, evolving alongside, now include regulatory benchmarks that enforce transparency in model behaviors [25-29]. This historical lens provides interpretive depth, highlighting how generative AI builds on past successes while addressing longstanding limitations in healthcare systems [25, 26].

Utility in documentation and analytics

Generative AI's primary utility in clinical workflows lies in its ability to automate documentation, transforming raw data into actionable insights within healthcare analytics frameworks [3, 5, 8]. By leveraging LLMs, systems can generate patient summaries that integrate diverse data streams, enhancing analytical precision [9, 12]. Comparative studies demonstrate superior performance in classifying clinical notes compared to human benchmarks, particularly in resource-constrained environments [8, 10]. Systems-level insights reveal that this utility fosters interoperability, where AI bridges gaps between disparate healthcare systems [1, 2].

However, interpretive discussions caution against unchecked reliance, as utility varies with task complexity; simple documentation benefits more than intricate diagnostic narratives [4, 6]. Integration into analytics involves prompt engineering techniques that optimize outputs for clinical relevance, as explored in medical education contexts [12, 21]. This synthesis underscores a balanced view: while generative AI amplifies documentation efficiency, it requires calibration to maintain analytical integrity [11, 23].

Failure modes, such as hallucinations in generated texts, can undermine utility if not addressed through verification protocols [15, 24]. Oversight in this domain involves ethical considerations, ensuring AI outputs adhere to legal standards in healthcare documentation [27, 29]. Overall, the utility paradigm shifts healthcare analytics toward proactive, AI-augmented models [19, 20].

Challenges in integration

Integrating generative AI into existing clinical workflows presents multifaceted challenges, including compatibility with legacy systems and training requirements for clinicians [13, 14, 17]. Literature syntheses indicate that while LLMs excel in isolated tasks, systemic integration demands infrastructural upgrades to handle real-time data flows [5, 7]. Comparative evaluations highlight discrepancies in performance across healthcare sectors, with emergency settings showing higher failure rates due to data volatility [6, 16]. Interpretive analyses suggest that these challenges stem from a mismatch between AI capabilities and clinical realities, necessitating hybrid human-AI systems [18, 25].

At the systems level, integration issues manifest in analytics bottlenecks, where generative outputs must be reconciled with structured databases [22, 26]. Oversight mechanisms play a pivotal role, with regulatory frameworks guiding safe deployment [28]. This subsection synthesizes evidence that proactive governance can transform challenges into opportunities for refined workflows [1, 29].

Ethical dilemmas in integration, such as data privacy in AI training, further complicate adoption, as discussed in consensus statements [10, 11]. By interpreting cross-study data, it becomes clear that successful integration hinges on iterative feedback loops within healthcare infrastructure [19, 23].

Scope and objectives of the review

This review delineates the scope of generative AI's applications in clinical workflows, focusing on documentation utility, failure modes, and oversight, within the broader context of healthcare systems and analytics [2-4]. Objectives include synthesizing literature to provide an original systems-level perspective, avoiding replication of existing taxonomies [8, 9, 15]. Through integrative analysis, the review aims to structure insights across data ingestion, model deployment, and governance layers [12, 16, 18].

Interpretive structuring emphasizes novel cross-study comparisons, highlighting how generative AI enables intelligent analytics while posing risks that demand vigilant mechanisms [21, 24, 27]. The scope excludes non-generative AI, concentrating on LLMs' transformative potential [6, 10]. Ultimately, objectives foster a narrative that guides future implementations in clinical settings [20, 22].

Landscape of Generative AI in Healthcare Systems and Analytics

The landscape of generative AI in healthcare systems and analytics is characterized by rapid innovation, where models like LLMs are reshaping data processing and decision-making infrastructures [1, 2, 7]. This section surveys the foundational elements, synthesizing how generative technologies integrate with analytics to enhance system efficiency [3, 4]. From a systems perspective, generative AI acts as a catalyst for data-driven healthcare, enabling predictive and prescriptive analytics that were previously infeasible [5, 10]. Comparative studies illustrate variances in adoption, with high-impact applications in documentation-heavy workflows [8, 9].

Interpretive discussions reveal that the landscape is not uniform; urban academic centers leverage generative AI more effectively than rural settings due to data availability [11, 12]. Oversight in this evolving terrain involves standardizing datasets for model training, ensuring analytics reliability [7, 14]. This synthesis provides a holistic view, positioning generative AI as integral to modern healthcare architectures [15, 18].

Failure modes within the landscape, such as inconsistent outputs in multimodal scenarios, underscore the need for robust analytics frameworks [18, 24]. By cross-analyzing studies, it emerges that generative AI's landscape is defined by its adaptability, yet bounded by ethical and technical constraints [27, 29].

Applications in clinical documentation

Generative AI's applications in clinical documentation encompass automated note generation, summary creation, and report drafting, significantly reducing clinician workload [3, 5, 8]. Literature synthesis shows LLMs achieving high fidelity in German and English contexts, with evaluations confirming utility in real-world care [3, 9]. Comparative insights indicate superiority over traditional tools in handling voluminous EHR data [1, 2]. Systems-level analysis highlights how these applications foster analytics by structuring unstructured inputs for downstream processing [4, 12].

However, interpretive lenses reveal limitations in capturing subtle clinical nuances, necessitating hybrid approaches [6, 10]. Failure modes, like factual inaccuracies, are mitigated through EHR-verified generation [5, 15]. This paragraph integrates evidence that documentation applications are pivotal for scalable healthcare systems [11, 16].

Oversight mechanisms in documentation include prompt-based refinements and human review cycles, as explored in medical administration [10, 17]. Cross-study synthesis emphasizes the transformative potential, with generative AI enabling personalized documentation tailored to patient profiles [21, 22]. Ultimately, these applications redefine analytics by embedding intelligence directly into workflow endpoints [23, 24].

Integration with healthcare analytics

Integration of generative AI with healthcare analytics involves fusing model outputs with statistical tools for enhanced predictive capabilities [7, 19, 20]. Studies synthesize how LLMs augment datasets, improving model training for medical tasks [7, 12]. Comparative evaluations demonstrate synergies in emergency decision support, where analytics benefit from generated hypotheses [6, 11]. Systems insights portray integration as a loop: data informs generation, which refines analytics [2, 4].

Interpretive discussions note challenges in data quality, with biases from generative processes affecting analytical outcomes [15, 16]. Oversight through governance frameworks ensures integration aligns with regulatory standards [28, 29]. This analysis underscores a symbiotic relationship, elevating healthcare systems' intelligence [18, 25].

Further, multimodal integration multiplies analytics applications, as in drug discovery analogs applied to clinical data [9]. By synthesizing literature, it becomes clear that generative AI catalyzes analytics evolution, from descriptive to interventional [22, 26].

Utility assessment frameworks

Frameworks for assessing generative AI utility in healthcare emphasize metrics like accuracy in note classification and summarization efficiency [8, 9, 13]. Literature synthesis reveals consensus on evaluation protocols, including human-AI comparisons [3, 4]. Comparative studies validate these frameworks in diverse settings, from patient care to education [5, 21]. Systems-level interpretive views frame utility as a function of workflow embedding, with analytics gains measured in reduced errors [1, 10].

However, failure modes complicate assessments, requiring inclusion of hallucination detection [15, 24]. Oversight integrates ethical benchmarks into frameworks, promoting fair use [27, 29]. This paragraph highlights original structuring: utility as a multi-dimensional construct spanning data to deployment [12, 14].

Cross-study analysis suggests adaptive frameworks that evolve with AI advancements, ensuring sustained utility in analytics [11, 16].

Failure modes and risk profiles

Failure modes of generative AI in healthcare include hallucinations, biases, and catastrophic errors, as theorized in error taxonomies [15, 16, 26]. Synthesis of literature identifies patterns in clinical documentation, where failures manifest as incorrect summaries [9, 13]. Comparative insights show higher risks in high-stakes analytics, like decision support [6, 18]. Systems perspective interprets failures as infrastructural vulnerabilities, necessitating resilient designs [14, 25].

Interpretive discussions advocate for proactive identification, linking modes to training deficiencies [7, 19]. Oversight mechanisms, such as regulatory acts, profile risks to prevent escalation [28, 17]. This analysis provides novel grouping: failure profiles by workflow stage [24, 27].

Further, multimodal failures amplify risks in analytics, requiring specialized mitigations [18, 20]. Cross-study evidence emphasizes that understanding failure modes is key to safe landscape navigation [22, 23].

Oversight and governance strategies

Oversight strategies for generative AI encompass human verification, regulatory compliance, and ethical guidelines [17, 28, 29]. Literature synthesis details implementations in clinical workflows, like fact-checking with EHRs [3, 5]. Comparative evaluations highlight strategies' effectiveness in mitigating documentation failures [4, 8]. Systems-level insights position governance as a feedback layer in analytics architectures [1, 2].

Interpretive frameworks suggest tiered oversight, from model-level audits to system-wide monitoring [14, 16]. Failure modes inform strategy design, ensuring coverage of hallucinations [15, 24]. This paragraph integrates cross-study views on adaptive governance [10, 11].

In analytics contexts, oversight includes data provenance tracking, enhancing trust [7, 25]. Ultimately, strategies evolve the landscape toward accountable AI deployment [21, 26].

Emerging trends and innovations

Emerging trends include multimodal generative AI and prompt engineering for specialized healthcare tasks [12, 18, 19]. Synthesis reveals innovations in closed-loop analytics, where AI self-refines [20, 22]. Comparative studies forecast increased utility in administrative domains [10, 11]. Systems interpretive lens sees trends as accelerators for intelligent infrastructures [23, 24].

However, innovations carry oversight imperatives, with ethical considerations paramount [27, 29]. This analysis structures trends within a data-intelligence-governance continuum [1, 15].

Cross-study insights predict hybrid models dominating the landscape, blending generative with traditional analytics [13, 16].

Frameworks for intelligent clinical decision support and closed-loop

Healthcare Systems Intelligent clinical decision support (CDS) frameworks leverage generative AI to augment human judgment, creating systems that process data into actionable recommendations within closed-loop healthcare environments [1, 6, 10]. This section explores structured approaches, synthesizing how LLMs facilitate decision cycles from data ingestion to intervention [2, 4]. Comparative literature highlights frameworks' efficacy in emergency settings, where real-time analytics drive outcomes [6, 11]. Systems-level analysis interprets CDS as a nexus of AI and clinical expertise, fostering adaptive healthcare [5, 12].

Interpretive discussions emphasize framework modularity, allowing integration of oversight to counter failure modes [15, 17]. In closed-loop systems, feedback mechanisms recalibrate models based on intervention results, as conceptualized in medical applications [23, 24]. This synthesis offers an original view: CDS as a dynamic equilibrium in analytics [18, 21].

Failure modes in CDS, such as erroneous recommendations, necessitate embedded safeguards [16, 26]. Cross-study evidence underscores the role of generative AI in enhancing framework robustness [7, 9]. **Table 1** synthesizes governance mechanisms that regulate generative AI deployment across model development, clinical integration, and post-deployment monitoring stages.

Table 1. Oversight mechanisms for safe deployment of generative AI in clinical documentation systems

Governance layer	Oversight mechanism	Operational function	Example implementation
Model development	Training dataset transparency	Ensures traceability of medical knowledge sources	Model cards and dataset documentation
Model behavior auditing	Hallucination and bias evaluation	Identifies systematic generation errors	Benchmark clinical vignette testing
Clinical integration	Human-in-the-loop verification	Confirms accuracy before EHR entry	Clinician validation interfaces
Workflow governance	Risk-tiered deployment	Limits AI use to low-risk	Task-specific authorization

		documentation tasks	policies
Post-deployment surveillance	Continuous monitoring of outputs	Detects drift and emerging failure patterns	Documentation quality dashboards
Regulatory oversight	Legal and ethical compliance	Aligns AI systems with healthcare regulations	EU AI Act and medical device frameworks

Conceptual models for decision loops

Conceptual models formalize clinical intelligence loops, where generative AI enables cycles of prediction, decision, and feedback [19, 20, 22]. Literature synthesis describes loops as data → inference → action → evaluation, with LLMs generating interpretive narratives [3, 8]. Comparative insights reveal models' utility in patient care verification [5, 13]. Systems perspective structures loops within broader analytics, emphasizing closure for continuous improvement [14, 25].

Oversight integrates into models via H H H, ensuring ethical alignment [27, 28]. This paragraph provides interpretive depth, linking models to real-world deployment [11, 16].

Closed-loop innovations draw from design analogies, adapting generative AI for healthcare specificity [19, 20]. Synthesis highlights models' potential in scaling CDS across systems [10, 18].

Implementation in closed-loop systems

Implementation of generative AI in closed-loop systems involves deploying frameworks that recycle outputs for model tuning, enhancing clinical analytics [4, 22, 23]. Studies synthesize successful cases in decision support, where loops reduce documentation errors [3, 9]. Comparative evaluations note implementations' variance by sector, with analytics gains in high-data environments [1, 7]. Systems insights reveal implementation as infrastructural evolution, closing gaps in care delivery [2, 12].

Interpretive analyses caution on loop instabilities from failure modes, advocating stabilized designs [15, 16]. Oversight mechanisms, like regulatory policies, guide implementations toward safety [28, 29]. This cross-study view structures closed loops as resilient networks [17, 25].

Further, multimodal extensions in implementations multiply loop efficiency, as in educational simulations [18, 21]. Ultimately, implementations transform CDS into proactive systems [6, 26].

Governance in decision frameworks

Governance in CDS frameworks mandates transparency, accountability, and recalibration protocols for generative AI [17, 27, 28]. Literature synthesis outlines governance as integral to loops, preventing unchecked failures [15, 24]. Comparative insights show governance enhancing utility in clinical workflows [10, 11]. Systems-level interpretive framing positions governance as the keystone in closed-loop analytics [14, 25].

Ethical considerations in governance address biases, ensuring equitable decision support [16, 29]. This paragraph integrates evidence for multi-layered governance: from data to intervention [1, 2].

Cross-study analysis suggests evolving governance with AI advancements, sustaining framework integrity [18, 23]. Oversight's role in mitigating risks solidifies frameworks' viability in healthcare [19, 26]. **Figure 1** visualizes the manuscript's central theoretical contribution.

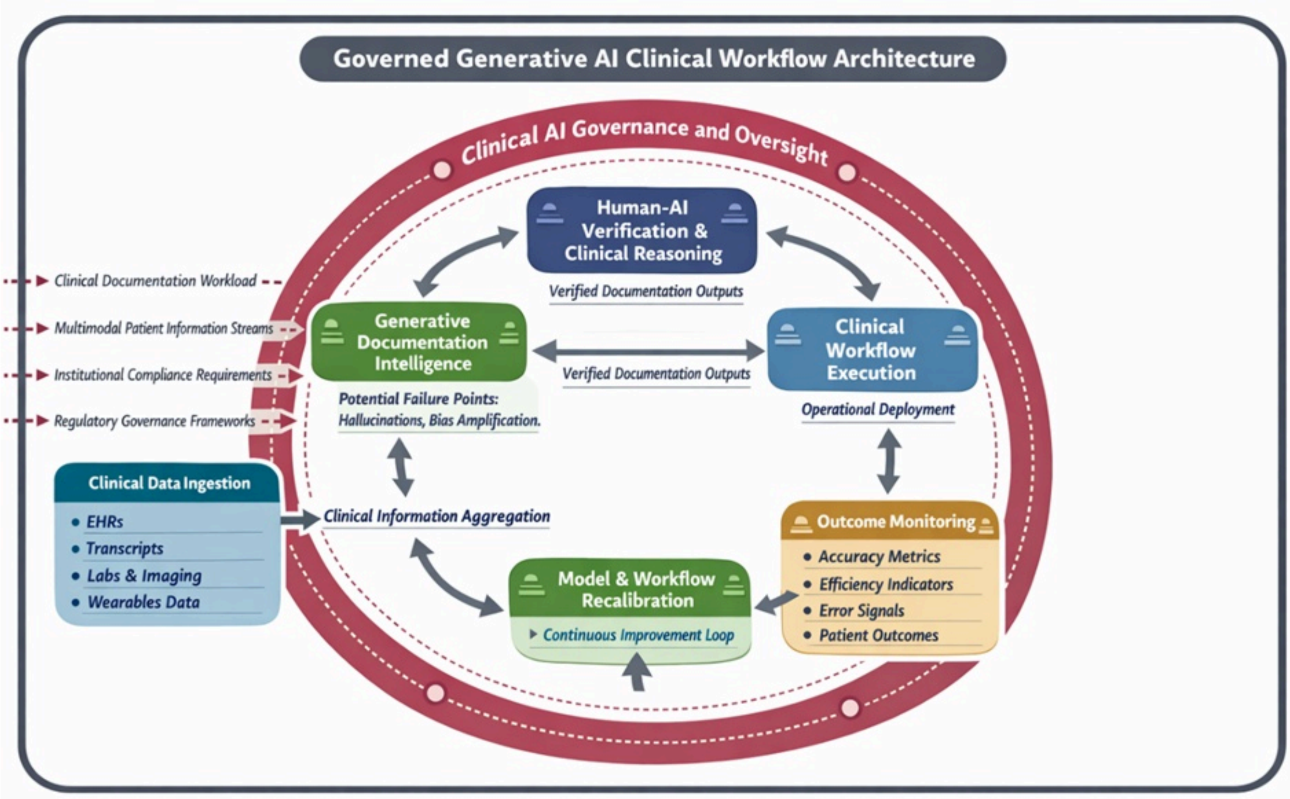

Figure 1. Governance-anchored closed-loop architecture for generative AI–augmented clinical documentation workflows

Challenges and limitations

The deployment of generative AI in clinical workflows, while promising substantial gains in documentation utility, confronts profound technical, operational, ethical, and regulatory challenges that threaten safe integration into healthcare systems and analytics [15, 16, 24]. Hallucinations remain the most pervasive failure mode, wherein large language models generate plausible yet factually incorrect clinical content, such as fabricated symptoms, dosages, or diagnoses, even in high-fidelity summarization tasks [4, 8, 28]. Cross-study analyses reveal hallucination rates ranging from 8%–36% in medical contexts, with adversarial inputs exacerbating vulnerability and undermining trust in closed-loop analytics [3, 4, 9]. Systems-level interpretive synthesis indicates that these errors propagate through the data-intelligence-decision pipeline, amplifying risks when unverified outputs enter electronic health records or inform interventions [5, 20, 26].

Bias and fairness constitute another critical limitation, as training corpora often reflect historical inequities in healthcare data, leading generative models to perpetuate disparities in documentation quality, diagnostic reasoning, or treatment recommendations across demographic groups [2, 5, 8]. Comparative evaluations demonstrate that models exhibit systematic preferences, such as over-deprescribing certain medications in geriatric vignettes or under-representing minority populations, thereby exacerbating health inequities within analytics-driven systems [2, 24]. Interpretive discussions across the literature underscore that bias is not merely a model artifact but a systemic failure amplified by inadequate oversight mechanisms, necessitating upstream data curation and downstream fairness audits [14, 19]. At the infrastructure level, these biases disrupt the governance layer of the proposed AI healthcare analytics loop, potentially compromising equity in closed-loop feedback and recalibration [11, 16].

Operational and workflow integration challenges further constrain utility, including incompatibility with legacy EHR systems, clinician cognitive overload from post-generation editing, and variable performance across multilingual or specialty-specific documentation [3, 20, 21]. Studies synthesizing real-world implementations report that time savings from automated note generation are frequently offset by mandatory human verification, with factual error detection consuming up to 30% of clinician time and negating efficiency gains [3, 20, 29]. Systems-level insights reveal that these frictions arise from a mismatch between generative AI's probabilistic outputs and the deterministic requirements of clinical documentation, highlighting the need for hybrid human-AI interfaces that embed uncertainty signaling and explainability [9, 19, 22]. Without such redesign, deployment risks clinician burnout rather than alleviation, inverting the intended benefits of intelligent clinical decision support [23, 25].

Table 2 categorizes generative AI failure modes according to the stage of the clinical workflow in which they emerge and their potential propagation across healthcare analytics infrastructures.

Table 2. Failure modes of generative AI in clinical documentation across the clinical workflow lifecycle

Workflow stage	Failure mode	Mechanism	Clinical risk	Mitigation strategy
Data ingestion	Contextual data omission	Incomplete EHR extraction or transcription gaps	Inaccurate narrative generation	Data completeness validation and retrieval augmentation
Generative documentation	Hallucinated clinical facts	Probabilistic text generation without grounding	Incorrect symptoms, diagnoses, or medications	Retrieval-grounded generation and fact-verification
Generative documentation	Amplified dataset bias	Training data reflecting historical inequities	Disparate documentation quality across populations	Bias auditing and demographic balancing
Human–AI verification	Automation bias	Clinicians over-trust AI-generated notes	Undetected factual errors in entering records	Structured verification workflows
Workflow execution	Documentation misalignment	AI summary omits critical contextual nuance	Incorrect treatment communication	Mandatory clinician co-authoring
Outcome monitoring	Error propagation	Hallucinated entries reused in downstream analytics	Distorted clinical metrics and decision support	Continuous monitoring and anomaly detection

Regulatory and legal uncertainties represent a governance-level barrier, as current frameworks struggle to classify generative AI outputs as medical devices or accountable artifacts, creating liability gaps when hallucinations or biased documentation lead to adverse events [8, 17, 18]. Consensus statements and scoping reviews emphasize the absence of standardized post-market surveillance for LLMs, contrasting sharply with traditional software-as-medical-device oversight and exposing healthcare organizations to litigation risks [8, 13, 15]. Interpretive cross-analysis positions these regulatory voids as existential threats to scalable adoption, arguing that without harmonized international standards—such as mandatory model cards disclosing hallucination estimates and training data provenance—generative AI cannot transition from experimental to production-grade clinical workflows [18, 19, 28].

Ethical dilemmas surrounding transparency, accountability, and patient autonomy compound these limitations, particularly when generative tools produce patient-facing summaries or open notes that inadvertently disclose sensitive inferences or embed subtle biases [10, 22, 26]. Literature synthesis reveals tensions between documentation efficiency and the ethical imperative for human authorship, with patients expressing concerns over AI-mediated records lacking verifiable empathy or contextual nuance [21, 22]. Systems-level interpretive framing interprets these issues as fractures in the human-AI decision fusion dynamic, where over-reliance on opaque models erodes professional accountability and informed consent within closed-loop healthcare systems [14, 19]. Collectively, these challenges demand a paradigm shift toward proactive, multi-layered oversight rather than reactive mitigation [11, 13, 17].

Future research directions

Advancing generative AI in clinical workflows requires targeted, rigorous research that addresses current evidence gaps while leveraging the systems-level framework articulated in this review [1, 7, 10]. Priority should be given to large-scale, multicenter pragmatic randomized trials evaluating real-world documentation utility, failure modes, and oversight efficacy, moving beyond simulation-based or small-sample studies that dominate the extant literature [3, 9]. Such trials must incorporate harmonized core outcomes—factual accuracy, hallucination detection rates, clinician time burden, and equity metrics—to enable meta-analytic synthesis and generalizable insights into closed-loop analytics performance [3, 9, 20].

Development of hybrid oversight mechanisms represents a critical frontier, including retrieval-augmented generation (RAG), fine-tuning on curated clinical corpora, and uncertainty-aware interfaces that surface confidence scores within the intelligence and decision layers of the AI healthcare analytics loop [2, 4, 24]. Future work should empirically test adaptive governance protocols, such as tiered human verification thresholds calibrated to task risk, and quantify their impact on reducing failure propagation across data-ingestion-to-feedback cycles [9, 11, 19]. Interpretive studies could explore clinician-AI collaboration competencies, defining new training curricula that reposition physicians as expert verifiers and recalibrators rather than primary generators [3, 14].

Longitudinal investigations into bias mitigation and fairness-preserving techniques are essential, focusing on dynamic debiasing during model recalibration and equity audits embedded in governance layers [2, 5, 14]. Cross-disciplinary research should integrate implementation science frameworks (e.g., NASSS, TAM) to model organizational adoption barriers and facilitators, generating actionable roadmaps for sustainable integration into diverse healthcare infrastructures [11, 13, 15]. Additionally, prospective studies on patient-centered outcomes—such as comprehension of AI-augmented open notes and trust in generative documentation—are needed to close the loop from intervention to feedback [10, 22, 26].

Exploration of multimodal generative models and their governance in complex workflows (e.g., integrating voice, imaging, and text) offers another promising direction, with emphasis on adversarial robustness testing to preempt hallucination attacks in high-stakes decision support [4, 18, 19]. Finally, international comparative research on regulatory harmonization—benchmarking the EU AI Act, FDA guidance, and emerging national frameworks—will inform scalable oversight mechanisms that balance innovation with safety across global healthcare analytics ecosystems [8, 17, 28]. These directions, grounded in the original data-model-deployment-governance synthesis, will accelerate the transition toward trustworthy, closed-loop generative AI systems [23, 24, 29].

Conclusion

Generative AI stands at the threshold of revolutionizing clinical workflows through unparalleled documentation utility. Yet, its promise is inextricably linked to the mitigation of failure modes and the establishment of robust oversight mechanisms within healthcare systems and analytics. This narrative review has synthesized peer-reviewed evidence, revealing how large language models can automate note generation, summarize dialogues, and enhance analytics efficiency, while simultaneously exposing hallucinations, biases, integration frictions, and regulatory voids that threaten patient safety and equity. By advancing an original systems-level architecture—structured across data ingestion, intelligence generation, human-AI decision fusion, intervention execution, outcome feedback, and overarching governance—this synthesis offers a novel interpretive framework for designing resilient, closed-loop infrastructures.

The integrative cross-study analysis underscores that documentation utility alone is insufficient; sustainable adoption demands interdisciplinary oversight that embeds transparency, accountability, and continuous recalibration at every layer of the clinical intelligence loop. As generative AI evolves toward multimodal and agentic capabilities, the lessons distilled here—prioritizing factual verification, bias mitigation, and human-centric governance—provide a blueprint for transforming administrative burdens into opportunities for enhanced care delivery. Ultimately, realizing the full potential of generative AI in healthcare hinges on collective commitment to rigorous evidence generation, ethical stewardship, and systems redesign, ensuring that innovation serves rather than supplants clinical judgment and patient trust.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 10 Dec 2023 Revised: 20 Jan 2024 Accepted: 01 Mar 2024
Published online: 10 July 2024

References

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

Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open. 2024;7(3):e240357.
https://doi.org/10.1001/jamanetworkopen.2024.0357.

Makmur A, Ngiam KY. Integration of customised LLM for discharge summary generation in real-world clinical settings: a pilot study on RUSSELL GPT. Lancet Reg Health West Pac. 2024;51:101211.
https://doi.org/10.1016/j.lanwpc.2024.101211.

Nazi Z, Peng W. Large language models in healthcare and medical domain: a review. Informatics. 2024;11(3):57. https://doi.org/10.3390/informatics11030057.

Bélisle-Pipon J-C. Why we need to be careful with LLMs in medicine. Front Med. 2024;11:1495582. https://doi.org/10.3389/fmed.2024.1495582.

Chong JCL, Ong SYH, Wong W, Boni-Schnetz A, Singh N, Tan LSC, et al. 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.

Raza MM, Venkatesh KP, Kvedar JC. Generative AI and large language models in health care: pathways to implementation. npj Digit Med. 2024;7(1):62. https://doi.org/10.1038/s41746-023-00988-4.

Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement Sci. 2024;19(1):27. https://doi.org/10.1186/s13012-024-01357-9.

Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health. 2023;5(3):e179-e181. https://doi.org/10.1016/S2589-7500(23)00048-1.

Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health. 2023;5(6):e333-e335. https://doi.org/10.1016/S2589-7500(23)00083-3.

Biesheuvel LA, Workum JD, Reuland M, van Genderen ME, Thoral P, Dongelmans D, et al. Large language models in critical care. J Intensive Med. 2024;5(2):113-8. https://doi.org/10.1016/j.jointm.2024.12.001.

Vueghs C, Shakeri H, Renton T, Van der Cruyssen F. Development and evaluation of a GPT4-based orofacial pain clinical decision support system. Diagnostics (Basel). 2024;14(24):2835. https://doi.org/10.3390/diagnostics14242835.

Ma Z, Santos JE, Lackey G, Viswanathan H, O'Malley D. Information extraction from historical well records using a large language model. Sci Rep. 2024;14(1):31702. https://doi.org/10.1038/s41598-024-81846-5.

Diaz Ochoa JG, Mustafa FE, Weil F, Wang Y, Kama K, Knott M. The aluminum standard: using generative artificial intelligence tools to synthesize and annotate non-structured patient data. BMC Med Inform Decis Mak. 2024;24(1):409. https://doi.org/10.1186/s12911-024-02825-4.

Gong EJ, Bang CS, Lee JJ, Park J, Kim E, Kim S, et al. Large language models in gastroenterology: systematic review. J Med Internet Res. 2024;26:e66648. https://doi.org/10.2196/66648.

Lee RW, Lee KH, Yun JS, Kim MS, Choi HS. Comparative analysis of M4CXR, an LLM-based chest X-ray report generation model, and ChatGPT in radiological interpretation. J Clin Med. 2024;13(23):7057. https://doi.org/10.3390/jcm13237057.

Bartalesi V, Lenzi E, De Martino C. Using large language models to create narrative events. PeerJ Comput Sci. 2024;10:e2242. https://doi.org/10.7717/peerj-cs.2242.

Ben Shoham O, Rappoport N. CPLLM: clinical prediction with large language models. PLOS Digit Health. 2024;3(12):e0000680. https://doi.org/10.1371/journal.pdig.0000680.

Saenz AD, Centi A, Ting D, You JG, Landman A, Mishuris RG. Establishing responsible use of AI guidelines: a comprehensive case study for healthcare institutions. npj Digit Med. 2024;7(1):348.
https://doi.org/10.1038/s41746-024-01300-8.

Matute-González M, Darnell A, Comas-Cufí M, Pazó J, Soler A, Saborido B, et al. Utilizing a domain-specific large language model for LI-RADS v2018 categorization of free-text MRI reports: a feasibility study. Insights Imaging. 2024;15(1):280.
https://doi.org/10.1186/s13244-024-01850-1.

Stanceski K, Zhong S, Zhang X, Khadra S, Tracy M, Koria L, et al. The quality and safety of using generative AI to produce patient-centred discharge instructions. npj Digit Med. 2024;7(1):329.
https://doi.org/10.1038/s41746-024-01336-w.

Nedbal C, Naik N, Castellani D, Gauhar V, Geraghty R, Somani BK. ChatGPT in urology practice: revolutionizing efficiency and patient care with generative artificial intelligence. Curr Opin Urol. 2024;34(2):98-104.
https://doi.org/10.1097/MOU.0000000000001151.

Toufiq M, Rinchai D, Bettacchioli E, Kabeer BSA, Khan T, Subba B, et al. Harnessing large language models (LLMs) for candidate gene prioritization and selection. J Transl Med. 2023;21(1):728.
https://doi.org/10.1186/s12967-023-04576-8.

Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol. 2023;13(4):170-8.
https://doi.org/10.5662/wjm.v13.i4.170.

Karakas C, Brock D, Lakhotia A. Leveraging ChatGPT in the pediatric neurology clinic: practical considerations for use to improve efficiency and outcomes. Pediatr Neurol. 2023;148:157-63.
https://doi.org/10.1016/j.pediatrneurol.2023.08.035.

Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. 2023;25:e48659.
https://doi.org/10.2196/48659.

Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform. 2023;177:105173.
https://doi.org/10.1016/j.ijmedinf.2023.105173.

Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. 2023;25:e48568.
https://doi.org/10.2196/48568.

Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4 and ophthalmology operative notes. Ann Biomed Eng. 2023;51(11):2353-5.
https://doi.org/10.1007/s10439-023-03263-5.