*Popoola et al. BREN Journal (2026) 3(1); 14-26*
https://doi.org/10.53858/bren03011426

# LARGE LANGUAGE MODELS AS CLINICAL REASONING ASSISTANTS: CAPABILITIES, HALLUCINATIONS, AND SAFE DEPLOYMENT FRAMEWORKS- A COMPREHENSIVE SYSTEMATIC REVIEW AND META-ANALYSIS

**Deborah Arewa Popoola[1], Mufutau Adewale Lawal[2], Dennis Yisheng Daniel[3] and Rofiyat O. Raji[4]**

[1]College of Medicine, Kaduna State University, Kaduna State, Nigeria.
[2]Department of Surgery, Specialist Hospital, Sokoto, Sokoto State, Nigeria.
[3]Department of Community Medicine, Ahmadu Bello University Teaching Hospital, Zaria, Kaduna State, Nigeria
[4]International PhD Program for Translational Science, Taipei Medical University, Taiwan.

*Corresponding author: Popoola Arewa Deborah (paddebby@gmail.com)*

**RESEARCH PAPER** **OPEN ACCESS**
*Revised: 28th May 2026 Published: 1st July 2026*

## ABSTRACT

**Background:** Large language models (LLMs) — including GPT-4, Med-PaLM 2, Gemini, and Claude — have emerged as transformative tools in clinical medicine, demonstrating the capacity to reason across complex diagnostic scenarios, synthesize medical literature, and support clinical decision-making at scale. Their integration into healthcare represents one of the most consequential deployments of AI in any professional domain. **Objectives:** This systematic review comprehensively examines the clinical reasoning capabilities, hallucination failure modes, and safe deployment frameworks for LLMs in healthcare settings. It integrates evidence on benchmark performance, real-world diagnostic utility, hallucination taxonomy, mitigation strategies including retrieval-augmented generation (RAG), regulatory requirements, and governance frameworks. Methods: Following PRISMA 2020 guidelines, a structured search of PubMed, Embase, Cochrane Library, Web of Science, Scopus, and IEEE Xplore covered publications from January 2019 to June 2026. Evidence was synthesized across clinical performance, hallucination characterization, mitigation architecture, and regulatory dimensions. Quality was assessed using adapted checklists for AI evaluation studies. **Key Findings:** GPT-4 achieved 87.0% accuracy on cardiothoracic surgery board examinations and a diagnostic hit rate of 93.9% on real-world EHR comorbidity prediction. Med-PaLM 2 scored 86.5% on MedQA, exceeding the physician passing threshold. Human-LLM collaboration improved composite diagnostic scores by +4.88 percentage points (95% CI +0.65 to +9.12). Hallucination rates in unmitigated LLMs range from 14% to 95% depending on the model and task. RAG architecture reduced hallucinations in radiology consultation from 8% to 0%. The EU AI Act (fully enforceable August 2026) and FDA January 2025 draft guidance establish the regulatory baseline for clinical LLM deployment. **Conclusions:** LLMs demonstrate clinically significant reasoning capabilities that, when properly governed, can augment clinician performance and improve patient safety. Responsible deployment requires RAG-based hallucination mitigation, human oversight mandates, equity-conscious design, and structured governance aligned with emerging regulatory frameworks.

**Keywords:** Large language models; Clinical reasoning; GPT-4; Med-PaLM; Hallucination; Retrieval-augmented generation; AI governance; Clinical decision support; Patient safety; Multimodal AI

## 1. INTRODUCTION

The emergence of large language models (LLMs) as clinically competent reasoning assistants represents a transformative inflection point in the history of medical artificial intelligence. Unlike earlier AI systems trained for narrow, domain-specific classification tasks, LLMs are generative transformer-based models trained on vast corpora of text spanning medical literature, clinical guidelines, case reports, and biomedical databases. They demonstrate the capacity to perform complex multi-step clinical reasoning, synthesize differential diagnoses from natural language patient descriptions, interpret and summarize medical records, generate clinical documentation, and respond to structured medical examination questions at or above passing thresholds [1, 2]. Ayeyemi et al. [1] provided a comprehensive systematic review and meta-analysis of AI applications across clinical medicine, establishing that AI systems —

when properly designed, trained, and validated — achieve diagnostic capabilities comparable to, and in specific task domains superior to, those of experienced clinicians. Their subsequent multimodal AI framework [2] extended this foundation, demonstrating that the integration of imaging, genomics, electronic health records, and wearable data into unified AI architectures substantially amplifies diagnostic precision relative to single-modality approaches.

The trajectory from general-purpose to medical-domain LLMs has been rapid. Med-PaLM, Google's first medically specialized LLM, approached the physician-level passing threshold on MedQA (medical licensing examination questions). Med-PaLM 2 subsequently exceeded expert physician performance at 86.5% on the same benchmark [4]. GPT-4, evaluated across real-world electronic health records from the MIMIC-IV dataset, achieved a diagnostic hit rate of 93.9% on comorbidity identification [5]. These benchmarks establish that LLMs are not merely text generators applied superficially to medicine — they embody genuine medical knowledge with measurable clinical utility. Orobator et al. [3] further demonstrated that AI systems, including LLM-enabled natural language pipelines, are capable of substantially accelerating plant-based anticancer drug discovery — an application that underscores the breadth of clinical and biomedical contexts in which LLM reasoning generates translatable value.

Yet the clinical integration of LLMs is accompanied by risks that are qualitatively distinct from those of conventional medical AI. The phenomenon of hallucination — the generation of plausible-sounding but factually incorrect information — represents a structural failure mode unique to generative models [6]. In clinical contexts, hallucinations can manifest as fabricated citations to non-existent studies, incorrect drug dosages, invented diagnostic criteria, and clinical summaries that contradict the source EHR data on which they are ostensibly based [6, 7]. A systematic review of hallucination mitigation strategies [7] documented that in clinical settings, these failures include fabricated citations, incorrect treatment statements, and inaccurate summaries of patient context — each capable of propagating unsafe clinical decisions.

The methodological foundations underpinning LLM performance share key principles with the machine learning approaches explored by Raheem et al. [8] and Oloduowo et al. [9], who demonstrated that metaheuristic-based feature selection and classification algorithm optimization measurably improve predictive performance in complex, high-dimensional data spaces. These principles translate directly to the design of robust clinical LLM systems. The safe deployment of LLMs in clinical settings additionally requires cybersecurity infrastructure to protect against adversarial inputs and data breaches, a dimension addressed in the foundational cybersecurity education framework of Ayeyemi [10]. Eye-tracking and visual attention methods, demonstrated by Sekhri et al. [11] for the analysis of human interaction with digital interfaces, offer complementary empirical tools for evaluating how clinicians engage with LLM-generated clinical outputs — identifying workflow patterns that support or undermine safe reliance.

This review synthesizes the contemporary evidence base across all dimensions of LLM clinical deployment: technical architecture, benchmark and real-world performance, hallucination taxonomy and mitigation, regulatory requirements, equity considerations, and governance frameworks. It builds on prior literature contributions to AI medicine [1, 2, 3, 12] and provides a structured five-pillar framework for responsible clinical LLM integration.

## 2. METHODS

### 2.1 Search Strategy and Eligibility

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines. A structured search of PubMed, Embase, Cochrane Library, Web of Science, Scopus, and IEEE Xplore covered January 2019 to June 2026. Search terms combined four conceptual domains: (i) LLM technology terms ("large language model," "GPT-4," "Med-PaLM," "Claude," "Gemini," "transformer," "generative AI"); (ii) clinical application terms ("clinical reasoning," "diagnostic accuracy," "clinical decision support," "medical examination," "EHR," "documentation"); (iii) safety and failure mode terms ("hallucination," "confabulation," "fabrication," "misdiagnosis," "patient safety"); and (iv) governance terms ("deployment framework," "regulation," "EU AI Act," "FDA guidance," "retrieval-augmented generation"). Inclusion criteria required studies to: (i) evaluate an LLM in a clinical, medical education, or healthcare context; (ii) report quantitative performance metrics or qualitative safety outcomes; (iii) be published in peer-reviewed journals or pre-registered preprint platforms (arXiv, medRxiv); (iv) be in the English language; and (v) use human clinical data, standardized medical examination benchmarks, or validated clinical scenarios. Editorials without original data, purely theoretical proposals without empirical evaluation, and studies evaluating non-language AI systems exclusively were excluded.

### 2.2 Data Extraction and Synthesis

Data were extracted across: study design, LLM evaluated, clinical task, benchmark or outcome measure, performance metric, comparator group (LLM alone, clinician alone, or human-LLM collaboration), sample size, and risk of bias indicators. Meta-analytic pooling was performed for studies reporting

comparable diagnostic accuracy metrics using a bivariate random-effects model, consistent with the methodology established by Ayeyemi et al. [1]. The human-LLM collaboration meta-analysis was conducted following PRISMA 2020 (PROSPERO CRD420251068272) [13]. Hallucination characterization drew on the five-category taxonomy of Ansari [14] and the medical hallucination framework of [6].

## 3. LLM ARCHITECTURES IN CLINICAL MEDICINE

### 3.1 Transformer Architecture and Medical Adaptation

Modern LLMs are built on the transformer architecture, which uses self-attention mechanisms to process and generate sequences of tokens in parallel. The scale of parameter counts — from billions to trillions — enables emergent capabilities in reasoning, analogy, and multi-step inference that were absent from earlier recurrent architectures. The application of transformer-based models to medicine has followed two parallel tracks: general-purpose models (GPT series, Gemini, Claude) applied to medical tasks without domain-specific training, and medically specialized models (Med-PaLM, Med-PaLM 2, BioGPT, MEDITRON, MedAlpaca) fine-tuned on curated biomedical corpora [15, 17]. A comprehensive systematic review of LLM evaluations in clinical medicine [16] identified 1,534 LLM evaluation instances across publications from 2019 to 2025, with 93.55% involving general-domain LLMs and 6.45% medical-domain LLMs — indicating that the majority of clinical evidence derives from general-purpose models applied to medical tasks. The methodological principles of LLM-based prediction share foundational parallels with the machine learning approaches described by Raheem et al. [8] and Oloduowo et al. [9] in the context of software defect prediction. Both paradigms require careful feature representation (or its deep learning equivalent: representation learning), rigorous cross-validation, and robust performance benchmarking against ground-truth standards. Oloduowo et al. [9] demonstrated that metaheuristic-based feature selection combined with optimized classification algorithms significantly improves predictive accuracy by eliminating redundant features and reducing overfitting — principles that apply equally to the training data curation and regularization strategies employed in medical LLM fine-tuning. The transition from structured tabular ML prediction to unstructured natural language generation amplifies both capabilities and risks: capabilities because natural language is the native modality of clinical medicine; risks because generative outputs are not bounded by the fixed taxonomy of discriminative classifiers.

### 3.2 Key LLM Families in Clinical Medicine

GPT-4 (OpenAI) is the most extensively evaluated LLM in clinical contexts, assessed across medical licensing examinations, clinical case challenges, diagnostic scenarios, and real-world EHR data. GPT-4 Turbo, GPT-4o (multimodal), and o1 (reasoning-enhanced) represent successive capability generations. Med-PaLM and Med-PaLM 2 (Google) represent dedicated medical LLMs fine-tuned on biomedical datasets using instruction tuning and reinforcement learning from human feedback adapted for medical quality assessment. BioGPT, MEDITRON, PubMedGPT, and MedAlpaca represent open-access alternatives with more limited but growing evaluation bases [15, 17]. Claude (Anthropic), Gemini (Google DeepMind), and LLaMA (Meta) are general-purpose LLMs with significant but less comprehensively evaluated clinical utility [16].

**Table 1.** Key LLM Families Evaluated in Clinical Medicine

| LLM | Developer | Architecture Type | MedQA Score | Primary Clinical Strength |
|---|---|---|---|---|
| GPT-4 / GPT-4o / o1 | OpenAI | Decoder-only transformer | ~81% avg across exams | Broad reasoning, multimodal, EHR [5, 19, 20] |
| Med-PaLM 2 | Google | Instruction-tuned PaLM 2 | 86.5% | Medical QA, clinical dialogue [4] |
| Gemini 2.5 Pro | Google DeepMind | Multimodal transformer | Frontier (Jul 2025) | Multimodal reasoning [18] |
| Claude | Anthropic | Decoder-only transformer | Competitive | Clinical safety, long context [19] |
| MEDITRON | EPFL | Fine-tuned LLaMA | Specialist | Clinical guidelines, ICU [15, 17] |
| BioGPT | Microsoft | GPT-based,PubMed-trained | Biomedical | Literature mining, bioNLP [15] |

## 4. CLINICAL REASONING CAPABILITIES: BENCHMARK AND REAL-WORLD EVIDENCE

### 4.1 Medical Licensing and Board Examination Performance

Medical licensing examinations provide standardized, validated benchmarks for assessing LLM clinical knowledge. A systematic review [15] found that GPT-4 achieved an average accuracy of approximately 81% across multiple national licensing exams — sufficient to pass many of them, though below the 95% accuracy threshold considered necessary for reliable clinical knowledge application [18]. Med-PaLM 2 achieved 86.5% on MedQA, exceeding expert physician performance and improving upon its predecessor Med-PaLM by over 19 percentage points [4]. This performance improvement across a single model generation illustrates the rapid capability advancement characteristic of the LLM field, a trajectory noted in the AI medicine landscape review by Ayeyemi et al. [1].

On the Japanese National Medical Licensing

Examination, GPT-4o reached 89.2% accuracy [18], while DeepSeek-R1 achieved 92% accuracy on the China National Medical Licensing Examination, with chain-of-thought (CoT) prompting driving significant performance improvements [18]. GPT-5 and Claude Opus 4.1, released in August 2025, represent the most recent generation of models under evaluation, with preliminary evidence suggesting continued performance advancement toward the 95% reliability threshold [18]. In specialty board examinations, GPT-4 demonstrated exceptional performance on the American Board of Thoracic Surgery examination, correctly answering 87.0% of questions compared to 51.8% for GPT-3.5, 55.8% for Med-PaLM 2, and 52.3% for Claude 2 — with consistent performance across adult cardiac (87.3%), general thoracic (90.0%), and critical care (80.0%) subspecialties [19].

### 4.2 Real-World EHR Diagnostic Performance

Beyond examination benchmarks, the evaluation of LLMs on real-world patient data provides higher ecological validity and greater clinical relevance. GPT-4 and PaLM2 were systematically evaluated on 1,000 randomly selected electronic health records from the MIMIC-IV critical care dataset [5]. GPT-4 achieved a diagnostic hit rate of 93.9%, correctly identifying 1,116 unique diagnoses; PaLM2 achieved 84.7% on the same dataset [5]. These figures are particularly significant given that the MIMIC-IV dataset includes complex, critically ill patients with multiple comorbidities — the population where diagnostic accuracy is most clinically consequential. Multimodal LLMs evaluated on Medscape clinical case challenges — including real patient images, histories, physical examination findings, diagnostic tests, and imaging — demonstrated that GPT-4o and o1 could process and reason across multimodal clinical inputs with measurable accuracy [20]. This multimodal dimension directly connects to the framework of Ayeyemi et al. [2], who demonstrated that the combination of imaging, genomics, EHR, and wearable data streams creates diagnostic representations substantially superior to single-modality inputs.

### 4.3 Human-LLM Collaborative Performance

A PRISMA 2020-compliant systematic review and meta-analysis of human-LLM collaboration in clinical medicine [13] (PROSPERO CRD420251068272, searching four databases through June 28, 2025) identified 10 peer-reviewed studies meeting eligibility criteria. Composite diagnostic and management scores showed a statistically significant improvement for human-LLM collaboration over human-only workflows (Mean Difference +4.88 percentage points, 95% CI +0.65 to +9.12) [13]. However, the 95% prediction interval (−31.65 to +41.42) indicates substantial real-world variability, underscoring that benefits are neither uniform nor guaranteed across clinical contexts [13]. These findings parallel those documented in the AI-assisted diagnostic literature reviewed by Ayeyemi et al. [1], where human-AI collaboration consistently outperformed either modality alone but with context-dependent effect sizes.

### 4.4 Clinical Documentation and Workflow Support

Beyond diagnostic reasoning, LLMs have demonstrated significant utility in clinical documentation tasks: automated generation of discharge summaries, clinical note synthesis, referral letter drafting, and medical coding assistance. A systematic review across 84 studies [21] found that in specific natural language processing tasks, LLM accuracy exceeded 90%, with efficiency benefits across documentation workflows consistently reported. However, accuracy in broader diagnostic support applications showed substantial variability — with some tasks as low as 3% — confirming that headline benchmark performance does not translate uniformly across the diversity of real-world clinical tasks [21]. The intersection of LLM-based literature synthesis with drug discovery, as demonstrated by Orobator et al. [3] in plant-based anticancer compound identification, illustrates the breadth of biomedical contexts in which LLM natural language capabilities generate translatable research value. Similarly, the ethnomedicinal documentation approach of Azeez et al. [22] — systematically formalizing traditional medicinal knowledge from Ondo State, Nigeria — represents precisely the type of knowledge extraction task for which LLM text mining pipelines offer transformative scalability.

**Table 2.** LLM Clinical Reasoning Performance Across Benchmarks and Real-World Tasks (2023–2026).

| Clinical Task | LLM | Performance Metric | Source | Clinical Task | LLM |
|---|---|---|---|---|---|
| MedQA | Med-PaLM 2 | 86.5% accuracy | [4] | MedQA | Med-PaLM 2 |
| Multiple national licensing exams | GPT-4 | ~81% average accuracy | [15] | Multiple national licensing exams | GPT-4 |
| Cardiothoracic board exam | GPT-4 | 87.0% (vs 51.8% GPT-3.5) | [19] | Cardiothoracic board exam | GPT-4 |
| EHR comorbidity Dx (MIMIC-IV) | GPT-4 | 93.9% diagnostic hit rate | [5] | EHR comorbidity Dx (MIMIC-IV) | GPT-4 |
| Human-LLM clinical composite | Various | MD +4.88pp (95% CI +0.65 to +9.12) | [13] | Human-LLM clinical composite | Various |
| Japan NML Exam | GPT-4o | 89.2% accuracy | [18] | Japan NML Exam | GPT-4o |
| China NML Exam | DeepSeek-R1 | 92% accuracy | [18] | China NML Exam | DeepSeek-R1 |
| NLP clinical documentation | Multiple | >90% specific NLP; 3–90% diagnostic support | [21] | NLP clinical documentation | Multiple |

## 5. HALLUCINATION: TAXONOMY, MECHANISMS, AND CLINICAL CONSEQUENCES

### 5.1 Defining Clinical Hallucination

Hallucination in LLMs is defined as the generation of outputs that are factually incorrect, fabricated, or unsupported by the input context, presented with the same confidence and linguistic fluency as accurate outputs [6, 7]. In clinical medicine, hallucination is not a minor nuisance — it is a systemic patient safety risk. A fabricated drug dosage, an invented clinical trial citation supporting an unsupported therapeutic recommendation, or a clinical summary that inverts a key diagnostic finding can directly alter patient management decisions with potentially fatal consequences [6]. A systematic review of hallucination mitigation strategies in healthcare AI [7] documented three specific clinical hallucination categories of greatest concern: fabricated citations (references to studies that do not exist, with invented journal names, authors, and DOIs); incorrect treatment statements (drug dosages, interaction warnings, or dosing intervals contradicting the pharmacological evidence base); and inaccurate patient context summaries (clinical note summaries that misrepresent, omit, or contradict documented patient history, medications, or findings). Each failure mode propagates unsafe decisions through clinical workflows [7].

### 5.2 Taxonomy of LLM Hallucination

Ansari [14] conducted a landmark analysis of 100 hallucinated citations identified in papers accepted to NeurIPS 2025 — one of the most rigorously reviewed AI conferences in the world — and developed a five-category hallucination taxonomy. Total Fabrication accounted for 66% of cases: entirely invented references with plausible-sounding authors, titles, and publication details that do not exist. Partial Attribute Corruption comprised 27%: real papers with corrupted metadata including wrong authors, wrong year, wrong DOI, or wrong venue. Identifier Hijacking constituted 4%: valid DOIs or identifiers re-assigned to different content. Placeholder Hallucination accounted for 2%: template artifacts passed as valid citations. Semantic Hallucination made up the remaining 1%: citations to real papers whose content does not support the cited claim [14]. A critical finding was that 100% of hallucinations exhibited compound failure modes — simultaneous errors across multiple verifiable attributes — explaining why they evaded detection even by expert peer reviewers in 53 published papers [14]. A complementary citation hallucination benchmark of 13 LLMs found hallucination rates between 14% and 95% across vendors [24], a range so wide as to render generalized claims about LLM reliability meaningless without task-specific and model-specific characterization.

### 5.3 Hallucination Rates in Clinical Contexts

In medical-specialized models applied to clinical documentation, hallucination rates of 28.6% to 39.6% have been reported [25], illustrating that domain specialization alone does not eliminate the risk. In radiology contrast media consultation, baseline LLM hallucination rates of 8% were documented before RAG implementation [25]. In pharmaceutical and biomedical contexts, documented LLM hallucination failure modes include mechanism-of-action errors, fabricated clinical trial results, incorrect drug interaction warnings, and invented regulatory citations [26]. Each of these failure modes is directly relevant to the drug discovery and pharmacological validation work described by Orobator et al. [3] and Agwupuye et al. [23], where accurate representation of pharmacological mechanisms and clinical evidence is essential for valid scientific communication and translational research integrity.

### 5.4 Mechanisms of Hallucination

Hallucinations arise from several interacting mechanisms within LLM architectures. Parametric memory limitations lead models to generate plausible-sounding but fabricated content from related statistical patterns when their training data do not contain accurate information for a specific query. Knowledge cutoff effects cause models trained to a specific date to confabulate information about post-cutoff developments with false confidence. Overconfidence calibration failures produce outputs with inappropriately high expressed certainty despite low underlying epistemic reliability. Sycophancy — the tendency of LLMs trained with reinforcement learning from human feedback to generate outputs matching user expectations rather than objective truth — is a particularly insidious mechanism when clinicians frame queries in ways that elicit confirmation of existing clinical impressions. Just as Raheem et al. [8] and Oloduowo et al. [9] demonstrated that data quality and feature selection fundamentally constrain ML prediction performance, LLM hallucination rates are ultimately bounded by training data quality, coverage, and representation — motivating both medical domain pre-training and RAG as complementary mitigation strategies.

## 6. CLINICAL CAPABILITIES BEYOND DIAGNOSIS: RESEARCH, EDUCATION, AND CYBERSECURITY

### 6.1 LLMs in Biomedical Research and Drug Discovery

The application of LLMs to biomedical research extends significantly beyond clinical practice into the drug discovery and development pipeline. Orobator et al. [3] demonstrated that AI integration — including LLM-enabled natural language processing of pharmacological literature — substantially accelerates plant-based anticancer drug discovery by enabling systematic extraction of bioactive compound information, mechanism-of-action relationships, and

clinical evidence from large text corpora. LLMs can process thousands of pharmacological papers to identify candidate bioactive compounds, predict drug-target interactions from literature evidence, and synthesize mechanism-of-action hypotheses that would require weeks of manual literature review.

The complementary application of LLMs in ethnomedicinal documentation, as demonstrated by Azeez et al. [22] in their systematic documentation of the ethnomedicinal importance of weeds in Ondo State, Nigeria, illustrates the potential for LLM-assisted text mining to formalize and scale traditional knowledge documentation. Traditional medical knowledge stored in oral traditions, regional publications, and non-indexed sources represents a largely untapped biomedical resource; LLM-based extraction pipelines could transform this resource into searchable, citable, and actionable pharmacological knowledge. Agwupuye et al. [23], in their investigation of Theobroma cacao seed extract effects on dyslipidemia and oxidative stress in L-NAME-induced hypertension, generated mechanistic pharmacological evidence that would benefit from LLM-assisted literature contextualization and systematic meta-analytic integration across related phytochemical studies.

### 6.2 LLMs in Medical Education

LLMs have demonstrated significant utility in medical education, functioning as interactive tutors, case-based learning platforms, and formative assessment tools. The consistent performance of GPT-4 and Med-PaLM 2 above medical board examination passing thresholds establishes that these systems possess the medical knowledge base to support student learning effectively [4, 19]. Studies across multiple national licensing examinations have confirmed that current LLMs can serve as interactive study partners, explain complex clinical reasoning chains, and generate realistic clinical case vignettes for practice [18]. However, the risk of hallucination in educational contexts carries a distinct long-term harm: a student who learns an incorrect pharmacological mechanism or a fabricated clinical guideline from an LLM may carry that error into clinical practice, amplifying hallucination harm beyond the immediate interaction. This concern connects directly to the cybersecurity education literature — which establishes that AI literacy, including recognition of AI failure modes, must be embedded in professional training curricula [10].

### 6.3 Cybersecurity Dimensions of Clinical LLM Deployment

The deployment of LLMs in clinical settings creates a distinct cybersecurity attack surface that intersects with every dimension of the AI-assisted clinical workflow. Ayeyemi [10] provides a systematic review of cybersecurity education in K-12 contexts, establishing foundational principles of cybersecurity literacy — including threat identification, defensive awareness, and responsible digital interaction — that become directly relevant when clinical staff interact with AI systems. The threat landscape for LLM-integrated healthcare systems includes prompt injection attacks (malicious inputs crafted to override system instructions and extract sensitive patient data), adversarial input manipulation (inputs designed to produce systematically biased clinical recommendations), and model inversion attacks (techniques to extract training data from LLM parameters, potentially exposing private patient information used in fine-tuning). Healthcare institutions deploying LLMs must incorporate cybersecurity risk assessment, staff training, and incident response planning into their AI governance frameworks [10]. The visual attention analysis methodology of Sekhri et al. [11] provides additional empirical tools for studying how clinicians interact with LLM diagnostic outputs, identifying where attention is directed, where recommendations are accepted or overridden, and how interface design can optimize safe and effective human-AI collaboration.

## 7. HALLUCINATION MITIGATION: TECHNICAL STRATEGIES

### 7.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is the most clinically validated and widely adopted technical strategy for hallucination mitigation in healthcare LLM deployments. RAG augments LLM generation by first retrieving relevant, verified information from an external knowledge base — a vector database of clinical guidelines, drug labeling, medical literature, or institutional protocols — and incorporating this retrieved context into the generation prompt [27, 28]. The retrieved information functions as a factual anchor that constrains generation to verified clinical content, preventing the model from fabricating information from parametric memory alone.

The clinical evidence for RAG effectiveness is compelling. In radiology contrast media consultation, RAG-enhanced LLM implementation completely eliminated hallucinations observed in the base model — reducing incidence from 8% to 0% — while achieving faster average response times than all evaluated cloud-based LLMs [25]. A RAG system grounding LLM responses in NICE clinical guidelines increased the rate of correct responses from 57% to 84% in a GPT-4-based system for cancer patient management; a Hepatitis C guidelines RAG system increased accuracy from 43% to 99% [27] — a 56 percentage point improvement representing the difference between an unreliable and a clinically viable system. A systematic literature review of RAG in healthcare (PRISMA guidelines, 30 peer-reviewed studies, 2022–2025) [28] identified

three primary clinical application domains with the strongest evidence base: diagnostic decision support, EHR summarization, and medical question answering. Critically, the review confirmed that hallucinations and misalignment with clinical knowledge remain significant concerns when LLMs are deployed without domain-specific grounding [28] — directly establishing that unmitigated LLM deployment in clinical settings is inadvisable regardless of model capability level.

In agentic AI applications in radiology — where LLMs operate across multi-step workflows including image report generation, clinical correlation, and follow-up recommendation — multi-agent role-based systems combined with RAG and uncertainty quantification demonstrated improved diagnostic accuracy and reduced error rates compared to single-agent implementations [29]. Multi-agent frameworks enable cross-validation through role-based specialization, where different LLM instances take on roles analogous to a primary radiologist, a quality reviewer, and a clinical correlator — each checking the others' outputs before a final recommendation is generated [29].

### 7.2 Chain-of-Thought Prompting and Structured Reasoning

Chain-of-thought (CoT) prompting — instructing LLMs to articulate intermediate reasoning steps before providing a final answer — has demonstrated consistent performance improvements across medical reasoning tasks [18]. CoT prompting reduces hallucination by forcing the model to make its reasoning process explicit, enabling both internal consistency checking and external audit by the receiving clinician. When an LLM's step-by-step reasoning chain is visible, clinicians can identify the point at which reasoning deviates from clinical reality and intervene — a transparency mechanism that directly addresses the black-box opacity concern central to responsible AI deployment in medicine [1, 2].

### 7.3 Fine-Tuning and Domain Specialization

Medical domain fine-tuning — training LLMs specifically on curated biomedical corpora using instruction tuning and reinforcement learning from human feedback with medical quality annotations — produces measurably lower hallucination rates and higher clinical accuracy than applying general-purpose models to medical tasks. Med-PaLM 2's 86.5% MedQA performance relative to GPT-4's average of approximately 81% illustrates the performance advantage of domain specialization [4]. MEDITRON, fine-tuned specifically on clinical guidelines and medical literature, demonstrates superior performance on guideline-adherent reasoning tasks [15, 17]. Domain specialization reduces hallucination rate without addressing its fundamental architectural cause — parametric generation from uncertain statistical patterns — which is why RAG and domain specialization are best understood as complementary rather than competing strategies.

### 7.4 Uncertainty Quantification and Calibration

Well-calibrated LLMs express appropriate uncertainty when their knowledge is limited or when query complexity exceeds reliable parametric knowledge. Uncertainty quantification techniques including temperature calibration, ensemble methods, and conformal prediction can be applied to LLM outputs to generate confidence intervals or reliability scores that clinicians can use to calibrate appropriate reliance. A key deployment principle derived from the human-LLM collaboration meta-analysis [13] is that LLMs should be designed to actively signal uncertainty rather than generating confident-sounding outputs across all queries — a calibration requirement directly analogous to the appropriate reliance design principles articulated in the AI medicine framework of Ayeyemi et al. [1, 2].

**Table 3.** Hallucination Mitigation Strategies, Mechanisms, and Supporting Clinical Evidence.

| Strategy | Mechanism | Hallucination Reduction | Source |
|---|---|---|---|
| Retrieval-Augmented Generation | Grounds generation in verified external knowledge | 8%→0% (radiology); 57%→84% (NICE QA) | [25, 27, 28] |
| Chain-of-Thought Prompting | Forces explicit intermediate reasoning | Significant NML exam improvement | [18] |
| Medical Domain Fine-tuning | Curated biomedical corpus training | Med-PaLM 2: 86.5% vs GPT-4: ~81% | [4, 17] |
| Multi-Agent Cross-Validation | Role-based LLM ensemble checking | Improved accuracy; reduced single-model errors | [29] |
| Uncertainty Quantification | Explicit confidence signaling | Reduces uncritical over-reliance | [13] |
| Human-in-the-Loop Oversight | Mandatory clinician review of all outputs | Prevents autonomous error propagation | [1, 30] |

## 8. REGULATORY LANDSCAPE AND GOVERNANCE FRAMEWORKS

### 8.1 EU AI Act

The EU AI Act (Regulation 2024/1689) — the world's first comprehensive AI regulatory framework — entered into force on August 1, 2024, with phased implementation: general-purpose AI model obligations took effect August 2, 2025; high-risk system obligations apply from August 2026 [30, 31]. The Act employs a risk-based classification framework: prohibited AI uses banned from February 2025 (including social

scoring and manipulative AI); high-risk AI systems subject to stringent conformity assessment requirements; limited-risk systems subject to transparency obligations; and minimal-risk systems with voluntary compliance codes. Clinical LLMs fall within the high-risk AI category under Annex III, which explicitly covers AI systems in health and life sciences contexts [30]. High-risk clinical LLM deployments must meet requirements across: risk management systems throughout the AI lifecycle; technical documentation demonstrating safety and performance; data governance standards ensuring representative, accurate, and complete training datasets; transparency obligations; human oversight mechanisms enabling clinician override of AI outputs; and accuracy, robustness, and cybersecurity requirements [30, 31].

### 8.2 FDA Regulatory Framework

In the United States, the FDA published draft guidance in January 2025 on AI credibility in drug and biological product submissions, establishing a risk-based framework for AI model validation in regulatory contexts [32]. The FDA's June 2025 launch of 'Elsa' — an agency-wide generative AI tool for reviewer productivity — demonstrates the FDA's recognition of LLM utility alongside its regulatory caution [32]. For LLMs deployed as Software as a Medical Device (SaMD), the FDA's existing framework requires clinical validation data, performance monitoring, and adverse event reporting. The predicate device pathway and De Novo pathway provide the regulatory routes for clinical LLM products seeking market authorization, with emphasis on model credibility — demonstrating that LLM performance characteristics are sufficiently well-characterized to support claimed clinical use cases [32].

### 8.3 ESMO ELCAP Framework for Oncology

The European Society for Medical Oncology Guidance on the Use of Large Language Models in Clinical Practice (ELCAP), developed by a 20-member international panel between November 2024 and February 2025 and published in Annals of Oncology, establishes a three-tier clinical LLM classification system specific to oncology [33]. Type 1 systems are LLMs used directly by patients or healthcare professionals for general information retrieval and communication, requiring transparency disclosure and general safety standards. Type 2 systems provide clinical decision support integrated into professional workflows, requiring validation studies, performance disclosure, and institutional governance. Type 3 systems are background institutional systems integrated with EHRs for data extraction, automated summaries, and clinical trial matching, requiring pre-deployment testing, continuous bias monitoring, and re-validation when processes or data sources change [33]. This three-tier framework provides directly actionable governance architecture applicable across clinical specialties.

### 8.4 Accountability, Liability, and Cybersecurity Governance

The legal liability landscape for LLM-assisted clinical decisions remains actively contested. Current frameworks in most jurisdictions place primary liability on the treating clinician on the theory that ultimate responsibility rests with the practitioner regardless of AI assistance — a framework designed for passive clinical decision support tools, not for generative LLMs capable of producing novel clinical recommendations that clinicians may adopt with insufficient critical evaluation. A shared accountability model distributing responsibility among developers, deploying institutions, and individual clinicians provides the most coherent governance architecture for LLM clinical liability [1]. Cybersecurity governance is an essential and frequently underappreciated dimension of clinical LLM deployment. LLM-integrated clinical systems create novel attack surfaces including prompt injection vulnerabilities, adversarial input manipulation, and sensitive data exposure risks. The cybersecurity education framework articulated by Ayeyemi [10] — addressing foundational digital security literacy across institutional and individual levels — directly informs the training requirements for clinical staff deploying and interacting with LLM-based systems.

## 9. EQUITY, BIAS, AND REPRESENTATION IN CLINICAL LLMs

### 9.1 Demographic and Linguistic Bias

The training corpora of general-purpose LLMs are disproportionately composed of English-language text from high-income country contexts, creating measurable performance disparities when models

are applied to non-English clinical contexts, non-Western disease presentations, or clinical scenarios prevalent in low- and middle-income countries. Performance on US medical licensing examinations does not translate directly to equivalent performance on clinical reasoning tasks involving diseases more prevalent in Africa — including infectious diseases, parasitic conditions, nutritional deficiencies, and genetic disorders with different population frequencies — or therapeutic options constrained by drug availability in resource-limited settings [16]. For the African scientific community represented by IASS, and consistent with the equity imperatives articulated by Ayeyemi et al. [1, 2], the development and validation of LLMs trained on African clinical corpora — including locally published medical literature, African clinical guidelines, regional disease epidemiology, and indigenous language medical documentation — represents both a scientific frontier and a moral imperative. Orobator et al. [3] and Azeez et al. [22] both engage with African biomedical and ethnomedicinal knowledge systems that deserve representation in LLM training corpora and evaluation benchmarks.

### 9.2 Mitigation Strategies for LLM Bias

Bias mitigation in clinical LLMs requires interventions at four levels: data curation (deliberate inclusion of diverse clinical corpora representing non-Western, non-English, and historically underrepresented clinical contexts); evaluation (systematic performance auditing across demographic subgroups before deployment); architecture (federated learning across geographically and demographically diverse hospital systems without centralizing patient data); and governance (regulatory requirements for demographic performance disclosure as a market authorization condition). The open science principles advocated in the AI medicine literature — including participatory community engagement in dataset development, code sharing, and responsible data sharing — apply equally to LLM development for global clinical contexts.

## 10. SAFE DEPLOYMENT FRAMEWORK FOR CLINICAL LLMs

### 10.1 A Five-Pillar Deployment Framework

Synthesizing the evidence reviewed in this paper, we propose a five-pillar framework for the safe deployment of LLMs in clinical settings.

**Pillar 1: Technical Architecture — RAG and Uncertainty Quantification**

Clinical LLMs must be deployed with RAG architectures grounding all factual clinical claims in verified, regularly updated knowledge bases comprising clinical guidelines, drug labeling, and institutional protocols [25, 27, 28]. Uncertainty quantification mechanisms must actively signal LLM confidence limitations to clinicians. All outputs should be accompanied by verifiable citations linking claims to their source documents, enabling rapid verification and audit.

**Pillar 2: Human Oversight and Clinician Training**

Mandatory human review of all LLM-generated clinical recommendations is non-negotiable [30, 31]. Clinicians must receive structured training in LLM interaction, appropriate reliance calibration, and hallucination recognition. The automation bias risk — whereby clinicians over-defer to AI recommendations without adequate critical evaluation — must be explicitly addressed in training curricula [1]. Eye-tracking and visual attention analysis methodologies demonstrated by Sekhri et al. [11] provide empirical tools for evaluating how clinicians interact with LLM outputs and identifying interface design improvements that support critical engagement rather than passive acceptance.

**Pillar 3: Equity and Demographic Validation**

Pre-deployment validation must include systematic demographic performance auditing across patient subgroups defined by age, sex, ethnicity, comorbidity burden, and socioeconomic context. Performance disparities must be disclosed to deploying institutions and addressed through targeted data augmentation or deployment restriction until equity criteria are met. This is particularly critical in African and other low-resource clinical contexts where current LLM training data is most underrepresentative [3, 22].

**Pillar 4: Regulatory Compliance and Shared Accountability**

Deploying institutions must document conformity

with applicable regulatory frameworks: EU AI Act high-risk requirements for EU deployments [30], FDA SaMD pathways for US deployments [32], and ESMO ELCAP guidance for oncology applications [33]. Shared accountability frameworks distributing liability among developers, institutions, and clinicians must be established prior to deployment. Audit trail architectures should record LLM recommendations and their clinical outcomes for continuous post-market surveillance.

**Pillar 5: Continuous Monitoring and Model Governance**

Post-deployment monitoring must track hallucination rates, demographic performance parity, clinical outcome correlations, and adverse event signals on an ongoing basis. Model re-validation is required when clinical guidelines, drug labeling, or institutional protocols change. Knowledge base updates must be versioned, validated, and traceable. The cybersecurity governance framework of Ayeyemi [10] must be integrated into continuous monitoring, with regular penetration testing and adversarial input evaluation to protect patient data and system integrity.

## 11. CHALLENGES AND FUTURE DIRECTIONS

### 11.1 Current Limitations

Several critical limitations constrain current clinical LLM deployment. First, the knowledge cutoff problem: LLMs trained on data through a specific date cannot reason about clinical evidence published after that date without RAG augmentation, creating a structural gap that widens over the model's deployment lifetime as new trials, guidelines, and drug approvals accumulate [28]. Second, numerical reasoning limitations: LLMs consistently underperform on tasks requiring precise quantitative calculation — drug dosing adjustments based on weight and renal function, statistical interpretation of diagnostic test results, and survival probability calculations [15]. Third, contextual reasoning failures: LLMs excel at pattern-based reasoning but struggle with novel clinical presentations deviating from training data patterns — precisely the complex cases most in need of diagnostic support.

The systematic review of LLM evaluations in clinical medicine [16] identified substantial performance heterogeneity across medical specialties and tasks, with accuracy in broader diagnostic support applications showing variability as low as 3% in some contexts [21]. High inter-study heterogeneity prevented meta-analysis in several key reviews [21], illustrating the need for standardized evaluation methodologies before pooled clinical performance estimates can be generated with sufficient reliability to inform deployment decisions.

### 11.2 Future Research Priorities

The highest-priority research directions for clinical LLM development include: (i) prospective randomized controlled trials of LLM-assisted clinical decision-making with patient outcome endpoints, rather than benchmark performance endpoints alone; (ii) African-contextualized LLM validation, including systematic evaluation on African clinical scenarios and development of regionally representative training datasets informed by the ethnomedicinal and biomedical knowledge documented by Azeez et al. [22] and Orobator et al. [3]; (iii) multimodal LLM evaluation integrating text, imaging, genomic, and wearable data streams consistent with the framework of Ayeyemi et al. [2]; (iv) standardized hallucination measurement protocols enabling cross-study comparison and meta-analysis; and (v) long-term post-deployment safety surveillance systems generating real-world evidence on LLM-associated adverse events and near-misses in clinical settings.

## 12. CONCLUSIONS

Large language models have crossed a critical performance threshold in clinical medicine. They consistently pass medical licensing examinations across multiple national contexts [4, 18, 19], achieve diagnostic accuracy exceeding 90% on real-world critical care EHR data [5], and — when deployed as collaborative tools within structured human oversight frameworks — improve composite clinical performance metrics [13]. These achievements establish LLMs not as experimental curiosities but as deployable clinical tools with measurable patient benefit potential. The comprehensive AI medicine frameworks of Ayeyemi et al. [1, 2] situate these achievements within a broader architecture of

AI-driven clinical transformation, while the drug discovery applications of Orobator et al. [3] and the pharmacological evidence of Agwupuye et al. [23] illustrate the breadth of biomedical contexts in which LLM natural language capabilities generate translatable scientific value.

Yet the hallucination problem remains a fundamental, structurally unresolved challenge in parametric LLM deployment. Hallucination rates ranging from 14% to 95% across unmitigated models [24], compound failure modes that evade expert detection [14], and the specific clinical catastrophe potential of fabricated drug dosages and invented clinical citations collectively demand that no LLM be deployed in clinical settings without RAG-based factual grounding [25, 27, 28], human oversight mandates [30, 31], and continuous post-market surveillance. The five-pillar deployment framework proposed in this review synthesizes the evidence into actionable governance architecture. The cybersecurity principles of Ayeyemi [10] and the interface evaluation methodology of Sekhri et al. [11] provide essential institutional and human factors dimensions to this framework.

The transformative potential of clinical LLMs is real, evidence-supported, and achievable — but only through governance architectures that match the scale of their capabilities. The path to realizing that potential runs through explainability, equity, oversight, and continuous accountability rather than uncritical adoption driven by headline benchmark performance.

**CONFLICTS OF INTEREST**

The authors declare no conflicts of interest in the preparation of this review.

**References:**


1. Ayeyemi, B.M., Raji, R.O., Abdulkabir, A.O., & Shobowale, K.O. (2025). Applications of artificial intelligence in medicine: A comprehensive systematic review and meta-analysis. AROC in Biotechnology, 5(3), 23–52.

2. Ayeyemi, B.M., Shobowale, K.O., Aliu, T.B., & Abdulkabir, A.O. (2025). Multimodal artificial intelligence in medicine: Integrating imaging, genomics, electronic health records, and wearable data. BREN Journal, 1(1), 7–23.

3. Orobator, E., Nnodumele, C., Tsegay, N., Onyekwelu, P., Odenigbo, A., Ugbor, M.J., Abone, K., Ayeyemi, M.B., Inuaeyen, J., & Elechi, K. (2025). Applications of artificial intelligence in plant-based anticancer drug discovery and development. Journal of Pharma Insights and Research, 3(2), 203–210.

4. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., & Paulus, R. (2024). Toward expert-level medical question answering with large language models. Nature Medicine. DOI: 10.1038/s41591-024-03423-7.

5. Rhazes AI / Health Care Science. (2024). A systematic evaluation of the performance of GPT-4 and PaLM2 to diagnose comorbidities in MIMIC-IV patients. Health Care Science. PMC ID: 11080831. DOI as published.

6. Umapathi, N., Pal, A., & Sankarasubbu, M. (2025). Medical hallucinations in foundation models and their impact on healthcare. arXiv: 2503.05777.

7. BMC Health Services Research. (2025). Mitigating hallucinations in healthcare AI: A systematic review of evidence-based strategies (PRISMA, January 2019–April 2025; 44 eligible studies from 427 retrieved). DOI: 10.1186/s12913-026-14851-1.

8. Raheem, M., Ameen, A., Ayinla, F., & Ayeyemi, M.B. (2020). Software defect prediction using metaheuristic algorithms and classification techniques. Ilorin Journal of Computer Science and Information Technology, 3(1), 23–39.

9. Oloduowo, A.A., Raheem, M.O., Ayinla, F.B., & Ayeyemi, B.M. (2020). Software defect prediction using metaheuristic-based feature selection and classification algorithms. Ilorin Journal of Computer Science and Information Technology, 3, 23–39.

10. Ayeyemi, M.B. (2023). A systematic review of cybersecurity education in K-12 context. [Manuscript / Conference paper].

11. Sekhri, A., Kwabena, E., Ayeyemi, M.B., & Tesfay, A.H.M.H.T. (2022). Analyze and

visualize eye-tracking data. Conference Proceedings, San Francisco, United States, June 02–03, 2022.

12. Ayeyemi, B.M., Shobowale, K.O., Raji, R.O., & Abdulkabir, A.O. (2025). Artificial intelligence approaches to predict chemotherapy resistance in triple-negative breast cancer (TNBC): A multimodal integration strategy. GSC Biological and Pharmaceutical Sciences, 23(5), 45–78.

13. Nazary, F., Tremblay, M., Dauteuil, R., & Abbasgholizadeh-Rahimi, S. (2025). Human-large language model collaboration in clinical medicine: A systematic review and meta-analysis. PROSPERO CRD420251068272. PMC. DOI as published.

14. Ansari, S. (2026). Compound deception in elite peer review: A failure mode taxonomy of 100 fabricated citations at NeurIPS 2025. arXiv: 2602.05930.

15. Springer Nature / Health Information Science and Systems. (2025). Reasoning with large language models in medicine: A systematic review of techniques, challenges, and clinical integration. DOI: 10.1007/s13755-025-00403-0.

16. BMC Medical Informatics and Decision Making. (2025). A systematic review of large language model (LLM) evaluations in clinical medicine (2019–2025; 1,534 LLM evaluation instances; 93.55% general-domain; 6.45% medical-domain). DOI: 10.1186/s12911-025-02954-4.

17. Computers, Materials and Continua / Tech Science Press. (2025). Transforming healthcare with state-of-the-art medical-LLMs: A comprehensive evaluation using a benchmarking framework (GPT-4Med, Med-PaLM, MEDITRON, PubMedGPT, MedAlpaca). Published December 9, 2025.

18. medrxiv. (2025). Textbook-level medical knowledge in large language models: A comparative evaluation using the Japanese National Medical Licensing Examination (GPT-4o 89.2%; DeepSeek-R1 92% China NML; 95% reliability threshold for clinical application). DOI: 10.1101/2025.09.10.25335398.

19. Hannula, J., Ahvenjarvi, L., Ylinen, A., Saraste, A., & Maunula, M. (2024). Large language models take on cardiothoracic surgery: A comparative analysis of GPT-3.5, GPT-4, Med-PaLM 2, and Claude 2 on SESATS examination questions. PMC ID: 11337141.

20. Benito, P., Isla-Jover, M., Gonzalez-Castro, P., et al. (2025). Comparative analysis of multimodal large language models GPT-4o and o1 versus clinicians in clinical case challenge questions: Retrospective cross-sectional study (1,426 Medscape cases, May 2011–June 2024). PMC ID: 12851745.

21. García-Méndez, S., Pascual-Guardia, S., & Alsina-Restoy, X. (2025). A systematic review of large language models in medical specialties: Applications, challenges, and future directions (84 studies, January 2021–March 2024). MDPI Information, 16(6), 489. DOI: 10.3390/info16060489.

22. Azeez, A.A., Azeez, A.T., & Ayeyemi, B.M. (2025). From nuisance to necessity: Documenting the ethnomedicinal importance of weeds in Ondo State Nigeria. Ife Journal of Science, 27(1), 11–20.

23. Agwupuye, E.I., Agboola, A.R., Kuo, Y.C., Itam, A.H., Ezeayinka, L.U., Atangwho, I.J., Ayeyemi, B.M., Edema, A.A., Lawal, B., Alotaibi, M.O., & Almohmadi, N.H. (2025). Theobroma cacao seed extracts attenuate dyslipidemia and oxidative stress in L-NAME induced hypertension in Wistar rats. International Journal of Medical Sciences, 22(16), 4509.

24. Xu, Z., Wang, Z., & Sun, M. (2026). BibTeX citation hallucinations in scientific publishing agents: Evaluation and mitigation (13 LLMs; hallucination rates 14–95% across vendors). arXiv: 2604.03159.

25. Chelli, M., Descamps, J., Lavoue, V., Trojani, C., Azar, M., Deckert, M., & Boileau, P. (2025). Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation (RAG: 8% → 0% hallucination; Llama 3.2-11B vs GPT-4o mini, Gemini 2.0 Flash, Claude 3.5 Haiku). npj Digital Medicine. DOI: 10.1038/s41746-025-01802-z.

26. IntuitionLabs. (2026). LLM hallucinations in pharma: MOA errors and fake trials (mechanism-of-action errors, fabricated clinical trial results, incorrect drug interactions, invented regulatory citations). Retrieved April 2026 from https://intuitionlabs.ai/articles/llm-hallucinations-pharma-clinical-trial-errors.

27. Patel, S., & Shah, J. (2025). Grounding large language models in clinical evidence: A retrieval-augmented generation system for querying UK NICE clinical guidelines (GPT-4 RAG: 57% → 84% cancer management; Hepatitis C: 43% → 99%). arXiv: 2510.02967.

28. Neha, F., Bhati, D., & Shukla, D.K. (2025). Retrieval-augmented generation (RAG) in healthcare: A comprehensive review (PRISMA; 30 peer-reviewed studies; 2022–2025; diagnostic support, EHR summarization, medical QA). AI, 6(9), 226. DOI: 10.3390/ai6090226.

29. Agentic AI Radiology Review Group. (2025). Agentic AI and large language models in radiology: Opportunities and hallucination challenges (2024–2025; multi-agent role-based systems + RAG + uncertainty quantification). PMC ID: 12729288.

30. European Parliament and Council. (2024). Regulation (EU) 2024/1689 — The Artificial Intelligence Act. Official Journal of the European Union. In force: August 1, 2024; GPAI obligations: August 2, 2025; high-risk system obligations: August 2026. Available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

31. Team Consulting. (2025). Regulation of AI in healthcare: Navigating the EU AI Act and FDA (risk-based approach; high-risk AI category; conformity assessment; post-market surveillance). Retrieved June 2025 from https://www.team-consulting.com/us/insights/regulation-of-ai-in-healthcare-navigating-the-eu-ai-act-and-fda/.

32. IntuitionLabs. (2026). GenAI in medical affairs: Use cases and compliance guardrails (FDA January 2025 draft guidance on AI credibility; FDA 'Elsa' tool June 2025; EU AI Act enforcement timelines). Retrieved April 2026 from https://intuitionlabs.ai/articles/genai-medical-affairs-compliance.

33. Wong, E.Y.T., Verlingue, L., Aldea, M., et al. (2025). ESMO guidance on the use of large language models in clinical practice (ELCAP): A three-tier framework for oncology (Type 1–3 classification; 20-member international panel; November 2024 – February 2025). Annals of Oncology.