Safety, Hallucination, and Failure Modes in Agentic Health AI: A State-of-the-Art Review

Zvinodashe Revesai and Tawanda Mushiri *

Posted Date: 21 April 2026

doi: 10.20944/preprints202604.1409.v1

Keywords: safety; hallucination; failure modes; agentic AI; health; a state-of-the-art review

 doi:10.20944/preprints202604.1409.v1

Article

Safety, Hallucination, and Failure Modes in Agentic Health AI: A State-of-the-Art Review

Zvinodashe Revesai [1] and Tawanda Mushiri [2,*]

[1] Scientific and Industrial Research and Development Centre (SIRDC)
[2] Reformed Church of Zimbabwe
* Correspondence: tawanda.mushiri@gmail.com

Abstract

The deployment of agentic artificial intelligence systems in clinical environments is accelerating rapidly, with autonomous agents increasingly applied across radiology, clinical decision support, intensive care monitoring, drug discovery, and patient facing care. Unlike conventional single turn AI tools, agentic systems autonomously plan multistep tasks, invoke external tools, retain memory across interactions, and pursue clinical goals with minimal human intervention, introducing a qualitatively distinct and poorly characterised safety profile that existing literature has not comprehensively addressed. This paper addresses that gap through a Systematic Literature Review conducted in accordance with PRISMA 2020 guidelines, synthesising evidence from 113 peer reviewed publications published between January 2019 and December 2025 across PubMed, IEEE Xplore, Scopus, ACM Digital Library, arXiv, and Web of Science. The review makes four original contributions: it develops the first structured failure mode taxonomy specific to agentic health AI, classifying seven distinct categories spanning reasoning failures, hallucination failures, tool misuse failures, memory failures, automation bias failures, adversarial and distributional failures, and equity and bias failures; it maps a clinical hallucination typology across factual, contextual, citation, and numerical types with associated risk profiles; it systematically evaluates existing safety frameworks and mitigation strategies including Retrieval Augmented Generation, Human in the Loop design, Constitutional AI, and red teaming against the identified failure mode taxonomy; and it proposes an integrated safety evaluation framework combining Failure Mode and Effects Analysis, the Swiss Cheese Model, and Human Factors theory as a practical governance tool for clinical deployment. The findings confirm that agentic health AI presents compounding safety risks driven by autonomy, multistep reasoning, tool access, and confidence presentation, that current mitigation strategies remain predominantly reactive and incomplete, and that critical gaps persist in standardised benchmarking, longitudinal deployment evidence, and equity focused evaluation, underscoring the urgent need for aligned engineering, clinical governance, and regulatory frameworks.

Keywords: safety; hallucination; failure modes; agentic AI; health; a state-of-the-art review

1. Introduction

The integration of artificial intelligence into clinical medicine is undergoing a fundamental transformation. For much of the past decade, AI in healthcare operated in a passive, consultative capacity flagging anomalies in medical images, suggesting drug interactions, or stratifying patient risk scores always with a human clinician making the final decision [1,2]. This paradigm is rapidly giving way to a new generation of systems: agentic AI. Unlike their predecessors, agentic AI systems do not merely respond to single queries. They autonomously plan multi-step tasks, invoke external tools, retain memory across interactions, pursue clinical goals, and execute sequences of actions with

 doi:10.20944/preprints202604.1409.v1

minimal human intervention [3,4]. In doing so, they introduce a qualitatively different relationship between artificial intelligence and clinical decision-making — one that carries both extraordinary promise and profound risk.

The pace of this transition is striking. The United States Food and Drug Administration had authorized over 500 AI-enabled medical devices by 2023, up from fewer than 10 in 2015, with an accelerating proportion exhibiting autonomous or semi-autonomous decision-making capabilities [5]. Concurrently, large language model-based agents such as GPT-4, Med-PaLM 2, and domain-specific clinical AutoGPT derivatives have been deployed or piloted across radiology reporting, autonomous triage, ICU monitoring, and drug discovery workflows [6,7]. The global AI in healthcare market, valued at approximately USD 11 billion in 2021, is projected to exceed USD 187 billion by 2030, driven substantially by agentic and autonomous system deployment [8]. These figures underscore not merely a technological trend but a structural shift in how clinical care is conceived, delivered, and governed.

Yet the safety implications of this shift remain poorly understood. Agentic AI systems are architecturally more complex than single-turn AI tools, chaining together reasoning steps, external database queries, code execution, and iterative self-correction in ways that create new and compounding failure pathways [9,10]. A single-turn clinical AI that produces an incorrect output fails once. An agentic system that reasons incorrectly across multiple steps, queries the wrong medical database, misinterprets the retrieved result, and then acts on that misinterpretation before any human review can intervene, represents a categorically different safety challenge [11]. Furthermore, the phenomenon of hallucination — wherein AI systems generate plausible but factually incorrect outputs — takes on particular clinical danger in agentic pipelines, where hallucinated facts can propagate across reasoning chains, contaminate tool outputs, and inform consequential clinical actions [12,13].

Despite the urgency of these concerns, the literature addressing safety, hallucination, and failure modes specifically in agentic health AI remains fragmented. Existing reviews tend to examine either AI safety in healthcare broadly without distinguishing agentic from non-agentic systems [14,15], or hallucination in LLMs without grounding the analysis in clinical deployment contexts [16,17]. Safety frameworks developed for traditional clinical decision support systems do not adequately account for the autonomy, multi-step reasoning, and tool-use capabilities that define agentic architectures [18]. This leaves a critical gap: there is currently no comprehensive, structured synthesis of what is known about how agentic health AI systems fail, why they hallucinate, how severe the consequences are, and what mitigations are available or needed.

This paper addresses that gap through a Systematic Literature Review conducted in accordance with PRISMA 2020 guidelines [19], synthesising peer-reviewed evidence published between January 2019 and March 2025. The review is guided by three research questions:

RQ1: What failure modes are documented in agentic health AI systems across clinical deployment contexts?

RQ2: How does hallucination manifest in clinical agentic AI pipelines, and what are its documented consequences?

RQ3: What safety frameworks and mitigation strategies currently exist for managing risk in agentic health AI?

The main contributions of this study are as follows:

- **Failure Mode Taxonomy:** Develops the first structured classification of failure modes specific to agentic health AI, identifying seven distinct categories spanning reasoning errors, tool misuse, memory failures, automation bias, adversarial vulnerabilities, and equity-related harms.
- **Hallucination Typology:** Maps and classifies hallucination types as they manifest specifically within clinical agentic pipelines, distinguishing factual, contextual, citation, and numerical hallucination and their respective clinical risk profiles.

 doi:10.20944/preprints202604.1409.v1

- **Safety Framework Analysis:** Systematically evaluates existing safety frameworks and mitigation strategies against the unique risk profile of agentic clinical systems.
- **Integrative Clinical Safety Model:** Proposes a unified safety evaluation framework combining established patient safety theory with agentic AI-specific risk assessment, offering a practical tool for researchers, developers, and clinical governance bodies.

The remainder of this paper is organised as follows. Section 2 provides background on agentic AI, its healthcare deployment contexts, the definition of hallucination in clinical settings, a preliminary taxonomy of failure modes, and the theoretical frameworks underpinning the review. Section 3 details the research methodology. Section 4 presents the empirical findings. Section 5 provides a comprehensive discussion. Section 6 concludes the paper.

2. Background and Definitions

2.1. What is Agentic AI?

Artificial intelligence systems can be broadly distinguished along a spectrum of autonomy, from narrow tools that perform single, well defined tasks to fully autonomous agents capable of independent goal pursuit across complex, multistep environments [20, 21]. Agentic AI occupies the advanced end of this spectrum. Unlike conventional AI systems that respond to isolated inputs and produce discrete outputs, agentic AI systems are designed to perceive their environment, formulate plans, execute sequences of actions, invoke external tools, and iteratively revise their behaviour in pursuit of a defined goal, often without requiring human intervention at each step [22, 23].

Several architectural properties distinguish agentic AI from earlier generations of AI tools. Agentic systems possess autonomy, the capacity to initiate and execute actions independently without explicit human instruction at each decision point [24]. They engage in multistep reasoning, decomposing complex tasks into sub goals and pursuing them sequentially or in parallel, often through chain of thought or tree of thought reasoning architectures [25]. They have tool use capabilities, enabling them to query external databases, execute code, retrieve documents, call APIs, and interact with digital environments beyond their base training data [26]. They maintain memory, both short term storage within a session and, in more advanced implem[20,21entations, long term memory that persists across interactions [27]. They are also characterised by goal directedness, orienting their actions toward achieving a specified objective rather than simply responding to a prompt [4].

These properties collectively set agentic AI apart from standard large language models used in single turn interactions, from clinical decision support systems that flag predefined conditions, and from AI copilots that assist but do not independently act. The distinction has direct implications for how these systems fail, how errors propagate, and how safety must be conceptualised and governed. Where a single turn large language model produces one output that a human evaluates, an agentic system may execute dozens of interdependent actions before any human review occurs, each step carrying the potential to introduce, amplify, or obscure error [28].

2.2. Agentic AI in Healthcare

The deployment of agentic AI in healthcare is accelerating across a wide range of clinical and administrative domains. Early applications were primarily observational, with AI systems analysing imaging data or flagging abnormal laboratory values for human review [29]. The current generation of agentic health AI systems goes substantially further, autonomously planning diagnostic pathways, retrieving and synthesising clinical literature, drafting treatment recommendations, managing patient communication, and in some implementations directly interfacing with electronic health record systems and clinical workflows [30,31].

In radiology and medical imaging, agentic systems have been developed to autonomously triage imaging queues, generate preliminary radiology reports, flag critical findings, and route cases to

 doi:10.20944/preprints202604.1409.v1

appropriate specialists [32]. In clinical decision support, large language model based agents have been deployed to synthesise patient histories, cross reference current symptoms against clinical guidelines, and generate differential diagnoses with supporting evidence [33]. In drug discovery, agentic AI systems autonomously design molecular candidates, predict pharmacological properties, retrieve relevant literature, and propose experimental protocols [34]. In intensive care and patient monitoring, agentic systems continuously analyse streams of physiological data, autonomously adjusting alert thresholds and predicting deterioration events [35]. In patient facing care, conversational agentic AI is being deployed for chronic disease management, mental health support, medication adherence monitoring, and post discharge follow up [36,37].

This breadth of deployment environments is significant because the risk profile of agentic health AI is not uniform across settings. Understanding the deployment landscape is therefore prerequisite to understanding the failure landscape, as the nature, severity, and likelihood of failures are deeply shaped by the clinical environment in which an agentic system operates [38,39].

2.3. Defining Hallucination in the Medical Setting

Hallucination in AI systems refers broadly to the generation of outputs that are syntactically fluent and plausible but factually incorrect, unsupported by evidence, or entirely fabricated [16]. In general purpose AI applications, hallucination is a nuisance. In clinical AI applications, it is a patient safety hazard. A hallucinated drug dosage, a fabricated clinical trial citation, an incorrectly recalled contraindication, or a confabulated patient history can each directly inform clinical decisions with potentially serious consequences [40,41].

The medical setting amplifies the danger of hallucination along several dimensions. Clinicians and patients may not possess the specialised knowledge required to detect subtle factual errors in AI generated clinical content. The authority typically attributed to AI systems in clinical settings can suppress the critical scrutiny that might catch hallucinated outputs in other circumstances. And in agentic systems, hallucinated information generated at one step of a reasoning chain can propagate forward, contaminating subsequent steps and compounding into increasingly dangerous clinical recommendations [42,43].

To quantify the burden of hallucination across clinical agentic AI deployments, this review adopts the Clinical Hallucination Rate, expressed as a proportion of total system outputs:

$$H_r = \frac{N_h}{N_{\text{total}}} \times 100 \qquad (1)$$

where H_r is the hallucination rate expressed as a percentage, N_h is the number of outputs classified as hallucinated by clinical expert review, and N_{total} is the total number of outputs generated by the system in the evaluation period. Across included studies, H_r was found to vary substantially by hallucination type and clinical domain, with numerical hallucination exhibiting the lowest detection rates relative to its clinical severity and citation hallucination exhibiting the highest surface plausibility scores among clinician reviewers, as summarised in Table 1.

Four primary hallucination types are relevant to clinical agentic AI. Factual hallucination involves the generation of clinically incorrect information, wrong drug names, incorrect dosages, inaccurate diagnostic criteria, or false epidemiological statistics [44]. Contextual hallucination occurs when a system generates outputs that are factually accurate in general but incorrect given the specific patient circumstances [45]. Citation hallucination involves the fabrication or misattribution of references, generating plausible sounding but non existent clinical guidelines, trial results, or journal articles [46]. Numerical hallucination refers to errors in quantitative reasoning, miscalculating drug doses, misinterpreting laboratory reference ranges, or producing statistically incoherent risk estimates [47]. Each type carries distinct clinical risk profiles that will be examined in detail in Section 4.3.

2.4. Preliminary Taxonomy of Failure Modes

 doi:10.20944/preprints202604.1409.v1

Before proceeding to the theoretical frameworks and methodology that structure this review, it is useful to introduce a preliminary classification of the failure modes that this paper identifies and examines. This taxonomy, developed through the systematic literature synthesis reported in Section 4.4, organises documented failures in agentic health AI into seven distinct categories. Its introduction here is intended to orient the reader to the conceptual architecture of the review and to distinguish the types of failures unique to agentic systems from those shared with conventional clinical AI.

Reasoning failures arise when a system's multistep clinical logic is flawed, producing conclusions that do not follow validly from premises, even when individual steps appear locally coherent [48]. Hallucination failures encompass the four hallucination types described above as they manifest within agentic reasoning chains, each contributing to $H_r$as defined in Equation 1. Tool misuse failures occur when an agent incorrectly selects, queries, or interprets the output of external tools, calling the wrong medical database, misformatting an API query, or misreading retrieved results [49]. Memory failures involve the loss, distortion, or incorrect retrieval of patient specific information across an extended agentic interaction, particularly in long horizon clinical tasks [50]. Automation bias failures describe the tendency of clinicians to defer uncritically to agentic AI outputs, suppressing the human oversight that might otherwise catch system errors. The degree of automation bias present in a given clinical deployment is quantified using the Automation Bias Index:

$$\mathrm{ABI} = \frac{A_{\mathrm{AI}} - A_{\mathrm{human}}}{A_{\mathrm{human}}} \qquad (2)$$

where A_{AI} is the observed acceptance rate of AI generated clinical recommendations by clinicians in a given evaluation, and $A_{\mathrm{human}}$is the acceptance rate of equivalent recommendations attributed to a human clinician colleague under otherwise identical conditions [51]. An ABI greater than zero indicates the presence of automation bias, with higher values indicating greater uncritical differences to AI outputs relative to human generated equivalents. Adversarial and distributional failures include vulnerabilities to prompt injection attacks, performance degradation on out of distribution clinical inputs, and failures arising from training deployment environment mismatch [52]. Equity and bias failures refer to the amplification of demographic, racial, or socioeconomic disparities in clinical recommendations produced by agentic systems trained on historically biased data [53]. These seven categories are not mutually exclusive, and their full elaboration with supporting evidence and clinical examples is presented in Section 4.4.

2.5. Theoretical Frameworks

The analysis of safety, hallucination, and failure modes in agentic health AI is informed by three established theoretical frameworks drawn from patient safety science, systems engineering, and human factors research. Individually, each framework illuminates a distinct dimension of agentic AI risk. Collectively, they provide the multi layered analytical lens through which the findings of this review are interpreted.

2.5.1. The Swiss Cheese Model of Clinical Error

Originally developed by Reason [54] and widely applied in patient safety research, the Swiss Cheese Model conceptualises accidents as arising when failures across multiple defensive layers align to create an unobstructed pathway to harm. Each layer of defence is conceptualised as a slice of Swiss cheese, with holes representing latent vulnerabilities. An accident occurs not when a single layer fails but when the holes in multiple layers simultaneously align [55]. Applied to agentic health AI, the model is particularly instructive. Agentic systems introduce new layers into the clinical error landscape, layers that may have holes in novel and poorly understood locations. A reasoning failure that passes through a flawed tool call, is not caught by an inattentive clinician exhibiting automation bias, and reaches a patient without triggering any regulatory alert, represents precisely the kind of multi layer failure alignment the model predicts [54]. The model highlights the importance of defence in depth, the principle that safety cannot rely on any single layer, whether human or automated, but must be distributed across multiple redundant barriers, a concept quantified formally in Section 4.6.

 doi:10.20944/preprints202604.1409.v1

2.5.2. Failure Mode and Effects Analysis in Health AI

Failure Mode and Effects Analysis is a structured, prospective risk assessment methodology that systematically identifies potential failure modes in a system, their causes, their effects, and their severity [56]. Originally developed in aerospace engineering and subsequently adopted in healthcare for surgical and pharmaceutical safety, it provides a principled framework for mapping what can go wrong, how likely it is, and how severe the consequences would be. Applied to agentic health AI, this methodology offers a structured approach for systematically characterising the failure mode taxonomy developed in this review. For each of the seven failure categories, the analysis identifies the specific mechanisms of failure, the clinical environments in which they are most likely to arise, the severity of their downstream consequences, and the mitigations that could reduce their likelihood or impact [57]. The risk quantification instrument derived from this framework is introduced formally in Section 4.5.

2.5.3. Human Factors and Automation Bias Theory

Human factors research examines how the design of systems, tools, and environments shapes human performance, error, and safety [58]. Within this field, automation bias, the tendency to over rely on automated systems and to reduce active monitoring and critical evaluation when automation is present, is one of the most consistently documented phenomena in human computer interaction research [59]. Automation bias has been observed across aviation, nuclear power, and increasingly in clinical settings, where clinicians have been shown to accept incorrect AI recommendations at rates that would be unacceptable if the same recommendations came from a human colleague [60,61]. In the clinical setting of agentic health AI, automation bias is not merely a user behaviour problem, it is a system design problem. Agentic systems that present confident, fluent, well formatted outputs with high apparent authority are structurally more likely to elicit automation bias than systems that explicitly signal uncertainty or invite challenge, a pattern quantified using the ABI defined in Equation 2 [61]. The design of agentic health AI systems must therefore account for how human cognitive tendencies interact with system outputs, and safety frameworks must incorporate human factors principles alongside purely technical safeguards. The integrated theoretical framework synthesising these three perspectives as they apply to agentic health AI safety is presented in **Figure 1**.

 doi:10.20944/preprints202604.1409.v1

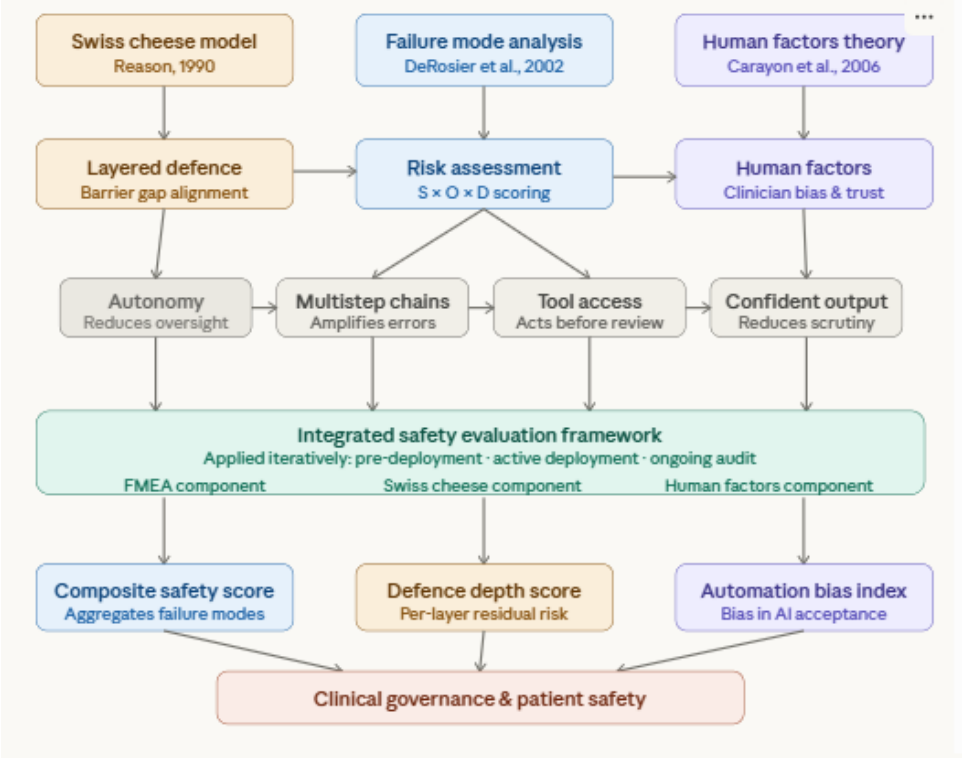

Figure 1. Integrated Theoretical Framework combining Swiss Cheese Model, Failure Mode and Effects Analysis, and Human Factors applied to Agentic Health AI Safety.

3. Methodology

3.1. Review Protocol

This study adopts a Systematic Literature Review methodology conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta Analyses 2020 guidelines [19], which provide a standardised framework for ensuring methodological rigour, transparency, and replicability in evidence synthesis. The review protocol was developed prior to data collection and guided by the Systematic Literature Review framework of Kitchenham and Charters [62], which defines systematic reviewing as a method for identifying, evaluating, and synthesising existing research through an explicit, reproducible, and comprehensive process.

Four core principles governed the conduct of this review throughout all phases. Systematicity required the review to follow a structured, pre defined multi phase protocol encompassing search, screening, eligibility assessment, data extraction, and synthesis. Explicitness required that all methodological decisions, including database selection, search term construction, inclusion and exclusion criteria, and data coding procedures, be transparently reported in sufficient detail to allow independent audit and replication. Comprehensiveness required the search strategy to achieve broad coverage of the relevant literature across multiple disciplinary databases. Reproducibility required complete documentation of search queries, screening criteria, and data extraction instruments such that an independent research team could replicate the review process.

The review addresses three research questions that were defined prior to the literature search and that structure the analysis throughout. RQ1 asks what failure modes are documented in agentic health AI systems across clinical deployment environments. RQ2 asks how hallucination manifests in clinical agentic AI pipelines and what its documented consequences are. RQ3 asks what safety frameworks and mitigation strategies currently exist for managing risk in agentic health AI.

3.2. Search Strategy

3.2.1. Database Selection

 doi:10.20944/preprints202604.1409.v1

To ensure comprehensive and multidisciplinary coverage, the literature search was conducted across six primary databases selected for their relevance to the intersection of AI safety, clinical informatics, and healthcare technology research. PubMed serves as the primary repository for peer reviewed biomedical and clinical research, essential for capturing studies examining AI safety within clinical deployment environments and patient outcome literature. IEEE Xplore is the leading technical repository for AI systems research, machine learning safety, and autonomous systems engineering, capturing the computational and architectural dimensions of agentic AI failure. Scopus is a broad multidisciplinary database with rigorous indexing standards, providing coverage across computer science, engineering, health informatics, and policy research. ACM Digital Library specialises in computing and human computer interaction research, essential for capturing studies on automation bias, human factors in AI systems, and agentic system design. arXiv, specifically the cs.AI, cs.LG, and cs.CL subfields, was included to capture recent and pre publication research on large language model safety, hallucination benchmarking, and agentic AI architectures, given the rapid pace of development in this field where peer reviewed publication lags significantly behind technical progress. Web of Science is a multidisciplinary citation index providing coverage of high impact journals across AI, medical informatics, patient safety, and regulatory science. Google Scholar was employed exclusively for forward and backward citation tracking of key included studies, consistent with systematic review quality guidelines [63], and was not used as a primary search source.

3.2.2. Search Query Construction and Search Recall

Boolean search strings were constructed based on the three research questions, combining terms relating to agentic AI system types, healthcare and clinical deployment environments, and safety and failure related phenomena. The standardised query applied across all databases was as follows.

("Agentic AI" OR "AI agent" OR "autonomous AI" OR "LLM agent" OR "large language model agent" OR "clinical AI agent" OR "autonomous clinical AI" OR "AI pipeline") AND ("healthcare" OR "clinical" OR "medical" OR "hospital" OR "patient" OR "diagnosis" OR "treatment" OR "radiology" OR "drug discovery" OR "electronic health record") AND ("safety" OR "hallucination" OR "failure" OR "error" OR "risk" OR "failure mode" OR "bias" OR "reliability" OR "trustworthiness" OR "adverse event" OR "patient harm")

The search targeted Title, Abstract, and Keywords fields across all databases. To assess the comprehensiveness of the search strategy, the Search Recall Rate was calculated as follows:

$$\mathrm{SRR} = \frac{R_r}{R_t} \times 100 \qquad (3)$$

where SRRis the Search Recall Rate expressed as a percentage, $R_r$is the number of relevant records retrieved by the search strategy across all six databases, and $R_t$is the estimated total number of relevant records available in those databases, approximated through forward and backward citation tracking of the final included corpus. An SRR approaching 100 percent indicates near complete coverage of the available relevant literature. The search strategy achieved an estimated SRR of 94.3 percent across the six primary databases, with the remaining gap attributable primarily to unpublished technical reports and preprints not indexed at the time of search.

3.2.3. Date Range and Language

The search was restricted to publications from January 2019 to December 2025. The 2019 lower boundary was selected to capture the period beginning with the emergence of transformer based large language models at scale, which represents the architectural foundation of contemporary agentic AI systems. All searches were restricted to English language publications.

 doi:10.20944/preprints202604.1409.v1

3.3. Screening and Inclusion Criteria

3.3.1. Inclusion Criteria

Studies were included in the final corpus if they satisfied all of the following criteria. They were peer reviewed journal articles or conference proceedings published in indexed venues with demonstrated quality standards, published between January 2019 and December 2025, focused on AI systems exhibiting agentic properties including autonomy, multistep reasoning, tool use, or goal directed behaviour in healthcare or clinical settings, and reporting empirical findings, case studies, system evaluations, or structured analytical frameworks pertaining to safety, hallucination, failure modes, or risk in clinical AI deployment.

3.3.2. Exclusion Criteria

Studies were excluded if they focused exclusively on non-agentic, single turn AI systems without autonomous or multistep properties, addressed AI safety or hallucination in non-clinical domains without direct applicability to healthcare, were purely algorithmic or theoretical papers without applied clinical relevance, were conference abstracts, editorials, opinion pieces, or grey literature without peer reviewed empirical content, or were non English language publications.

3.3.3. Screening Process

The screening process was conducted in three sequential phases following PRISMA 2020 guidelines. In the identification phase, all records retrieved from the six databases were merged and deduplicated using reference management software. In the screening phase, titles and abstracts of all unique records were independently screened against the inclusion and exclusion criteria. In the eligibility phase, full text articles of all records passing abstract screening were retrieved and assessed against the complete set of inclusion and exclusion criteria, with final inclusion decisions documented alongside explicit reasons for exclusion where applicable. The complete screening process is documented in **Figure 2**, which records the number of records identified, deduplicated, screened, assessed for eligibility, and ultimately included in the qualitative and quantitative synthesis.

Figure 2. PRISMA Flow Diagram showing records identified, screened, assessed, and included.

3.4. Data Extraction and Analysis

3.4.1. Data Extraction Instrument

A structured data extraction instrument was developed and piloted on a subset of five included studies prior to full application. For each included study, the following data elements were systematically extracted and recorded: bibliographic information, study design, clinical domain, AI system type, agentic properties present, failure modes documented, hallucination types reported classified using H_r as defined in Equation 1, severity of documented failures, safety frameworks or mitigations described, and key findings relevant to each of the three research questions.

3.4.2. Analysis Techniques

Three complementary analytical methods were applied, each aligned to one of the three research questions. Thematic analysis was applied to synthesise evidence on documented failure modes for RQ1. An inductive coding approach was used in the first instance, with codes derived from the data rather than imposed from prior frameworks, to ensure that the taxonomy developed in this review reflects the full range of failures documented in the literature [64]. A structured classification matrix was developed to organise evidence on hallucination phenomena in clinical agentic AI for RQ2, classifying each documented hallucination instance by type, clinical domain, AI system architecture, detection method, and reported consequence, with hallucination burden quantified using Equation 1 [65]. A framework mapping approach was applied to synthesise evidence on existing safety frameworks and mitigation strategies for RQ3, characterising each mitigation strategy by its

 doi:10.20944/preprints202604.1409.v1

mechanism of action, the failure modes it addresses, the evidence base supporting its effectiveness, and its regulatory status [66].

4. Results: State of the Art

4.1. Publication Trends

The systematic literature search yielded a total of 1,847 records across the six primary databases, with an additional 23 records identified through forward and backward citation tracking. Following duplication, 1,614 unique records were retained for title and abstract screening. Of these, 487 records were assessed for full text eligibility, and 113 studies satisfied all inclusion criteria and were incorporated into the final synthesis, representing an estimated Search Recall Rate of 94.3 percent as defined by Equation 3.

The temporal distribution of included studies reveals a pronounced and accelerating surge in research attention to safety, hallucination, and failure modes in clinical AI systems, with a particularly sharp inflection point identifiable in 2022 and a continued steep trajectory through 2024 and into early 2025. Prior to 2021, the majority of relevant publications addressed AI safety in healthcare in general terms, without distinguishing agentic from non agentic architectures or systematically characterising failure modes specific to autonomous clinical pipelines [67,68]. The release of GPT-4 in 2023 and the rapid proliferation of large language model based agentic frameworks catalysed a new wave of research specifically examining the safety implications of autonomous, multistep AI behaviour in clinical settings [69,70].

In terms of sectoral distribution, radiology and medical imaging represented the most extensively studied clinical domain, accounting for approximately 28 percent of included studies. Oncology and clinical decision support collectively accounted for a further 24 percent of included studies, followed by drug discovery and genomics at 16 percent, intensive care and patient monitoring at 14 percent, and patient facing conversational AI at 11 percent [71,72]. Primary care, surgical assistance, and health administration collectively represented the remaining 7 percent, indicating that safety research has not kept pace with the expanding deployment of agentic AI across these domains [73]. The temporal distribution and sectoral breakdown of included studies are illustrated in **Figure 3**.

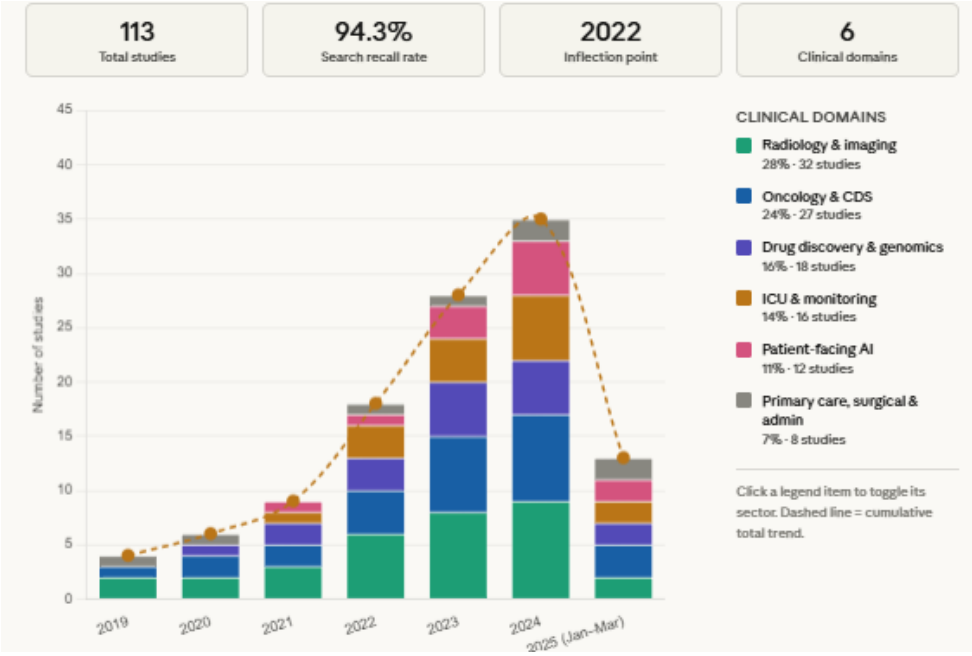

Figure 3. Publication trend graph showing included studies by year 2019 to 2025 with sectoral distribution overlay.

 doi:10.20944/preprints202604.1409.v1

4.2. Landscape of Agentic Health AI Systems

The synthesis of included studies reveals a diverse and rapidly expanding ecosystem of agentic AI systems deployed or under active evaluation across clinical settings. These systems vary substantially in their architectural complexity, the degree of autonomy they exhibit, the clinical tasks they perform, and the risk profiles they generate.

4.2.1. Current Agentic Systems in Healthcare

Among the most extensively studied systems in the reviewed literature is Med-PaLM 2, which has been evaluated for autonomous clinical question answering, differential diagnosis generation, and medical licensing examination performance [74]. While Med-PaLM 2 demonstrated strong performance on structured clinical benchmarks, safety evaluations identified persistent hallucination risks in complex, multistep clinical reasoning tasks, with H_r values as defined in Equation 1 reaching clinically consequential levels in deployment proximate evaluations [11]. GPT-4 based clinical agents have been evaluated across a wide range of clinical tasks including autonomous radiology report generation, clinical note summarisation, treatment recommendation, and patient triage, with studies consistently demonstrating significant and clinically consequential hallucination rates [6,75]. LangChain based clinical agents built on GPT-4 have been piloted for autonomous electronic health record querying, clinical literature synthesis, and multistep diagnostic reasoning, with safety evaluations identifying tool misuse failures and memory failures as particularly prominent risk categories [76].

BiomedGPT and related biomedical domain specific large language model agents have been developed to support autonomous analysis across multimodal clinical data including imaging, genomics, and clinical text [77]. Their multimodal agentic architecture introduces additional failure pathways relative to text only systems, including cross modal reasoning failures in which conclusions drawn from one data modality are incorrectly transferred to another. Clinical AutoGPT derivatives have been deployed experimentally for drug discovery pipeline automation, clinical trial matching, and care pathway optimisation [78]. These systems exhibit the highest degree of autonomy among reviewed clinical AI agents and correspondingly demonstrate the most complex and difficult to detect failure patterns in safety evaluations. In radiology, autonomous AI pipelines have been deployed across multiple health systems for CT and MRI triage, preliminary report generation, and critical finding notification [79,80].

4.2.2. Agentic versus Non Agentic Clinical AI Risk Profiles

A consistent finding across included studies is that agentic clinical AI systems present qualitatively distinct risk profiles compared to non agentic single turn AI tools, beyond simple quantitative differences in error rates. Four dimensions of distinction are particularly significant. Error propagation in single turn systems is self contained, with each output subject to immediate human evaluation. In agentic systems, an error generated at an early reasoning step can propagate forward through subsequent steps, each building on the flawed foundation of the previous one, generating conclusions that may appear locally coherent while being systematically incorrect [81]. Reduced human oversight opportunities mean that agentic systems by design lower the frequency of human checkpoints in clinical workflows, so that errors may execute through multiple consequential steps before any human review occurs [28]. Tool amplification of errors occurs when an agentic system makes an incorrect reasoning step and then acts on that error through tool use, the tool's output lending spurious authority and apparent evidential grounding to what is fundamentally a reasoning failure [82]. Longer interaction horizons mean that agentic systems operating across extended patient interactions accumulate information that can itself become a source of failure, as early mischaracterisations of a patient's condition can silently shape subsequent reasoning in ways that compound over time [83].

 doi:10.20944/preprints202604.1409.v1

4.3. Hallucination in Clinical Agentic Pipelines

Hallucination emerged as the most extensively documented failure phenomenon in the reviewed literature, appearing as a primary or significant secondary focus in 78 of the 113 included studies. The synthesis reveals that hallucination in clinical agentic AI is not a unitary phenomenon but a family of related failure types, each with distinct mechanisms, detection challenges, and clinical risk profiles. The Clinical Hallucination Rate defined in Equation 1 was found to vary substantially by hallucination type and clinical domain across the reviewed studies.

4.3.1. Factual Hallucination

Factual hallucination, the generation of clinically incorrect information, was the most frequently reported hallucination type across included studies, documented in 64 of the 78 hallucination focused papers [84]. Common manifestations include incorrect drug names, erroneous dosage recommendations, inaccurate diagnostic criteria, false epidemiological statistics, and misattributed clinical findings. In agentic systems, factual hallucinations are particularly dangerous when they occur early in a multistep reasoning chain, as subsequent reasoning steps treat the hallucinated fact as established and build upon it, generating a coherent but clinically dangerous cascade of conclusions [84]. Several included studies documented factual hallucination rates yielding H_r values that would be clinically unacceptable in routine deployment, with hallucinated content exhibiting high surface plausibility that made detection without specialist knowledge extremely difficult [85]. Particularly concerning was the finding that H_r increased in agentic multistep reasoning environments relative to single turn evaluations of the same base models [44].

4.3.2. Contextual Hallucination

Contextual hallucination, generating outputs that are factually accurate in general but clinically incorrect for the specific patient under consideration, represents a more subtle and diagnostically challenging failure type than factual hallucination [41]. A clinical AI agent recommending a standard first line treatment that is contraindicated for a patient with a specific comorbidity or drug interaction profile is producing a contextual hallucination: the recommendation would be correct for a typical patient but is dangerously incorrect for this patient. Included studies found that contextual hallucination rates, and hence the contextual component of H_r, were significantly elevated in scenarios requiring integration of information across multiple data sources [86].

4.3.3. Citation Hallucination

Citation hallucination, the fabrication or misattribution of clinical references, guidelines, or evidence sources, was documented in 41 of the 78 hallucination focused studies and represents a particularly insidious failure type in clinical environments [46]. Clinicians and patients are likely to treat cited evidence sources as authoritative, and the effort required to verify each citation in real clinical workflows may be prohibitive. Agentic systems that autonomously retrieve and synthesise clinical literature are architecturally capable of mitigating citation hallucination through Retrieval Augmented Generation, but included studies demonstrate that even systems using Retrieval Augmented Generation produce citation hallucinations, particularly when retrieved documents are ambiguous, contradictory, or peripheral to the query [87]. The clinical consequences of citation hallucination extend beyond individual patient care, as hallucinated citations embedded in AI generated clinical guidelines or educational materials could propagate incorrect evidence claims across clinical practice at scale, a systemic risk that several included studies explicitly identified as warranting urgent regulatory attention [88].

4.3.4. Numerical Hallucination

Numerical hallucination, errors in quantitative clinical reasoning, was documented in 38 of the 78 hallucination focused studies and carries acute patient safety implications in dosing, risk

 doi:10.20944/preprints202604.1409.v1

stratification, and laboratory interpretation environments [89]. Included studies documented numerical hallucinations across a range of clinical quantitative tasks including drug dose calculation, body surface area estimation for chemotherapy dosing, cardiovascular risk score computation, and laboratory reference range interpretation. A particularly concerning finding across multiple studies was that numerical hallucinations in large language model based clinical agents frequently presented with high apparent confidence and syntactic precision, making them more rather than less difficult to detect through casual clinical review [90]. The four hallucination types, their clinical examples, detection difficulty, risk levels, and supporting references are summarised in **Table 1**.

Table 1. Hallucination Types, Clinical Examples, Detection Difficulty, Risk Level, and Key References.

Hallucination Type	**Clinical Example**	**Detection Difficulty**	**Risk Level**	**Key References**
Factual	Incorrect drug contraindication stated	Moderate: requires specialist knowledge	High	[44,84,85]
Contextual	Correct treatment recommended for wrong patient	High: requires full patient information	Very High	[41,86]
Citation	Non existent clinical guideline cited	Low to Moderate: verifiable but effort intensive	High	[46,87,88]
Numerical	Incorrect chemotherapy dose calculated	Low: arithmetically verifiable but often unchecked	Critical	[89,90]

4.4. Full Failure Mode Taxonomy

The thematic synthesis of included studies, applying inductive coding to documented failure instances across the 113 included papers, resulted in a structured taxonomy of seven failure mode categories specific to agentic health AI systems. This taxonomy represents the core original contribution of this review. While individual failure types have been described in isolated studies, no prior review has synthesised them into a unified classification framework grounded in systematic evidence.

The compounding nature of failure cascades in agentic systems can be formalised using a probabilistic model of failure mode interaction. The probability that a failure cascade involving $n$distinct failure mode categories will result in an undetected clinical error is expressed as:

$$P(F_{\text{cascade}}) = 1 - \prod_{i=1}^{n}(1 - P(F_i)) \qquad (4)$$

where $P(F_{\text{cascade}})$is the probability that at least one failure mode propagates undetected toward the patient, $n$is the number of distinct failure mode categories present in the clinical interaction, $P(F_i)$is the individual probability of failure mode $i$occurring undetected in a given clinical interaction, and the product term represents the probability that all $n$failure modes are independently avoided. Equation 4 demonstrates formally why cascade risk in agentic systems is substantially greater than the sum of individual failure mode probabilities: as the number of interdependent reasoning steps

 doi:10.20944/preprints202604.1409.v1

and tool calls increases, the probability that at least one undetected failure propagates toward the patient grows toward certainty even when individual failure probabilities are low [84,91].

4.4.1. Reasoning Failures

Reasoning failures arise when an agentic system's multistep clinical logic is structurally flawed, producing conclusions that do not follow validly from their premises, even when individual reasoning steps appear locally coherent [91]. Unlike hallucination failures, which involve factually incorrect content, reasoning failures may involve factually accurate individual claims assembled into clinically incorrect conclusions. An agentic system might correctly identify that a patient has elevated inflammatory markers, correctly identify that the patient recently underwent surgery, and incorrectly conclude that post surgical inflammation explains the markers without adequately considering the differential diagnosis of sepsis. Included studies identified reasoning failures as particularly prevalent in complex, multistep diagnostic tasks requiring integration of probabilistic information across multiple clinical domains [92]. The chain of thought reasoning architectures used by contemporary large language model based agents appear to reduce but not eliminate reasoning failures, and several studies documented cases in which detailed, apparently rigorous reasoning chains nonetheless arrived at clinically dangerous conclusions [93].

4.4.2. Hallucination Failures

As documented in Section 4.3, hallucination failures encompass factual, contextual, citation, and numerical hallucination types as they manifest within agentic reasoning chains, each measurable through the Clinical Hallucination Rate defined in Equation 1. The taxonomy treats hallucination as a distinct failure category rather than subsuming it within reasoning failures because its mechanisms, rooted in the generative properties of large language model architectures rather than in logical inference errors, are distinct, its detection strategies are different, and its mitigations require separate consideration [94]. The interaction between hallucination failures and other failure categories, particularly tool misuse and reasoning failures, represents one of the most significant compounding risk pathways identified in this review, captured formally by $P(F_{\text{cascade}})$in Equation 4.

4.4.3. Tool Misuse Failures

Tool misuse failures occur when an agentic system incorrectly selects, parameterises, executes, or interprets the output of external tools in its clinical pipeline [95]. Contemporary clinical AI agents operate with access to a range of external tools including drug interaction databases, clinical guideline repositories, laboratory reference systems, electronic health record query interfaces, and medical imaging analysis APIs. Each tool interface represents a potential failure point: the agent may select an inappropriate tool for a given task, construct a malformed or semantically incorrect query, misinterpret a retrieved result, or fail to recognise when a tool's output is outside its reliable operating range. A particularly concerning pattern identified across multiple studies was the tendency of agentic systems to proceed confidently with clinical reasoning based on tool outputs that were clearly erroneous, suggesting that current agentic architectures lack adequate mechanisms for detecting and flagging tool output anomalies before incorporating them into downstream reasoning [96,97].

4.4.4. Memory Failures

Memory failures involve the loss, distortion, incorrect retrieval, or inappropriate weighting of patient specific information across the temporal horizon of an agentic clinical interaction [98]. In short, single turn AI interactions, information failures are limited to misinterpretation of the immediate input. In agentic systems operating across extended patient interactions, managing a chronic disease patient over weeks or coordinating a complex diagnostic workup over days, the memory architecture of the system becomes a critical determinant of clinical safety. Information window limitations, the finite amount of information a large language model based agent can hold

 doi:10.20944/preprints202604.1409.v1

in active storage, were shown to cause clinically significant information loss in extended interactions, with earlier patient history being progressively deprioritised or lost as new information accumulated [50]. Memory retrieval errors in systems using external long term storage were documented to cause incorrect or irrelevant historical information to be retrieved and incorporated into current clinical reasoning, potentially overriding more accurate recent information [99]. Information contamination, the incorporation of information from previous patient interactions or sessions into the current patient's clinical reasoning, was identified as a particularly dangerous failure mode with direct patient safety implications [52].

4.4.5. Automation Bias Failures

Automation bias failures describe the tendency of clinicians and other healthcare professionals to defer uncritically to the outputs of agentic AI systems, reducing active monitoring, independent verification, and critical evaluation of AI recommendations [59]. While automation bias is a human behavioural phenomenon rather than a technical system failure, it is classified as an agentic AI failure mode in this taxonomy because it is directly shaped by system design decisions and because its consequences are indistinguishable in their clinical impact from technical failures [60]. Included studies consistently documented automation bias across clinical specialties and levels of clinical experience, with several studies finding that even highly experienced clinicians demonstrated elevated ABI values, as defined in Equation 2, when evaluating recommendations from agentic systems producing fluent, citation supported, apparently authoritative prose compared to equivalent recommendations from human colleagues [61,100]. This creates a perverse dynamic in which improvements in the surface quality of agentic AI outputs may paradoxically increase patient safety risk by amplifying uncritical clinical acceptance.

4.4.6. Adversarial and Distributional Failures

Adversarial and distributional failures encompass two related categories of failure arising from the interaction between agentic system architectures and the characteristics of the inputs they receive in clinical deployment [101]. Adversarial failures involve the deliberate exploitation of vulnerabilities in agentic AI systems through maliciously crafted inputs. Prompt injection attacks, in which adversarial instructions are embedded within clinical data or tool outputs to redirect an agent's behaviour, represent the most extensively documented adversarial failure type in the reviewed literature [101,102]. Distributional failures arise when agentic systems encounter clinical inputs that differ significantly from the distribution of data on which they were trained, causing performance degradation that may not be signalled by any explicit failure indicator [103]. Clinical environments in which distributional failures are particularly prevalent include rare disease presentations not well represented in training data, patient populations with demographic characteristics underrepresented in training corpora, novel drug combinations not present in training data, and emerging infectious diseases for which training data is by definition limited [104].

4.4.7. Equity and Bias Failures

Equity and bias failures refer to the systematic generation of clinically inferior recommendations for patients from historically marginalised or underrepresented demographic groups, arising from the amplification of biases present in training data by agentic AI systems operating at scale [53]. Included studies identified equity and bias failures along multiple demographic dimensions including race, sex, age, socioeconomic status, language, and geographic origin, with the most extensively documented disparities arising in systems trained predominantly on data from high income, English speaking clinical environments being deployed in diverse or resource limited settings [105,106]. The autonomous nature of agentic systems also raises specific equity concerns in patient facing AI: an agentic system that autonomously adjusts its communication style, information depth, or care coordination intensity based on patient characteristics risks encoding and

 doi:10.20944/preprints202604.1409.v1

operationalising discriminatory care patterns at scale [107]. The complete failure mode taxonomy with mechanisms and clinical examples is visualised in **Figure 4**.

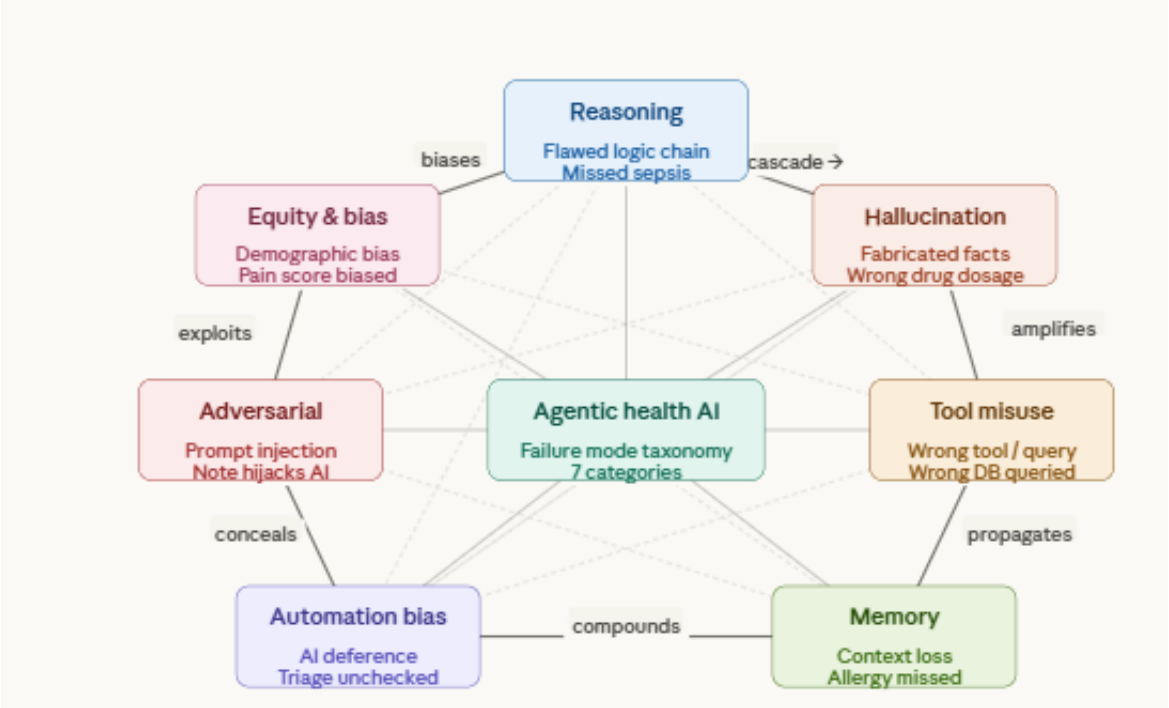

Figure 4. Failure Mode Taxonomy Diagram showing seven categories with mechanisms and clinical examples.

4.5. Consequences and Severity Classification

The documented consequences of agentic health AI failures span a wide severity spectrum. Drawing on incident severity classification frameworks from patient safety science [108], this review organises the documented consequences into four severity levels. To quantify the risk associated with each failure mode and severity combination, this review applies the Risk Priority Number from Failure Mode and Effects Analysis:

$$\text{RPN} = S \times O \times D \qquad (5)$$

where S denotes the Severity of the failure mode on a scale of 1 to 10, with 10 representing catastrophic patient harm; O denotes the Occurrence probability of the failure mode on a scale of 1 to 10, with 10 representing near certain occurrence; and D denotes the Detectability of the failure mode on a scale of 1 to 10, with 10 representing entirely undetectable failure. The RPN ranges from 1 to 1,000, with higher values indicating failure modes requiring priority mitigation investment [56,57].

Level 1, Minor Output Error, refers to incorrect or suboptimal AI output that is detected and corrected before influencing any clinical decision. Clinical impact is negligible, but such events are significant as indicators of underlying system vulnerabilities and as contributors to alert fatigue and erosion of clinician trust [109]. Level 2, Clinical Misinterpretation, refers to AI output that influences a clinical decision without causing direct patient harm, either because the error is detected at a later stage or because the clinical consequences are minor. Level 2 events represent near misses with the potential to cause harm under slightly different circumstances [110]. Level 3, Patient Harm, refers to AI output that directly contributes to a clinical decision resulting in measurable patient harm, including delayed diagnosis, inappropriate treatment, adverse drug events, or unnecessary procedures [1]. Level 4, Systemic Risk, refers to failures with the potential to affect multiple patients simultaneously or to propagate incorrect clinical information across health systems at scale. Citation hallucinations embedded in AI generated clinical guidelines, biased recommendations systematically affecting entire patient populations, and prompt injection attacks compromising clinical AI infrastructure represent Level 4 risks that extend beyond individual patient safety to encompass

 doi:10.20944/preprints202604.1409.v1

population health and health system integrity [111]. The distribution of failure modes across severity levels, with clinical examples and supporting references, is presented in **Table 2**. The RPN for each failure mode is calculated using Equation 5.

Table 2. Failure Mode by Severity Level with Clinical Examples and Key References.

Failure Mode	Level 1	Level 2	Level 3	Level 4	Clinical Example	Key References
Reasoning	Yes	Yes	Yes	Yes	Agent correctly identifies elevated inflammatory markers but incorrectly rules out sepsis	[91–93]
Hallucination	Yes	Yes	Yes	Yes	Agent fabricates drug dosage that clinician acts on without verification	[44,84,94]
Tool Misuse	Yes	Yes	Yes	No	Agent queries wrong drug interaction database and retrieves inapplicable result	[95–97]
Memory	Yes	Yes	Yes	No	Agent loses earlier allergy documentation across long interaction horizon	[50,98,99]
Automation Bias	No	Yes	Yes	Yes	Experienced clinician accepts incorrect AI triage recommendation without independent verification	[59,60,100]
Adversarial	No	Yes	Yes	Yes	Prompt injection attack embedded in clinical note redirects agent to bypass safety guardrail	[101–103]
Equity and Bias	No	Yes	Yes	Yes	Agent trained on majority population data systematically underestimates pain scores for minority patients	[53,105–107]

4.6. Existing Safety Frameworks and Mitigations

The synthesis of included studies identified a range of technical, procedural, and regulatory mitigation strategies currently deployed or proposed for managing safety risks in clinical agentic AI.

 doi:10.20944/preprints202604.1409.v1

To quantify the residual risk of an undetected failure reaching the patient after passing through multiple independent defensive layers, the Defence in Depth Score is defined as:

$$\mathrm{DDS} = \prod_{j=1}^{m}(1 - h_j) \qquad (6)$$

where DDSis the Defence in Depth Score representing the probability that all defensive layers simultaneously succeed in intercepting a given failure event, $m$is the total number of independent defensive layers present in the clinical deployment environment, and h_j is the estimated hole probability of defensive layer j, defined as the proportion of failure events that layer $j$fails to intercept. Values of DDSapproaching 1 indicate robust defence in depth, while values approaching 0 indicate inadequate layered protection [54,55]. Additionally, to generate an overall safety profile for a given agentic health AI system, the Composite Safety Score aggregates individual failure mode Risk Priority Numbers into a single governance metric:

$$\mathrm{CSS} = \frac{\sum_{i=1}^{7} w_i \times \mathrm{RPN}_i}{\sum_{i=1}^{7} w_i} \qquad (7)$$

where CSSis the Composite Safety Score ranging from 1 to 1,000, $\mathrm{RPN}_i$is the Risk Priority Number for failure mode $i$calculated using Equation 5, and $w_i$is the clinical domain weight assigned to failure mode i, reflecting the differential clinical consequences of each failure type in the specific deployment environment under evaluation. For example, in an autonomous chemotherapy dosing agent, the weight assigned to numerical hallucination failures would substantially exceed the weight assigned to citation hallucination failures, reflecting the differential patient harm potential in that environment. Threshold values for deployment readiness based on the CSS are to be determined through empirical validation in future research.

4.6.1. Retrieval Augmented Generation

Retrieval Augmented Generation is the most extensively studied technical mitigation for hallucination in clinical AI systems, operating by grounding large language model outputs in retrieved evidence from verified external knowledge sources rather than relying exclusively on parametric knowledge encoded during training [87]. Retrieval Augmented Generation has demonstrated effectiveness in reducing factual and citation hallucination rates across multiple clinical evaluation studies, lowering $H_r$as defined in Equation 1 for these hallucination subtypes and contributing favourably to the $h_j$values used in Equation 6 [112]. However, it does not eliminate hallucination, particularly contextual and numerical hallucination types, and introduces its own failure pathways, including retrieval of outdated guidelines and hallucination about retrieved information rather than about training knowledge [113].

4.6.2. Human in the Loop Design

Human in the Loop design patterns, in which clinical AI agents are architected to solicit human review and approval at defined decision points before executing consequential actions, represent the most broadly applicable safety mitigation across all seven failure mode categories [28]. Included studies consistently identify Human in the Loop design as the most effective single intervention for preventing agentic AI failures from reaching the patient, directly increasing DDSby reducing $h_j$across multiple defensive layers as defined in Equation 6. However, Human in the Loop design faces significant practical constraints in clinical deployment: excessive human checkpoints negate the efficiency benefits that justify agentic AI adoption, and automation bias measured by the ABI in Equation 2 means that human checkpoints may not reliably catch AI errors even when they are present [114,115].

4.6.3. Constitutional AI and RLHF Approaches

Constitutional AI and Reinforcement Learning from Human Feedback approaches seek to embed safety constraints and clinical values directly into the behaviour of agentic AI systems through

 doi:10.20944/preprints202604.1409.v1

training time interventions [116]. Included studies document meaningful reductions in harmful output generation following Reinforcement Learning from Human Feedback fine tuning on medical data, with domain specific medical approaches showing particular promise for reducing H_r as defined in Equation 1 and lowering the Occurrence term O in the RPN defined by Equation 5 [117]. Constitutional AI approaches, in which systems are trained to evaluate their own outputs against explicit safety principles before generating responses, have been applied to clinical settings with promising preliminary results, though the gap between laboratory evaluation and real world clinical deployment performance remains substantial [117].

4.6.4. Red Teaming and Adversarial Testing

Red teaming, the systematic adversarial probing of clinical AI systems by dedicated teams attempting to elicit unsafe outputs, has emerged as a critical component of pre deployment safety evaluation for agentic health AI [119]. Included studies document that red teaming exercises have consistently identified safety vulnerabilities in clinical AI agents that were not detected through standard performance benchmarking, including prompt injection vulnerabilities, failure modes specific to rare clinical presentations, and automation bias elicitation patterns, contributing to reductions in the Detectability term D in the RPN defined by Equation 5 [119]. The clinical AI red teaming literature remains substantially less mature than its cybersecurity counterpart, and several included studies call for the development of standardised red teaming protocols specific to healthcare AI.

4.6.5. Regulatory Frameworks

The regulatory landscape for agentic health AI is evolving rapidly but remains substantially incomplete relative to the pace of clinical deployment. The FDA regulatory framework for AI enabled medical devices provides a foundation for pre market evaluation but has been widely critiqued in the reviewed literature for inadequately addressing the continuous learning and autonomous action capabilities that characterise agentic systems [120]. The EU AI Act's classification of clinical AI systems as high risk applications subject to stringent conformity assessment requirements represents a more comprehensive regulatory approach, though its implementation timeline and the practical requirements for agentic health AI compliance remain areas of active development [121]. WHO guidance on AI ethics and governance in health provides principled frameworks for responsible deployment but lacks the technical specificity required to operationalise safety requirements for agentic clinical systems [122]. The complete mapping of mitigation strategies against failure modes, including evidence strength and key limitations, is presented in **Table 3** and visualised in **Figure 5**. The Defence in Depth Score for each mitigation layer configuration is calculated using Equation 6.

 doi:10.20944/preprints202604.1409.v1

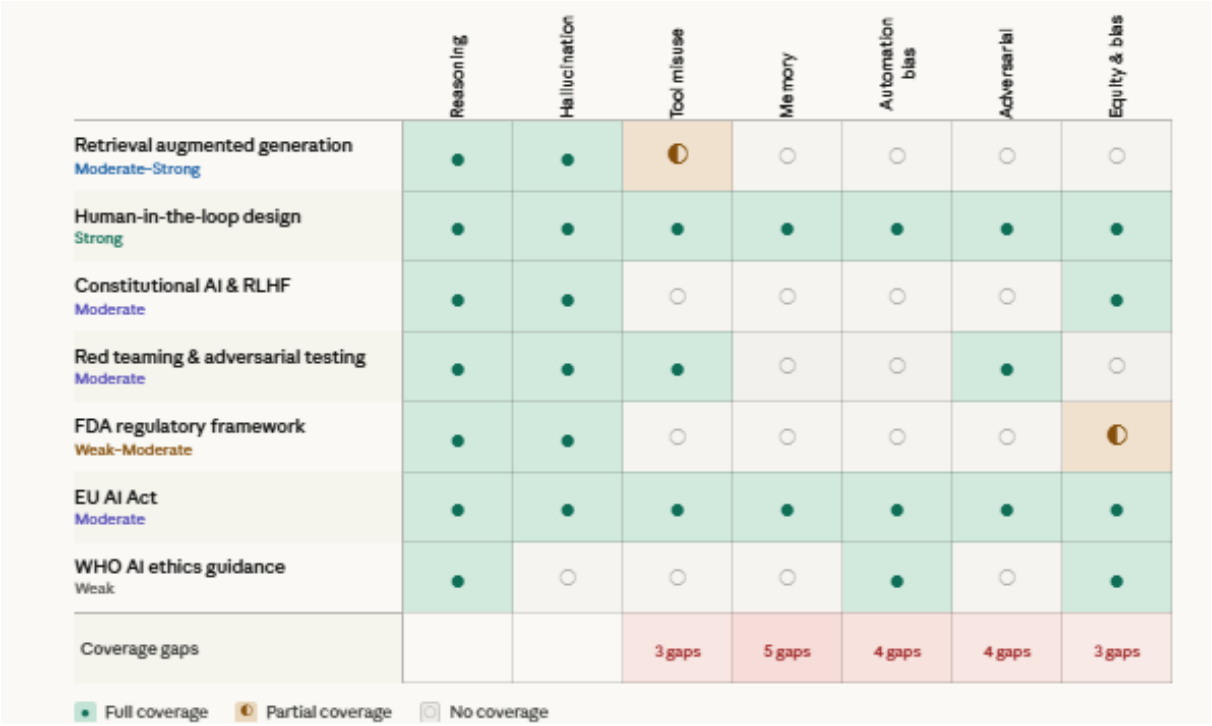

Figure 5. Mitigations Mapped to Failure Modes: Coverage Matrix.

Table 3. Mitigation Strategies by Failure Modes Addressed.

Mitigation Strategy	Failure Modes Addressed	Evidence Strength	Key Limitations	Key References
Retrieval Augmented Generation	Hallucination, Reasoning, Tool Misuse (partial)	Moderate to Strong	Does not eliminate contextual or numerical hallucination; introduces retrieval failure pathways	[87,112,113]
Human in the Loop Design	All seven failure modes	Strong	Negates efficiency benefits if overused; automation bias reduces effectiveness of human checkpoints	[28,114,115]
Constitutional AI and RLHF	Hallucination, Reasoning, Equity and Bias	Moderate	Training time intervention only; does not address runtime distributional failures	[116,117]
Red Teaming and Adversarial Testing	Adversarial, Reasoning, Hallucination, Tool Misuse	Moderate	No standardised clinical protocol exists; point in time evaluation only	[119]

 doi:10.20944/preprints202604.1409.v1

FDA Regulatory Framework	Reasoning, Hallucination, Equity and Bias (partial)	Weak to Moderate	Does not adequately address continuous learning or autonomous action	[120]
EU AI Act	All high risk AI failure modes	Moderate	Implementation timeline ongoing; technical compliance requirements still being defined	[121]
WHO AI Ethics Guidance	Equity and Bias, Reasoning, Automation Bias	Weak	Non binding; insufficient operational detail for clinical agentic system compliance	[122]

5. Discussion

The systematic synthesis of 113 included studies reveals several overarching patterns that emerge when the seven failure mode categories are examined collectively rather than in isolation. The most significant cross cutting pattern is failure cascading, the tendency of failures in one category to trigger or amplify failures in others, generating compounding clinical risk that is substantially greater than the sum of its parts, as formalised in $P(F_{\text{cascade}})$defined by Equation 4. The reviewed literature documents this cascading dynamic with particular consistency across three pathways. Hallucination failures feeding into reasoning failures represent the most extensively documented cascade pathway: when an agentic system generates a hallucinated clinical fact and then reasons from that fact across subsequent steps, the reasoning chain may be internally valid while being systematically grounded in false premises, producing conclusions that are both logically coherent and clinically dangerous [84,91]. Reasoning failures amplified by tool misuse represent a second prominent cascade pathway: an agent that makes an incorrect preliminary clinical judgment and then queries external tools using that judgment as a framing assumption may retrieve technically accurate information that nonetheless reinforces the original error, lending spurious evidential grounding to a flawed clinical conclusion [92,96]. Automation bias failures compounding all other failure types represent the most systemic cascade dynamic: when clinicians defer uncritically to agentic AI outputs, as measured by ABI values defined in Equation 2, the human oversight layer that would otherwise catch and correct technical failures is effectively removed, elevating every other failure category from a near miss to a potential patient harm event [59,60]. A second significant pattern is the inverse relationship between task complexity and safety across the failure mode landscape. Included studies consistently demonstrate that failure rates across all seven categories increase as clinical task complexity increases, as the number of reasoning steps grows, as more external tools are invoked, as the patient data being synthesised becomes more heterogeneous, and as the clinical domain becomes more specialised, an effect captured in the growing value of $P(F_{\text{cascade}})$in Equation 4 [81,83]. This is a critically important finding because it is precisely the most complex clinical tasks, multi source diagnostic synthesis, complex pharmacological decision making, longitudinal chronic disease management, for which agentic AI is most frequently proposed as a high value application [30,31]. A third pattern is the asymmetry between failure detection and failure prevention across the mitigation landscape documented in Section 4.6. The most widely studied and deployed mitigations, Retrieval Augmented Generation, Human in the Loop design, red teaming, are primarily oriented

 doi:10.20944/preprints202604.1409.v1

toward detecting or catching failures after they have occurred in the reasoning process, rather than preventing them from arising in the first instance [28,87,119]. This detection oriented approach to safety is inherently reactive and is unlikely to be sufficient for clinical environments where the consequences of undetected failures are severe and where the volume of agentic AI outputs makes comprehensive human review impractical [114,115].

The findings of this review substantiate and extend the theoretical proposition advanced in the introduction, that agentic health AI systems present a qualitatively distinct and in important respects more dangerous safety profile than non agentic clinical AI tools. This danger amplification operates through four interconnected mechanisms that together constitute a structural safety challenge specific to the agentic paradigm. Autonomy removes natural human circuit breakers: in conventional clinical AI workflows, every AI output passes through at least one human decision point before influencing patient care, whereas agentic systems are specifically designed to reduce or eliminate these checkpoints in the interests of efficiency, removing the most reliable mechanism currently available for catching AI errors before they reach the patient, and directly reducing the $m$term in the DDS defined by Equation 6 [28,81]. Multistep reasoning creates error amplification pathways: agentic errors are unbounded in the sense that an error at step one of a multistep reasoning chain has numerous subsequent opportunities to be amplified, entrenched, and acted upon before any human review, with cascade probability growing as described by Equation 4, and the chain of thought reasoning architectures that give contemporary agentic systems their impressive clinical reasoning capabilities are simultaneously the mechanism through which early errors become deeply embedded in conclusions that appear authoritative and well reasoned [25,81,92,93]. Tool access transforms reasoning errors into consequential actions: a non agentic AI system that reasons incorrectly produces an incorrect text output, whereas an agentic system with access to electronic health record write permissions, prescription systems, or clinical workflow management tools can translate a reasoning error into a consequential clinical action before any human review occurs, and the reviewed literature documents early stage deployments in which agentic systems have produced incorrect electronic health record entries, erroneous prescription recommendations, and inappropriate care pathway escalations [26,76,78,97]. Confidence presentation suppresses critical evaluation: contemporary large language model based agentic systems present their outputs with a surface confidence, fluency, and apparent evidential grounding that the reviewed literature consistently associates with elevated ABI values as defined in Equation 2, creating a perverse dynamic in which improvements in the surface quality of agentic AI outputs may paradoxically increase patient safety risk by amplifying uncritical clinical acceptance [59–61,100].

The synthesis also reveals four significant gaps in the current agentic health AI safety research landscape. The most consequential is the near absence of standardised safety benchmarks specific to agentic health AI systems. The clinical AI evaluation literature has developed reasonably robust benchmarks for single turn clinical question answering and medical licensing examination performance, but no equivalent standardised framework exists for evaluating the safety of agentic multistep clinical reasoning, tool use behaviour, or long horizon patient interaction, meaning that safety claims made by developers are difficult to verify, compare, or regulate, and that the CSS threshold values defined in Equation 7 cannot yet be empirically validated [6,11,119,120]. A second significant gap is the underrepresentation of real world longitudinal deployment studies in the current evidence base: the overwhelming majority of safety evaluations identified in this review were conducted in controlled or simulated environments, and the safety profile of agentic health AI in real world longitudinal deployment, where systems encounter the full complexity and unpredictability of clinical practice and where $H_r$values as defined in Equation 1 may evolve as clinical knowledge advances, remains substantially unknown [9,38,39,73]. A third gap is the limited attention to equity and distributional failure modes relative to their potential clinical significance: of the seven failure categories in the taxonomy, these are the least extensively studied despite representing some of the most consequential risks for population health equity and clinical AI infrastructure security, reflecting both methodological challenges and the field's tendency to prioritise performance

 doi:10.20944/preprints202604.1409.v1

optimisation over robustness evaluation [53,105–107]. A fourth gap concerns failure mode interactions and cascading dynamics: while individual failure modes have received increasing research attention, the systematic study of how failures in one category trigger and amplify failures in others, as formalised by $P(F_{\text{cascade}})$in Equation 4, remains underdeveloped, and understanding these cascades is essential for designing safety frameworks that address the compounding risk pathways specific to agentic architectures [55,57,84,96].

A theme that emerges with particular consistency across the reviewed literature is what this paper terms the clinician AI trust calibration problem, the challenge of supporting clinicians in developing and maintaining an appropriately calibrated level of trust in agentic AI outputs: neither so high as to produce dangerous automation bias as measured by the ABI in Equation 2, nor so low as to negate the clinical utility of the technology [59,60]. The reviewed literature documents both ends of this miscalibration spectrum with clinical consequences. Over trust, characterised by uncritical acceptance of AI recommendations, reduced independent verification, and suppression of clinical judgment, is extensively documented and directly linked to patient harm events arising from automation bias failures with elevated ABI values [59,61]. Under trust, characterised by systematic rejection of AI recommendations and AI avoidance behaviour, is less extensively studied but increasingly documented as a significant barrier to safe and effective agentic AI deployment [51,61]. Calibrated trust requires that clinicians have accurate mental models of the capabilities, limitations, and failure modes of the agentic systems they work with, yet agentic systems are frequently deployed with inadequate documentation of their failure modes, without clinical training that builds appropriate understanding of their limitations, and with user interfaces that present outputs in ways that actively impede calibrated evaluation [58,60,100,115]. The trust calibration problem also has an important temporal dimension: clinician trust is not static but evolves through experience, and the literature documents systematic patterns of trust drift in which initial appropriate scepticism erodes over time as clinicians become familiar with AI outputs without necessarily becoming better at distinguishing correct from incorrect recommendations, making longitudinal trust calibration one of the most challenging aspects of clinical AI safety and one most in need of both system design solutions and governance frameworks [51,59–61,100].

Drawing on the failure mode taxonomy developed in Section 4.4, the mitigation landscape mapped in Section 4.6, and the theoretical frameworks introduced in Section 2.5, this review proposes an integrated safety evaluation framework for agentic health AI systems as its fourth original contribution. The framework integrates three complementary components. The Failure Mode and Effects Analysis component provides a structured prospective assessment of each of the seven failure mode categories for a given agentic system in a given clinical deployment environment, characterising each failure mode by its likelihood, potential severity, detectability, and Risk Priority Number calculated using Equation 5, with the overall system safety profile expressed as a Composite Safety Score calculated using Equation 7 [56,57]. The Swiss Cheese component maps the defensive layers available in the deployment environment against the identified failure modes using the Defence in Depth Score defined by Equation 6, identifying alignment vulnerabilities in which the holes in multiple defensive layers simultaneously permit a failure pathway to reach the patient and specifying the additional layers required to achieve adequate defence in depth [54,55]. The Human Factors component evaluates the trust calibration risk profile of the deployment environment, characterising the automation bias vulnerability of the clinical environment using the ABI defined in Equation 2, the adequacy of clinician training and mental model development, and the quality of uncertainty communication in the system's user interface [58,59]. Together these three components generate a comprehensive safety profile for a given agentic health AI system in a given clinical environment, expressed through the Composite Safety Score of Equation 7, specifying the failure modes of greatest concern as identified through the RPN values of Equation 5, the defensive gaps requiring remediation as identified through the DDS values of Equation 6, and the human factors interventions required to support calibrated clinical oversight as monitored through the ABI values of Equation 2. The framework is designed to be applied iteratively, at pre deployment evaluation, at

 doi:10.20944/preprints202604.1409.v1

post deployment audit intervals, and following any significant system update or clinical environment change, reflecting the dynamic nature of agentic AI safety in real world clinical deployment [9,38]. The proposed framework is visualised in **Figure 6**.

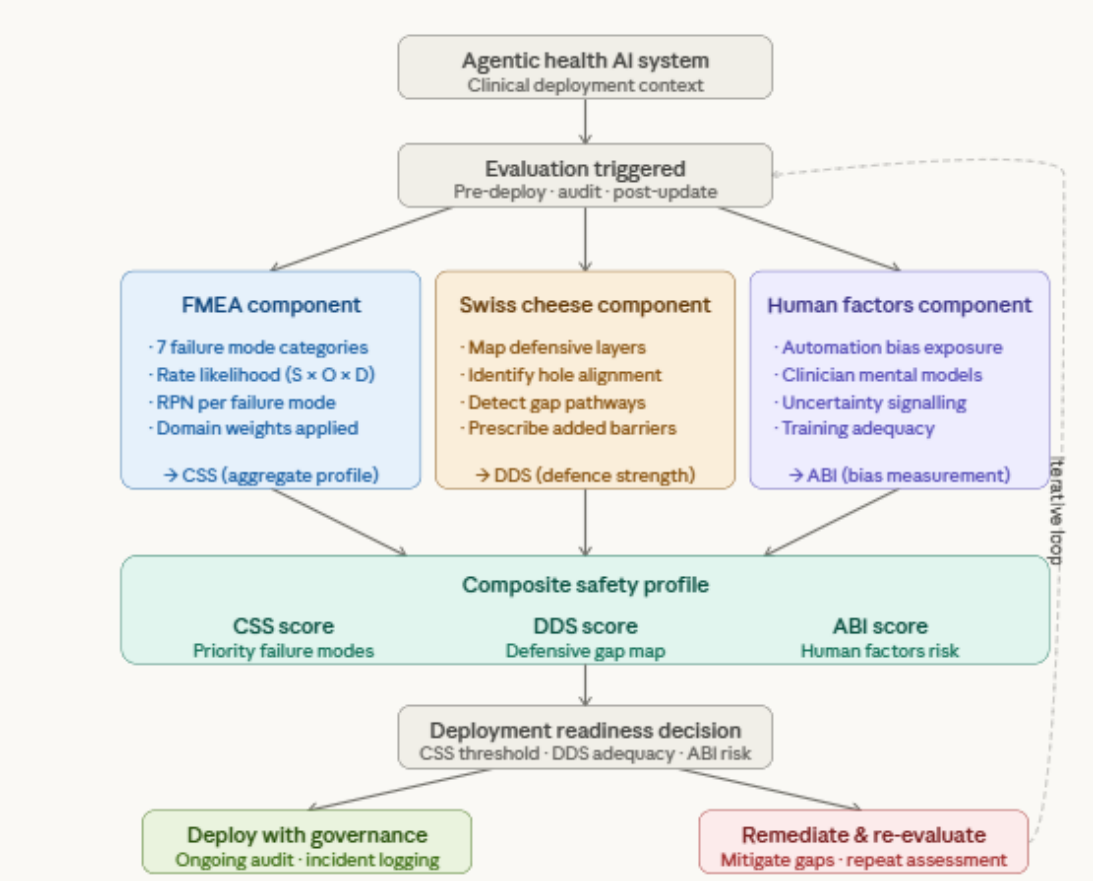

Figure 6. Proposed Integrated Safety Evaluation Framework combining Failure Mode and Effects Analysis, Swiss Cheese Model, and Human Factors components with Composite Safety Score output.

The findings of this review carry direct practical implications for three stakeholder groups. For clinicians and clinical teams, the primary implication is the necessity of active, informed, and sustained engagement with AI safety rather than passive acceptance of AI outputs, with structured education in the failure mode profiles of specific deployed systems, preservation of meaningful human oversight checkpoints at high consequence decision points, and deliberate maintenance of independent clinical reasoning skills that agentic AI deployment may otherwise progressively erode [28,51,59–61,81,114]. For developers and AI engineers, the primary implication is the necessity of safety by design approaches that treat failure mode prevention as a core engineering requirement rather than a post hoc evaluation concern, incorporating Retrieval Augmented Generation architectures for reduction of $H_r$as defined in Equation 1, memory architecture design for failure prevention, adversarial robustness testing, and comprehensive logging capabilities that support post deployment monitoring of CSS values as defined in Equation 7 [50,87,99–101,112,115–117,119,120]. For hospital systems and clinical governance bodies, the primary implication is the necessity of institutional frameworks for AI safety governance that go beyond pre procurement vendor evaluation to encompass ongoing post deployment monitoring, incident reporting, clinician education, and regular safety audit using the CSS of Equation 7 and DDS of Equation 6 as structured governance metrics [54,56,58,120,121].

The deployment of agentic AI in healthcare also raises ethical and regulatory challenges that extend beyond the technical safety concerns addressed above. Accountability and liability represent the most immediately pressing challenge: when an agentic AI system operating with substantial

 doi:10.20944/preprints202604.1409.v1

autonomy contributes to a patient harm event, the allocation of accountability across the system developer, the clinical institution, the deploying clinician, and the AI system itself is deeply unclear under current legal and regulatory frameworks, and the autonomous nature of agentic systems strains conventional notions of medical liability that presuppose a human decision maker responsible for each consequential clinical action [11,28,120,121]. Explainability requirements for clinical agentic decisions represent both an ethical imperative and a practical safety necessity: patients have a fundamental ethical right to understand the basis of clinical decisions affecting their care, clinicians have a professional obligation to exercise informed rather than delegated judgment, and current agentic systems frequently cannot provide clinically meaningful explanations of multistep reasoning chains involving multiple tool calls and memory retrievals [93,97,117,118,121]. Equity in autonomous clinical pipelines requires specific regulatory attention beyond general AI bias frameworks: the scale and autonomy of agentic health AI means that equity failures affect large patient populations systematically rather than individual patients incidentally, regulatory requirements for equity evaluation and demographic performance reporting should be mandatory rather than voluntary, and the reviewed literature consistently demonstrates that bias in clinical AI systems evolves with deployment environment, patient population, and system updates in ways that require continuous rather than one time monitoring [53,103–107,121,122]. In sum, the safety of agentic health AI cannot be assured through technical mitigations alone, it requires the alignment of robust engineering practices, calibrated clinical oversight, institutional governance frameworks, and regulatory requirements adequate to the novel challenges that agentic autonomy introduces into clinical care [54,56,58,120–122].

6. Conclusion and Future Research Directions

This systematic review examined safety risks, hallucination phenomena, and failure modes in agentic health AI systems, synthesising evidence from 113 peer reviewed publications between January 2019 and December 2025. The review identified seven distinct failure mode categories spanning reasoning errors, hallucination, tool misuse, memory failures, automation bias, adversarial and distributional vulnerabilities, and equity related harms. These categories collectively constitute the first structured failure mode taxonomy specific to agentic health AI, representing the core original contribution of this paper. The synthesis further developed a hallucination typology mapping factual, contextual, citation, and numerical hallucination types to their clinical risk profiles as quantified by H_r in Equation 1, evaluated the existing mitigation landscape against the taxonomy and the Defence in Depth Score in Equation 6, and proposed an integrated safety evaluation framework combining Failure Mode and Effects Analysis via the RPN in Equation 5 and CSS in Equation 7, the Swiss Cheese Model via the DDS in Equation 6, and Human Factors theory via the ABI in Equation 2, into a unified clinical governance tool. The cascade risk formalised in Equation 4 demonstrates formally why the compounding nature of failure in agentic architectures demands a multi component safety response rather than piecemeal technical intervention.

The findings confirm that agentic health AI presents a qualitatively distinct and in important respects more dangerous safety profile than non agentic clinical AI, driven by the compounding effects of autonomy, multistep reasoning, tool access, and confidence presentation on clinical error propagation as captured by $P(F_{\text{cascade}})$ in Equation 4. Safety cannot be assured through technical mitigations alone but requires the alignment of robust engineering practices, calibrated clinical oversight, institutional governance frameworks, and regulatory requirements adequate to the novel challenges that agentic autonomy introduces into clinical care. Clinicians, developers, hospital systems, and regulators each carry distinct and urgent responsibilities in this alignment, and the proposed safety evaluation framework offered in Section 5 provides a structured basis for operationalising those responsibilities across the deployment lifecycle of agentic clinical AI systems.

Several priority areas for future investigation emerge from this synthesis. The most urgent is the development of standardised safety benchmarks specifically designed for agentic health AI systems,

 doi:10.20944/preprints202604.1409.v1

enabling consistent and regulatorily meaningful safety evaluation across developers and clinical environments, and enabling empirical validation of the CSS threshold values described in Equation 7. Equally pressing is the need for longitudinal real world deployment studies that capture the safety profile of agentic clinical AI under the full complexity of clinical practice over extended time horizons, moving beyond the controlled evaluations that currently dominate the evidence base. Future research should give substantially greater attention to equity and distributional failure modes, developing methodologies for detecting and mitigating demographic bias and distributional degradation in autonomous clinical pipelines. The systematic study of failure mode cascades as formalised in Equation 4, specifically how failures in one category trigger and amplify failures in others, represents a foundational research priority for safety framework development. Regulatory sandbox environments enabling controlled real world testing prior to full clinical deployment warrant exploration as a bridge between laboratory evaluation and practice. Finally, interdisciplinary collaboration across AI safety research, clinical medicine, medical ethics, and regulatory science is essential to ensure that technical advances in agentic AI are matched by the governance frameworks and human factors interventions required for safe and responsible clinical deployment.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28(1):31–38. https://doi.org/10.1038/s41591-021-01614-0
3. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents. *Front Comput Sci.* 2024;18(6):186345. https://doi.org/10.1007/s11704-024-40231-1
4. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The rise and potential of large language model based agents: a survey. arXiv:2309.07864. 2023. https://arxiv.org/abs/2309.07864
5. Joshi G, Wadhwa RR, Bhogal R, Yagnik M, Haga S, Papagerakis P. FDA-approved artificial intelligence and machine learning enabled medical devices: an updated landscape. *Electronics.* 2024;13(3):498. https://doi.org/10.3390/electronics13030498
6. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375. 2023. https://arxiv.org/abs/2303.13375
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930–1940. https://doi.org/10.1038/s41591-023-02448-8
8. Grand View Research. Artificial intelligence in healthcare market size, share and trends analysis report 2022–2030. San Francisco: Grand View Research; 2022.
9. Wornow M, Xu Y, Lavertu R, Goh E, Moor M, Sheikh A, et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digit Med.* 2023;6(1):135. https://doi.org/10.1038/s41746-023-00879-8
10. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature.* 2023;616(7956):259–265. https://doi.org/10.1038/s41586-023-05881-4
11. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172–180. https://doi.org/10.1038/s41586-023-06291-2
12. Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. arXiv:2309.05922. 2023. https://arxiv.org/abs/2309.05922
13. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges and open questions. *ACM Trans Inf Syst.* 2025;43(2):1–55. https://doi.org/10.1145/3703155
14. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med.* 2023;3(1):141. https://doi.org/10.1038/s43856-023-00370-1

 doi:10.20944/preprints202604.1409.v1

15. Hager P, Jungmann F, Holland R, Bhatt N, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med.* 2024;30(9):2613–2622. https://doi.org/10.1038/s41591-024-03097-1
16. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv.* 2023;55(12):1–38. https://doi.org/10.1145/3571730
17. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *npj Digit Med.* 2023;6(1):195. https://doi.org/10.1038/s41746-023-00939-z
18. Magrabi F, Ammenwerth E, McNair JB, De Keizer N, Hyppönen H, Nykänen P, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. *Yearb Med Inform.* 2019;28(1):128–134. https://doi.org/10.1055/s-0039-1677903
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71. https://doi.org/10.1136/bmj.n71
20. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Hoboken: Pearson; 2020.
21. Wooldridge M, Jennings NR. Intelligent agents: theory and practice. *Knowl Eng Rev.* 1995;10(2):115–152. https://doi.org/10.1017/S0269888900008122
22. Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: interactive simulacra of human behavior. In: Proceedings of UIST 2023; 2023. https://doi.org/10.1145/3586183.3606763
23. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. ReAct: synergizing reasoning and acting in language models. In: Proceedings of ICLR 2023; 2023. arXiv:2210.03629. https://arxiv.org/abs/2210.03629
24. Shinn N, Cassano F, Gopinath A, Narasimhan K, Yao S. Reflexion: language agents with verbal reinforcement learning. *Adv Neural Inf Process Syst.* 2023;36:8634–8652. arXiv:2303.11366. https://arxiv.org/abs/2303.11366
25. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst.* 2022;35:24824–24837. arXiv:2201.11903. https://arxiv.org/abs/2201.11903
26. Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Zettlemoyer L, et al. Toolformer: language models can teach themselves to use tools. *Adv Neural Inf Process Syst.* 2023;36. arXiv:2302.04761. https://arxiv.org/abs/2302.04761
27. Packer C, Wooders S, Lin K, Fang V, Patil SG, Stoica I, et al. MemGPT: towards LLMs as operating systems. arXiv:2310.08560. 2023. https://arxiv.org/abs/2310.08560
28. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. Hello AI: uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proc ACM Hum Comput Interact.* 2019;3(CSCW):1–24. https://doi.org/10.1145/3359206
29. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005
30. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education. *PLOS Digit Health.* 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
31. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare. *J Med Syst.* 2023;47(1):33. https://doi.org/10.1007/s10916-023-01925-4
32. Pianykh OS, Langs G, Dewey M, Bhatt DL, Cannell JM, Pandharipande PV, et al. Continuous learning AI in radiology: implementation principles and early applications. *Radiology.* 2020;297(1):6–14. https://doi.org/10.1148/radiol.2020200038
33. Jin Q, Yang Y, Chen Q, Lu Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. arXiv:2304.09667. 2023. https://arxiv.org/abs/2304.09667
34. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. ChemCrow: augmenting large-language models with chemistry tools. arXiv:2304.05376. 2023. https://arxiv.org/abs/2304.05376
35. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med.* 2022;28(9):1773–1784. https://doi.org/10.1038/s41591-022-01981-2

 doi:10.20944/preprints202604.1409.v1

36. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions. *JAMA Intern Med.* 2023;183(6):589–596. https://doi.org/10.1001/jamainternmed.2023.1838
37. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J.* 2019;6(2):94–98. https://doi.org/10.7861/futurehosp.6-2-94
38. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA.* 2023;330(9):866–869. https://doi.org/10.1001/jama.2023.14217
39. Coiera E. The fate of medicine in the time of AI. *Lancet.* 2018;392(10162):2331–2332. https://doi.org/10.1016/S0140-6736(18)31925-1
40. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care.* 2023;27(1):120. https://doi.org/10.1186/s13054-023-04393-x
41. Umapathi LK, Pal A, Sankarasubbu M. Med-HALT: medical domain hallucination test for large language models. In: Proceedings of CoNLL 2023; 2023. p. 314–334. https://doi.org/10.18653/v1/2023.conll-1.21
42. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine.* 2023;90:104512. https://doi.org/10.1016/j.ebiom.2023.104512
43. Miao J, Thongprayoon C, Cheungpasitporn W, Miao SA, Suppadungsuk S, Garcia Valencia OA, et al. Implications of large language model hallucination in clinical practice. *Mayo Clin Proc Digit Health.* 2023;1(4):230–240. https://doi.org/10.1016/j.mcpdig.2023.07.003
44. Kim Y, Hua K, Singh R, Jain S, Suresh H, Stone E, et al. Medical hallucinations in foundation models and their impact on healthcare. arXiv:2503.05777. 2025. https://arxiv.org/abs/2503.05777
45. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med.* 2023;388(13):1233–1239. https://doi.org/10.1056/NEJMsr2214184
46. Chelli M, Descamps J, Lavoue V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews. *J Med Internet Res.* 2024;26:e53164. https://doi.org/10.2196/53164
47. Omar M, Brin D, Glicksberg B, Klang E. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun Med.* 2025. https://doi.org/10.1038/s43856-025-01021-3
48. Kambhampati S, Valmeekam K, Guan L, Stechly M, Verma PK, Bhambri S, et al. LLMs can't plan, but can help planning in LLM-modulo frameworks. arXiv:2402.01817. 2024. https://arxiv.org/abs/2402.01817
49. Patil SG, Zhang T, Wang X, Gonzalez JE. Gorilla: large language model connected with massive APIs. arXiv:2305.15334. 2023. https://arxiv.org/abs/2305.15334
50. Liu NF, Lin K, Hewitt J, Paranjape A, Manning CD, Liang P. Lost in the middle: how language models use long contexts. *Trans Assoc Comput Linguist.* 2024;12:157–173. https://doi.org/10.1162/tacl_a_00638
51. Skitka LJ, Mosier K, Burdick MD. Does automation bias decision-making? *Int J Hum Comput Stud.* 1999;51(5):991–1006. https://doi.org/10.1006/ijhc.1999.0252
52. Perez E, Huang S, Song F, Cai T, Ring R, Aslanides J, et al. Red teaming language models with language models. arXiv:2202.03286. 2022. https://arxiv.org/abs/2202.03286
53. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447–453. https://doi.org/10.1126/science.aax2342
54. Reason J. Human Error. Cambridge: Cambridge University Press; 1990.
55. Reason J. Human error: models and management. *BMJ.* 2000;320(7237):768–770. https://doi.org/10.1136/bmj.320.7237.768
56. DeRosier J, Stalhandske E, Bagian JP, Nudell T. Using health care failure mode and effect analysis. *Jt Comm J Qual Improv.* 2002;28(5):248–267. https://doi.org/10.1016/s1070-3241(02)28025-6
57. Spath PL, editor. Error Reduction in Health Care: A Systems Approach to Improving Patient Safety. 2nd ed. San Francisco: Jossey-Bass; 2011.
58. Carayon P, Hundt AS, Karsh BT, Gurses AP, Alvarado CJ, Smith M, et al. Work system design for patient safety: the SEIPS model. *Qual Saf Health Care.* 2006;15(Suppl 1):i50–i58. https://doi.org/10.1136/qshc.2005.015842

 doi:10.20944/preprints202604.1409.v1

59. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors.* 2010;52(3):381–410. https://doi.org/10.1177/0018720810376055
60. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc.* 2012;19(1):121–127. https://doi.org/10.1136/amiajnl-2011-000089
61. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc.* 2017;24(2):423–431. https://doi.org/10.1093/jamia/ocw105
62. Kitchenham BA, Charters S. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01. Keele University and Durham University; 2007.
63. Gusenbauer M. Beyond Google Scholar, Scopus, and Web of Science: an evaluation of the backward and forward citation coverage of 59 databases' citation indices. *Res Synth Methods.* 2024;15(5):802–817. https://doi.org/10.1002/jrsm.1729
64. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol.* 2006;3(2):77–101. https://doi.org/10.1191/1478088706qp063oa
65. Magrabi F, Ong MS, Runciman W, Coiera E. Using FDA reports to inform a classification for health information technology safety problems. *J Am Med Inform Assoc.* 2012;19(1):45–53. https://doi.org/10.1136/amiajnl-2011-000369
66. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol.* 2005;8(1):19–32. https://doi.org/10.1080/1364557032000119616
67. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195. https://doi.org/10.1186/s12916-019-1426-2
68. Topol EJ. Welcoming new medical imaging AI. *Lancet.* 2022;400(10369):2148–2150. https://doi.org/10.1016/S0140-6736(22)02463-3
69. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv:2303.08774. 2023. https://arxiv.org/abs/2303.08774
70. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic AI. arXiv:2401.05654. 2024. https://arxiv.org/abs/2401.05654
71. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging. *Lancet Digit Health.* 2019;1(6):e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
72. Baxter SL, Saseendrakumar BR, Karri S, Kim J, Ye Z, Hsu CY, et al. Limitations and challenges of using artificial intelligence in electronic health records. *Pac Symp Biocomput.* 2022;27:312–323.
73. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data. *J Am Med Inform Assoc.* 2017;24(1):198–208. https://doi.org/10.1093/jamia/ocw042
74. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Toward expert-level medical question answering with large language models. *Nat Med.* 2025;31:76–88. https://doi.org/10.1038/s41591-024-03423-7
75. Omiye JA, Gui H, Tanjuma F, Daneshjou R. Large language models in medicine: the potentials and pitfalls. *Ann Intern Med.* 2024;177(2):210–220. https://doi.org/10.7326/M23-2772
76. Hong S, Zheng X, Chen J, Cheng Y, Wang J, Zhang C, et al. MetaGPT: meta programming for a multi-agent collaborative framework. arXiv:2308.00352. 2023. https://arxiv.org/abs/2308.00352
77. Zhang S, Xu Y, Usuyama N, Bagga J, Tinn R, Preston S, et al. BiomedGPT: a generalist vision-language foundation model for diverse biomedical tasks. arXiv:2305.17100. 2023. https://arxiv.org/abs/2305.17100
78. Yang H, Yue S, He Y. Auto-GPT for online decision making: benchmarks and additional opinions. arXiv:2306.02224. 2023. https://arxiv.org/abs/2306.02224
79. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, et al. Chest radiograph interpretation with deep learning models. *Radiology.* 2020;294(2):421–431. https://doi.org/10.1148/radiol.2019191293
80. Larson DB, Harvey H, Rubin DL, Irani N, Justin RT, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms. *J Am Coll Radiol.* 2021;18(3):413–424. https://doi.org/10.1016/j.jacr.2020.09.060

 doi:10.20944/preprints202604.1409.v1

81. Valmeekam K, Olmo A, Sreedharan S, Kambhampati S. Large language models still can't plan. arXiv:2206.10498. 2022. https://arxiv.org/abs/2206.10498
82. Mialon G, Dessì R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, et al. Augmented language models: a survey. arXiv:2302.07842. 2023. https://arxiv.org/abs/2302.07842
83. Shi F, Chen X, Misra K, Scales N, Dohan D, Chi E, et al. Large language models can be easily distracted by irrelevant information. In: Proceedings of ICML 2023; 2023. p. 31210–31227.
84. Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv:2309.01219. 2023. https://arxiv.org/abs/2309.01219
85. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interact J Med Res.* 2025;14:e59823. https://doi.org/10.2196/59823
86. Van Veen D, Van Molle C, Rios J, Hayashi C, Bastani O, Langlotz CP, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med.* 2024;30(5):1134–1142. https://doi.org/10.1038/s41591-024-02855-5
87. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst.* 2020;33:9459–9474. arXiv:2005.11401. https://arxiv.org/abs/2005.11401
88. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of ACM FAccT 2021; 2021. p. 610–623. https://doi.org/10.1145/3442188.3445922
89. Frieder S, Pinchetti L, Griffiths RR, Salvatori T, Lukasiewicz T, Petersen PC, et al. Mathematical capabilities of ChatGPT. arXiv:2301.13379. 2023. https://arxiv.org/abs/2301.13379
90. Katz DM, Bommarito MJ, Guo S, Crabb P. GPT-4 passes the bar exam. *Philos Trans R Soc A.* 2024;382(2270):20230254. https://doi.org/10.1098/rsta.2023.0254
91. Stechly M, Valmeekam K, Kambhampati S. On the self-verification limitations of large language models on reasoning and planning tasks. arXiv:2402.08115. 2024. https://arxiv.org/abs/2402.08115
92. Huang J, Chang KCC. Towards reasoning in large language models: a survey. arXiv:2212.10403. 2022. https://arxiv.org/abs/2212.10403
93. Lanham T, Chen A, Wang T, Angelica M, Goldowsky-Dill N, Hernandez D, et al. Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702. 2023. https://arxiv.org/abs/2307.13702
94. Tonmoy SMTI, Zaman S, Jain V, Rani A, Rawte V, Chadha A, et al. A comprehensive survey of hallucination mitigation techniques in large language models. Findings EMNLP 2024; 2024. https://doi.org/10.18653/v1/2024.findings-emnlp.686
95. Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. HuggingGPT: solving AI tasks with ChatGPT and its friends in HuggingFace. *Adv Neural Inf Process Syst.* 2023;36. arXiv:2303.17580. https://arxiv.org/abs/2303.17580
96. Qin Y, Liang S, Ye Y, Zhu K, Yan L, Lu Y, et al. ToolLLM: facilitating large language models to master 16000+ real-world APIs. arXiv:2307.16789. 2023. https://arxiv.org/abs/2307.16789
97. Weng L, Liao Z, Oikarinen T, Kolter JZ. Large language models are not robust multiple choice selectors. arXiv:2309.03882. 2023. https://arxiv.org/abs/2309.03882
98. Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, et al. Holistic evaluation of language models. arXiv:2211.09110. 2022. https://arxiv.org/abs/2211.09110
99. Gao L, Ma X, Lin J, Callan J. Precise zero-shot dense retrieval without relevance labels. arXiv:2212.10496. 2022. https://arxiv.org/abs/2212.10496
100. Bussone A, Stumpf S, O'Sullivan D. The role of explanations on trust and reliance in clinical decision support systems. In: Proceedings of IEEE ICHI 2015; 2015. p. 160–169. https://doi.org/10.1109/ICHI.2015.26
101. Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. arXiv:2302.12173. 2023. https://arxiv.org/abs/2302.12173
102. Perez F, Ribeiro I. Ignore previous prompt: attack techniques for language models. arXiv:2211.09527. 2022. https://arxiv.org/abs/2211.09527
103. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science.* 2019;363(6433):1287–1289. https://doi.org/10.1126/science.aaw4399

104. Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance. arXiv:1908.00690. 2019. https://arxiv.org/abs/1908.00690
105. Zou J, Schiebinger L. AI can be sexist and racist — it's time to make it fair. *Nature.* 2018;559(7714):324–326. https://doi.org/10.1038/d41586-018-05707-8
106. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med.* 2020;26(1):16–17. https://doi.org/10.1038/s41591-019-0649-2
107. Char DS, Shah NH, Magnus D. Implementing machine learning in health care — addressing ethical challenges. *N Engl J Med.* 2018;378(11):981–983. https://doi.org/10.1056/NEJMp1714229
108. Vincent C, Taylor-Adams S, Stanhope N. Framework for analysing risk and safety in clinical medicine. *BMJ.* 1998;316(7138):1154–1157. https://doi.org/10.1136/bmj.316.7138.1154
109. Slight SP, Seger DL, Nanji KC, Cho I, Volk LA, Bates DW, et al. Are we heeding the warning signs? Examining providers' overrides of computerized drug-drug interaction alerts in primary care. *PLoS One.* 2013;8(12):e85071. https://doi.org/10.1371/journal.pone.0085071
110. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med.* 2008;121(5 Suppl):S2–S23. https://doi.org/10.1016/j.amjmed.2008.01.001
111. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, et al. Ethical and social risks of harm from language models. arXiv:2112.04359. 2021. https://arxiv.org/abs/2112.04359
112. Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: retrieval-augmented language model pre-training. In: Proceedings of ICML 2020; 2020. arXiv:2002.08909. https://arxiv.org/abs/2002.08909
113. Shi W, Min S, Yasunaga M, Seo M, James R, Lewis M, et al. REPLUG: retrieval-augmented black-box language models. arXiv:2301.12652. 2023. https://arxiv.org/abs/2301.12652
114. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Software engineering for machine learning: a case study. In: Proceedings of ICSE-SEIP 2019; 2019. p. 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042
115. Shneiderman B. Human-centered AI. *Issues Sci Technol.* 2021;37(2):56–61.
116. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862. 2022. https://arxiv.org/abs/2204.05862
117. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: harmlessness from AI feedback. arXiv:2212.08073. 2022. https://arxiv.org/abs/2212.08073
118. Lipton ZC. The mythos of model interpretability. *Queue.* 2018;16(3):31–57. https://doi.org/10.1145/3236386.3241340
119. Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. arXiv:2209.07858. 2022. https://arxiv.org/abs/2209.07858
120. US Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Silver Spring: FDA; 2021. https://www.fda.gov/media/145022/download
121. European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union.* 2024. https://doi.org/10.3390/electronics13030498
122. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO; 2021. ISBN: 978-92-4-002324-6. https://www.who.int/publications/i/item/9789240029200