International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	GENERATIVE ARTIFICIAL INTELLIGENCE AND LARGE LANGUAGE MODELS IN EMERGENCY MEDICINE: A NARRATIVE REVIEW
DOI	https://doi.org/10.31435/ijitss.3(51).2026.6076
RECEIVED	10 May 2026
ACCEPTED	03 August 2026
PUBLISHED	10 August 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

GENERATIVE ARTIFICIAL INTELLIGENCE AND LARGE LANGUAGE MODELS IN EMERGENCY MEDICINE: A NARRATIVE REVIEW

Klaudia Kwolek *(Corresponding Author, Email: kwolek.klaudia2@gmail.com)*
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0003-2145-7528

Wiktoria Laskowska
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0000-8116-0038

Marcel Pilarek
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0004-4114-5078

Sebastian Ożga
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0003-1337-7800

Monika Roszkowska
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0007-7925-8395

Michał Furtak
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0006-5949-5118

Szymon Janczarski
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0004-0881-1083

Dominika Sarna
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0001-4849-8093

Łukasz Wąs
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0003-2962-5303

Piotr Tryczyński
Medical University of Silesia in Katowice, Poland
ORCID ID: 0009-0001-8997-3225

ABSTRACT

The integration of generative artificial intelligence (AI) and large language models (LLMs) into healthcare has accelerated dramatically, with emergency medicine emerging as a particularly dynamic yet challenging domain for clinical deployment. This comprehensive narrative review synthesizes contemporary literature to examine current clinical applications, critical safety considerations, implementation scalability and governance, and future research priorities surrounding generative AI in acute care settings. We examine the current clinical applications, critical safety considerations, implementation scalability and governance, and future research priorities for generative AI in acute care settings. Current emergency medicine applications span automated documentation via ambient clinical intelligence systems, clinical decision support for triage and diagnostic prediction, patient communication tools including discharge summary generation, and multilingual data extraction supporting care transitions. Despite promising efficiency gains - including documented reductions in clinician burnout from 50.6% to 29.4% and modeled documentation time savings of up to 7.1 hours per shift cycle - substantial safety concerns persist, with hallucination rates ranging from 26% to 36% across automated pipelines and systematic misclassification in high-acuity triage tasks. We address automation bias (26% increased risk) and data privacy and governance risks, and identify algorithmic equity as a critical research priority. We conclude that while generative AI holds transformative potential to reduce clerical burden and augment clinical reasoning, its successful deployment in emergency medicine requires rigorous attention to clinical safety, health equity, workflow integration, and human-factors considerations.

KEYWORDS

Generative Artificial Intelligence, Large Language Models, Emergency Medicine, Clinical Decision Support, Ambient Clinical Intelligence, Patient Safety

CITATION

Klaudia Kwolek, Wiktoria Laskowska, Marcel Pilarek, Sebastian Ożga, Monika Roszkowska, Michał Furtak, Szymon Janczarski, Dominika Sarna, Łukasz Wąs, Piotr Tryczyński. (2026) Generative Artificial Intelligence and Large Language Models in Emergency Medicine: a Narrative Review. *International Journal of Innovative Technologies in Social Science*. 3(51). doi: 10.31435/ijitss.3(51).2026.6076

1. Introduction

Emergency medicine is characterized by extreme time pressure, diagnostic uncertainty, high patient acuity, and substantial documentation burdens - conditions that present both opportunities and challenges for artificial intelligence (AI) applications. Generative artificial intelligence (GenAI) and large language models (LLMs) can produce fluent, contextually appropriate clinical text and have been investigated for a range of healthcare applications (Clusmann et al., 2023; Singhal et al., 2023). GenAI systems have demonstrated the ability to integrate diverse data types and to generate structured medical documentation and patient-facing instructions (Tu et al., 2024; Clusmann et al., 2023).

Documentation burden is associated with clinician burnout, and ambient clinical intelligence (ACI) systems have been evaluated in multi-site implementations; studies have reported reductions in burnout and improvements in clinician well-being (Marquis et al., 2026; You et al., 2025). LLMs have also been investigated for clinical decision support tasks including triage severity scoring, diagnostic feature extraction from unstructured text, and generation of patient discharge instructions (Lederer et al., 2026; Schipper et al., 2026; Huang et al., 2024).

However, translating these technical capabilities into safe, equitable, and effective clinical tools faces substantial challenges. High factual error rates and context errors have been consistently reported (Wang et al., 2026; Fuster-Casanovas et al., 2026).

This review provides a synthesis of the emerging literature on GenAI and LLMs in healthcare, with a focused examination of emergency medicine applications. We examine the technical foundations underpinning these systems, critically appraise current clinical applications, delineate the safety and ethical challenges constraining deployment, analyze implementation scalability and governance considerations, and articulate a future research agenda that prioritizes rigorous prospective evaluation, health equity, and the preservation of

clinical reasoning and oversight. This review focuses on the clinical deployment, safety challenges, and implementation strategies for GenAI in emergency medicine, without a detailed re-examination of the underlying model architectures.

2. Methodology

The present study utilizes a comprehensive narrative review methodology to synthesize the interdisciplinary landscape of GenAI and LLM clinical integration within the high-acuity, time-pressured context of modern emergency medicine workflows. This research design was selected as the most appropriate framework for addressing the multidimensional and rapidly evolving nature of the topic, which necessitates the integration of evidence from diverse fields, including clinical acute care, biomedical ethics, computer engineering, implementation science, and medical law. Unlike a systematic review, which is typically constrained to answering a narrow clinical question through quantitative aggregation, the narrative approach facilitates an interpretative and thematic synthesis. This allows for a holistic evaluation of how technological innovations drive broader health system transformations and produce complex operational, socio-technical, and ethical implications. The primary objective of this methodology is to provide a cohesive narrative that connects clinical accuracy and documentational efficiency with structural safety, data privacy, and equity-focused governance.

To achieve this objective, relevant literature was identified and selected based on its capacity to inform the technical, clinical, and regulatory nodes of generative AI deployment in acute care environments. The evidence base evaluated within this review encompasses a broad matrix of study designs, including systematic reviews and meta-analyses, prospective randomized controlled trials and blinded crossover designs, cross-sectional mixed-methods pilot survey studies, retrospective observational and comparative reader studies, and single-arm prospective observational implementations embedded directly in hospital workflows. This empirical foundation was further contextualized by architectural and goal-driven risk assessment case studies, clinical accuracy and readability audits, prospective multi-model safety evaluations, viewpoint analyses, and international consensus guidelines on trustworthy AI deployment. Specific emphasis was placed on literature tracking the clinical deployment of ambient clinical intelligence scribes for automated electronic health record documentation, LLM-driven decision support for Emergency Severity Index triage scoring, and patient-facing communication tools such as automated discharge instructions, layperson summaries, and multimodal video narratives.

The gathered evidence was organized and critically interpreted through thematic analysis rather than quantitative pooling, focusing on the core functional layers and safety constraints reflected throughout this manuscript. Literature was systematically classified into major themes spanning automated documentation quality, clinical risks in high-acuity environments - specifically clinical hallucination and context error rates - automation bias and cognitive over-reliance, and information security infrastructure. Finally, the synthesis integrates these operational realities to automate the assessment of full product lifecycle regulations, articulate future research priorities, emphasize the escalation from single-site pilots to multi-center pragmatic validation, and optimize the deployment of proactive, governance-aware defense models to ensure equitable healthcare outcomes.

3. Results

3.2 Current Clinical Applications in Emergency Medicine

3.1.1. Automated Documentation and Ambient Clinical Intelligence

GenAI has been applied to automate clinical documentation through ambient scribe technologies that capture audio recordings of patient-clinician interactions and leverage LLMs to generate structured drafts of electronic health record notes, including history of presenting illness, physical examination findings, and management plans (Palm et al., 2025; Marquis et al., 2026; Akhlaghi et al., 2026).

Implementation data from multi-center deployments reveal varied adoption patterns. In a large-scale implementation involving 1,430 clinicians across ambulatory, inpatient, and emergency environments, ambient documentation technology was utilized by 48.5% of clinicians for at least half of their patient visits at one health system, and by 43.5% for most or all notes at another (You et al., 2025). An observational trial evaluating an ACI system across 248 emergency department (ED) encounters found that its use yielded modelled documentation time savings of 7.1 hours for average-speed typists (a 40.8% efficiency increase) and 4.9 hours for rapid typists (a 35.3% efficiency increase) (Akhlaghi et al., 2026). Adoption was stratified by professional seniority: emergency medicine trainees exhibited higher adoption (48%) compared to attending

consultants (22.5%), with trainees authoring 74.6% of all AI-generated entries (Akhlaghi et al., 2026). In a cross-sectional survey, 64.3% of emergency physicians reported overall satisfaction with ACI technology, and 50% indicated a preference for AI scribes over traditional in-person medical scribes (Marquis et al., 2026).

3.1.2. EHR Note Generation Quality and Structural Characteristics

Comparisons of LLM-generated notes with human-authored documentation show differences across several quality domains. In head-to-head specialty assessments, AI-generated ambient notes were rated significantly more thorough ($p < 0.001$) and better organized ($p = 0.03$) than human expert-drafted reference charts (Palm et al., 2025). Human-authored notes received significantly higher ratings for accuracy ($p = 0.05$), succinctness ($p < 0.001$), and internal consistency ($p = 0.004$) (Palm et al., 2025).

In an ED cohort study, 1,143 manual alterations were made across AI-generated notes; the sections requiring the highest frequencies of modification were the history of presenting illness (23.3%), management plan (22.9%), and clinical impression (16.7%) (Akhlaghi et al., 2026). Survey responses from emergency physicians indicated that ACI was considered useful for compiling the history of present illness, but less helpful for documenting physical examinations (23.1% helpfulness rating) and medical decision-making (35.7% helpfulness rating) (Marquis et al., 2026).

In evaluations of note accuracy, 31% of AI ambient notes were found to contain clinical hallucinations compared to 20% of expert human notes ($p = 0.01$) (Palm et al., 2025). In an evaluation of ChatGPT-4, errors were categorized as omissions of necessary clinical facts (86.3%), unprompted additions (10.5%), and incorrect data points (3.2%); the total number of errors per case ranged from 5.7 to 64.7 (Kernberg et al., 2024). Inverse correlations between note accuracy and both original transcription length and the number of scorable clinical data elements were observed, suggesting that performance may be lower for more complex, information-dense encounters (Kernberg et al., 2024).

3.1.3. Triage and Diagnostic Predictive Modeling

GenAI has also been investigated for clinical decision support, including ED triage scoring and diagnostic data extraction for downstream machine learning pipelines (Lederer et al., 2026; Schipper et al., 2026). In triage evaluations, LLMs were asked to assign Emergency Severity Index scores based on chief complaint, vital signs, and pain score. Among twelve evaluated models, Claude Sonnet 4.5, ChatGPT-5 Instant, and Claude Opus 4.1 achieved the highest concordance rates with human-assigned scores, ranging from 62.37% to 62.89%, with average absolute deviations between 0.37 and 0.40 points (Lederer et al., 2026). Lower-performing models, including Gemini 2.5 Flash and ChatGPT-4o Mini, demonstrated lower concordance (43.81% and 45.36%, respectively) (Lederer et al., 2026). Systematic misclassification patterns were observed: certain models, such as ChatGPT-o3 (40.72%) and ChatGPT-4.1 (35.05%), showed a propensity for over-triage, while others, including Grok (45.36%) and Gemini 2.5 Flash (44.85%), tended toward under-triage (Lederer et al., 2026).

3.1.4. Patient-Clinician Communication and Clinician Well-Being

Associations between ambient documentation technology and clinician well-being have been examined in multi-site surveys (You et al., 2025). After 42 days of access to ambient documentation technology at one academic center, the proportion of clinicians meeting standardized burnout criteria decreased from 50.6% to 29.4% ($P < .001$); the reduction was sustained at 84 days (30.7%, $P < .001$) (You et al., 2025). At another health system, the proportion of clinicians reporting a positive impact of their documentation workflow on personal well-being increased from 1.6% at baseline to 32.3% ($P < .001$) after 60 days of software use (You et al., 2025). In qualitative free-text responses, some clinicians reported enhanced professional joy and improved interaction with patients. User feedback also noted limitations, including reduced utility for pediatric physical examinations and psychiatric evaluations (You et al., 2025).

3.1.5. Discharge Documentation and Patient Education

AI-generated discharge instructions and patient-facing educational materials are an active area of investigation. In a randomized blind survey comparing GPT-4-generated after-visit summaries with standard institutional education resources, lay respondents assigned significantly more favorable ratings ($P = 0.01$) to AI-generated return precautions; for other subsections (diagnoses, procedures, treatments, post-ED medication changes), no statistically significant differences were observed between the two types of materials (Huang et al., 2024). Linguistic analyses indicate that although AI-generated, patient-centered letters can improve objective comprehension scores compared to standard discharge forms, plain-language prompts may omit key clinical objectives or produce text with readability levels that remain challenging for some lay populations (Ayre et al., 2026).

Multimodal systems have been developed that synthesize structured data (vital sign trends, laboratory findings, prescriptions, International Classification of Diseases, 10th Revision medical histories) and unstructured radiology impressions into plain-language video narratives delivered by a virtual avatar (Luo et al., 2025). In structured evaluations, board-certified emergency physicians rated AI-generated video summaries significantly higher ($P < 0.001$) than standard written documents on completeness (4.1 vs. 3.1), correctness (3.9 vs. 3.5), and patient accessibility (3.4 vs. 2.9) (Luo et al., 2025). Physicians reported that 60% of these video instructions were appropriate for review with patients under supervision, and 40.9% were considered suitable as supplements to standard text forms; in 20.5% of cases, the AI-generated instructions showed underperformance in correctness, particularly for complex therapeutic regimens (Luo et al., 2025).

3.1.6. Interprofessional Communication and Care Transitions

LLMs have been evaluated for interprofessional communication during care transitions and for multilingual support in clinical natural language processing (Nadalini et al., 2026; Schipper et al., 2026). In a phase 1 pilot study of an electronic health record (EHR)-integrated discharge summary generator (GPT-4o and GPT-4.1) across 10 hospital departments, the system generated 379 summary drafts for 253 admissions (Nadalini et al., 2026). Clinicians incorporated LLM-generated content into 58.5% of final hospital documentation, with verifiable AI content retained in 29.1% of finalized discharge letters. Adoption varied by department: the emergency department showed the lowest text adoption, with 11 total generations, 6 copy clicks, and a median Bilingual Evaluation Understudy-4 score of 0.02 (Nadalini et al., 2026). The authors suggested that because the triage note may be the initial or sole point of clinical documentation at the time of the ED encounter, the LLM often had insufficient in-hospital documentation to generate clinically detailed discharge narratives. Automated user action logging showed that despite 86.9% of pilot users self-reporting subjective time savings, objective time-in-workflow was significantly higher for cases using the LLM compared to non-LLM cases ($P = 0.000$); the authors noted that safety-oriented user instructions may have contributed to this increase (Nadalini et al., 2026).

3.2 Critical Challenges and Safety Considerations

3.2.1. Accuracy Limitations and Hallucinations

The most pervasive and clinically consequential limitation of current GenAI systems is their propensity to produce factually incorrect, misleading, or entirely fabricated outputs - commonly termed hallucinations - with a degree of fluency and confidence that can obscure their falsehood. Literature audits reveal that up to 49.6% of chatbot responses in domains susceptible to health misinformation can be classified as problematic, encompassing both somewhat problematic (30.0%) and highly problematic (19.6%) content delivered with unwarranted certainty and an authoritative tone (Tiller et al., 2026). Systematic evaluations of clinical documentation pipelines consistently report hallucination rates ranging between 26% and 36%, underscoring a profound dissociation between the high structural and linguistic quality of generated text and its factual correctness (Wang et al., 2026). Qualitative error taxonomies in automated clinical note generation identify context or meaning errors as the most prevalent category (38.7%), followed by terminology or numeric errors (19.4%), critical omissions of clinically relevant data (15.6%), and unsupported additions - content with no basis in the source conversation or patient record (7.5%) (Fuster-Casanovas et al., 2026). Beyond factual errors, LLMs frequently fail to follow complex clinical guidelines, exhibit marked performance degradation when encountering unfamiliar or rare clinical presentations, and display a tendency toward sycophancy - prioritizing outputs that align with perceived user beliefs or expectations rather than adhering to objective medical evidence (Reis et al., 2026; Tiller et al., 2026). Reference and citation validation queries further expose severe structural limitations, with a documented median reference completeness score of only 40% and a high prevalence of entirely fabricated bibliographic entries that appear superficially credible (Tiller et al., 2026).

3.2.2. Clinical Risks in High-Acuity Settings

In high-acuity emergency medicine and critical care environments, the consequences of these non-deterministic and variably accurate outputs are particularly severe, as errors propagate rapidly through tightly coupled clinical workflows (Armoundas & Singh, 2026; Nagaraja et al., 2026). Architectural risk modeling traces how adversarial inputs or systemic model errors can cascade across clinical systems to cause life-threatening harm. The misdiagnosis or delayed recognition of critical, time-sensitive conditions - such as acute stroke, sepsis, or emergent malignancies - carries catastrophic clinical impact, directly delaying or precluding life-saving interventions and violating established regulatory safety standards (Nagaraja et al., 2026). Other documented clinical risks include the potential execution of unauthorized medical procedures, such as unintended radiation exposure or unnecessary surgical interventions resulting from unvalidated downstream

agent tasks, and corrupted medication recommendations wherein critical allergy flags are ignored, unsafe pharmaceutical dosages are generated, or dangerous drug-drug interactions are overlooked (Nagaraja et al., 2026). Moreover, error rates and model behavior are not static: they compound when models undergo continuous modification from live localized data streams, and the phenomenon of "off-policy AI learning" - wherein models learn from historical data that no longer reflects active clinical policies or current best practices - can lead to severe performance drift and baseline degradation over time (Armoundas & Singh, 2026).

3.2.3. Automation Bias and Health Equity

Human-AI interaction dynamics introduce critical sociotechnical vulnerabilities that can undermine clinical safety even when the underlying algorithms perform with high technical accuracy. Foremost among these is automation bias, defined as the clinician's tendency to over-rely on automated outputs as a heuristic replacement for vigilant, independent information seeking and critical appraisal (Goddard et al., 2011; Wang et al., 2026). Automation bias manifests in two primary forms: errors of commission, wherein users follow incorrect automated advice against their own better judgment, and errors of omission, wherein users fail to take necessary clinical action because they were not explicitly prompted to do so by the automated system (Goddard et al., 2011). Cognitive resources are highly sensitive to environmental mediators: high workloads, task complexity, and severe time constraints - conditions endemic to emergency medicine - significantly accelerate a user's reliance on external computer-generated prompts and reduce the cognitive bandwidth available for critical verification (Goddard et al., 2011). Empirical meta-analytic data indicate that prominent on-screen displays of automated recommendations significantly increase the likelihood of clinicians reversing their own correct unaided decisions in favor of erroneous software advice, with a pooled risk ratio of 1.26 denoting a 26% increased risk of incorrect clinical decision-making under faulty decision support (Goddard et al., 2011).

This risk is compounded by the "collaboration paradox," wherein human-AI teams do not universally outperform standalone human or standalone AI workflows, and in certain configurations may underperform both (Wang et al., 2026). Conflict between AI-generated advice and a clinician's independent judgment triggers cognitive dissonance and confirmation biases, leading to failure modes wherein clinicians either anchor on incorrect algorithmic suggestions and fail to override them, or conversely, inappropriately override high-performing AI outputs, diluting the potential benefit of the automated assistance (Wang et al., 2026). Furthermore, the increasingly human-like, empathetic, and conversationally fluent tone of modern LLM interfaces poses an insidious sociotechnical hazard by encouraging users to anthropomorphize or "humanize" the underlying models, building unwarranted trust and emotional rapport that obscures the models' propensity for hidden hallucinations and their fundamental lack of genuine understanding or clinical judgment (Pal et al., 2025; Wang et al., 2026).

3.2.4. Data Privacy and Governance

Data privacy and information governance constitute substantial implementation barriers, particularly when clinical workflows rely on continuous audio recording of patient encounters or require the uploading of sensitive, identifiable clinical information to external cloud-based LLM services (Armoundas & Singh, 2026; Fuster-Casanovas et al., 2026). Deployed ambient intelligence and automated documentation workflows often involve large-scale data harvesting that risks exposing protected health information without explicit, fully informed patient consent (Pal et al., 2025; Armoundas & Singh, 2026). Because the default configurations of major commercial LLM services are not designed to persist user content or expose traditional retrieval endpoints compliant with healthcare data regulations, healthcare organizations are forced to engineer complex, bespoke pipelines to extract, process, and archive conversational metadata and generated clinical text (Radanliev et al., 2026). This continuous retention of sensitive data streams increases institutional vulnerability to malicious data breaches, identity theft, and unauthorized third-party access, demanding strict compliance with data protection regimes such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) through robust encryption, stringent purpose limitations, and comprehensive, tamper-evident audit logs (Radanliev et al., 2026; Rajpurkar et al., 2022). Data governance is further challenged by secondary data reuse across separate projects or research initiatives, which fundamentally complicates the execution of valid informed consent and introduces latent safety hazards when human AI trainers or system developers routinely access and review generated conversational threads as part of quality assurance or model improvement processes (Pal et al., 2025; Rajpurkar et al., 2022).

3.3 Future Directions and Research Priorities

3.3.1. Multi-Center Validation and Implementation Scalability

A primary research priority identified across real-world implementation studies is the transition from single-institution pilot projects - which possess limited external validity due to localized data overfitting and idiosyncratic workflows - to rigorous multi-center validation studies (Lekadir et al., 2025; Artsi et al., 2025). Multi-site evaluations are deemed essential to characterize model generalizability, technical interoperability, and local clinical validity across heterogeneous healthcare settings, varying clinical workflows, information technology infrastructures, and diverse patient populations (Lekadir et al., 2025; Artsi et al., 2025). Implementation science priorities focus on systematically mapping how cross-setting variations in medical equipment, data formats, documentation conventions, and clinical protocols influence algorithmic performance, as baseline technical accuracy in a controlled research environment does not automatically translate into population-level clinical benefit (Ooi & Periasamy, 2026; Lekadir et al., 2025). Additionally, the literature formalizes an inverse relationship between intervention complexity and evidence strength: while narrow, single-modality diagnostic tools possess a relatively established empirical backing, complex, multi-domain care coordination workflows - precisely the applications toward which GenAI is rapidly expanding - represent a substantial research gap requiring long-term, methodologically sophisticated evaluation (Ooi & Periasamy, 2026). At a macro-fiscal level, the capacity of agentic and GenAI architectures to mitigate healthcare expenditure growth remains entirely unvalidated, indicating a need for large-scale longitudinal health economic evaluation strategies rather than reliance on micro-level efficiency extrapolations (Ooi & Periasamy, 2026).

3.3.2. Patient-Centered Evaluation and Healthcare Equity

Explicit methodological gaps persist in current evaluation frameworks, which traditionally prioritize clinician-judged accuracy and technical performance while systematically omitting systematic assessments of how patients and their families interpret, understand, and emotionally respond to AI-generated clinical documentation and communication (Han et al., 2026). Authors strongly recommend the development and validation of dual-perspective evaluation frameworks that balance traditional clinical precision metrics with rigorous measurement of patient understanding, readability, reassurance, and overall communication quality (Han et al., 2026). Future research directions for patient-facing AI tools include the refinement of self-triage symptom checkers through advanced deep learning architectures to address their current high risk-aversion, inconsistent sorting logic, and well-documented tendency to misclassify true medical emergencies as non-urgent conditions, potentially delaying critical care (Chenais et al., 2023). In designing these digital patient portals, the literature stresses the paramount importance of proactively addressing equity-related research needs; specific evaluation strategies must propose and test alternative or guided access solutions for vulnerable, high-acuity patient cohorts who face digital literacy gradients, language barriers, or age-related technological access barriers, such as the oldest-old population (Ooi & Periasamy, 2026; Chenais et al., 2023).

3.3.3. Governance, Regulatory Compliance, and Ethical Infrastructure

The clinical deployment of workflow-embedded generative models necessitates structural governance and regulatory research that proactively aligns with rapidly evolving legal frameworks, such as the EU AI Act, which imposes strict conformity assessments, risk classification, and post-market monitoring requirements for high-risk clinical software tools (Ooi & Periasamy, 2026; Lekadir et al., 2025). The literature calls for the clear, prospective pre-specification of individual and collective liability, robust accountability mechanisms, and fair compensation structures for patients who suffer harm as a result of AI-related errors (Lekadir et al., 2025). Ethical research priorities emphasize the critical importance of identifying and managing systemic, human, and institutional biases that are frequently overlooked by purely computational or statistical debiasing methods, which cannot address the upstream structural inequities embedded in training data generation (Chenais et al., 2023). Authors highlight that current training datasets often flatten or omit complex social, behavioral, and demographic factors, risking the replication and automation of historic triage errors and outcome disparities across genders, ethnicities, and socioeconomically marginalized older adult cohorts (Ooi & Periasamy, 2026; Chenais et al., 2023). Systematic reviews of existing reporting standard frameworks reveal wide variability in development rigor, scope, and practical applicability, prompting strong recommendations to align medical machine learning artifacts with Findable, Accessible, Interoperable, and Reusable (FAIR) principles to enhance transparency, reproducibility, and collaborative progress (Shiferaw et al., 2025). Finally, an emerging but increasingly urgent research priority involves quantifying and mitigating the substantial environmental impact and carbon footprint of large-scale computational infrastructure, with recommendations to develop energy-efficient algorithms, model pruning and distillation techniques, and localized edge computing solutions that reduce reliance on energy-intensive cloud data centers (Lekadir et al., 2025; Shiferaw et al., 2025).

4. Discussion

The synthesis of recent literature reveals a paradigm shift in emergency department informatics, characterized by the progressive evolution of GenAI from superficial administrative utilities toward interactive clinical augmentation systems. Automated documentation via ACI currently represents the most mature and implementation-ready application within acute care settings, driven by its capacity to alleviate the severe clerical burden contributing directly to the epidemic of physician burnout (Marquis et al., 2026; Akhlaghi et al., 2026; You et al., 2025). The documented reductions in burnout and modeled time savings represent outcomes of substantial operational significance. The generational divide in ACI adoption suggests that integrating these tools into residency curricula may accelerate acceptance (Akhlaghi et al., 2026).

However, a clear and consequential demarcation exists between administrative tasks - where LLMs demonstrate strong organizational and communicative capabilities (Palm et al., 2025; Huang et al., 2024) - and direct clinical decision-support applications. Yet as the focus transitions to clinical triage and diagnostic prediction, serious reliability limitations emerge. Even the highest-performing architectures achieve only moderate concordance with human-assigned triage scores, while exhibiting systematic over-triage or under-triage tendencies that carry direct patient safety implications (Lederer et al., 2026). This disparity explains why low-risk administrative and communication workflows are being prioritized for real-world rollout, while diagnostic tools remain largely confined to retrospective and pilot validation (Chenais et al., 2023; Artsi et al., 2025).

The reviewed evidence consistently highlights a fundamental structural tension between operational optimization and clinical risk management. While ACI systems deliver substantial workflow efficiency and measurable improvements in clinician well-being, these favorable metrics stand in stark contrast to persistent hallucination rates documented in the literature (Wang et al., 2026). The fluent, empathetic tone of modern LLMs encourages anthropomorphization, building unwarranted trust that obscures hidden inaccuracies (Pal et al., 2025; Wang et al., 2026).

In the zero-tolerance safety environment of emergency medicine, where minor discrepancies can cascade into life-threatening harm, this productivity–safety trade-off represents the defining challenge for future implementation. The mandatory human verification requirement introduces a “vigilance tax” that can paradoxically increase workflow time - as documented in the discharge summary pilot where objective time-in-workflow paradoxically increased due to mandatory validation (Nadalini et al., 2026). This tension is further exacerbated by human factor vulnerabilities including automation bias (Goddard et al., 2011) and the “collaboration paradox” wherein human-AI teams fail to outperform standalone workflows (Wang et al., 2026). Moreover, the high-acuity risks catalogued in Section 3.2.2 - including misdiagnosis cascades, unauthorized procedures, and medication errors - underscore the potentially catastrophic consequences of unreliable decision support in emergency settings. Navigating this dichotomy requires rigorous, multi-center prospective validation and continuous post-deployment auditing, not reliance on localized, unmonitored pilots.

The clinical integration of GenAI is constrained by significant data privacy and information governance challenges. As detailed in Section 3.2.4, continuous audio recording and cloud-based data processing expose protected health information to risks of breach and unauthorized access, demanding strict compliance with regulations such as GDPR and CCPA through robust encryption and audit logs (Radanliev et al., 2026; Rajpurkar et al., 2022). Addressing these multifaceted risks requires governance frameworks that promote transparency and reproducibility; the literature emphasizes alignment with FAIR principles (Shiferaw et al., 2025). Moreover, as discussed in Section 3.3.3, evolving regulatory frameworks like the EU AI Act impose post-market monitoring requirements, underscoring the need for continuous auditing infrastructures that enforce transparency and preserve clinical autonomy rather than reliance on singular pre-market clearances (Ooi & Periasamy, 2026; Lekadir et al., 2025).

The literature reveals a growing recognition that evaluation of GenAI must extend beyond technical accuracy to incorporate patient-centered and real-world implementation outcomes. Section 3.3.2 highlights systematic omissions in current frameworks, which typically prioritize clinician-judged accuracy while neglecting patient comprehension, readability, and emotional impact (Han et al., 2026). Similarly, Section 3.3.1 underscores the need for multi-center prospective trials that assess not only clinical validity but also workflow integration and scalability (Lekadir et al., 2025; Artsi et al., 2025). Future evaluation frameworks should therefore capture both operational metrics - such as documentation time and clinician well-being - and patient-centered measures of communication quality, readability, and understanding (Han et al., 2026). The field must progress from retrospective in silico studies toward prospective, multi-center pragmatic trials that evaluate clinical validity and real-world scalability (Lekadir et al., 2025; Artsi et al., 2025). This approach prioritizes safety, equity, and the preservation of clinical reasoning and oversight as primary endpoints.

5. Conclusions

Generative AI and large language models are rapidly advancing in emergency medicine, with the most mature applications centered on automated documentation and ambient clinical intelligence. These tools have demonstrated measurable reductions in clinician burnout and substantial modeled documentation time savings, addressing a critical driver of occupational distress. In parallel, LLM-generated patient communication materials and discharge summaries show promise in improving accessibility and interprofessional information transfer, though their adoption varies by clinical context and requires careful clinician oversight.

Despite these efficiency gains, the evidence reveals persistent and clinically significant safety limitations. Hallucination rates in generated clinical text remain high, and systematic misclassification in triage tasks - manifesting as both over-triage and under-triage - poses direct risks in high-acuity environments. Automation bias and the collaboration paradox further undermine the reliability of human-AI teams, as clinicians may over-rely on erroneous outputs or dismiss high-performing algorithmic advice under cognitive load. These human-factors vulnerabilities are compounded by data privacy challenges inherent in cloud-based processing and continuous audio recording, demanding rigorous compliance with evolving regulatory frameworks.

Realizing the transformative potential of GenAI in emergency care will require moving beyond single-institution pilots to multi-center prospective evaluations that prioritize clinical safety, health equity, and patient-centered outcomes. Future work must address implementation barriers through structural governance, transparent reporting aligned with FAIR principles, and proactive mitigation of algorithmic bias. The responsible integration of generative AI into emergency care depends on a balanced approach that preserves clinical reasoning and oversight while harnessing automation to reduce clerical burden, with patient well-being as the guiding endpoint.

All authors have read and agreed with the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

Funding statement: No external funding was received to perform this review.

Statement of data availability: The data supporting the findings of this study are available within the article's bibliography.

REFERENCES

1. Akhlaghi, H., Freeman, S., Sun, K., Nie, N., Ding, J., Chen, L., Pham, E., Morrissey, B., & Karro, J. (2026a). Evaluation of an AI scribe tool in the emergency department: A single-arm observational study. *Emergency Medicine Australasia, 38*(3), e70272. https://doi.org/10.1111/1742-6723.70272
2. Akhlaghi, H., Freeman, S., Sun, K., Nie, N., Ding, J., Chen, L., Pham, E., Morrissey, B., & Karro, J. (2026b). Evaluation of an AI scribe tool in the emergency department: A single-arm observational study. *Emergency Medicine Australasia, 38*(3), e70272. https://doi.org/10.1111/1742-6723.70272
3. Armoundas, A. A., & Singh, J. P. (2026). Total product lifecycle regulatory considerations and recommendations for generative AI-enabled medical devices. *European Heart Journal—Digital Health, 7*(3), ztag019. https://doi.org/10.1093/ehjdh/ztag019
4. Artsi, Y., Sorin, V., Glicksberg, B. S., Korfiatis, P., Nadkarni, G. N., & Klang, E. (2025). Large language models in real-world clinical workflows: A systematic review of applications and implementation. *Frontiers in Digital Health, 7*, 1659134. https://doi.org/10.3389/fdgth.2025.1659134
5. Ayre, J., Shao, L., & Dunn, A. G. (2026). Why we need patients and community at the center of AI health communication research. *Journal of Medical Internet Research, 28*, e97577. https://doi.org/10.2196/97577
6. Chenais, G., Lagarde, E., & Gil-Jardiné, C. (2023). Artificial intelligence in emergency medicine: Viewpoint of current applications and foreseeable opportunities and challenges. *Journal of Medical Internet Research, 25*, e40031. https://doi.org/10.2196/40031
7. Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J.-N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G. P., Wagner, S. J., & Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine, 3*(1), 141. https://doi.org/10.1038/s43856-023-00370-1
8. Fuster-Casanovas, A., Vidal-Alaball, J., Alonso, C., Catalina, M., Heinisch, D. H., Domínguez-Alonso, J. A., Hamud, I. J., Acosta-Rojas, R., Torres-Mercado, A. A., Baró, R., Tebé, M. C., Castaño, A., & Reguant, L. S. (n.d.). *Evaluating patient and professional satisfaction and documentation time reduction through AI-driven automatic clinical note generation in primary care: Proof-of-concept study.*

9. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089
10. Han, B., Barnes, T., Reddy, C. D., & Shin, A. Y. (2026). Evaluating large language model–generated clinical summaries through a dual-perspective framework: Retrospective observational study. *JMIR AI, 5*, e85221. https://doi.org/10.2196/85221
11. Huang, T., Safranek, C., Socrates, V., Chartash, D., Wright, D., Dilip, M., Sangal, R. B., & Taylor, R. A. (2024). Patient-representing population's perceptions of GPT-generated versus standard emergency department discharge instructions: Randomized blind survey assessment. *Journal of Medical Internet Research, 26*, e60336. https://doi.org/10.2196/60336
12. Kernberg, A., Gold, J. A., & Mohan, V. (2024). Using ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters: Comparative study. *Journal of Medical Internet Research, 26*, e54419. https://doi.org/10.2196/54419
13. Lederer, T. G., Herring, W. C., Ammar, L. A., Abella, B. S., Apakama, D. J., Abbott, E. E., & Shekhar, A. C. (2026). Large language models (LLM) for emergency department triage based on vital signs. *Emergency Care and Medicine, 3*(1), 9. https://doi.org/10.3390/ecm3010009
14. Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., . . . Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388*, e081554. https://doi.org/10.1136/bmj-2024-081554
15. Luo, L., Chen, E., Zhang, X., Acosta, J. N., Jin, B. T., Gunturkun, F., Rose, C., Preiksaitis, C., Suffoletto, B., Rajpurkar, P., & Kim, D. A. (2026). ED-Explain: Personalized video instructions for patients discharged from the emergency department. *Pacific Symposium on Biocomputing, 31*, 280–293. https://doi.org/10.1142/9789819824755_0020
16. Marquis, T., Kopp, M., Anderson, J. S., Napoli, A. M., Brown, L. L., & Berlyand, Y. (2026). AI-powered ambient scribe technology experiences among emergency physicians: Cross-sectional, mixed methods pilot survey study. *JMIR Formative Research, 10*, e80401. https://doi.org/10.2196/80401
17. Nadalini, N., Mehri, T., Hoekman, A. H., Kagialari, K., & Doornberg, J. N. (n.d.). *Phase 1 implementation of LLM-generated discharge summaries showing high adoption in a Dutch academic hospital.*
18. Nagaraja, N., & Bahsi, H. (2026). *Goal-driven risk assessment for LLM-powered systems: A healthcare case study* (arXiv:2603.03633). arXiv. https://doi.org/10.48550/arXiv.2603.03633
19. Ooi, D. R., & Periasamy, B. (2026). *Agentic AI for ageing healthcare systems in advanced economies: A structured review of evidence, institutional barriers, and a sociotechnical implementation roadmap* (SSRN Scholarly Paper No. 6376138). Social Science Research Network. https://doi.org/10.2139/ssrn.6376138
20. Pal, A., Wangmo, T., Bharadia, T., Ahmed-Richards, M., Bhanderi, M., Kachhadiya, R., Allemann, S., & Elger, B. (2025). Generative AI/LLMs for plain language medical information for patients, caregivers and general public: Opportunities, risks and ethics. *Patient Preference and Adherence, 19*, 2227–2249. https://doi.org/10.2147/PPA.S527922
21. Palm, E., Manikantan, A., Mahal, H., Belwadi, S. S., & Pepin, M. E. (2025). Assessing the quality of AI-generated clinical notes: Validated evaluation of a large language model ambient scribe. *Frontiers in Artificial Intelligence, 8*, 1691499. https://doi.org/10.3389/frai.2025.1691499
22. Radanliev, P., Santos, O., & Maple, C. (2026). Threats and vulnerabilities in artificial intelligence and agentic AI models. *Frontiers in Artificial Intelligence, 9*, 1731566. https://doi.org/10.3389/frai.2026.1731566
23. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine, 28*(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0
24. Reis, F., Agha-Mir-Salim, L., Hickstein, R., Reis, M., Piper, S. K., Balzer, F., & Boie, S. D. (2026). Disclaimers and referral patterns for medical advice across urgency levels: Large language model evaluation study. *Journal of Medical Internet Research, 28*, e84668. https://doi.org/10.2196/84668
25. Schipper, A., Belgers, P., O'Connor, R. D., Van De Wouw, L., Builtjes, L., Bosma, J. S., Kusters, R., Kurstjens, S., Rutten, M., & Van Ginneken, B. (2026). Large language model automated extraction of clinical signs and symptoms from emergency department reports for machine learning prediction models: Development and validation study. *JMIR Medical Informatics, 14*, e81500. https://doi.org/10.2196/81500
26. Shiferaw, K. B., Roloff, M., Balaur, I., Welter, D., Waltemath, D., & Zeleke, A. A. (n.d.). *Guidelines and standard frameworks for artificial intelligence in medicine: A systematic review.*
27. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

28. Tiller, N. B., Marcon, A. R., Zenone, M., Kidd, K. E., Jeukendrup, A. E., Master, Z., & Caulfield, T. (2026). Generative artificial intelligence-driven chatbots and medical misinformation: An accuracy, referencing and readability audit. *BMJ Open, 16*(4), e112695. https://doi.org/10.1136/bmjopen-2025-112695
29. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Palepu, A., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., . . . Natarajan, V. (2024). Towards generalist biomedical AI. *NEJM AI, 1*(3). https://doi.org/10.1056/AIoa2300138
30. Wang, G., Zhang, K., Jiang, J., Wang, C., Bi, H., Liang, H., Qi, Z., Huang, Y., Li, Y., & Yang, X. (2026). Human–large language model collaboration in clinical medicine: A systematic review and meta-analysis. *npj Digital Medicine, 9*(1), 195. https://doi.org/10.1038/s41746-026-02382-2
31. You, J. G., Dbouk, R. H., Landman, A., Ting, D. Y., Dutta, S., Wang, J. C., Centi, A. J., Macfarlane, M., Bechor, E., Letourneau, J., Choo-Kang, G., Kim, E. H., Magee, C., Lang, B. J., Angelo, L., Olin, J., Frits, M., Iannaccone, C., Rui, A., . . . Mishuris, R. G. (2025). Ambient documentation technology in clinician experience of documentation burden and burnout. *JAMA Network Open, 8*(8), e2528056. https://doi.org/10.1001/jamanetworkopen.2025.28056