Safety, Bias, and Hallucination Mitigation in Generative Artificial Intelligence Clinical Applications: A Systematic Review of Current Evidence

Josh Khorsandi1, Brian Mansoury1, Justin Kahen1, Aria Damavandi1, Michael Kahen1, Moez Khorsandi2
1Kirk Kirkorian School of Medicine at UNLV
²Clinical Professor of Surgery, Western University of Health Sciences

KIRK KERKORIAN SCHOOL OF MEDICINE | UNI

Introduction

Generative Artificial Intelligence (GenAI) models are rapidly transitioning from prototype tools to active participants in clinical workflows. Despite their promise in documentation, triage, and diagnostic reasoning, concerns regarding hallucinations, biased outputs, and unreliable reasoning remain a major barrier to safe clinical adoption. This systematic review synthesizes current evidence on safety risks associated with GenAI in medicine and evaluates the effectiveness of emerging mitigation strategies

Objective

To evaluate safety risks of generative AI in clinical settings and assess the effectiveness of strategies to mitigate hallucinations and bias.

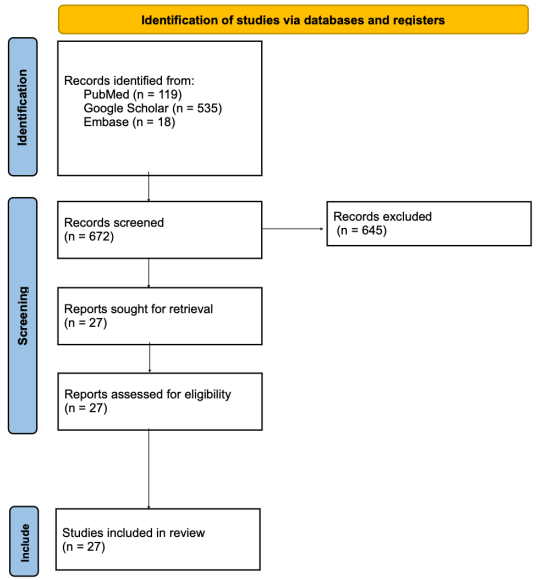
Methods

A systematic search of PubMed, Scopus, Web of Science, and IEEE Xplore was conducted for studies published between January 2015 and October 2025. In total, 672 records were identified across databases (including PubMed, Google Scholar, and Embase). After screening titles and abstracts, 645 records were excluded, and 27 reports were sought for full-text retrieval. All 27 reports were assessed for eligibility and subsequently included in the final review. Eligible studies empirically evaluated the clinical use, safety, or mitigation strategies for hallucinations or biased outputs in generative AI (GenAI) models. Two independent reviewers conducted screening, data extraction, and methodological quality assessment in accordance with PRISMA guidelines. Identified mitigation strategies were categorized into model-level, workflow-level, and human-in-the-loop interventions.

Methods cont.

Inclusion and Exclusion Criteria

Inclusion	Exclusion
•Studies on generative AI in clinical/healthcare settings •Evaluation of safety risks (hallucinations, bias) •Assessment of mitigation strategies •Empirical studies (with data/results)	•Non-clinical or non-healthcare AI applications •No evaluation of safety or bias •No mitigation strategies discussed •Editorials, opinions, or purely theoretical papers

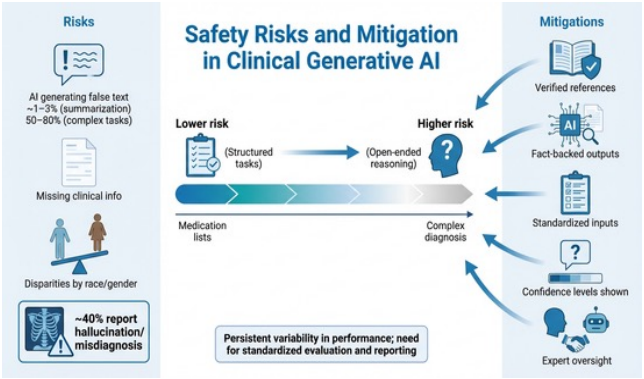


Results

Bias has been documented across domains; approximately 40% of radiology-reporting studies reported hallucinations or misdiagnosis, with broader disparities linked to race and gender.

Results cont.

Effective mitigation strategies include grounding outputs in verified clinical sources, retrieval-augmented generation, structured prompting, explicit uncertainty expression, and human-in-the-loop review to reduce error propagation.



Future Perspectives

Standardized Safety Evaluation
Develop consistent frameworks to measure hallucinations, bias, and clinical reliability across GenAI applications.

Transparent Uncertainty & Oversight
Improve model transparency through uncertainty reporting and stronger human-in-the-loop review.

Safer Real-World Integration
Advance scalable mitigation strategies, including retrieval grounding and structured prompting, for routine clinical use.

Conclusions

Current evidence demonstrates that hallucinations and biased outputs remain substantial risks in GenAI-supported clinical practice, but several emerging mitigation approaches show promise for improving reliability. Standardized evaluation frameworks and transparent reporting of model uncertainty are urgently needed to support safe, equitable, and scalable integration of GenAI into healthcare.