



SAFETY AND SECURITY OF GENERATIVE AI IN HEALTHCARE: A RISK-AWARE FRAMEWORK FOR CLINICAL LLM DEPLOYMENT

Dr. Siddhant Maske, Dr. Supriya khatke

Department of Computer Science, Government Holkar Science College, Indore

ABSTRACT

Generative artificial intelligence (GenAI), particularly large language models (LLMs), is rapidly transforming healthcare by supporting clinical documentation, medical information retrieval, patient communication, clinical decision support, medical education, research, and administrative workflows. Despite these opportunities, the deployment of LLMs in healthcare introduces significant safety and security risks because clinical decisions directly affect patient well-being. Unlike conventional software systems, LLMs can generate plausible but incorrect information, exhibit hallucinations, reproduce biases, expose sensitive information, respond unpredictably to adversarial prompts, and provide outputs that may be difficult for clinicians to verify. This paper proposes a **Risk-Aware Clinical LLM Deployment Framework (RAC-LDF)** that integrates clinical safety, cybersecurity, privacy protection, evidence verification, human oversight, model monitoring, and governance into the lifecycle of generative AI systems. A conceptual literature-based methodology is used to identify major risk categories and map corresponding mitigation mechanisms. The proposed framework contains six major layers: data and privacy protection, model assurance, clinical evidence grounding, secure inference, human oversight, and continuous governance. Recent evidence demonstrates that LLMs can achieve strong performance on medical knowledge tasks while still producing clinically inappropriate or potentially harmful outputs, highlighting the difference between benchmark performance and safe clinical deployment. Current guidance from the World Health Organization, National Institute of Standards and Technology, and medical-device regulators further emphasizes human autonomy, transparency, accountability, safety, equity, cybersecurity, and lifecycle monitoring. The paper concludes that clinical LLM deployment should follow a risk-stratified approach in which autonomy is proportional to clinical risk. High-risk applications should require stronger validation, human review, auditability, and continuous post-deployment monitoring. The proposed framework provides a practical foundation for healthcare organizations seeking to adopt generative AI while reducing patient-safety, privacy, cybersecurity, and governance risks.

Keywords: Generative AI, Large Language Models, Healthcare AI, Clinical Safety, Cybersecurity, Patient Privacy, Risk Management, Artificial Intelligence Governance, Clinical Decision Support, Responsible AI

Received: 27/08/2026

Revised: 22/09/2026

Accepted: 12/10/2026

INTRODUCTION

Artificial intelligence has become an increasingly important component of modern healthcare. Machine-learning systems have been applied to medical imaging, disease prediction, clinical decision support, drug discovery, patient monitoring, and health-system management. The emergence of generative artificial intelligence has expanded these applications by enabling systems to generate natural-language responses, summarize clinical information, retrieve medical knowledge, produce documentation, and interact directly with healthcare professionals and patients.

Large language models represent an important category of generative AI. Models such as Med-PaLM and other healthcare-oriented LLMs have demonstrated substantial capabilities in medical question answering and clinical knowledge tasks. Singhal et al. (2023), for example, demonstrated that medically adapted language models could achieve strong performance across several medical question-answering benchmarks. However, their study also



showed that high benchmark performance did not eliminate clinically important problems involving factuality, harmful content, bias, and missing information.

This distinction is critical. In healthcare, an incorrect response is not merely an information-quality problem. It may influence diagnosis, treatment, medication use, triage, or patient behavior. Consequently, clinical LLM deployment requires considerably stronger safety mechanisms than ordinary consumer applications.

Recent research has expanded the understanding of these risks. A 2026 *Nature* review systematically mapped LLM safety and security hazards across design, data, model, inference, and clinical-environment stages, emphasizing that safety cannot be treated as a single model-level property.

Healthcare LLMs also create cybersecurity challenges. Potential risks include data leakage, prompt injection, malicious manipulation of retrieved information, model exploitation, insecure application programming interfaces, and attacks against connected clinical systems. A 2024 review specifically identified privacy risks, data poisoning, bias propagation, hallucinations, and cybersecurity threats as important concerns for generative AI in healthcare.

The World Health Organization (WHO) similarly emphasizes that AI used in health should protect human autonomy, promote safety, ensure transparency, foster accountability, support equity, and remain responsive and sustainable. Therefore, the central question addressed in this paper is:

How can healthcare organizations deploy clinical LLMs while systematically managing patient-safety, privacy, cybersecurity, reliability, explainability, and governance risks?

To address this question, this study proposes a **Risk-Aware Clinical LLM Deployment Framework (RAC-LDF)**.

LITERATURE REVIEW

2.1 Generative AI in Healthcare

Generative AI has potential applications across clinical, operational, educational, and research settings. These include:

- clinical documentation;
- patient communication;
- medical summarization;
- clinical decision support;
- medical education;
- drug discovery;
- research assistance;
- administrative automation.

WHO guidance on large multimodal models identifies healthcare, scientific research, public health, and drug development as important application areas while emphasizing the need for governance and risk mitigation.

However, the flexibility of generative AI also creates uncertainty because a general-purpose model may be used for tasks that were not explicitly anticipated during development.

2.2 Clinical Accuracy and Hallucination

One of the most important safety concerns associated with LLMs is hallucination: the generation of information that is incorrect, unsupported, or fabricated but presented in a convincing manner.

Singhal et al. (2023) showed that medically adapted LLMs could substantially improve clinical question-answering performance. Nevertheless, human evaluation revealed remaining limitations related to scientific alignment, harmful content, bias, and missing information.



Research involving LLM analysis of unstructured electronic medical records has also examined accuracy, consistency, and hallucination, demonstrating why clinical validation cannot rely exclusively on automated metrics (Shah, 2024). A 2026 systematic review of hallucination mitigation strategies identified approaches including retrieval-augmented generation, knowledge graphs, self-reflection, specialized evaluation metrics, human-in-the-loop methods, specialized training, and red teaming.

2.3 Privacy and Confidentiality

Healthcare data are highly sensitive. Clinical LLM applications may process:

- electronic health records;
- laboratory results;
- medical images;
- genomic information;
- medication histories;
- patient communications;
- insurance information.

Sending such information to external AI systems may create privacy and confidentiality risks.

Recent research has identified privacy attacks and data leakage through LLM-based healthcare applications and emphasized the tension between privacy protection and model utility.

Potential mitigation strategies include:

- data minimization;
- de-identification;
- encryption;
- access control;
- local/private deployment;
- privacy-preserving machine learning;
- secure API management;
- audit logging.

2.4 Cybersecurity Risks

Clinical LLM systems may introduce a new attack surface into healthcare information systems.

Important threats include:

Prompt Injection

An attacker may place malicious instructions inside user input or retrieved documents.

Data Poisoning

Manipulated training or retrieval data may influence model behavior.

Unauthorized Data Access

Poorly controlled integrations may expose protected health information.

Tool Misuse

An LLM connected to external tools may perform unauthorized actions.

Model Manipulation

Attackers may attempt to influence system behavior through adversarial inputs.

Supply-Chain Risks

Third-party models, APIs, libraries, plugins, and datasets may introduce vulnerabilities.

These risks demonstrate that clinical LLM security must extend beyond the language model itself.

2.5 Bias and Health Equity



AI systems may reproduce biases contained in training data. If an LLM performs differently across demographic or socioeconomic groups, unequal healthcare outcomes may result.

WHO emphasizes inclusiveness and equity as fundamental principles for healthcare AI.

Therefore, clinical LLM evaluation should consider:

- demographic variation;
- language variation;
- socioeconomic differences;
- geographic context;
- health-system differences;
- accessibility.

2.6 Explainability and Accountability

Clinicians need to understand the basis and limitations of AI-generated recommendations.

A system that provides:

"The patient should receive treatment X."

without evidence or uncertainty information is difficult to evaluate.

A safer system should provide:

- evidence;
- source information;
- confidence/uncertainty;
- relevant clinical factors;
- contraindications;
- alternative considerations;
- clear indication that the output is AI-generated.

WHO explicitly identifies transparency, explainability, intelligibility, responsibility, and accountability as core principles for AI in health.

RESEARCH METHODOLOGY / MATERIALS AND METHODS

3.1 Research Design

This study uses a **conceptual, literature-based research methodology**.

The research synthesizes peer-reviewed studies, international health guidance, AI risk-management frameworks, and regulatory materials to identify major risks associated with clinical LLM deployment.

The methodology consists of four stages:

1. Identification of clinical LLM risks.
2. Classification of risks by healthcare AI lifecycle stage.
3. Mapping of risks to mitigation mechanisms.
4. Development of a risk-aware clinical deployment framework.

3.2 Literature and Guidance Sources

The conceptual framework was informed by research and guidance from:

- peer-reviewed biomedical journals;
- *Nature* and *npj Digital Medicine*;
- World Health Organization;
- National Institute of Standards and Technology;
- U.S. Food and Drug Administration;
- International Medical Device Regulators Forum;
- clinical AI safety literature.



NIST's Generative AI Profile extends the AI Risk Management Framework specifically to risks associated with generative AI and supports risk management throughout the AI lifecycle.

3.3 Risk Classification

The identified risks were classified into six major categories:

Risk Category	Examples	Potential Consequence
Clinical Safety	Hallucination, incorrect advice	Patient harm
Privacy	Data leakage, unauthorized access	Confidentiality breach
Cybersecurity	Prompt injection, poisoning	System compromise
Fairness	Bias, unequal performance	Health inequity
Governance	Lack of accountability	Regulatory risk
Operational	Model drift, integration failure	Workflow disruption

PROPOSED RISK-AWARE CLINICAL LLM DEPLOYMENT FRAMEWORK

The proposed **RAC-LDF** contains six integrated layers.

4.1 Layer 1: Data and Privacy Protection

This layer protects patient information before it reaches the LLM.

Key mechanisms include:

- data minimization;
- de-identification;
- encryption;
- role-based access control;
- secure data storage;
- consent management;
- privacy auditing.

The system should process only information necessary for the intended clinical task.

4.2 Layer 2: Model Assurance

The model should be evaluated before deployment.

Evaluation should include:

- clinical accuracy;
- factuality;
- hallucination rate;
- bias;
- robustness;
- calibration;
- reliability;
- harmful-output testing.

A model should not be deployed merely because it performs well on general language benchmarks.

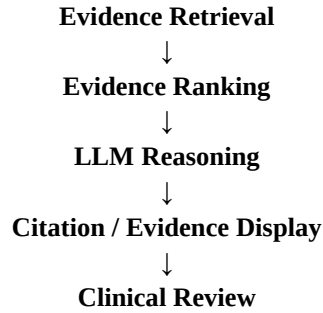
4.3 Layer 3: Evidence-Grounded Clinical Reasoning

The proposed framework recommends retrieval-augmented generation for clinical applications where current evidence is required.

The workflow is:

Clinical Query





This can reduce dependence on potentially outdated internal model knowledge.

4.4 Layer 4: Secure Inference

Every clinical interaction should pass through a security layer.

The security layer should detect:

- malicious prompts;
- suspicious inputs;
- unauthorized requests;
- sensitive information leakage;
- unsafe tool calls;
- anomalous behavior.

External tools should operate under least-privilege access.

4.5 Layer 5: Human Oversight

Human oversight should depend on clinical risk.

Low-Risk Applications

Examples:

- administrative summarization;
- formatting;
- appointment communication.

AI may operate with limited supervision.

Moderate-Risk Applications

Examples:

- clinical documentation;
- literature synthesis;
- preliminary decision support.

Clinician review should be incorporated into workflow.

High-Risk Applications

Examples:

- treatment recommendations;
- medication decisions;
- diagnosis-related support;
- emergency triage.

Human clinical authorization should be required before consequential action.

4.6 Layer 6: Continuous Governance

Clinical LLMs should not be considered permanently validated after deployment.

Continuous monitoring should examine:



- model performance;
- hallucination;
- safety incidents;
- cybersecurity incidents;
- demographic performance;
- model drift;
- user feedback;
- unexpected behavior.

The FDA and IMDRF's Good Machine Learning Practice principles emphasize lifecycle-oriented development and maintenance of AI/ML-enabled medical technologies.

RESULTS / FINDINGS

Because this study is a conceptual framework study rather than a prospective clinical trial, the results are presented as **literature-derived findings** rather than fabricated numerical experimental outcomes.

5.1 Finding 1: Accuracy Alone Is Insufficient

Evidence indicates that LLMs can perform strongly on medical knowledge benchmarks while still producing clinically problematic outputs.

Singhal et al. (2023) demonstrated this difference through human evaluation of medically oriented language models. Their work supports the need for multidimensional evaluation rather than accuracy-only assessment.

5.2 Finding 2: Hallucination Is a Major Clinical Safety Risk

Clinical hallucinations may involve:

- fabricated medical facts;
- incorrect treatment recommendations;
- inaccurate patient summaries;
- fabricated references;
- incorrect interpretation of clinical information.

Recent systematic evidence indicates that RAG, knowledge graphs, human oversight, self-reflection, and red-teaming can form complementary hallucination-mitigation strategies.

5.3 Finding 3: Security Must Be Integrated Across the AI Lifecycle

The recent *Nature* review of clinical LLM safety and security categorizes hazards across design, data, model, inference, and environment stages.

This supports the proposed framework's lifecycle-oriented approach.

5.4 Finding 4: Privacy and Utility Require Balance

Strong privacy controls can reduce the amount of information available to an AI system and potentially reduce clinical usefulness. Conversely, unrestricted access to clinical data can increase privacy risks.

Therefore, healthcare organizations require controlled data-access mechanisms rather than simply choosing between "maximum privacy" and "maximum performance."

5.5 Finding 5: Human Oversight Remains Essential

WHO guidance emphasizes that humans should remain in control of healthcare systems and medical decisions.

This supports risk-based human oversight rather than unrestricted clinical autonomy.

DISCUSSION



The findings indicate that safe clinical LLM deployment requires a shift from **model-centric evaluation** toward **system-level risk management**.

A model can be highly capable while the overall healthcare application remains unsafe.

For example, even a clinically competent model may become unsafe if:

- incorrect patient data are supplied;
- the retrieval database contains outdated information;
- a malicious prompt manipulates the model;
- the model has excessive access to hospital systems;
- clinicians misunderstand the model's confidence;
- monitoring is absent after deployment.

Therefore:

Clinical AI Safety $\neq$ Model Accuracy

Instead:

Clinical AI Safety = Model Quality + Data Quality + Security + Human Oversight + Governance

This distinction is central to the proposed framework.

6.1 Risk-Aware Deployment

The appropriate degree of autonomy should depend on the consequences of an incorrect decision.

A simple conceptual risk score can be expressed as:

$$R = P(H) \times S(H) \times ER = P(H) \times S(H) \times E$$

where:

- $P(H)$ = probability of harmful outcome;
- $S(H)$ = severity of harm;
- EE = exposure or frequency of system use.

Higher-risk applications should receive stronger validation and human control.

6.2 Clinical Evidence Grounding

LLMs should not be treated as authoritative sources of clinical truth.

Instead, evidence-grounded systems should retrieve current information from approved clinical knowledge sources and clearly distinguish:

Retrieved Evidence

from

Model-Generated Interpretation

This distinction improves transparency and allows clinicians to independently inspect important claims.

6.3 Security as a Clinical Safety Issue

Cybersecurity and clinical safety should not be treated as separate domains.

For example, a successful attack that changes retrieved clinical information could indirectly cause a clinical error.

Therefore:

Cybersecurity failure $\rightarrow$ Information integrity failure $\rightarrow$ Clinical safety risk

This means security teams and clinical governance teams should collaborate during system design and monitoring.

RECOMMENDATIONS / IMPLICATIONS

7.1 For Healthcare Organizations

Healthcare institutions should establish a formal AI governance committee involving:

- clinicians;



- pharmacists;
- nurses;
- data scientists;
- cybersecurity professionals;
- legal experts;
- ethicists;
- hospital administrators.

AI systems should undergo documented risk assessment before deployment.

7.2 For AI Developers

Developers should prioritize:

1. evidence-grounded generation;
2. uncertainty estimation;
3. robust hallucination testing;
4. adversarial testing;
5. privacy protection;
6. secure API design;
7. audit logs;
8. model cards and system documentation.

7.3 For Clinicians

Clinicians should receive training regarding:

- model limitations;
- hallucinations;
- verification procedures;
- appropriate use;
- data privacy;
- escalation mechanisms.

AI outputs should be interpreted as decision support rather than automatically accepted clinical truth.

7.4 For Regulators

Regulatory frameworks should emphasize the entire lifecycle of clinical AI.

Evaluation should cover:

Development → Validation → Deployment → Monitoring → Updating → Retirement

The FDA's current digital-health guidance ecosystem includes clinical decision-support and cybersecurity-related materials, illustrating the movement toward lifecycle-oriented oversight.

7.5 Proposed Clinical AI Safety Checklist

Before deployment, healthcare organizations should verify:

Checkpoint	Required
Clinical validation	Yes
Privacy assessment	Yes
Cybersecurity assessment	Yes
Bias assessment	Yes
Hallucination testing	Yes
Human oversight	Yes



Checkpoint	Required
Audit logging	Yes
Incident response	Yes
Model monitoring	Yes
User training	Yes

LIMITATIONS

This study has several limitations.

First, the proposed framework is conceptual and has not been evaluated through a prospective clinical deployment.

Second, the rapid development of generative AI means that model capabilities and vulnerabilities can change quickly.

Third, healthcare environments differ considerably across countries, hospitals, specialties, and regulatory systems.

Therefore, a single risk threshold may not be appropriate for all applications.

Fourth, some security risks may emerge only after integration with hospital information systems and external tools.

Fifth, human oversight itself can fail if clinicians become excessively dependent on AI recommendations or do not have sufficient time to verify outputs.

Finally, this paper does not claim that the proposed framework eliminates all clinical AI risks. Instead, it provides a structured approach for identifying, reducing, monitoring, and governing those risks.

FUTURE RESEARCH

Future studies should empirically evaluate the proposed framework using real-world or carefully controlled clinical scenarios.

9.1 Clinical LLM Safety Benchmarks

Researchers should develop standardized benchmarks measuring:

- factuality;
- hallucination;
- clinical harm;
- bias;
- robustness;
- privacy;
- cybersecurity.

9.2 Real-Time Risk Detection

Future systems could automatically identify when an AI response has unusually high uncertainty or potential clinical risk.

9.3 Privacy-Preserving Clinical LLMs

Research should explore:

- federated learning;
- differential privacy;
- secure local inference;
- confidential computing.

9.4 Adversarial Testing



Clinical LLMs should undergo systematic red-team testing for:

- prompt injection;
- data poisoning;
- jailbreaks;
- information leakage;
- unsafe tool use.

9.5 Multimodal Clinical AI

Future research should evaluate systems combining:

- text;
- medical imaging;
- laboratory results;
- genomic data;
- electronic health records.

9.6 Long-Term Monitoring

Longitudinal studies are needed to determine whether clinical performance changes after model updates, new data, changing clinical guidelines, and changes in patient populations.

CONCLUSION

Generative AI and large language models have substantial potential to improve healthcare delivery, clinical documentation, medical education, research, and decision support. However, clinical deployment presents risks that are qualitatively different from those associated with ordinary consumer-facing AI.

The central challenge is not simply whether an LLM can produce medically knowledgeable responses. The more important question is whether the **complete AI-enabled healthcare system can operate safely, securely, transparently, and accountably within clinical workflows.**

This paper proposed the **Risk-Aware Clinical LLM Deployment Framework (RAC-LDF)**, consisting of six interconnected components: data and privacy protection, model assurance, evidence-grounded reasoning, secure inference, human oversight, and continuous governance.

The framework recognizes hallucination, privacy leakage, cybersecurity attacks, bias, inadequate explainability, model drift, and excessive automation as interconnected risks rather than isolated technical problems.

Recent evidence supports this system-level perspective. Clinical LLM research has demonstrated both substantial capabilities and persistent safety limitations, while contemporary WHO, NIST, FDA, and international regulatory guidance increasingly emphasizes lifecycle risk management, human oversight, transparency, safety, and accountability.

The recommended deployment strategy is therefore **risk-proportional clinical AI**. Low-risk administrative applications may permit greater automation, whereas high-risk clinical applications should require stronger validation, evidence grounding, clinician review, access controls, auditability, and continuous monitoring.

Ultimately, responsible clinical GenAI deployment should not aim for unrestricted automation. Its goal should be to create **safe human-AI collaboration in which generative AI enhances healthcare professionals' capabilities while preserving patient safety, privacy, clinical accountability, and human control.**



REFERENCES

1. Clusmann, J., Freyer, O., Ostermann, M., Ferber, D., Ghaffari Laleh, N., Hilgers, L., Kolbinger, F. R., Schneider, C. V., Downing, A., Wekenborg, M. K., Gilbert, S., Foersch, S., Truhn, D., Wiest, I. C., & Kather, J. N. (2026). Safety and security of large language models in healthcare. *Nature*, 656, 577–589. <https://doi.org/10.1038/s41586-026-10687-1>
2. European Commission. (2019). *Ethics guidelines for trustworthy AI*. High-Level Expert Group on Artificial Intelligence.
3. Food and Drug Administration. (2025). *Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations*. U.S. Department of Health and Human Services.
4. International Medical Device Regulators Forum. (2025). *Good machine learning practice for medical device development: Guiding principles*. IMDRF.
5. Ning, Y., Zhang, Z., Wang, H., et al. (2024). Generative artificial intelligence and ethical considerations in health care: A scoping review and ethics checklist. *The Lancet Digital Health*, 6(11), e848–e856. [https://doi.org/10.1016/S2589-7500\(24\)00143-2](https://doi.org/10.1016/S2589-7500(24)00143-2)
6. National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
7. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile*. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
8. Shah, S. V. (2024). Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8), e2425953. <https://doi.org/10.1001/jamanetworkopen.2024.25953>
9. 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, 172–180. <https://doi.org/10.1038/s41586-023-06291-2>
10. Teo, Z. L., Quek, C. W. N., Wong, J. L. Y., & Ting, D. S. W. (2024). Cybersecurity in the generative artificial intelligence era. *Asia-Pacific Journal of Ophthalmology*, 13(4), 100091. <https://doi.org/10.1016/j.apjo.2024.100091>
11. World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization. <https://doi.org/10.1016/j.speech.2021.05.001>
12. World Health Organization. (2023). *WHO calls for safe and ethical AI for health*. World Health Organization.
13. World Health Organization. (2025). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. World Health Organization. <https://www.who.int/publications/i/item/9789240084759>
14. World Health Organization. (2024). *Artificial intelligence for health: Supporting countries to deploy responsible AI technologies to accelerate equitable health for all*. World Health Organization.
15. Zhou, Q., et al. (2026). Recent advances in defending the privacy attacks of large language models for healthcare applications: A concise review. [Review article].