# Hallucinations of Large Language Models in Medical Environments: A Systematic Review of Risks, Detection, and Mitigation

*Shengxuan* Huang[1*]

[1]School of Computer Science, University of Technology Sydney (UTS), Sydney, Australia

**Abstract.** Large language models (LLMs) are demonstrating transformative potential in medical informatics, assisting with tasks ranging from diagnostic reasoning to patient communication. However, their propensity to generate confident yet unfounded outputs—termed hallucinations—poses significant risks to patient safety and clinical accountability. This paper presents a systematic literature review of research from 2023 to 2025, analyzing the risks, benchmarks, detection paradigms, and mitigation strategies associated with medical hallucinations. The paper synthesizes our findings into a novel evaluative framework, $CR^2$ (Capability × Reliability × Cost × Clinical Risk), designed to guide risk-aware adoption. Our analysis confirms that hallucination is a structural property of autoregressive text generation under uncertainty. Consequently, we argue that hybrid control—integrating retrieval grounding, verification mechanisms, calibrated generation, and human oversight—constitutes the most credible path toward trustworthy deployment. The review concludes by identifying critical open challenges, including the need for harm-weighted evaluation, multilingual generalisability, and operational governance mechanisms.

## 1 Introduction

The integration of large language models (LLMs) into clinical workflows marks a significant shift in healthcare delivery, offering unparalleled efficiency in processing complex medical information. Models like ChatGPT, Med-PaLM 2, and Gemini now power clinical chatbots, decision-support systems, and documentation tools, achieving performance on professional exams that rivals human experts [1]. However, this promise is tempered by a fundamental flaw: their remarkable linguistic fluency can mask a deep-seated epistemic brittleness. These models can fabricate medical entities, cite non-existent evidence, or assemble individually true facts into dangerously misleading clinical advice [2].

In clinical domains, where verifiability and traceability are non-negotiable, such hallucinations introduce profound safety, liability, and governance concerns. Documented instances of models inventing medical terminology or providing confidently incorrect recommendations highlight a critical gap between linguistic plausibility and factual validity

* Corresponding author's email: Shengxuan.Huang-1@student.uts.edu.au

[3]. The pressing question, therefore, is not whether hallucinations occur, but how to effectively bound and govern them to ensure safe utility.

This systematic review addresses three core research questions (RQs):

• RQ1: What are the salient risks of LLM hallucinations across diverse medical environments?

• RQ2: How can these hallucinations be effectively detected, quantified, and characterized?

• RQ3: Which mitigation strategies show the greatest promise for real-world clinical integration?

By synthesizing recent literature, we highlight areas of consensus and contention, and propose the $CR^2$ framework to bridge the often-disconnected considerations of model capability, reliability, operational cost, and clinical risk. Our contribution is threefold: (1) a systematic analysis of the current landscape of medical hallucinations, (2) the introduction of the $CR^2$ framework for risk-aware evaluation, and (3) the translation of technical insights into practical deployment playbooks for clinical settings.

## 2 Background and Related Work

### 2.1 Defining the Spectrum of Medical Hallucinations

In the context of LLMs, "hallucination" describes the generation of coherent text that is unfaithful to source knowledge or unverifiable against authoritative evidence. We adopt a pragmatic, evidence-based definition crucial for medicine: an output is hallucinatory if it is unfaithful to available evidence or unverifiable against trusted sources like clinical guidelines or drug databases. This conceptualization distinguishes true hallucinations from simple factual errors by three characteristics: epistemic confidence (presented with unwarranted certainty), contextual plausibility (appearing reasonable within the clinical scenario), and verifiability failure (contradicting established medical knowledge).

In medical practice, hallucinations manifest across a graduated spectrum:

• Factual Errors: Representing basic failures in knowledge recall (e.g., incorrect drug-dosage pairings, diagnostic criteria).

• Reference Errors: Involving fabricated or irrelevant citations to seemingly legitimate sources.

• Reasoning Gaps: Where individually correct facts are combined through flawed clinical logic to reach dangerous conclusions.

• Cross-Modal Inconsistencies: Creating contradictions between generated text and accompanying data (e.g., imaging, lab results).

The literature increasingly treats hallucination not as a rare bug but as a structural feature of next-token prediction trained on uncertain and conflicting corpora [3]. This architectural perspective explains why scaling alone is insufficient and why systematic interventions are necessary at multiple levels of the AI stack.

### 2.2 A Multi-Dimensional Risk Taxonomy

The harms of medical hallucinations extend beyond factual inaccuracy to multi-faceted risks that vary significantly across clinical contexts:

1. Clinical Risk: Direct patient harm from misdiagnosis or inappropriate treatment plans that may alter care pathways.

2. Patient-Safety Risk: Immediate physical consequences from errors in medication dosing, procedure timing, or follow-up intervals, particularly critical in high-acuity settings.
3. Trust and Reputation Risk: Gradual erosion of clinician and patient confidence through recurrent inaccuracies, potentially causing automation bias or complete rejection of beneficial AI tools.
4. Ethical/Regulatory Risk: Complicated accountability under emerging AI governance frameworks such as WHO guidance and regional regulations like the EU AI Act, creating both clinical liability and compliance challenges [4].

Empirical studies consistently demonstrate that scale does not guarantee factuality in medical domains. On reasoning-heavy medical QA tasks, hallucination rates routinely reach 15-40% depending on task complexity, with larger models sometimes producing more sophisticated and therefore more dangerous forms of error [5]. This underscores the need for specialized approaches beyond general-purpose model improvement.

### 2.3 The Evolution of Benchmarks and Evaluation

Evaluation frameworks have evolved substantially from merely measuring multiple-choice accuracy (e.g., MedQA [5]) to more clinically meaningful assessments. Newer benchmarks like MultiMedQA incorporate expert assessment of factuality and potential harm [1], while others focus specifically on reasoning vulnerabilities (Med-HALT) or provide fine-grained error labels for precise characterization. Continuous detection frameworks (e.g., CHECK) represent another advancement, aligning model outputs with dynamically updated clinical knowledge bases to assess factual consistency against ground truth [6].

Despite these improvements, significant gaps remain. Most notable is the under-representation of extended clinical dialogue and longitudinal reasoning scenarios that characterize real-world practice. Additionally, modest inter-annotator agreement on nuanced cases reflects the context-dependent nature of "clinical correctness," where appropriateness often depends on patient-specific factors invisible to the model. Consequently, rationale-level and harm-weighted metrics are gaining traction as more clinically relevant evaluation approaches that better capture the potential impact of errors.

## 3 Detection Paradigms

### 3.1 Self-Consistency and Sampling

SelfCheckGPT and similar approaches probe model stability by sampling multiple responses to the same query and flagging divergence as a proxy for hallucination risk [7]. These methods operate on the theoretical foundation that well-grounded responses should demonstrate semantic stability across multiple sampling runs, while uncertain models exhibit higher output variance.

• Strengths: Methodologically simple and model-agnostic, requiring no additional training or external resources. Particularly effective for identifying uncertainty in factual recall tasks.

• Limitations: Inherently recall-oriented with lower precision when multiple valid reasoning paths or paraphrases exist—a common scenario in clinical medicine where different justification patterns may all lead to correct conclusions. Additionally, consistently wrong answers evade detection, creating false negatives in systematically misguided models.

### 3.2 Retrieval-Optimised Training

Methods such as HALO integrate retrieval during training and apply contrastive learning objectives that penalise inconsistency with retrieved evidence [8]. These approaches essentially teach models to align their internal representations with external knowledge sources, creating what might be termed "epistemically constrained" generation.

• Strengths: Improves first-pass reliability without heavy runtime cost, making them suitable for latency-sensitive applications.

• Limitations: Performance remains sensitive to evidence freshness; fast-moving topics (e.g., emerging clinical trial results, new treatment guidelines) can rapidly outdate training caches. The effectiveness also depends heavily on the quality and coverage of the retrieval corpus, presenting challenges for rare diseases or specialized subdomains with limited literature.

### 3.3 Knowledge-Graph Reinforcement

Approaches like Re-KGR cross-verify entities and relations against structured biomedical ontologies (e.g., SNOMED CT, UMLS, DrugBank) [9], often yielding higher precision on terminology-dense tasks such as oncology staging or pharmacovigilance. By grounding generation in formal knowledge representations, these methods can effectively prevent certain categories of entity-level hallucinations.

• Strengths: High precision for structured medical knowledge and relationship verification.

• Limitations: Practical constraints include ontological coverage gaps, mapping quality between natural language expressions and canonical concepts, and ontology drift as medical knowledge evolves. The computational overhead of real-time knowledge graph querying must also be considered in clinical deployment scenarios where response time affects usability.

### 3.4 Clinician-Guided Detection

Embedding expert rationales as reference reasoning chains represents a promising hybrid approach that improves detectors' ability to flag clinically unsafe leaps even when individual facts are technically correct. These systems shift focus from what fact is wrong to why the clinical conclusion is unsafe—a crucial advancement for addressing the most pernicious forms of medical hallucination.

• Strengths: Captures nuanced clinical reasoning patterns that purely automated methods miss.

• Limitations: Gains are task-dependent and generally positive only when high-quality rationale data exist, though creating such datasets requires significant clinical expertise and annotation effort. Recent work has explored using clinician feedback not just as training data but as reinforcement learning signals, creating models that learn to avoid reasoning patterns previously flagged as unsafe by medical experts.

### 3.5 Persistent Failure Modes and Open Challenges

Across all detection methods, several challenging failure modes persist. Soft hallucinations—outputs whose individual sentences are factually correct but collectively imply unsafe advice (e.g., emphasising rare complications while omitting first-line therapy)—remain particularly difficult to detect automatically. Multimodal cases (text contradicting imaging findings) also

present significant challenges due to alignment barriers between different data representations and domain-specific uncertainty characteristics. Additionally, most detectors struggle with temporal reasoning, failing to identify when model responses rely on outdated clinical guidelines or superseded treatment recommendations, creating a significant safety gap in rapidly evolving medical fields.

## 4 Mitigation Strategies

### 4.1 Retrieval-Augmented Generation (RAG)

RAG grounds answers in curated sources (e.g., PubMed, ClinicalTrials.gov, UpToDate) [5], with studies commonly reporting double-digit percentage reductions in hallucination rates compared to non-retrieval baselines. The effectiveness of RAG systems in clinical settings depends critically on several design choices: retrieval must be query-aware to identify relevant evidence, sources must be trustworthy and up-to-date to prevent grounding in erroneous information, and citations should be surfaced to users for provenance verification and clinical validation.

In medical contexts, specialized RAG implementations have emerged that incorporate domain-specific retrieval strategies, such as prioritizing guideline documents over general literature or implementing specialty-aware evidence selection. However, challenges remain in handling evidence conflicts where reputable sources provide contradictory recommendations, managing retrieval latency in time-sensitive clinical scenarios, and addressing cases where relevant evidence is absent or insufficient for novel clinical presentations.

### 4.2 Verifier-Enhanced Pipeline Designs

Verifier-RAG and frameworks like CHECK add explicit verification steps during or after generation [6], rejecting or revising unsupported claims. These pipelines typically deliver the largest reductions in obvious factual errors but trade off increased latency and require comprehensive schema coverage for free-text evidence interpretation.

The most effective clinical implementations often employ multi-stage verification, with initial fast checks for obvious contradictions followed by more computationally intensive reasoning validation for borderline cases. Practical deployment requires careful calibration of rejection thresholds—too aggressive rejection creates unusable systems, while too permissive approval undermines safety benefits. Several recent implementations have incorporated clinician feedback loops, allowing verification models to improve over time based on expert corrections and thereby adapting to the specific needs and standards of different clinical environments.

### 4.3 Reinforcement Learning with Structured Knowledge

Reward shaping via biomedical ontologies and knowledge graphs encourages verifiable statements and reduces speculative jumps [9]. By explicitly rewarding pathway transparency and evidence alignment, these approaches can significantly improve the factual reliability of model outputs, particularly for structured tasks like medication recommendation or diagnostic criteria application.

Benefits concentrate most strongly on terminology and relationship correctness, with more limited impact on open-ended counselling tasks or rapidly evolving evidence domains where knowledge graphs may be incomplete. Successful implementations often combine

multiple reward signals, balancing factual accuracy with clinical appropriateness and communication quality to create more comprehensively reliable systems.

### 4.4 Hybrid Control and Risk-Adaptive Triage

Operationally robust clinical stacks increasingly combine retrieval + verification + human review in adaptive architectures that respond to contextually assessed risk. A practical pattern is tiered verification: high-confidence, evidence-linked answers flow through with minimal latency; borderline cases trigger stronger retrieval and computational verification; high-risk scenarios (e.g., pediatric dosing, oncology recommendations) automatically escalate to clinician review regardless of model confidence.

This design performs particularly well under the $CR^2$ lens by dynamically balancing reliability gains with operational cost and clinical risk considerations. Several healthcare systems have implemented such approaches successfully, though they require careful workflow integration and clear responsibility assignment between AI systems and clinical staff. The effectiveness of these systems depends on sophisticated risk stratification heuristics that go beyond simple confidence scores to incorporate contextual factors including clinical specialty, acuity level, and potential harm severity.

## 5 Systematic Review Methodology

### 5.1 Search Strategy and Selection Process

As shown in figure 1, the paper conducted a comprehensive systematic search of four major databases (PubMed, arXiv, IEEE Xplore, ACM Digital Library) for works published between January 2023 and May 2025, covering the period of most intensive research on medical LLM hallucinations. Our search strategy used structured Boolean queries with key term variations: (LLM OR "large language model" OR "foundation model") AND (hallucination OR factuality OR "factual accuracy" OR reliability) AND (medical OR clinical OR healthcare OR biomedical).

The paper established explicit inclusion criteria requiring: (1) primary focus on medical/biomedical applications; (2) explicit treatment of hallucination detection or mitigation as a central theme; (3) measurable outcomes or expert-graded evaluations with reported results; and (4) empirical validation or systematic analysis rather than purely conceptual contributions. After removing duplicates, our initial search yielded 287 records. Through title/abstract screening and subsequent full-text assessment, we identified 11 studies that met all inclusion criteria for final synthesis. The PRISMA-style flow diagram below details the literature selection process.

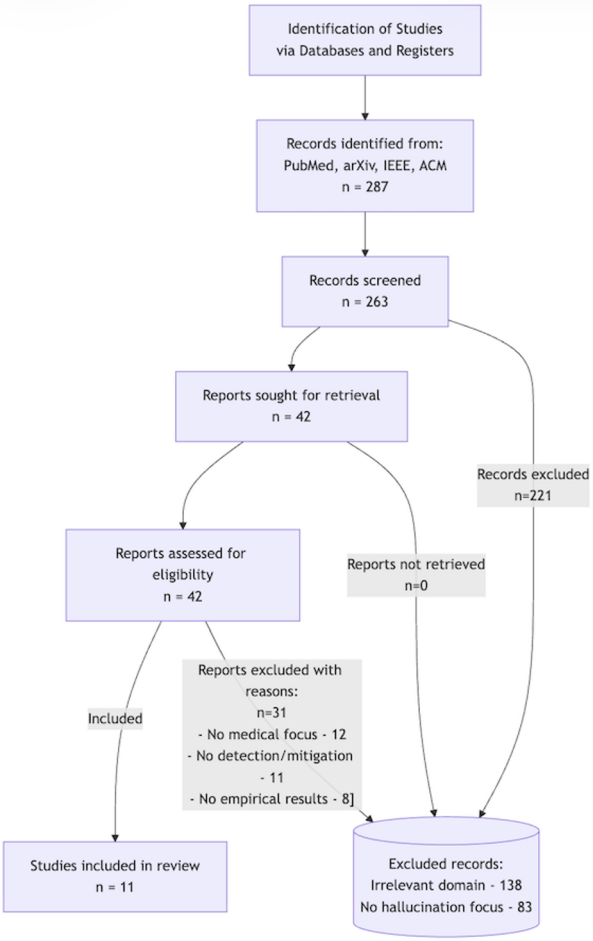


**Fig. 1.** Search strategy and selection process diagram

## 5.2 Data Extraction and Synthesis Framework

Each included study was systematically coded along four primary analytical axes:

1. Source Reliability: Evidence provenance, authority structure, and verification mechanisms.
2. Detection Design: Intrinsic (model-based) vs. extrinsic (external verification) approaches and their implementation details.
3. Mitigation Mechanism: Retrieval, reinforcement, architectural, verification, or hybrid methods and their evaluated effectiveness.
4. Clinical Applicability: Deployment constraints, safety considerations, integration requirements, and scalability factors.

We employed a qualitative synthesis methodology to compare and contrast findings across studies, identify convergent themes and divergent results, and articulate the $CR^2$ framework as an organizing principle for evaluating the trade-offs in different approaches.

### 5.3 Reproducibility Assurance and Validity Considerations

To ensure coding reliability, a double-blind coding procedure was implemented on a stratified random subsample (30% of papers), which indicated substantial inter-rater agreement (Cohen's $\kappa = 0.72$). All discrepancies were resolved through consensus discussion with a third researcher. Numeric claims lacking appropriate baselines, confidence intervals, or sample sizes were treated qualitatively to avoid over-interpretation of potentially underpowered results.

We explicitly acknowledge several threats to validity inherent in this research domain, including publication bias toward positive results, heterogeneity of evaluation metrics across studies, and the scarcity of real-world clinical trials with patient outcome measures. We note these limitations throughout our synthesis to appropriately qualify findings and conclusions.

## 6 Results and Comparative Analysis

### 6.1 Prevalence Patterns and Performance Characteristics

Our synthesis of the included studies confirms that reasoning-heavy subsets of medical QA benchmarks (MedQA [5]/MedMCQA/PubMedQA) routinely show large error bands, with hallucination rates varying from 15% to over 40% depending on task complexity and model size. A consistent finding across multiple studies is that instruction-tuning and alignment methods improve stylistic quality and user-perceived helpfulness but only modestly boost factual accuracy, underscoring the need for explicit grounding and verification mechanisms beyond parametric scaling alone.

The analysis reveals that hallucinations are not uniformly distributed across clinical tasks. Procedural knowledge and dosage calculations show particularly high error rates, while factual recall from established sources demonstrates greater reliability. This pattern suggests that mitigation strategies should be tailored to specific task types rather than employing one-size-fits-all approaches.

### 6.2 Head-to-Head Comparison of Detection Paradigms

Our comparative analysis of detection methods reveals a clear precision-recall trade-off across approaches:

- Self-consistency (SelfCheckGPT [7]): Demonstrates high recall (sensitivity) but produces more false positives when legitimate clinical variation is misinterpreted as uncertainty. Most valuable as an initial triage mechanism.
- Re-KGR [9]: Shows high precision for terminology-dense tasks but is highly dependent on ontology coverage and mapping fidelity. Limited effectiveness for open-ended clinical reasoning.
- HALO [8]: Provides balanced gains by constraining the latent space with retrieved evidence but benefits attenuate significantly with stale evidence, particularly in fast-moving specialties.
- Clinician-guided detectors: Demonstrate improved focus on clinically unsafe reasoning patterns but remain limited by the availability and scalability of expert-annotated rationale datasets.

Across all methods, soft hallucination detection remains the most significant shared challenge, with current approaches failing to consistently identify clinically misleading but factually accurate outputs.

### 6.3 Mitigation Outcomes and Practical Trade-offs

Analysis of mitigation strategies reveals distinct performance profiles:

• Verifier pipelines achieve the largest relative error reductions (often 25-40% compared to baseline) but incur substantial latency costs and engineering complexity [6].

• RAG provides robust first-order gains (typically 15-25% improvement) with greater portability across clinical domains and more manageable implementation requirements [10].

• Hybrid, tiered designs attain the most operationally acceptable trade-offs by strategically allocating computational and human resources based on assessed risk, though they require more sophisticated system architecture and workflow integration.

Our synthesis indicates that the most effective implementations combine multiple strategies in a defense-in-depth approach, with RAG providing broad coverage against factual errors, verifiers catching remaining inconsistencies, and human oversight addressing the most challenging edge cases.

## 7 Discussion

### 7.1 The Fluency–Factuality Tension in Clinical Contexts

A central paradox emerging from our analysis is that models optimized for human-like fluency and coherence are often more likely to produce dangerous hallucinations when operating without external constraints [3]. This counterintuitive relationship stems from the fundamental objective of next-token prediction, which prioritizes linguistic plausibility over factual verification. In clinical settings, where plausibility is often mistaken for correctness, this creates particularly dangerous failure modes.

This structural understanding explains why external constraints like retrieval and verification consistently demonstrate greater impact on reducing clinically significant errors than approaches focused solely on scaling model parameters or refining training data. The most effective systems appear to be those that explicitly acknowledge and compensate for this inherent tension rather than attempting to eliminate it through model-centric approaches alone.

### 7.2 CR²: An Integrative Framework for Clinical Decision-Making

We propose the CR² framework as a practical tool for bridging the gap between technical performance metrics and clinical utility considerations. The framework mandates evaluating any medical LLM application through four interdependent lenses:

• Capability: Diagnostic accuracy, multimodal comprehension, and adaptability to complex, unstructured clinical queries.

• Reliability: Factual grounding, reproducibility across similar cases, interpretability of reasoning, and appropriate confidence calibration.

• Cost: Computational latency, infrastructure requirements, and ongoing maintenance of evidence stores and verification systems.

• Clinical Risk: Potential harm severity, affected population size, and regulatory scrutiny level.

CR² enables systematic risk-stratified implementation decisions. For instance, a high-capability model with medium reliability may be appropriately deployed for literature triage or documentation assistance, while being clearly unsuitable for bedside dosing calculations or diagnostic decision support. This structured approach prevents the common pitfall of

evaluating medical AI systems through single-dimensional metrics like overall accuracy that obscure critical risk trade-offs.

### 7.3 From Ethical Principles to Operational Governance

Translating high-level ethical guidelines into practical clinical systems requires concrete implementation mechanisms:

• Evidence-Linked Defaults: Designing systems that surface provenance for every clinical claim, enabling efficient verification and correcting misconceptions about AI infallibility.

• Comprehensive Audit Trails: Maintaining detailed logs of model versions, prompts, retrieval results, and subsequent edits to support accountability and continuous improvement.

• Context-Aware Escalation Policies: Implementing mandatory human review for predefined high-risk intents (e.g., pediatrics, pregnancy, oncology, dose calculations) regardless of model confidence levels.

• Uncertainty Calibration: Ensuring models express appropriately graded confidence levels that correspond to actual likelihood of correctness, preventing overreliance on uncertain recommendations.

Successful implementation must also address the practical constraints of healthcare environments, including EHR integration challenges, variable latency requirements across clinical settings, privacy-preserving retrieval mechanisms, and operational resilience protocols for handling verifier failures or knowledge base unavailability.

### 7.4 Limitations of the Current Research Landscape

Our review identifies several persistent gaps in the current research landscape. Significant English/Western dataset bias limits understanding of how these systems perform across global healthcare contexts with different practices and resource constraints. Widespread metric inconsistency (where accuracy measures poorly correlate with safety outcomes) complicates cross-study comparisons and clinical risk assessment. The field also shows a striking scarcity of prospective live trials in actual clinical environments, with most studies conducted in simulated settings that may not capture real-world performance.

Additionally, many studies evaluate answer correctness in isolation without considering downstream clinical consequences or workflow impacts. There remains significant ethics under-specification regarding responsibility allocation when AI-assisted decisions lead to patient harm. Addressing these limitations will require broader collaboration among AI researchers, clinical specialists, healthcare administrators, and policy stakeholders.

## 8 Deployment Playbooks for Clinical Settings

Based on our synthesis of the evidence, we propose eight actionable playbooks for deploying LLMs in clinical environments:

1. Triage-First Design: Begin implementation with low-stakes tasks like routing patient inquiries or summarizing non-critical data, deliberately prioritizing calibrated uncertainty expressions over maximal fluency to establish appropriate user expectations.

2. Evidence Linking by Default: Mandate that all system outputs surface relevant PubMed IDs, guideline sections, or other provenance information, reducing over-trust and expediting human verification when needed.

3. Tiered Verification Architecture: Implement automated flows for high-confidence, evidence-linked answers while automatically triggering stronger retrieval and human-in-the-loop review for borderline or high-risk cases identified through $CR^2$ analysis.
4. Guardrail Intents with Mandatory Escalation: Predefine and always escalate high-risk domains (e.g., pediatrics, pregnancy, oncology, mental health crises, dose calculations) to qualified clinicians, regardless of model confidence metrics.
5. Proactive Drift Monitoring: Establish scheduled protocols for re-indexing knowledge sources and running change-point detection when major clinical guidelines are updated, maintaining deprecation lists for outdated advice.
6. Privacy and Auditability by Design: Implement strict data retention windows, role-based access controls, and automated redaction policies for retrieval caches, complemented by comprehensive logging of all model interactions for post-hoc audit and improvement.
7. Human Factors Engineering: Train clinical users to critically interpret model uncertainty statements and provenance cues, while designing interfaces that actively discourage blind copy-pasting into clinical notes without verification.
8. Instrumented Prospective Piloting: Begin with closely monitored, low-risk deployments to collect real-world performance data and workflow impact assessments before considering escalation to any form of direct decision support.

These playbooks operationalize the $CR^2$ framework in clinical practice: evidence linking directly enhances Reliability; triage and escalation protocols reduce Clinical Risk; and tiered routing strategically manages Cost while preserving essential Capability for appropriate use cases.

## 9 Future Directions: An Actionable Agenda

Building on our analysis of current limitations and promising approaches, we outline a focused research agenda to advance the field:

1. Develop Harm-Weighted Metrics: Move beyond simple factuality checks to create evaluation frameworks that quantify the potential clinical severity of different error types, paired with rationale-level annotation to assess the safety of the underlying reasoning process.
2. Advance Uncertainty-Aware Training: Develop novel training paradigms that explicitly reward models for expressing well-calibrated uncertainty estimates and discourage overconfident speculation in ambiguous clinical situations.
3. Implement Comprehensive Temporal-Drift Evaluation: Establish robust benchmarks and methods to test and ensure model robustness as medical guidelines evolve, creating systems that can automatically detect and adapt to outdated knowledge.
4. Build Standardized Evidence Toolchains: Create portable, interoperable verification components based on standardized interfaces for key clinical knowledge sources (e.g., PubMed, drug databases, guideline repositories) to reduce deployment barriers and enable independent safety auditing.
5. Engineer Specialized Detectors for Soft Hallucinations: Pioneer methods specifically designed to score argumentative completeness and identify clinically misleading omissions or unbalanced risk-benefit presentations that current methods miss.
6. Ground Evaluation in Clinical Reality: Develop multimodal benchmarks where text generation is grounded in pixel-level imaging or waveform evidence, and create multilingual, culturally situated datasets to ensure models generalize equitably across global patient populations.
7. Conduct Prospective Clinical Trials: Prioritize carefully designed, low-risk clinical studies that measure both process efficiency and patient outcomes in real-world settings, establishing an evidence base for responsible escalation to decision support applications.

8. Establish Auditable Governance Artefacts: Create implementable, versioned specifications for prompts, retrieval policies, escalation rules, and access controls that translate ethical principles into operational, auditable practice.

## 10 Conclusion

In conclusion, this systematic review affirms that hallucination represents a structural vulnerability inherent to the architecture and training of large language models, not a peripheral issue that can be solved through scaling alone. The most credible path forward requires a defense-in-depth strategy of hybrid control, which synergistically combines the proactive grounding of Retrieval-Augmented Generation (RAG), the rigorous fact-checking of verifier pipelines, the appropriate confidence estimation of calibrated generation, and the indispensable judgment of human oversight for residual risk.

Evaluated through the $CR^2$ framework, risk-aware, tiered deployments can achieve an optimal balance of reliability, clinical safety, and operational cost. Therefore, the ultimate goal for medical AI is not the unattainable ideal of zero hallucinations, but rather the establishment of bounded, explainable, and auditable behaviour within clearly defined clinical guardrails. This shift from prevention to governance, coupled with the strategic deployment of hybrid control systems, provides the most promising path toward realizing the transformative potential of LLMs in healthcare while ensuring patient safety remains the foremost priority.

## References

1. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. "Large language models encode clinical knowledge." *Nature*, vol. 620, no. 7972, pp. 172-180, (2023)
2. Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang. "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge." *Cureus*, vol. 15, no. 8, p. e44138, (2023)
3. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys*, vol. 55, no. 12, pp. 1-38, (2023)
4. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization, (2021)
5. D. Jin, E. Pan, N. Oufattole, W. H. Weng, H. Fang, and P. Szolovits. "What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams." *Journal of the American Medical Informatics Association*, vol. 28, no. 11, pp. 2454-2461, (2021)
6. A. Pal, D. Karkhanis, S. Dooley, M. Roberts, and S. Naidu. "FactCheck: Assessing the Factual Accuracy of Language Models in Real-World Scenarios." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. (2024)
7. P. Manakul, A. Liusie, and M. J. F. Gales. "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004-9017. (2023)

8. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. "Retrieval Augmented Language Model Pre-Training." In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, pp. 3929-3938. PMLR, (2020)
9. M. Yasunaga, J. Leskovec, and P. Liang. "LinkBERT: Pretraining Language Models with Document Links." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4003-4017. (2022)
10. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, (2020)