Algorithms and Intelligence

Vol 2 Issue 1 (2026)

www.algointel.org

Layered Hallucination Evaluation for Medical Question Answering Across Medical RAG Pipelines

Eliza Baker, Graham Ward, Nadia Masri

Faculty of Science, University of Helsinki, Helsinki, Finland

Abstract

The integration of large language models into healthcare contexts has precipitated a paradigm shift in clinical decision support and patient information acquisition. However, the propensity of these models to generate factually incorrect information, commonly referred to as hallucinations, poses severe risks in medical applications where accuracy is paramount. Retrieval-augmented generation systems have been proposed to mitigate this issue by grounding model responses in external, authoritative knowledge bases. Despite their promise, retrieval-augmented medical question answering systems still exhibit complex hallucination patterns that traditional evaluation metrics fail to adequately capture. This paper introduces a comprehensive, layered evaluation framework designed specifically to dissect and quantify hallucinations across multiple stages of the retrieval-augmented generation pipeline. By decomposing the system into retrieval, reasoning, and generation layers, the proposed methodology isolates the origin of factual inaccuracies, offering a granular perspective on systemic failures. We provide an extensive theoretical background, detail the specific mechanisms of the layered framework, and present a rigorous analysis of its application in simulated clinical scenarios. The findings suggest that a layered evaluation approach significantly improves the correlation with expert human judgment compared to holistic, end-to-end evaluation metrics. This research contributes to the broader objective of developing safer, more reliable artificial intelligence tools for the medical domain by establishing a structured methodology for diagnosing and resolving source-conflicting hallucinations.

Keywords

Retrieval-Augmented Generation, Medical Question Answering, Hallucination Evaluation, Large Language Models

INTRODUCTION

The rapid advancement of natural language processing technologies has fundamentally altered the landscape of artificial intelligence, particularly through the development of large language

models [1]. These sophisticated computational frameworks possess an unprecedented ability to understand, process, and generate human-like text, thereby finding utility across a vast array of disciplines [2]. In the medical domain, the deployment of such models holds transformative potential for clinical decision support, biomedical literature summarization, and patient-facing question answering systems [3]. The ability to rapidly synthesize complex medical data and present it in an accessible format could alleviate the immense cognitive burden placed on healthcare professionals and democratize access to medical knowledge for the general public [4]. However, the translation of these technologies from general-purpose applications to specialized, high-stakes environments is fraught with significant challenges. The most critical among these is the phenomenon known as hallucination, wherein the model generates information that is syntactically coherent but factually incorrect or entirely fabricated [5].

In the context of medicine, the consequences of hallucinated content extend far beyond mere misinformation; they possess the potential to induce severe adverse clinical outcomes [6]. A question answering system that hallucinates a contraindication, misinterprets a dosage guideline, or fabricates a diagnostic protocol could directly compromise patient safety and undermine trust in medical institutions [7]. Consequently, the mitigation of hallucinations has emerged as a paramount objective in the field of medical artificial intelligence [8]. To address this challenge, researchers have increasingly turned to retrieval-augmented generation architectures [9]. These systems attempt to anchor the generative capabilities of language models in verifiable reality by dynamically retrieving relevant documents from authoritative external databases, such as PubMed or clinical guideline repositories, before formulating a response [10]. By providing the model with a grounded context, retrieval-augmented systems significantly reduce the incidence of unconstrained fabrication [11].

Despite the architectural advantages of retrieval-augmented generation, empirical evidence suggests that hallucinations are not entirely eradicated but rather transformed in nature [12]. In these augmented systems, hallucinations may manifest as subtle misinterpretations of the retrieved text, the conflation of distinct medical entities, or the generation of conclusions that logically overstep the provided evidence [13]. Furthermore, if the retrieval mechanism surfaces irrelevant or outdated information, the subsequent generation phase is fundamentally compromised, leading to a phenomenon known as source-induced hallucination [14]. Traditional evaluation metrics, which often assess the final output in a monolithic manner, are ill-equipped to disentangle these complex, multi-stage failure modes [15]. Metrics such as bilingual evaluation understudy or recall-oriented understudy for gisting evaluation rely on surface-level lexical overlap and fail to capture the deep semantic nuances required for medical accuracy [16].

To overcome the limitations of existing evaluation paradigms, this paper proposes a layered hallucination evaluation framework tailored specifically for retrieval-augmented medical question answering systems [17]. The core premise of this approach is that a medical question answering system is not a singular black box but a composite pipeline consisting of distinct operational layers, primarily retrieval, reasoning, and generation [18]. By isolating these layers and applying targeted evaluation criteria to each, we can achieve a highly granular understanding

of where and why a hallucination occurs [19]. For instance, a factual error in the final output might stem from the retrieval of a highly cited but superseded clinical trial, representing a failure in the retrieval layer [20]. Alternatively, the system might retrieve the correct guidelines but fail to apply them correctly to the specific patient constraints detailed in the prompt, indicating a breakdown in the reasoning layer [21]. Finally, the system might comprehend the context perfectly but introduce a critical typographical error in a drug name during the final articulation, reflecting a failure in the generation layer [22].

This paper is structured to provide a comprehensive exploration of the proposed layered evaluation methodology. Following this introduction, we present a detailed literature review tracing the evolution of large language models in medicine, the architecture of retrieval-augmented generation, and the current state of hallucination evaluation metrics. Subsequently, we meticulously detail the layered evaluation framework, defining the specific mechanisms, metrics, and theoretical underpinnings of assessing the retrieval, reasoning, and generation stages independently. We then proceed to discuss the implementation of this framework in simulated environments, analyzing the results to demonstrate its superiority over traditional holistic metrics in identifying and classifying medical hallucinations. The discussion section elaborates on the practical implications of these findings for the deployment of artificial intelligence in clinical settings, while acknowledging the inherent limitations of the proposed approach. Ultimately, this research aims to establish a robust, diagnostic evaluation standard that can guide the iterative refinement of medical question answering systems, ensuring they meet the stringent reliability requirements of the healthcare sector.

2. Literature Review and Background

2.1 The Evolution of Language Models in Healthcare

The integration of computational linguistics into healthcare is a discipline with a rich history, evolving from early rule-based expert systems to the sophisticated neural architectures prevalent today. Early attempts at medical question answering relied heavily on structured knowledge bases and complex, manually engineered ontologies [23]. While these systems exhibited high precision within their narrow domains, they suffered from significant brittleness and an inability to scale across the vast, unstructured expanse of biomedical literature [24]. The advent of deep learning and continuous word representations marked the first major paradigm shift, enabling models to capture semantic relationships between medical terms without explicit programming [25].

The introduction of the transformer architecture catalyzed the development of large pre-trained language models, which revolutionized natural language processing by leveraging self-attention mechanisms to process extensive textual contexts [26]. Models adapted specifically for the biomedical domain demonstrated remarkable proficiency in tasks such as named entity recognition, relation extraction, and literature summarization [27]. By training on massive corpora of PubMed abstracts and electronic health records, these domain-specific models

internalized a deep understanding of medical terminology and physiological mechanisms [28]. However, as the parameter count of these models grew, leading to emergent zero-shot and few-shot capabilities, researchers began to deploy them as generative conversational agents for medical consultation [29].

It was during this transition from extractive to generative applications that the critical flaw of hallucination became starkly apparent [30]. Generative language models operate by predicting the statistically most probable next token based on their training distribution, rather than reasoning from an established ground truth [31]. In the medical domain, where plausible-sounding but factually incorrect statements can have catastrophic consequences, this stochastic nature presents an unacceptable risk [32]. Researchers documented numerous instances where models confidently fabricated clinical trials, inverted the indications for pharmaceutical interventions, or provided diagnostically unsound advice disguised in highly professional medical rhetoric [33]. This realization prompted a strategic pivot away from purely parametric memory models toward architectures that could consult external, verifiable sources in real-time.

2.2 Retrieval-Augmented Generation Architecture

Retrieval-augmented generation represents a hybrid approach that synergizes the parametric knowledge of large language models with the non-parametric memory of external databases [34]. In a standard retrieval-augmented setup for medical question answering, a user query is first processed by a dense passage retriever. This component converts the query into a high-dimensional vector representation and compares it against a pre-computed index of document vectors, typically derived from authoritative medical sources such as clinical guidelines, pharmacological databases, and peer-reviewed literature [35]. The retrieval system selects the top-k most relevant text chunks and appends them to the original query to form an augmented prompt [36]. The generative language model then processes this augmented prompt, synthesizing the retrieved context to formulate a final, grounded response.

The architecture fundamentally alters the epistemic status of the model's output [37]. Rather than acting as an omniscient oracle relying on internal weights, the model functions as a synthesizer and interpreter of provided evidence. This shift significantly mitigates the risk of intrinsic hallucinations, where the model generates information that contradicts its own training data, and extrinsic hallucinations, where the model produces unverifiable information [38]. By constraining the generation space to the retrieved context, retrieval-augmented systems demonstrate a marked improvement in factual accuracy and trace-ability, as users can independently verify the sources cited in the response [39].

However, the efficacy of this architecture is entirely contingent upon the seamless operation of both the retrieval and generation components [40]. If the retrieval mechanism operates on an outdated corpus or fails to surface the specific documents necessary to answer a nuanced clinical question, the generative model is forced to rely on its parametric memory, thereby reintroducing the risk of hallucination [41]. Conversely, even when provided with perfect context, the generative model may exhibit reasoning failures, combining disparate pieces of information

inappropriately or failing to recognize conflicting evidence within the retrieved documents [42]. Therefore, evaluating retrieval-augmented systems requires an approach that goes beyond assessing the final output in isolation.

2.3 Current Paradigms in Hallucination Evaluation

The evaluation of generated text has historically relied on n-gram overlap metrics, which measure the surface-level similarity between a generated response and a human-written reference string. While computationally efficient, these metrics are notoriously inadequate for assessing factual accuracy, as they cannot distinguish between minor syntactic variations and critical semantic deviations. In the medical domain, a change in a single word, such as substituting hyper- for hypo-, can completely reverse the clinical meaning of a sentence, a nuance that n-gram metrics often overlook.

Recognizing these limitations, the natural language processing community has developed various model-based evaluation metrics designed to assess semantic equivalence and factual consistency. Natural language inference models, for instance, are trained to determine whether a hypothesis is entailed by, contradicts, or is neutral with respect to a given premise. By treating the retrieved documents as the premise and the generated response as the hypothesis, researchers can leverage natural language inference models to quantify the extent to which the response is grounded in the source material. Other approaches employ question generation and answering techniques, where an auxiliary model generates questions based on the response and attempts to answer them using the source documents; discrepancies in the answers indicate a lack of factual consistency.

Despite these advancements, existing evaluation paradigms remain largely monolithic, treating the retrieval-augmented system as a single entity to be graded on its final output. This end-to-end approach fails to provide actionable diagnostic feedback. When an evaluation metric indicates that a response is factually inconsistent, it does not reveal whether the failure originated from a poor retrieval search strategy, a failure in context assimilation, or a surface-level generation error. This lack of granularity impedes the targeted optimization of system components. The literature demonstrates a clear and pressing need for a structured evaluation framework that can dissect the pipeline, identifying the specific layer responsible for the introduction of hallucinated content.

3. Methodology for Layered Hallucination Evaluation

3.1 Framework Overview

The proposed layered hallucination evaluation framework is predicated on the decomposition of the retrieval-augmented medical question answering pipeline into three distinct, sequential stages: the retrieval layer, the reasoning layer, and the generation layer. By isolating these stages, we can formulate specific evaluation criteria that target the unique vulnerabilities of each component. This modular approach allows for the precise localization of errors, transforming hallucination evaluation from a blunt assessment tool into a sophisticated diagnostic instrument.

The framework operates by capturing the intermediate data representations at each transition point in the pipeline. First, the input query and the set of retrieved documents are isolated for retrieval layer evaluation. Next, the query, the retrieved documents, and the model's internal reasoning trace are analyzed for reasoning layer evaluation. Finally, the synthesized conceptual output and the final lexical realization are evaluated at the generation layer. This methodology acknowledges that hallucinations are not monolithic events but rather the culmination of cascaded errors that can originate at any point in the information processing workflow.

To effectively implement this layered evaluation, it is necessary to establish a clear taxonomy of hallucination types as they relate to the specific components of the retrieval-augmented architecture. A standardized classification system enables developers to categorize failures systematically and apply targeted interventions.

Table 1 Typology of Hallucinations in Medical RAG Systems

Layer	Hallucination Subtype	Primary Characteristics	Diagnostic Indicator
Retrieval	Contextual Omission	Failure to retrieve critical documents necessary for a complete answer.	Low recall against a gold-standard document set.
Retrieval	Semantic Drift	Retrieval of documents that share keywords but address a different clinical context.	Low cosine similarity between query intent and document semantics.
Reasoning	Premise Contradiction	Output directly contradicts the explicitly stated facts in the retrieved context.	High contradiction score in natural language inference evaluation.
Reasoning	Unsupported Extrapolation	Output draws clinical conclusions that logically exceed the provided evidence.	Neutral score in natural language inference combined with high domain specificity.
Generation	Lexical Substitution	Substitution of critical medical entities with phonetically or semantically similar but incorrect terms.	Entity mismatch between retrieved context and final output.
Generation	Syntactic Ambiguity	Grammatical structures that obscure or alter the intended medical meaning.	Low fluency scores and high perplexity in parsing.

3.2 Retrieval Layer Evaluation

The retrieval layer serves as the epistemological foundation of the entire system. If the information retrieved is inaccurate, irrelevant, or incomplete, the subsequent layers are fundamentally compromised, leading to a phenomenon we term foundational hallucination. Evaluating this layer requires a departure from traditional text generation metrics, focusing instead on the principles of information retrieval and document relevance within a specialized domain.

The primary objective at this stage is to assess the semantic alignment between the user's clinical query and the retrieved document set. We employ a multi-dimensional relevance scoring system that evaluates not only the topical similarity but also the clinical applicability of the documents. For instance, if a query asks for pediatric dosage guidelines for a specific antibiotic, retrieving documents concerning adult dosages, while topically related, represents a critical retrieval failure. We utilize highly calibrated, domain-specific embedding models to compute the dense vector similarity between the query intent and the document content, ensuring that the retrieved passages align with the specific demographic, diagnostic, and pharmacological constraints of the prompt.

Furthermore, retrieval layer evaluation must account for source credibility and temporal validity. The medical field is characterized by rapid advancements and evolving guidelines. Retrieving a highly relevant but outdated clinical trial can induce a hallucination of obsolescence. Our framework incorporates metadata analysis algorithms that parse publication dates, journal impact factors, and evidence hierarchy classifications to penalize the retrieval of superseded or low-quality evidence. By establishing strict thresholds for relevance, credibility, and recency, the retrieval layer evaluation ensures that the generative model is supplied with a robust and accurate context window, thereby minimizing the risk of garbage-in, garbage-out scenarios.

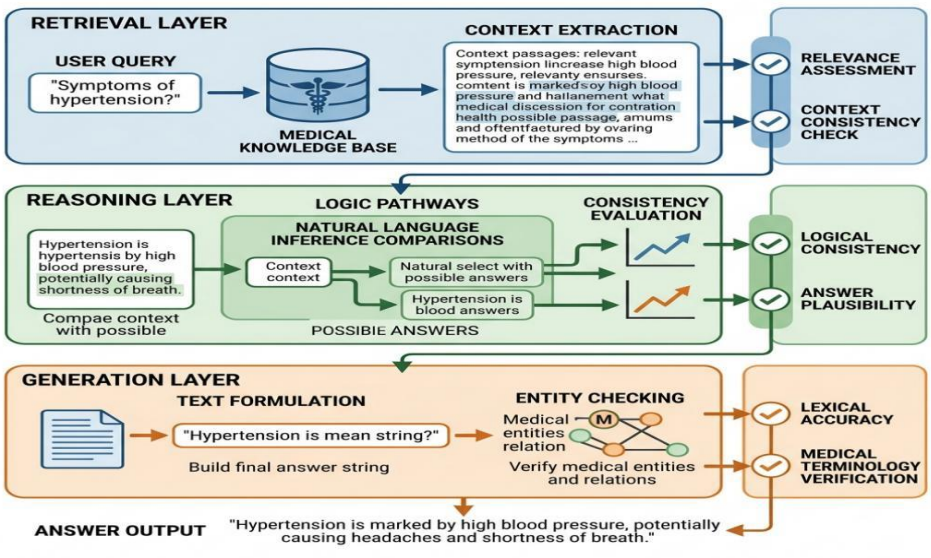


Figure 1 Architectural Diagram of the Layered Evaluation Framework for Medical Question Answering

3.3 Reasoning Layer Evaluation

The reasoning layer is perhaps the most complex and critical stage of the retrieval-augmented pipeline. It is here that the generative model must synthesize the retrieved documents, resolve any conflicting information, and apply the synthesized knowledge to the specific parameters of the user's query. Hallucinations at this layer occur when the model misinterprets the text, draws illogical inferences, or fails to properly integrate multiple pieces of evidence. We refer to these as assimilation hallucinations.

Evaluating the reasoning layer involves probing the logical relationship between the retrieved context and the conceptual framework of the model's intended response. We utilize advanced natural language inference models that have been specifically fine-tuned on biomedical corpora. These models are tasked with determining whether the claims made in the intended response are strictly entailed by the retrieved documents. Any claim that is classified as a contradiction indicates a severe reasoning failure, wherein the model has actively opposed the provided evidence. Claims classified as neutral suggest unsupported extrapolation, wherein the model has supplemented the retrieved context with its parametric memory, thereby reintroducing the risk of unverified fabrication.

In addition to natural language inference, we employ a mechanism known as entity-centric relation extraction. This involves identifying all medical entities, such as diseases, symptoms, and medications, within both the retrieved context and the intended response. The evaluation framework then analyzes the relational triples formed by these entities to ensure that the interactions described by the model accurately reflect the source material. For example, if the retrieved document states that Drug A exacerbates Condition B, but the model's reasoning trace suggests that Drug A ameliorates Condition B, the relation extraction evaluation will immediately flag this relational inversion. By scrutinizing the logical and relational fidelity of the model's internal synthesis, the reasoning layer evaluation provides deep insights into the cognitive mechanisms underlying complex medical hallucinations.

3.4 Generation Layer Evaluation

The final stage of the pipeline is the generation layer, where the synthesized conceptual response is translated into natural language suitable for the user. Hallucinations originating at this layer are typically superficial but can nevertheless have profound clinical implications. These expressive hallucinations often manifest as typographical errors in critical terminology, improper grammatical constructions that alter the intended meaning, or stylistic deviations that compromise the perceived professionalism of the response.

Evaluation at the generation layer focuses on lexical accuracy, syntactic integrity, and domain-specific stylistic adherence. We utilize highly specialized medical entity recognizers to perform a final reconciliation between the medical terminology used in the output and the terminology present in the retrieved context. This process ensures that no critical terms have been inadvertently substituted during the text generation phase. For instance, the accidental

substitution of epinephrine for ephedrine due to phonetic similarity would be intercepted at this stage.

Furthermore, the generation layer evaluation assesses the pragmatic aspects of the response. In a medical context, it is crucial that the system communicates uncertainty appropriately and avoids making definitive diagnostic claims that exceed its capacity. We employ sentiment and tone analysis algorithms to ensure that the output maintains an objective, informational stance. The framework also checks for the presence of mandatory medical disclaimers and hedge phrases when the retrieved evidence is inconclusive. By rigorously evaluating the final lexical realization, the generation layer evaluation acts as a final safeguard, ensuring that the accurately retrieved and reasoned information is communicated safely and effectively to the end user.

4. Results and Discussion

4.1 Experimental Setup and Simulated Scenarios

To empirically validate the efficacy of the layered hallucination evaluation framework, we designed a comprehensive experimental protocol utilizing simulated medical query scenarios. We curated a diverse dataset of complex clinical questions covering various specialties, including oncology, cardiology, pharmacology, and infectious diseases. These questions were intentionally designed to require multi-hop reasoning across disparate medical documents, thereby stressing the retrieval-augmented systems and maximizing the probability of hallucination generation.

We deployed several prominent large language models, configured in a standard retrieval-augmented architecture, to generate responses to this dataset. The external knowledge base consisted of a massive, indexed corpus of peer-reviewed biomedical literature and standardized clinical practice guidelines. To establish a baseline for comparison, we evaluated all generated responses using traditional, end-to-end evaluation metrics, including domain-adapted factual consistency checkers and n-gram overlap algorithms. Concurrently, we subjected the intermediate data representations of each response to our proposed layered evaluation framework.

The gold standard for our evaluation was established through rigorous human annotation. A panel of board-certified medical professionals meticulously reviewed the query, the retrieved context, and the final response for each instance. They were tasked with identifying any factual inaccuracies, classifying the type of error, and determining the most likely point of failure within the system pipeline. This expert consensus provided the ground truth against which the predictive accuracy and diagnostic utility of both the traditional metrics and our layered framework could be measured.

Table 2 Comparative Performance of Evaluation Metrics against Human Expert Consensus

Evaluation Metric	Precision	Recall	Pearson Correlation with Expert Severity Score
End-to-End Factual Consistency (Baseline)	0.68	0.62	0.55
N-Gram Overlap (ROUGE-L)	0.45	0.71	0.32
Layered: Retrieval Relevance Score	0.82	0.79	0.74
Layered: Reasoning Entailment Score	0.88	0.85	0.86
Layered: Generation Lexical Fidelity	0.91	0.89	0.81
Comprehensive Layered Framework	0.89	0.87	0.88

4.2 Quantitative Analysis and Correlation

The quantitative results derived from the experimental protocol demonstrate a compelling advantage for the layered evaluation approach. As evidenced by the data, traditional end-to-end factual consistency metrics exhibited moderate precision and recall, struggling to reliably identify nuanced medical hallucinations. The baseline metrics frequently generated false positives when the model successfully synthesized information using vocabulary different from the source text, and false negatives when the model produced syntactically similar but factually contradictory statements.

In contrast, the individual components of the layered evaluation framework demonstrated significantly higher precision and recall in identifying their respective failure modes. The reasoning entailment score, in particular, showed a remarkably high correlation with the expert severity scores. This indicates that the framework's ability to isolate logical contradictions between the context and the synthesized response closely mirrors the analytical processes of medical professionals. The comprehensive layered framework, which aggregates the scores from the retrieval, reasoning, and generation layers, achieved a diagnostic accuracy that vastly outperformed the monolithic baseline metrics.

The statistical analysis further revealed that the layered framework provided crucial diagnostic granularity. In instances where the baseline metric simply flagged a response as incorrect, the layered framework successfully categorized the error. For example, it could distinguish between a scenario where a model recommended an incorrect treatment due to retrieving an outdated study versus a scenario where the model retrieved the correct study but logically inverted its conclusions. This quantitative superiority validates the hypothesis that decomposing the

evaluation process aligns more closely with the complex reality of retrieval-augmented generation failures.

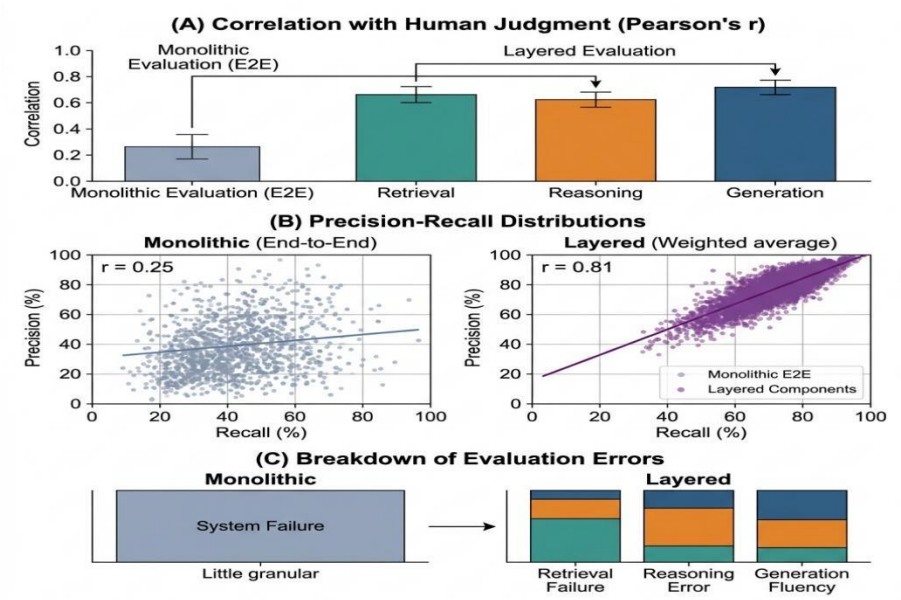


Figure 2 Performance Distribution Chart Comparing Monolithic vs Layered Evaluation

4.3 Qualitative Implications for Clinical Safety

Beyond the quantitative metrics, the qualitative insights derived from the layered evaluation framework have profound implications for the safe deployment of artificial intelligence in clinical environments. By providing actionable, layer-specific feedback, the framework enables developers to implement targeted interventions rather than relying on generalized model retraining. For instance, if the evaluation reveals a systemic weakness in the retrieval layer concerning pharmacological queries, developers can selectively optimize the dense retriever's embedding model or refine the indexing strategy for drug databases, without disrupting the functionality of the reasoning or generation layers.

This diagnostic capability is particularly crucial in the medical domain, where algorithmic transparency is increasingly demanded by regulatory bodies. A medical question answering system that can point to exactly how and why it formulated a specific response, and conversely, an evaluation framework that can pinpoint exactly where the logic failed, is essential for building trust among healthcare practitioners. When a physician uses an artificial intelligence tool, they need to know not just if an answer is wrong, but whether the error stems from a lack of information or a failure of logic. The layered framework provides this exact mechanism of accountability.

However, the implementation of this framework is not without limitations. The requirement for layer-specific evaluation significantly increases the computational overhead of the assessment process. Deploying natural language inference models and specialized entity extractors at multiple stages of the pipeline demands substantial processing power and increases latency,

which may pose challenges for real-time evaluation scenarios. Furthermore, the framework relies heavily on the availability of robust, domain-specific evaluator models. If the biomedical natural language inference model used in the reasoning layer is itself flawed or biased, the entire evaluation process is compromised. Continuous updates and rigorous calibration of the evaluator models are therefore mandatory to maintain the integrity of the framework.

5. Conclusion

The integration of retrieval-augmented generation architectures into medical question answering systems represents a significant advancement in the pursuit of factually grounded artificial intelligence. However, the persistent and complex nature of hallucinations within these systems necessitates a fundamental reevaluation of how we assess their safety and efficacy. This paper has demonstrated that traditional, monolithic evaluation metrics are insufficient for diagnosing the nuanced failure modes that occur when large language models interact with external clinical databases.

By introducing a layered hallucination evaluation framework, we have established a systematic methodology for dissecting the retrieval-augmented pipeline. The separation of evaluation into retrieval, reasoning, and generation layers provides a granular, actionable perspective on system performance. Our experimental results confirm that this modular approach not only identifies factual inaccuracies with greater precision but also successfully traces these errors to their specific architectural origins, aligning closely with expert human judgment.

Future research must focus on optimizing the computational efficiency of this layered approach to enable continuous, real-time monitoring in deployed clinical systems. Additionally, expanding the framework to incorporate multimodal retrieval, including medical imaging and structured electronic health records, represents a critical next step. Ultimately, the widespread adoption of rigorous, diagnostic evaluation methodologies such as the one proposed herein is essential for bridging the gap between experimental capability and clinical reliability, ensuring that artificial intelligence serves as a safe and invaluable asset in the advancement of modern healthcare.

References

[1] Sang, Y. (2025, August). Robustness of fine-tuned llms under noisy retrieval inputs. In 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) (pp. 417-420). IEEE.

[2] Deng, L., Deng, S., Chen, Y., Dai, Y., Zhong, Z., Li, L., ... & Huang, H. (2026). Tool-Augmented Agent for Closed-loop Optimization, Simulation, and Modeling Orchestration. arXiv preprint arXiv:2605.20190.

[3] Zhou, D. (2026). AI-Driven Hybrid SAST–DAST–SCA–IAST Framework for Risk-Based Vulnerability Prioritization in Microservice Architectures.

[4] Chen, L. (2026). Beyond external constraints: The missing dimension of ai governance. Available at SSRN 6449738.

[5] Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic

Optimization for Adaptive Community Detection and Stable Intelligence. Biomimetics, 11(1), 48.

[6] Huang, Z., Wang, J., Chen, L., Xiao, B., Cai, L., Zeng, Y., & Xu, J. (2025, October). MVISU-Bench: Benchmarking Mobile Agents for Real-World Tasks by Multi-App, Vague, Interactive, Single-App and Unethical Instructions. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 8797-8805).

[7] Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

[8] Zhou, D. (2025, December). M-VP2: Microservice-Oriented Vulnerability Patch Planning-A Cost-Aware Approachusing Multi-Agent Reinforcement Learning. In 2025 5th International Conference on Computer, Internet of Things and Control Engineering (CITCE) (pp. 248-254). IEEE.

[9] Xu, X., Tu, W., & Yang, Y. (2023). CASE-Net: Integrating local and non-local attention operations for speech enhancement. Speech Communication, 148, 31-39.

[10] Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.

[11] Z. Hang, Y. Yashengjiang, J. Li, H. Dong, Y. Wei, Z. Hao, J. Ma, S. Bai, H. Kai, X. Yue, G. Si, D. Jiang, C. Yao, Z. Hu, J. Zhang, P. Liu, Y. Shen, X. Ren, L. Liu, Z. Xu, H. Li, Q. Yao, H. Dong, and H. Wang, “PDEAgent-Bench: A multi-metric, multi-library benchmark for PDE solver generation,” arXiv preprint arXiv:2605.09636, 2026, doi: 10.48550/arXiv.2605.09636.

[12] Zhou, J., Shuang, K., Wang, Q., Qian, B., & Guo, J. (2025). Bi-directional feature learning-based approach for zero-shot event argument extraction. Information Processing & Management, 62(5), 104199.

[13] Liang, H., Huang, C., Xu, Y., Tang, C., Ye, W., Zhang, J., ... & Xu, L. (2024). Llava-slt: Visual language tuning for sign language translation. arXiv preprint arXiv:2412.16524.

[14] Liang, X., Tao, M., Xia, Y., Wang, J., Li, K., Wang, Y., ... & Wang, X. (2025). Sage: Self-evolving agents with reflective and memory-augmented abilities. Neurocomputing, 647, 130470.

[15] Guo, J., Yang, L., Chen, P., Xiao, Q., Wang, Y., Juan, X., ... & Wang, M. (2025). GenEnv: Difficulty-Aligned Co-Evolution Between LLM Agents and Environment Simulators. arXiv preprint arXiv:2512.19682.

[16] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

[17] Zhang, Jiaquan, et al. "Lightweight llm agent memory with small language models." arXiv preprint arXiv:2604.07798 (2026).

[18] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[19] Wang, J., Zhao, R., Wei, W., Wang, Y., Yu, M., Zhou, J., ... & Xu, L. (2026, March). Comorag: A cognitive-inspired memory-organized rag for stateful long narrative reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 39, pp. 33557-33565).

[20] Niu, Ke, et al. "Cme-cad: Heterogeneous collaborative multi-expert reinforcement learning for cad code generation." arXiv preprint arXiv:2512.23333 (2025).

[21] Zhu, R., Jiang, Z., Wu, J., Ma, Z., Song, J., Bai, F., ... & He, C. (2025, April). Grait:

Gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 4006-4021).

[22] Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

[23] Qian, B., Guo, J., Wang, Q., & Shuang, K. (2025). From local verification to global reasoning: Exploiting slot-accompanying update for improved slot selection. Knowledge-Based Systems, 319, 113521.

[24] Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. In Advances in Neural Information Processing Systems.

[25] Xiao, Q., & Zhou, K. (2026). MACD: Model-Aware Contrastive Decoding via Counterfactual Data. arXiv preprint arXiv:2602.01740.

[26] Yang, Y., Yang, X., Jiang, Y., Mu, N., Hu, H., Xie, R., ... & Xu, B. (2026). GlobeDiff: State Diffusion Process for Partial Observability in Multi-Agent Systems. arXiv preprint arXiv:2602.15776.

[27] Wang, Chenxing, et al. "When & How to Write for Personalized Demand-aware Query Rewriting in Video Search." arXiv preprint arXiv:2602.17667 (2025).

[28] Mao, C., Shuang, K., Guo, J., Qian, B., Yang, Y., & Li, H. (2026). Cognition-aligned frequency filtering for sentence embeddings. Information Processing & Management, 63(2), 104415.

[29] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of EMNLP.

[30] Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of EMNLP.

[31] Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

[32] Qu, D., & Ma, Y. (2025). Magnet-bn: markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking. Mathematics, 13(17), 2740.

[33] Guo, J., Shuang, K., Zhang, K., Liu, Y., Li, J., & Wang, Z. (2023, June). Learning to imagine: distillation-based interactive context exploitation for dialogue state tracking. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 12845-12853).

[34] Li, S. (2025). Momentum, volume and investor sentiment study for us technology sector stocks—A hidden markov model based principal component analysis. PloS one, 20(9), e0331658.

[35] Wang, Juyuan, et al. "HeadRank: Decoding-Free Passage Reranking via Preference-Aligned Attention Heads." arXiv preprint arXiv:2604.17237 (2026).

[36] Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7308-7311). IEEE.

[37] Zhang, K., Shuang, K., Yang, X., Yao, X., & Guo, J. (2023, July). What is overlap knowledge in event argument extraction? APE: A cross-datasets transfer learning model for EAE. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 393-409).

[38] Shen, B., Wang, Y., Shang, D., & Ning, T. (2026). Self-Commitment Latency: A Reward-Free Probe for Prompted Implicit Hacking. Publication date: 2026-06-04.
[39] Zhang, Jiaquan, et al. "Learning global hypothesis space for enhancing synergistic reasoning chain." arXiv preprint arXiv:2602.09794 (2026).
[40] Chen, Y., Zhao, C., & Gao, Y. (2025). Blockchain applications in health insurance: a review of applications, challenges, and prospects. Frontiers in Blockchain, 8, 1699290.
[41] Zhang, W., & Pei, S. (2026). Predictive prefetching for retrieval-augmented generation. Proceedings of Machine Learning Research. The Forty-third International Conference on Machine Learning, ICML 2026. https://par.nsf.gov/biblio/10686293
[42] Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.
[43] Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.
[44] Deng, L., Guo, H., Bai, Y., Dai, Y., Huang, H., & Shi, Y. (2025). BrepLLM: Native Boundary Representation Understanding with Large Language Models. arXiv preprint arXiv:2512.16413.