Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026223232

Factual Hallucination in Medical Large Language Models: Typology, Evaluation, and Systematic Governance

Zijiao Liu*

Beijing Institute of Graphic Communication, Beijing, China

**Corresponding author: Zijiao Liu.*

Abstract

Large language models (LLMs) have demonstrated transformative potential in clinical documentation generation, diagnostic assistance, and patient consultation. However, their tendency toward "hallucination"—generating semantically fluent but factually inconsistent content with established medical knowledge or input context—constitutes a core safety barrier to clinical deployment. This paper systematically reviews the typology, evaluation frameworks, underlying mechanisms, and mitigation strategies for medical LLM hallucinations. First, we propose a fine-grained hallucination classification framework based on medical knowledge graphs and establish a three-tier risk stratification scheme that references FDA medical AI software risk classification standards. Second, we analyze multi-factor causes from data, model, and inference dimensions, highlighting the unique characteristics of the medical domain. Third, we systematically review current evaluation benchmarks and methodologies, deeply analyzing the limitations of automated metrics and the potential of LLMs as evaluators. Regarding mitigation strategies, we propose a three-stage classification framework—"training-stage internal intervention—inference-stage external constraints—post-processing collaborative verification" —that provides in-depth analysis of each strategy's internal mechanisms and trade-offs. Finally, we discuss challenges in real-world clinical deployment, emphasizing that the "human-in-the-loop" paradigm remains an irreplaceable safety line.

Keywords

large language models, medical hallucination, retrieval-augmented generation, patient safety, clinical governance, risk stratification

1. Introduction

Large language models are rapidly advancing in medical applications, giving rise to specialized models such as BioBERT and Med-PaLM, demonstrating promise in clinical documentation, diagnostic reasoning, and patient education [1,2]. However, LLMs' tendency toward "hallucination"—generating semantically fluent but factually incorrect or entirely fabricated content—constitutes a core safety barrier to clinical translation [3,4].

Recent empirical studies have revealed the severity of this problem, yet reported hallucination rates vary dramatically: Asgari et al. found a hallucination rate of 1.47% in clinical text summarization [5]; oncology evaluations reported overall hallucination rates of 23%, with some open-source models reaching 48% [6]; under simulated scenarios with embedded false information, default-condition hallucination rates reached 82.7% [7]; the MHB benchmark constructed by Lu et al. showed that even the best-performing model had a 29.1% hallucination rate [8].

The roots of this variation require in-depth analysis. First, task difficulty varies significantly: clinical text summarization (1.47%) is an information compression task in which the source text provides clear factual anchors, whereas oncology Q&A (23%-48%) requires models to access extensive knowledge and perform reasoning at a much higher level. Second, testing paradigms differ: Grover et al.'s 82.7% hallucination rate came from “adversarial” testing involving active injection of false information, designed to probe the model's susceptibility to misinformation rather than to assess routine performance [7]. Third, model capabilities vary widely: the factuality gap between open-source and closed-source commercial models (e.g., GPT-4 series) can exceed 20% [6]. Fourth, evaluation stringency differs: some studies use lenient semantic similarity matching, whereas benchmarks like MHB require exact factual consistency [8]. These findings led the ECRI to rank inappropriate use of AI chatbots as the top health technology hazard for 2026 [9].

A key consensus is emerging: medical LLM hallucinations should be framed as a system-level risk management problem rather than a model-level defect that can be eliminated [4]. This paper systematically reviews current research progress on medical LLM hallucinations, with main contributions including: (1) introducing a fine-grained hallucination classification framework based on medical knowledge graphs and establishing a three-tier risk stratification referencing FDA risk classification standards; (2) deeply analyzing the unique characteristics of hallucination causes in the medical domain; (3) proposing a three-stage mitigation strategy classification framework with in-depth analysis of each strategy's internal mechanisms and trade-offs.

2. Typology, Manifestations, and Scenario-Based Risk Stratification of Medical LLM Hallucinations

2.1 Fine-Grained Hallucination Classification Based on Medical Knowledge Graphs

Existing research commonly distinguishes between factual and faithfulness hallucinations [4,10], but this binary classification is insufficient to guide the design of specific mitigation strategies. Drawing on the structured representations of medical knowledge graphs, this paper proposes a more granular classification framework.

First, Entity-Level Hallucinations refer to the model fabricating, substituting, or omitting medical entities. These can be further subdivided into: Entity Fabrication, which involves generating non-existent drugs, diseases, or anatomical structures. The “referential hallucination” documented by Kuculmez et al. falls into this category—LLMs fabricating clinical practice guidelines to support their recommendations [11]. Entity Substitution, by replacing correct entities with incorrect ones, such as substituting “insulin” for “metformin.” Entity Omission, which omits necessary entities in critical positions, such as missing core ingredients in a medication prescription.

Second, Relation-Level Hallucinations refer to incorrect representations of medical relationships between entities. These include Contraindication Relationship Errors, Causal Relationship Reversal, and Hierarchical Relationship Errors.

Third, Numerical and Attribute-Level Hallucinations involve specific clinical parameters. Numerical Errors include deviations in drug dosages, reference ranges for laboratory tests, time windows, and other quantitative information. Attribute Errors are incorrect information about drug formulations, routes of administration, frequencies of adverse reactions, and other attributes.

Fourth, Semantic and Faithfulness Hallucinations refer to outputs that deviate from the input context: Instruction Deviation is the failure to follow user-specified constraints (e.g., “please list only FDA-approved drugs”). Contextual Contradiction occurs when later dialogue contradicts earlier dialogue. Hypothetical Assumption Elaboration is a detailed elaboration on non-existent assumptions.

The advantage of this fine-grained classification framework is that different types of hallucinations may correspond to different causal mechanisms and thus require differentiated mitigation strategies. For example, entity-level hallucinations are more likely to stem from insufficient knowledge coverage in pretraining corpora and are amenable to mitigation via knowledge graph-augmented training. In contrast, faithfulness hallucinations arise more from attention diffusion during inference and are better addressed through prompt engineering or multi-agent verification.

2.2 Clinical Risk Stratification Referencing FDA Risk Classification Standards

Different medical scenarios have substantially different tolerances for hallucinations. This paper draws on the FDA's risk classification framework for medical AI software (21 CFR Part 860) to establish a three-tier risk stratification.

Errors in high-risk scenarios may lead to death or serious injury, the FDA's most stringent regulatory tier. These include: medication dose calculation, contraindication screening, emergency triage, and critical care diagnostic recommendations—direct clinical decisions. Chu et al.'s adversarial attack experiments showed that LLMs are highly susceptible to adversarial attacks that induce harmful hallucinations in clinical decision support, with simulated attack success rates reaching 88% [13]. In such scenarios, autonomous LLM output should be prohibited, and any generated content must undergo mandatory verification by clinical experts, with models requiring external validation before deployment consideration.

Errors in moderate-risk scenarios may cause non-serious injury, and the risk can be reduced through additional safeguards. These include: assisted diagnostic reasoning, treatment planning recommendations, and preliminary radiology report screening. Cheung et al. assembled 27 clinical experts with a mean of 25.9 years of practice to conduct a blinded evaluation of 4,795 LLM responses, finding that 3%-19% of responses were rated “unacceptable” by expert panels [14]. Jung concluded that existing evidence supports an assistive rather than autonomous role for LLMs [1]. These scenarios require LLM output to be verified by clinicians and should be equipped with specialized hallucination detection systems as safety nets.

Errors in low-risk scenarios do not pose significant patient risk. These include: clinical documentation draft generation, patient education material writing, medical literature summarization, and appointment scheduling assistance. Asgari et al. showed that, with prompt and workflow optimization, LLM error rates in documentation scenarios can be reduced to levels below human error rates [5]. These scenarios can rely primarily on inference-stage strategies with routine review.

The advantage of this stratification framework is its alignment with established medical device regulatory systems, providing operational risk threshold references for clinical deployment and a basis for technology strategy combinations across different risk levels.

3. Quantitative Evaluation of Hallucinations: Benchmarks, Methods, and Dilemmas

3.1 Major Evaluation Benchmarks

Specialized evaluation benchmarks for medical LLM hallucinations have developed rapidly. The MHB benchmark, constructed by Lu et al., systematically injects “hallucination traps” into multi-turn medical conversations—embedding fabricated physical examination values, non-existent diagnostic codes, and fabricated literature citations—validated by 60 licensed physicians to generate 4,695 samples. Results showed that even the best-performing models exhibited systematic defects when encountering fabricated data [8].

The MedHallu benchmark by Pandit et al. contains 10,000 Q&A pairs, generating hallucinated responses through controlled processes, with top models like GPT-4o achieving F1 scores of only 0.625 on hard-class detection tasks [15]. Gu et al. proposed MedVH for multimodal scenarios, covering six task types to probe both visual and textual hallucinations [16].

3.2 Limitations of Automated Metrics and Potential of LLM-as-an-Evaluator

While computationally efficient, automated NLP metrics exhibit systematic biases in medical contexts. There are three limitations of traditional metrics: First, lexical bias of BLEU/ROUGE—overemphasizing n-

gram overlap at the expense of clinical accuracy, potentially rating surface-fluent but factually incorrect responses as high-quality while penalizing responses that use synonyms for more accurate expression [16]. Second, the semantic vagueness of BERTScore—although capable of capturing semantic similarity, it is insensitive to precise matching of medical entities, unable to distinguish the critical difference between "metformin for type 2 diabetes" and "metformin for type 1 diabetes." Third, factual inconsistency metrics (e.g., QUEST, ALIGN) have made progress but are mostly based on general-domain natural language inference models, which exhibit unstable performance in medical reasoning.

Recent research has explored using more powerful LLMs (e.g., GPT-4) as evaluators to supplement or assist human evaluation. Preliminary evidence suggests LLMs can achieve moderate agreement with expert evaluation on specific tasks [14]. However, this approach faces a fundamental dilemma: using an LLM to evaluate LLMs effectively means "detecting hallucinations with hallucinations." If the evaluator model itself has medical knowledge gaps or factuality biases, its judgments will be unreliable. Cheung et al.'s Medieval framework offers a compromise: employing structured scoring systems that map response quality to physician career stages, using LLMs for preliminary screening and pre-annotation, followed by tiered human expert review, achieving excellent intraclass correlation coefficients >0.9 across 13 specialties [14].

Development of medical-specific faithfulness evaluation metrics is needed, incorporating structured constraints from medical knowledge graphs to enable multi-level fact verification at entity, relation, and attribute levels. Concurrently, a hybrid evaluation pipeline of "automated pre-screening → LLM-assisted annotation → tiered expert review" should be established to enhance scalability while maintaining evaluation quality.

3.3 The Evaluation Trilemma

Current evaluation faces a trilemma: insufficient scenario coverage—existing benchmarks focus on specific task formats, leaving many clinical scenarios unaddressed [8,15]; timeliness issues—static benchmarks fail to capture model performance in the context of continuously evolving medical knowledge [4]; unresolved clinician-engineer evaluation gap—engineers focus on statistical metrics while clinicians focus on patient safety consequences, and the mapping between them remains unclear.

4. Deep Causal Mechanisms of Hallucinations: Medical Domain Specificity

The causes of hallucinations span the entire AI lifecycle and can be understood from three dimensions. However, unlike general-domain hallucination analyses, this section emphasizes specific characteristics of the medical domain.

4.1 Data Layer: Structural Dilemmas of Medical Data

General-domain pretraining corpus quality issues are amplified in the medical domain. First, privacy constraints on medical data make high-quality annotated data extremely costly to obtain. Patient records are protected by regulations such as HIPAA and GDPR, limiting the scale of publicly available de-identified clinical corpora, forcing researchers to trade off between data quantity and privacy compliance.

Second, conflicts and the evolution of medical guidelines present challenges absent in general domains. Different professional societies may give inconsistent recommendations for the same clinical question (e.g., starting age for breast cancer screening), and the update cycle of guidelines (typically 3-5 years) structurally mismatches the knowledge cutoff of LLM training [17]. Models may learn "outdated mainstream views" rather than "current best practices."

Third, extreme long-tail distribution: data on rare diseases and special populations (pediatrics, pregnancy, elderly with multimorbidity) are extremely sparse. Model "knowledge" of these scenarios may come from a small number of case reports or from entirely fabricated accounts.

4.2 Model Layer: Structural Conflict Between Generation Objectives and Causal Reasoning

The limitations of the autoregressive architecture discussed in general domains manifest more severe consequences in medical scenarios. Maximum likelihood training optimizing statistical fluency may directly conflict with causal reasoning requirements in clinical decision-making: a statistically frequent "symptom-

diagnosis" co-occurrence may represent an incorrect causal association (e.g., "cough → lung cancer" holds statistically, but the vast majority of coughs are not lung cancer).

Causality requirements of medical reasoning are absent in general domains. Clinical decision-making requires interpretable causal chains ("because the patient has X, tests show Y, therefore diagnosis is Z, thus treatment W is recommended"). In contrast, autoregressive models fundamentally learn conditional probability distributions P(next token preceding context), not causal reasoning. When asked "why," the model's generated "explanation" may be post-hoc plausibility fabrication rather than genuine reasoning pathways [4].

Furthermore, expression of diagnostic uncertainty, acceptable in general dialogue, is critically important in medical contexts. Existing RLHF preference optimization tends to reward "confident, certain" responses while suppressing expressions of uncertainty, such as "I'm not sure" or "more information is needed" [19], leading models to reach definitive conclusions when the evidence is insufficient.

4.3 Inference Layer: Multi-Evidence Integration and Attention Decay

In complex reasoning requiring the integration of multiple variable evidence across steps (e.g., comprehensive diagnosis based on history, physical examination, and laboratory results), attention mechanisms diffuse as sequence length increases, leading to functional "forgetting" of critical abnormal findings that appear early in the sequence [4].

Under-specified prompts are common in medical scenarios—patient chief complaints are often incomplete, and clinicians need to ask follow-up questions to obtain critical information. However, LLMs tend to "fill in the gaps" rather than actively inquire when information is insufficient, a "helpful" fine-tuning preference that proves dangerous in medical contexts [18]. All the above causes exhibit cascading amplification. Privacy restrictions → reduced available corpora → insufficient long-tail knowledge coverage → model gap-filling → hallucination generation. This systematic characteristic is overlooked by general-domain hallucination research.

5. Hierarchical Mitigation Strategy System: Internal Mechanisms and Cross-Comparisons

Based on the preceding causal analysis, this paper proposes a three-stage classification framework of "training-stage internal intervention—inference-stage external constraints—post-processing collaborative verification." Unlike the simple listings in Sections 5.1-5.3, this section provides in-depth analysis of the internal mechanisms for each strategy type and conducts cross-comparisons both within and across categories.

5.1 Training Stage: Mechanisms and Trade-offs of Internal Intervention

Training-stage strategies improve internal knowledge representations by modifying model parameters or training objectives. Traditional RLHF optimizes human preferences (Helpful-Honest-Harmless), but "honesty" is defined differently in medical contexts than in general dialogue. Wang et al.'s HiMed-RL decomposes reward learning into three levels: token-level fluency (avoiding grammatical errors), concept-level factuality alignment (entity correctness), and semantic-level diagnostic consistency (plausibility of reasoning paths) [20]. The internal logic of this hierarchical design is that upper-level rewards become learnable signals only after lower-level rewards stabilize. Key trade-off: overly aggressive factuality alignment may cause models to refuse to answer in uncertain scenarios (reduced utility), while excessive conservatism fails to suppress hallucinations effectively.

Deng et al. use MedKA to transform the UMLS knowledge graph into Q&A corpora for fine-tuning [21]. The core mechanism converts knowledge graph triples (head entity-relation-tail entity) into natural language templates ("Q: What is the recommended treatment for X? A: Y"). This approach's advantage is maintaining consistency between fine-tuning supervision signals and the pre-training format. Inherent limitation: knowledge graphs themselves suffer from incompleteness and lag in updates. If injected knowledge is itself controversial (e.g., conflicting recommendations from different guidelines), the model learns inconsistent mappings.

HIME identifies hierarchical hallucination sensitivity (differences in encoding strength across network layers for different fact types) and applies targeted weight modifications [22]. Its technical principle: in causal

tracing experiments, encoding of specific facts often concentrates on a small number of neurons; modifying these neurons' weights can precisely correct erroneous facts without affecting others' knowledge. Key trade-off: scalability of editing operations—each erroneous fact requires one editing operation, making this method difficult to deploy at scale, given the tens of thousands of fact entries in clinical knowledge bases.

RLHF improves generation preferences (models are more likely to output factual content), knowledge graph-augmented training injects specific facts (models "know" more correct answers), and model editing corrects specific errors (precisely fixing known problems). Their input costs, scalability, and applicable scenarios differ significantly. Under resource constraints, the priority order should be: knowledge augmentation (low cost, high return) → RLHF (medium cost, improves generalization) → model editing (high cost, for critical error patching).

5.2 Inference Stage: Mechanisms and Vulnerabilities of External Anchoring

Inference-stage strategies apply external constraints during generation without modifying model parameters. RAG anchors generation to retrieved external documents, theoretically reducing hallucination rates to the lower bound of the retrieval source's factuality error. Jiang et al.'s MEGA-RAG integrates multi-source evidence with difference-aware refinement, reducing hallucination rates by over 40% [23]. Internal mechanism: RAG's effectiveness depends on the smooth functioning of three components—query parsing (converting user questions into effective retrieval queries), retrieval ranking (recalling relevant documents from knowledge bases), and generation fusion (integrating retrieved content with the model's internal knowledge). Failure at any component causes RAG to degenerate to pure generation.

RAG cannot resolve models' insufficient reasoning capabilities. If models cannot integrate multiple retrieved evidence pieces causally (as described in Section 4.2), even when correct answers are retrieved, they may selectively ignore or misinterpret retrieved content. More seriously, Johnson et al. found that RAG can increase hallucination rates under specific clinical text configurations—when retrieved documents contain outdated or inconsistent information, models may select the wrong one among multiple conflicting sources [24]. This suggests that RAG is not a universal solution for hallucinations; its effectiveness depends on retrieval knowledge source quality control and on models' multi-source evidence integration capabilities.

MedKP queries knowledge graphs in real time during generation, applying structured constraints to decoder logits. If the tokens the model is about to generate violate known relationships in the knowledge graph, their probabilities are suppressed [25]. The essential difference from RAG: RAG provides "reference text" from which models may deviate, whereas knowledge graph constraints provide "hard boundaries"—models cannot generate outputs violating known medical relationships. Applicability boundary: knowledge graph constraints require the target domain's knowledge to be formalizable as deterministic logical rules. This is effective for drug contraindications (clear rules), but inadequate for diagnostic reasoning that requires probabilistic judgments ("symptom X points to disease Y with 80% probability").

Grover et al. found that simply reminding users that "the input may contain inaccurate information" reduced average hallucination rates across six LLMs from 66% to 44% [7]. The primary mechanism is that prompts adjust the model's uncertainty threshold, making it more likely to acknowledge uncertainty rather than forcefully answering when unconfident. However, this mechanism's vulnerability is that Wang et al. found that paraphrasing the prompt (e.g., synonymous rewording) can cause substantial performance fluctuations, and prompt robustness is worse on complex reasoning tasks [18]. Prompt engineering cannot inject medical knowledge that models don't already possess; it can only adjust expression strategies for the knowledge they already possess.

RAG has the richest knowledge sources but suffers from multi-source evidence fusion reliability issues; knowledge graph constraints are most reliable (hard boundaries) but have limited coverage; prompt engineering is the lightest but most vulnerable. In practical deployment, priority should be: knowledge graph constraints (highest priority) → RAG (primary) → prompt engineering (fallback).

5.3 Post-Processing Stage: Mechanisms and Costs of Collaborative Verification

Post-processing strategies apply external verification after generation completion, detecting and correcting hallucinations through redundancy and cross-checking. Multi-agent systems deploy multiple role-specific

LLM agents (e.g., diagnosis agent, evidence verification agent, treatment plan agent), each generating outputs independently and reaching consensus through voting or debate. MedAide achieves state-of-the-art results across four medical benchmarks through syntactic constraint decomposition and rotating-agent collaboration [26]. Internal mechanism: effectiveness depends on two assumptions—(1) hallucination patterns across agents are mutually independent (errors made by one agent are unlikely to be simultaneously made by others); (2) consensus mechanisms can correctly identify correct outputs. The first assumption's validity in practice is questionable—if all agents share identical pretrained weights (e.g., multiple GPT-4 instances), their error patterns are highly correlated, significantly attenuating cross-validation effectiveness. Ideal multi-agent systems should use diverse models with different architectures and training data.

CHECK integrates structured clinical databases with information-theoretic classifiers [27]. Its detection principle: comparing each clinical assertion generated by the model against facts in knowledge bases to compute semantic consistency scores; simultaneously, analyzing information-theoretic features during generation (e.g., token prediction entropy, mutual information) to identify segments generated by "guessing" rather than "confidence." Garcia-Fernandez et al. reported that CHECK reduced Llama3.3-70B-Instruct hallucination rates from 31% to 0.3% [27]—a key deployment consideration: the trade-off between true positive rate and false positive rate. Lowering detection thresholds captures more hallucinations (high recall) but increases false positives on correct outputs (low precision), potentially causing the "crying wolf" effect, where clinicians ignore warnings. In clinical deployment, thresholds should be dynamically adjusted based on risk level—high-recall priority for high-risk scenarios (better false alarms than missed detections), with moderate precision for low-risk scenarios.

Multi-agent collaboration achieves robustness through redundancy but incurs linearly increasing computational costs with the number of agents; specialized detection systems are more computationally efficient but depend on the completeness and timeliness of external knowledge bases. The two approaches can complement each other: detection systems as first-line defense for efficient screening, with suspicious samples referred to multi-agent systems for in-depth verification.

Table 1: Comparison of the three-stage classification system for medical LLM hallucination mitigation strategies

Intervention Stage	Core Mechanism	Representative Methods	Technical Advantage	Fundamental Limitation	Deployment Cost	References
Training Stage	Modify internal representations	HiMed-RL, MedKA, HIME	Improves factuality at the source	Resource-intensive, knowledge update lag	High	[20-22]
Inference Stage	External anchoring	MEGA-RAG, MedKP, Prompt Engineering	Dynamic knowledge access, no retraining needed	Knowledge source quality dependence, prompt fragility	Medium	[7,23-25]
Post-Processing	Output verification	MedAide, CHECK	Adds redundant safety layer, decoupled from model	Computational overhead, increased latency	Medium-High	[12,26,27]

6. Discussion: Real-World Clinical Deployment Challenges and Governance Pathways

6.1 Synergistic Effects of Complementary Strategies and Deployment Plans

No single strategy can comprehensively address medical hallucinations. RLHF-aligned models (training stage), supplemented by knowledge graph constraints and RAG (inference stage), with post-processing verification by specialized hallucination-detection systems. Based on the FDA risk stratification framework, differentiated deployment plans are proposed: Deploy all three strategy layers with mandatory human verification. Set the detection system's miss rate to the minimum (prioritize capturing all potential hallucinations); any flagged output must undergo clinical expert review before being passed to users. In moderate-risk scenarios, combine inference-stage anchoring (RAG + knowledge graph constraints) with post-processing verification (detection systems). Use balanced detection thresholds, with flagged outputs

undergoing clinician sampling review (e.g., 10% random samples + all flagged samples). In Low-risk scenarios, rely primarily on inference-stage strategies (prompt engineering + RAG) with automated detection and routine review (e.g., daily aggregated review). Human review ratios can be reduced to 1%-5%.

6.2 Core Challenges in Real-World Clinical Deployment

Post-processing strategies (multi-agent collaboration, specialized detection) introduce additional latency (typically 2-10 seconds). In time-sensitive scenarios such as emergency triage, this latency may be unacceptable. A solution is to establish tiered response mechanisms: time-sensitive operations use only inference-stage strategies, while non-time-sensitive operations can undergo post-processing verification.

The effectiveness of RAG and knowledge graph constraints depends on continuous updates to the knowledge sources. Clinical knowledge undergoes paradigm updates every 3-5 years (e.g., adjustments in hypertension diagnostic thresholds), while drug information requires quarterly updates. This requires medical institutions to establish dedicated knowledge engineering teams—a significant barrier for most.

When clinicians make decisions based on LLM hallucinations leading to adverse events, how should legal responsibility be allocated? Existing legal frameworks lack clear precedents. One discussed approach is a "product liability + professional liability" hybrid framework: model providers are responsible for foreseeable, insufficiently warned systematic hallucinations; clinicians are responsible for negligence in failing to perform reasonable review processes [1].

6.3 "Human-in-the-Loop" Paradigm: Irreplaceable Safety Bottom LINE

The ECRI recommends treating AI outputs as supplementary information requiring critical review [9]. This requires cultivating clinicians' AI literacy—leveraging AI efficiency advantages while maintaining critical vigilance to identify potential hallucinations. Clinical trial evidence shows that experienced clinical experts can raise LLM qualification thresholds to acceptable levels [14], but this depends on two prerequisites: physicians are informed that outputs may contain hallucinations. Physicians possess sufficient domain expertise for independent judgment. Evaluating the effectiveness of the combined strategy in real-world clinical simulations, constructing dynamically updated evaluation benchmarks, and establishing clear accountability and regulatory frameworks. The goal of trustworthy medical AI is not to achieve unattainable zero hallucinations, but to construct systematic mechanisms to anticipate, detect, constrain, and manage hallucination risks.

7. Conclusion

This paper systematically reviews the typology, evaluation frameworks, causal mechanisms, and mitigation strategies for medical LLM hallucinations. Hallucinations are multifaceted phenomena with incidence rates ranging from 1.5% to over 80%, with variation attributable to multiple factors including task difficulty, testing paradigms, model capabilities, and evaluation stringency. Current models lack safety guarantees for autonomous clinical deployment. The causes of hallucinations in medical scenarios exhibit unique characteristics—privacy constraints, guideline conflicts, evolution, and structural conflicts with causal reasoning—that suggest general-domain mitigation strategies cannot be directly transferred.

Three-stage classification framework and mechanistic analysis demonstrate that: training-stage strategies improve knowledge representations at the source but are resource-intensive; inference-stage strategies provide dynamic anchoring but depend on external knowledge quality; post-processing strategies add safety redundancy but introduce latency. The three are complementary rather than substitutable, and should be differentially based on clinical risk levels. High-risk scenarios require mandatory human verification, moderate-risk scenarios require physician supervision, and low-risk scenarios can moderately rely on automation. The "human-in-the-loop" paradigm remains an irreplaceable safety bottom line.

Future research should prioritize evaluating the effectiveness of combined strategies in real-world clinical simulations, constructing dynamically updated evaluation benchmarks, and establishing clear accountability and regulatory frameworks.

References

[1] Jung, K. H. (2025). Large language models in medicine: clinical applications, technical challenges, and ethical considerations. Healthcare Informatics Research, 31(2), 114-124.

[2] Li, S., Wang, X., Chen, Y., Tian, M., Lin, P., Lai, M., & Jiang, L. (2026). Large language models for primary care ophthalmic education: a systematic review. *Frontiers in Medicine*, *13*, 1810098.

[3] Zhang, R. (2025). Classification and solution of large language model hallucination. In *ITM Web of Conferences*(Vol. 80, p. 01035). EDP Sciences.

[4] Passban, P., Matthews, A., Roosta, T., & Vankadaru, V. (2026). Rethinking Medical LLM Hallucinations: A System-Level Survey.

[5] Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ digital medicine*, *8*(1), 274.

[6] Yoon, S. M., Lyu, J., Djunadi, T. A., Song, J., Kim, H. S., Min, R. S., ... & Chae, Y. K. (2025). Navigating artificial intelligence (AI) accuracy: A meta-analysis of hallucination incidence in large language model (LLM) responses to oncology questions.

[7] Prandner, D., Wetzelhütter, D., & Hese, S. (2025, January). ChatGPT as a data analyst: an exploratory study on AI-supported quantitative data analysis in empirical research. In *Frontiers in Education* (Vol. 9, p. 1417900). Frontiers Media SA.

[8] Lu, J., Liu, J., Zheng, X., Yang, M., Wang, J., Wang, P., & Zhang, Y. (2026, March). MHB: Medical Hallucination Benchmark for Large Language Models in Complex Clinical Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 45, pp. 38971-38978).

[9] Ross, J. (2019). Be Aware of Health Technology Hazards. *Journal of PeriAnesthesia Nursing*, *34*(2), 435-438.

[10] Li, Y., Fu, X., Verma, G., Buitelaar, P., & Liu, M. (2025). Mitigating Hallucination in Large Language Models (LLMs): An Application-Oriented Survey on RAG, Reasoning, and Agentic Systems. *arXiv preprint arXiv:2510.24476*.

[11] Kuculmez, O., Usen, A., & Ahi, E. D. (2025). Referential hallucination and clinical reliability in large language models: a comparative analysis using regenerative medicine guidelines for chronic pain. *Rheumatology International*, *45*(10), 240.

[12] Salehi, S., Singh, Y., Horst, K. K., Hathaway, Q. A., & Erickson, B. J. (2025). Agentic AI and Large Language Models in Radiology: Opportunities and Hallucination Challenges. *Bioengineering*, *12*(12), 1303.

[13] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1-55.

[14] Aydin, S., Karabacak, M., Vlachos, V., & Margetis, K. (2024). Large language models in patient education: a scoping review of applications in medicine. *Frontiers in medicine*, *11*, 1477898.

[15] Pandit, S., Xu, J., Hong, J., Wang, Z., Chen, T., Xu, K., & Ding, Y. (2025, November). Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 2858-2873).

[16] Gu, Z., Chen, J., Liu, F., Yin, C., & Zhang, P. (2026). MedVH: Toward systematic evaluation of hallucination for large vision language models in the medical context. *Advanced Intelligent Systems*, *8*(1), 2500255.

[17] Li, H., Huang, J., Ji, M., Yang, Y., & An, R. (2025). Use of retrieval-augmented large language model for COVID-19 fact-checking: development and usability study. *Journal of medical Internet research*, *27*, e66098.

[18] Lee, J. T., Li, V. C. S., Wu, J. J., Chen, H. H., Su, S. S. Y., Chang, B. P. H., ... & Atun, R. (2025). Evaluation of performance of generative large language models for stroke care. *npj Digital Medicine*, *8*(1), 481.

[19] Garcia-Fernandez, C., Felipe, L., Shotande, M., Zitu, M., Tripathi, A., Rasool, G., ... & Valdes, G. (2025). Trustworthy ai for medicine: Continuous hallucination detection and elimination with check. *arXiv preprint arXiv:2506.11129*.

[20] Wang, Y., Gao, S., Liu, J., Jiang, S., Haoxiang, X., Zhang, X., ... & Liu, Z. (2026, March). Beyond n-grams: A hierarchical reward learning framework for clinically-aware medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 40, pp. 33719-33727).

[21] Deng, Y., Zhao, S., Miao, Y., Zhu, J., & Li, J. (2025). MedKA: A knowledge graph-augmented approach to improve factuality in medical Large Language Models. *Journal of Biomedical Informatics*, 104871.

[22] Akl, A., Khamis, A., Cheraghian, A., Wang, Z., Khalifa, S., & Wang, K. (2026). HIME: Mitigating Object Hallucinations in LVLMs via Hallucination Insensitivity Model Editing. *arXiv preprint arXiv:2602.18711*.

[23] Xu, S., Yan, Z., Dai, C., & Wu, F. (2025). MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. *Frontiers in Public Health*, *13*, 1635381.

[24] Scanlin, J., Cuesta, A., & Varsavsky, M. (2026). Representation Before Retrieval: Structured Patient Artifacts Reduce Hallucination in Clinical AI Systems. *medRxiv*, 2026-02.

[25] Wu, J., Wu, X., Zheng, Y., & Yang, J. (2025). Clinical pathway-aware large language models for reliable and transparent medical dialogue. *Journal of Biomedical Informatics*, 104942.

[26] Yang, D., Wei, J., Li, M., Liu, J., Liu, L., Hu, M., ... & Zhang, L. (2025). MedAide: information fusion and anatomy of medical intents via LLM-based agent collaboration. *Information Fusion*, 103743.

[27] Garcia-Fernandez, C., Felipe, L., Shotande, M., Zitu, M., Tripathi, A., Rasool, G., ... & Valdes, G. (2025). Trustworthy ai for medicine: Continuous hallucination detection and elimination with check. *arXiv preprint arXiv:2506.11129*.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access