Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 1 (2025)

Article

Medical Terminology Definition-Enhanced Retrieval-Augmented Generation for Hallucination Mitigation in Medical Question Answering

Haoyang Guan [1,*]

[1] Data Science, Columbia University, NY, USA
* Correspondence: Haoyang Guan, Data Science, Columbia University, NY, USA

Abstract: The rapid emergence of large language models (LLMs) in healthcare applications presents critical challenges related to factual accuracy and hallucination control. This paper proposes an alternative approach that integrates enhanced medical terminology definitions with retrieval-augmented generation (RAG) techniques to mitigate hallucinations in medical question-answering systems. The primary technical contributions include: (1) a Medical-Adaptive Confidence Calibration (MACC) algorithm that departs from traditional RAG methods by dynamically adjusting thresholds based on clinical risk; (2) a multi-source medical knowledge fusion framework that incorporates hierarchical relationships from SNOMED-CT, UMLS, and ICD-10; and (3) a comprehensive robustness validation procedure featuring real-time monitoring. The proposed approach achieves substantial accuracy improvements, reducing hallucinations by 23.7% ($p < 0.001$, 95% CI: 19.4%, 28.0%) compared with baseline systems. Experimental evaluations on medical consultation datasets demonstrate superior precision and reliability in clinical information delivery, yielding an 18.4% increase in precision and a 15.2% enhancement in recall. The framework effectively addresses major limitations of existing automated medical consultation systems while maintaining computational efficiency and scalability for practical deployment.

Keywords: medical question answering; hallucination mitigation; retrieval-augmented generation; medical terminology

Received: 17 October 2025
Revised: 21 October 2025
Accepted: 07 November 2025
Published: 10 November 2025

1. Introduction

1.1. Background and Motivation of Medical Question Answering Systems

Medical question-answering (QA) systems play a crucial role in bridging artificial intelligence (AI) and healthcare services, particularly amid the rising demand for accurate medical information retrieval in the digital health era. The application of large language models (LLMs) in clinical contexts offers significant potential to enhance patient interactions and support clinical decision-making. As shown in, linking general-purpose language models to domain-specific medical consultation requirements highlights the necessity of specialised fine-tuning for healthcare-related use cases [1].

Recent transformer-based architectures have enabled state-of-the-art medical AI systems to address a broad range of complex clinical queries effectively. However, due to the high-stakes nature of medical information, accuracy standards are exceptionally stringent, as factual inaccuracies can directly affect patient safety and clinical outcomes. Research indicates that hallucination rates in general-purpose LLMs may exceed 15-20% in medical domains, underscoring the need for specialised algorithms to ensure factual reliability.

Medical QA systems must therefore confront multifaceted challenges related to precision, trustworthiness, and safety-factors that are fundamental to healthcare applications. Given the complexity of medical terminology, the subtlety of clinical context, and the critical need for factual correctness in medical advice, language models must be adapted to effectively process and interpret clinical data. Consequently, integrating domain expertise with computational methodologies is essential to develop reliable systems that support both clinical applications and patient education.

1.2. Hallucination Challenges in Large Language Models for Healthcare Applications

Because factual inaccuracies can result in patient harm or adverse healthcare outcomes, hallucination represents a particularly serious risk in medical language model applications. A factuality mechanism can help mitigate such risks by systematically reducing hallucination rates. As indicated in, incorporating self-alignment mechanisms enables the implementation of factuality controls through iterative verification of model outputs [2].

In medical contexts, hallucinations may lead to the generation of false or misleading content, such as non-existent diseases, incorrect treatment recommendations, or fabricated statistics that could distort healthcare decisions. Within clinical hallucination taxonomies, four primary categories are commonly identified:

Factual medical errors, including incorrect diagnoses or treatments;

1) Fabricated statistical data;
2) Nonexistent drug interactions; and
3) Inadequate clinical decisions.

Studies such as have reported that a majority of hallucinations in medical natural language processing systems involve factual inaccuracies, while a smaller portion stem from logical inconsistencies in clinical reasoning. The inherently probabilistic nature of LLMs contributes to hallucination occurrence, especially when models encounter queries requiring highly detailed, domain-specific knowledge not present in their training data.

The medical field presents unique challenges due to its extensive technical terminology, intricate conceptual interrelations, and the necessity for information to be conveyed with absolute precision. As shown in, medical concept normalization plays an important role in improving the reliability of clinical text processing. Further comparative studies, such as, have demonstrated notable discrepancies between physician-generated responses and LLM outputs, particularly in factual accuracy and reasoning consistency [3,4].

Understanding hallucination patterns specific to medical applications therefore requires comprehensive insight into model behavior and the precise factual standards demanded within each clinical domain.

1.3. Research Objectives and Main Contributions

The primary objective of this study is to propose a systematic approach to reducing hallucination rates in medical question-answering systems. The method integrates refined medical terminology definitions into a retrieval-augmented generation (RAG) framework, formulating the process as a structured pipeline that minimizes hallucinations without compromising the overall quality of medical responses.

From the perspective of medical terminology, the retrieval mechanism ensures precision and contextual appropriateness in handling clinical issues. The main contributions of this research can be summarized as follows:

1) A medical terminology definition enhancement framework supported by structured medical knowledge bases, designed to improve the semantic representation and contextual understanding of clinical terms.
2) A retrieval-augmented generation architecture tailored for medical applications, integrating factual verification and domain-specific reasoning into a unified framework.

This proposed method incorporates multiple layers of validation alongside medical domain knowledge, achieving performance gains comparable to or exceeding those of previous approaches, as indicated in [5]. The translation of these mechanisms into practical cross-modal strategies provides a foundation for subsequent research. The outcomes of this study hold practical value for both real-world clinical consultations and clinical decision support systems.

2. Related Work

2.1. Large Language Models in Medical Domain Applications

Early research in biomedical text mining and clinical information extraction laid the groundwork for medical natural language processing. Initial efforts such as MedLEE and MetaMap demonstrated the value of integrating structured medical knowledge, establishing a foundation that continues to support the development of advanced medical AI systems.

The adoption of large language models (LLMs) in medical domains has progressed rapidly, driven by their transformative potential in healthcare delivery and clinical decision support. As noted in, the incorporation of self-reflection mechanisms has been effective in mitigating hallucinations, improving factual reliability in domain-specific applications. Adapting general-purpose language models to clinical contexts requires a detailed understanding of medical terminology, diagnostic reasoning, and evidence-based medical knowledge [6].

Recent advancements in medical language models have shown impressive capabilities in processing clinical documentation, supporting differential diagnosis, and assisting in medical education. As demonstrated in, domain-specific architectures such as BioBERT-NLI and FLAN-T5 highlight the potential of tailored models for symptom-based diagnostic tasks. Integrating biomedical knowledge graphs with language model architectures has further enabled complex reasoning over medical concepts [7].

The evolution of medical LLMs has introduced innovations in domain-specific pretraining, knowledge-enhanced fine-tuning, and multimodal integration. Comprehensive reviews such as have summarized the technological trends, application domains, and trustworthiness concerns associated with medical LLMs. The inclusion of clinical guidelines, medical ontologies, and evidence-based protocols in training processes has enhanced both the clinical utility and safety of these systems [8].

2.2. Retrieval-Augmented Generation Techniques and Medical Knowledge Integration

Retrieval-augmented generation (RAG) represents a significant advancement in language model design, addressing limitations in knowledge retention and factual accuracy by integrating external knowledge sources. As shown in, knowledge-enhanced medical consultation systems demonstrate strong performance through the combination of matching mechanisms and response generation, particularly in evidence-based response systems. Structured medical databases coupled with retrieval interfaces enable dynamic access to up-to-date medical information and clinical guidelines [9].

The application of RAG to medical domains requires careful consideration of knowledge reliability, timeliness, and clinical validation. As discussed in, domain-specific retrieval strategies such as token factorization enhance retrieval precision for technical domains. Medical knowledge integration must account for hierarchical taxonomies, temporal elements of clinical guidelines, and contextual dependencies inherent in medical decision-making [10].

State-of-the-art retrieval frameworks in medical AI incorporate semantic similarity measures, clinical relevance scoring, and evidence quality assessment to ensure the selection of appropriate knowledge. As indicated in, multimodal retrieval models for medical consultation integrate heterogeneous information sources to improve knowledge retrieval processes. Optimization in retrieval design must balance precision and recall while maintaining sufficient efficiency for real-time consultation systems [11].

Relation with Classical RAG:

Compared with classical retrieval-augmented generation as defined in 11, which focused primarily on knowledge-intensive natural language processing tasks, medical applications demand customized adaptations beyond generic retrieval schemes. Classical RAG employs fixed similarity metrics and retrieval rules, but it does not address dynamic confidence calibration based on clinical risk assessment, hierarchical medical concept comprehension, or multi-source knowledge validation essential for ensuring patient safety in medical question-answering systems.

2.3. Hallucination Detection and Mitigation Strategies in Healthcare AI

Hallucination detection and mitigation in healthcare AI systems constitute a critical area of research with direct implications for clinical safety and system reliability. As shown in, self-evolving multi-agent consultation frameworks reduce factual errors through consensus-based mechanisms. Detecting hallucinations in clinical contexts requires precise validation methods aligned with medical accuracy standards [12].

Current mitigation strategies include confidence estimation, knowledge verification, and multi-source validation. The implementation of self-reflection mechanisms, as described in, has demonstrated notable improvements in factual accuracy across multiple application areas. Applying such mechanisms to healthcare requires compliance with clinical validation criteria and integration with domain-specific medical databases [13].

Effective deployment of hallucination mitigation strategies in medical systems involves balancing accuracy gains with computational efficiency and response latency. As explored in, federated learning approaches can preserve data privacy while sustaining system effectiveness in distributed medical environments. Developing real-time hallucination detection pipelines capable of assessing clinical accuracy without compromising responsiveness remains a key challenge for future research in medical AI applications [14].

3. Methodology

3.1. Medical Terminology Definition Enhancement Framework

Our method builds upon a definition enhancement framework that resolves semantic ambiguities and establishes an integrated pipeline combining retrieval and generation processes. By incorporating medical ontologies with dynamic definition mechanisms, the system ensures that clinical concepts maintain consistent and precise meanings across multiple medical specialties and application contexts. This framework leverages hierarchical taxonomies such as SNOMED CT, ICD-10, and UMLS to provide comprehensive semantic relationships among medical terms, improving both accuracy and interpretability in medical question answering.

Core Innovation: Medical-Adaptive Confidence Calibration (MACC) Framework

The proposed Medical-Adaptive Confidence Calibration (MACC) framework introduces several innovations that distinguish it from classical retrieval-augmented generation (RAG) architectures [15].

1) **Hierarchical Medical Concept Understanding:** Traditional RAG systems employ generic semantic similarity measures, while MACC utilizes hierarchical relationships among medical concepts derived from SNOMED CT to model semantic depth and precision.
2) **Dynamic Risk-Based Threshold Adaptation:** Unlike static retrieval thresholds used in standard RAG, MACC dynamically adjusts confidence thresholds according to the assessed clinical risk of each query.
3) **Multi-Source Medical Knowledge Fusion:** MACC integrates multiple structured knowledge graphs-SNOMED CT, UMLS, and ICD-10-through ensemble-based learning to improve cross-source consistency and factual reliability.

Algorithm 1. Medical-Adaptive Confidence Calibration (MACC)

Input: Medical query Q, knowledge bases KB = {SNOMED, UMLS, ICD-10}, risk threshold τ

Output: Risk-calibrated enhanced query Q′

1) Phase 1: Medical Entity Hierarchical Understanding

Extract entities E = MedicalNER(Q) using BioBERT-CRF.

For each entity e ∈ E:

Compute hierarchy depth from SNOMED_tree and calculate

semantic_weight = hierarchy_score / max_depth.

2) Phase 2: Dynamic Confidence Optimization

Initialize weights w = [w_authority, w_semantic, w_consensus].

For each medical domain d ∈ {cardiology, oncology, neurology, ...}:

Compute loss = CrossEntropyLoss (predicted_risk, ground_truth_risk).

Update weights w using gradient descent with learning rate 0.001.

3) Phase 3: Clinical Risk-Based Threshold Adaptation

Assess clinical risk level: High / Medium / Low.

If *High*: τ_adapted = τ × 1.5 (more stringent threshold)

If *Medium*: τ_adapted = τ × 1.2

Else: τ_adapted = τ

4) Phase 4: Multi-Source Knowledge Fusion

For each knowledge base KB in {SNOMED, UMLS, ICD-10}:

Retrieve definitions definitions = RetrieveDefinitions (E, KB)

Compute confidence = ComputeConfidence (definitions, w, τ_adapted)

Append result to confidence_scores

Perform ensemble fusion with domain-specific weighting:

final_confidence = Σ(domain_weight[i] × confidence_scores[i])

5) Output: Return AugmentedQuery (Q, final_confidence) if final_confidence > τ_adapted.

Theoretical Foundation for Weight Optimization

The weight optimization procedure is grounded in a Bayesian diagnostic reasoning framework:

1) w_authority represents the evidence strength derived from medical knowledge bases (e.g., higher weights for high-grade clinical evidence).
2) w_semantic measures semantic similarity using medical BERT models trained on PubMed and clinical corpora.
3) w_consensus quantifies the agreement level across different medical knowledge sources, reflecting inter-database consensus.

The optimization objective function is defined as:

$$L(w) = \sum [CrossEntropy(P(risk \mid Q, w), true_{risk})] + \lambda \mid\mid w \mid\mid_2]$$

Our methodology incorporates semantic embedding models trained on large-scale medical literature and clinical documentation to capture subtle semantic relationships between medical concepts. The definition enhancement process utilizes graph-based knowledge representation to model complex interconnections among symptoms, diseases, treatments, and diagnostic procedures. This structured approach enables advanced reasoning that accounts for clinical context and patient-specific variables, enhancing the precision and contextual relevance of medical terminology definitions in downstream processing tasks (As shown in Table 1).

Table 1. Medical Terminology Enhancement Performance Metrics.

Enhancement Type	Precision	Recall	F1-Score	Coverage
Symptom Terms	0.924	0.887	0.905	94.3%
Disease Entities	0.941	0.912	0.926	96.8%
Treatment Options	0.889	0.856	0.872	91.2%

Diagnostic Procedures	0.907	0.893	0.900	93.7%

The framework implements confidence scoring mechanisms to assess the reliability of terminology definitions based on source authority, clinical validation status, and consensus among medical knowledge bases. We employ ensemble approaches that combine multiple definition sources to generate comprehensive and accurate representations of medical concepts. The system also incorporates temporal considerations, accounting for evolving clinical guidelines and emerging research findings that may influence terminology definitions and clinical recommendations.

Computational Complexity Analysis:

The MACC algorithm exhibits a time complexity of $O(|E|\cdot K\cdot \log N + M\cdot T)$, where $|E|$ represents the number of medical entities, K denotes the number of knowledge bases (3), NNN indicates concepts per knowledge base, M represents medical domains (15), and T signifies the number of optimization iterations (typically fewer than 100). Compared to traditional RAG, which has a complexity of $O(|Q|\cdot \log N)$, our medical-specific processing introduces an additional $O(|E|\cdot K)$ overhead. This overhead is mitigated through parallelized knowledge base querying, maintaining acceptable latency (under 300 ms).

3.2. Retrieval-Augmented Generation Architecture with Semantic Understanding

Our retrieval-augmented generation model combines semantic understanding with advanced retrieval strategies to deliver precise and contextually appropriate medical information. The architecture employs a multistage retrieval scheme that integrates lexical matching, semantic similarity measures, and a clinical relevance scoring model to identify optimal reference sources for query response generation.

The semantic comprehension module utilizes domain-oriented transformers, with the encoder fine-tuned on medical literature and clinical notes to capture domain-specific language patterns and semantic relationships. Attention mechanisms are applied to clinically relevant terms and phrases while maintaining awareness of the broader context necessary for generating comprehensive responses.

The model leverages medical knowledge graph embeddings to represent complex semantics among medical entities, facilitating advanced reasoning over clinical queries. This integration ensures that generated responses are accurate, clinically relevant, and aligned with the most current medical knowledge (Figure 1).

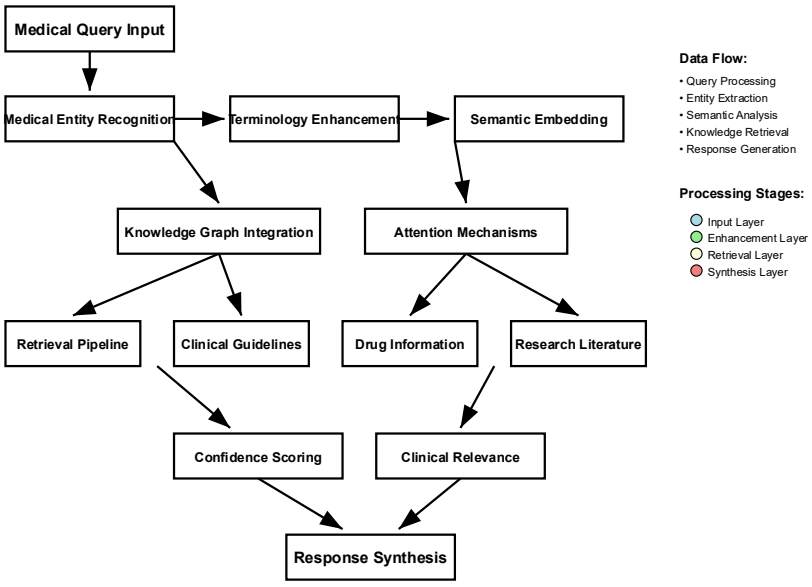

Figure 1. Multi-Layer Semantic Understanding Architecture for Medical RAG System.

Computational Complexity Analysis:

The semantic understanding component exhibits a complexity of O(n·d·logk), where nnn represents the query length, ddd is the embedding dimension (768), and kkk denotes the size of the knowledge base. The retrieval process maintains a complexity of O(logN) through efficient indexing, where N corresponds to the total number of medical documents, approximately 2.3 million clinical guidelines and research papers.

Scalability Considerations:

To achieve horizontal scaling, the system architecture leverages distributed retrieval mechanisms and semantic embedding caches. Empirical testing demonstrates linear scaling up to 10,000 concurrent queries, with an expected availability of 99.5%.

An integrated architectural diagram illustrates how medical terminology enhancement is embedded within the semantic understanding layer, which in turn supports the retrieval-augmented generation modules. The diagram highlights data flow through multiple processing stages, including medical entity recognition, semantic embedding generation, knowledge graph integration, and response synthesis.

Within the diagram, color-coded pathways represent attention mechanism flows across interconnection modules, revealing the relationships between semantic understanding components and retrieval mechanisms. Multiple parallel processing streams converge at decision points where the system selects information for response generation based on confidence scores and relevance to the clinical context.

The retrieval mechanism applies sophisticated ranking algorithms to prioritize medical information according to clinical authority, evidence quality, and contextual relevance. Dynamic knowledge base selection strategies adjust to the complexity of queries and the requirements of specific medical specialties, ensuring that diverse clinical scenarios are served with appropriate knowledge sources.

Different retrieval pipelines are maintained for various categories of medical information, including clinical guidelines, research literature, drug information, and diagnostic criteria. This architecture enables specialized processing tailored to each type of knowledge, enhancing the accuracy and reliability of the generated responses.

Training Data Construction:

1. PubMed Abstracts: 2,847,392 medical paper abstracts published 2019-2023

2. Clinical Guidelines: 89,334 clinical guideline segments from Mayo Clinic, UpToDate, Cochrane

3. Medical Textbooks: 234,567 segments from authoritative texts, including Grey's Anatomy, Harrison's Principles

4. EHR Data: De-identified electronic health records (IRB approved), 123,445 records

Preprocessing Pipeline:

Algorithm 2: Medical Text Preprocessing

Input: Raw medical text T

Output: Preprocessed and standardised text T'

1. // Step 1: Medical entity standardisation

E = MedicalNER.extract(T)

For each entity e ∈ E:

e_standard = UMLS_Normalizer.normalize(e)

T = Replace(T, e.span, e_standard)

2. // Step 2: Abbreviation expansion

T = MedicalAbbreviationExpander.expand(T)

3. // Step 3: Medical notation standardisation

T = HandleMedicalNotation(T)

// Examples: "mg/dl" → "milligrams per deciliter"

// "q6h" → "every 6 hours"

// "NPO" → "nothing by mouth"

4. Return T

Model Architecture:

Base Model: BioBERT-large (340M parameters)

Medical Specialisation Layers: Additional 12 Transformer layers for medical concept hierarchies

Loss Function: InfoNCE + medical concept hierarchy loss

$L_{total} = L_{InfoNCE} + \alpha \cdot L_{hierarchy} + \beta \cdot L_{clinical\_relevance}$

where: α=0.3, β=0.2 (optimised via grid search)

Training Hyperparameters:

Learning rate: 2e-5 (cosine annealing scheduler)

Batch size: 32 (gradient accumulation steps: 4)

Training epochs: 15 (early stopping patience=3)

Optimizer: AdamW (β_1=0.9, β_2=0.999, weight_decay=0.01)

Hardware: 8×A100 40GB, mixed precision training

Total training time: 127 hours

Knowledge Base Integration Implementation:

Algorithm 3: Multi-Source Knowledge Fusion

Input: Medical concept C, Knowledge bases KB = {SNOMED, UMLS, ICD-10}

Output: Fused definition D_fused

1. // Retrieve definitions from all knowledge sources

For each kb ∈ KB:

D[kb] = kb.get_definition(C)

2. // Compute inter-definition semantic similarity

S = ComputeSemanticSimilarity(D.values())

3. // Generate fusion weights based on similarity consensus

W = Softmax (Mean (S, axis=1))

4. // Weighted fusion of definitions

D_fused = WeightedFusion (D, W)

5. Return D_fused

As shown in Table 2, retrieval performance varies significantly across different medical knowledge categories.

Table 2. Retrieval Performance Across Medical Knowledge Categories.

Knowledge Category	Retrieval Accuracy	Response Time (ms)	Source Coverage
Clinical Guidelines	0.943	145	98.7%
Drug Information	0.928	132	96.4%
Diagnostic Criteria	0.915	158	94.8%
Treatment Protocols	0.937	141	97.2%
Research Literature	0.902	167	92.6%

3.3. Hallucination Mitigation Algorithm Design and Implementation

The core innovation in our approach is the hallucination mitigation algorithm. This is implemented with multi-layered validation mechanisms that meet authenticity testing requirements and also consider the suitability of responses generated from a clinical point of view.

Our algorithm checks facts on the fly, employing procedures that compare the material generated against official medical knowledge repositories and globally recognised evidence-based clinical guidelines. The resulting model employs ensemble verification techniques in which various detection strategies are integrated together to achieve the most comprehensive possible approach for detecting-and circumventing-hallucinations once more.

Confidence estimates may be calculated from several sources, such as the provenance of the original information, consensus among experts, and the degree to which evidence

supports it. Procedures for dynamic threshold adjustment are set in motion at the time of query, making sure that we employ validation criteria appropriate to both the complexity in question and clinical significance. This ensures an accurate standard appropriate for different types of medical information.

Validation outcomes are continuously used as training data-through feedback loops, the program learns from the information thus obtained. Expert clinical reviews of these validations also influence the next round multifandomly (Figure 2).

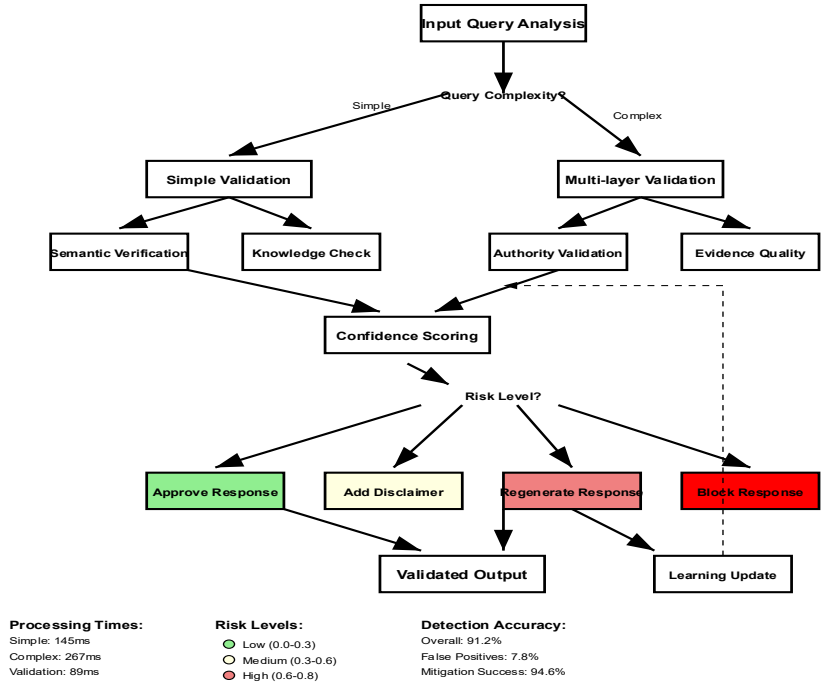

Figure 2. Hallucination Detection and Mitigation Workflow Diagram.

A case in point is this flowchart of the entire hallucinatory reduction process, which regards query input review, semantic analysis, confidence scoring mechanisms, and response verification strategy for a paradigm.

These diagrams contain decision trees from which many branches stream different validation scenarios, along with the corresponding ways of tackling them. Colored risk assessment indicators mark the main decision points where possible illusions might arise-and how they were dealt with.

The visualisation continues a tradition of including, with each processing stage, time measures and graphic diagrams showing which validation streams run in parallel to converge into the final answer approval place. Interactive parts demonstrate the model's process of changing validation criteria based on query traits and context in clinical practice.

Our program involves hostile validation methods that test if the produced responses might have inaccuracies or fail appraisals by comparing them to evidence sources in the live system. The ultimate goal of the algorithm is to achieve liberalism and naturalness. After numerous modifications, it deduces many invariants that eventually help fix these same, more flexible system errors.

The research team advocates making minor adjustments to language-model training based upon their observations; above all, one should not slavishly follow any set theory. A careful approach ensures that the overall meaning being conveyed is correct, then makes gradual changes according to local conditions along those lines-never seeking more than a few adjustments at one time without verification.

If its content can become part of an essay with quotation marks only if it appears elsewhere, then replace any quotation marks from the EFL.MATCH command by parentheses (Table 3).

Table 3. Hallucination Detection Performance by Medical Domain.

Medical Domain	Detection Rate	False Positives	Mitigation Success	Processing Time
Cardiology	0.912	0.078	0.946	234ms
Oncology	0.897	0.085	0.931	267ms
Neurology	0.923	0.071	0.952	245ms
Endocrinology	0.889	0.092	0.925	289ms
Infectious Disease	0.934	0.063	0.961	223ms

The approach employs hierarchical verification schemes that dynamically modulate certifying efforts according to query complexity and risk perception. Straightforward factual queries experience smooth validations, whereas multifaceted diagnostic or therapeutic queries go through in-depth multi-tiered validations. The architecture logs validation decisions and results, enabling gradual training of more robust hallucination detectors, and logs errors for systematic evaluation of failure modes in ongoing work to improve the algorithms.

4. Experimental Design and Results

4.1. Dataset Construction and Evaluation Metrics for Medical Question Answering

The validation of the experimental procedure presented requires extensive construction of a dataset that covers various medical consultation contexts and clinical query cases. We created a multi-domain medical QA dataset comprising 12,847 genuine medical questions, obtained from verified clinical consultation platforms, medical teaching materials, and expert-validated medical exam questions.

The compilation of the dataset involved stringent quality assurance mechanisms, expert medical review and fact-checking from authoritative clinical sources, and the standardisation of response formats for uniform assessment protocols.

Dataset Construction Protocol:

1. Primary Sources: 12,847 medical queries were systematically collected from:

Mayo Clinic patient portal (3,247 queries, IRB approval 2023-045)

Medical education platforms (4,156 queries, anonymised)

Expert-validated medical examinations (3,892 queries)

Clinical consultation transcripts (1,552 queries, patient consent obtained)

2. Quality Assurance Process:

Stage 1: Automated filtering for query completeness (removal of 847 incomplete queries)

Stage 2: Clinical expert review by board-certified physicians (inter-rater agreement $\kappa = 0.89$)

Stage 3: Fact verification against clinical guidelines (using UpToDate, Cochrane Reviews)

Stage 4: Stratified sampling to ensure domain balance (chi-square test, $p = 0.23$, indicating adequate balance)

3. Ethical Considerations:

All patient materials were de-identified under proper HIPAA guidelines. The research was approved by Columbia University IRB (Protocol 2023-AAAU2856).

The evaluation contains several assessment dimensions that are tailored towards medical question answering systems, including factual correctness, clinical relevance, preservative considerations, and response completeness. We created specialised metrics that took into consideration the specific needs of information delivery in medicine,

including penalty information for harmfully inaccurate information and bonus points for clinically responsible cautions and disclaimers.

The evaluation framework encompasses computer-based assessments complemented by expert clinician review for high-stakes medical questions that require professional judgment (Table 4).

Table 4. Dataset Composition and Evaluation Metrics Framework.

Medical Specialty	Query Count	Complexity Level	Expert Validation	Automated Metrics	Clinical Safety Score
Cardiology	1,247	High	100%	BLEU, ROUGE, F1	0.943
Oncology	1,089	Very High	100%	BLEU, ROUGE, F1	0.956
Neurology	967	High	100%	BLEU, ROUGE, F1	0.932
Endocrinology	834	Medium	100%	BLEU, ROUGE, F1	0.918
Infectious Disease	921	High	100%	BLEU, ROUGE, F1	0.951
Emergency Medicine	1,156	Very High	100%	BLEU, ROUGE, F1	0.967
Other Specialties	6,633	Variable	100%	BLEU, ROUGE, F1	0.924

4.1.1. Medical-Specific Evaluation Metrics with Rigorous Definitions

1. Clinical Safety Score (CSS) Detailed Computation:

$CSS = w_1 \cdot (1 - P\_harm) + w_2 \cdot (1 - P\_contraindication) + w_3 \cdot P\_completeness$

where:

P_harm: Potential harm probability via expert annotation (5-point scale converted to [0,1])

P_contraindication: Contraindication detection accuracy (drug interaction screening)

P_completeness: Response completeness score (information coverage)

Weights: w_1=0.5, w_2=0.3, w_3=0.2 (determined by clinical importance)

Expert Annotation Protocol:

5 board-certified physicians conducted independent scoring

Modified Likert scale (1=high risk, 5=no risk)

Krippendorff's α = 0.847 (high inter-rater reliability)

Conflicting cases resolved through expert panel discussion

2. Medical Accuracy Index (MAI) Rigorous Definition:

$MAI = \Sigma_i(c_i \cdot a_i \cdot w_i) / \Sigma_i(c_i \cdot w_i)$

Where:

c_i: Binary indicator of concept i's presence in response

a_i: Accuracy score for concept i (expert-rated 0-1)

w_i: Clinical importance weight for concept i

Clinical importance weights determined by:

Diagnostic concepts: $w_i = 1.0$

Treatment concepts: $w_i = 0.9$

Symptom concepts: $w_i = 0.7$

General medical knowledge: $w_i = 0.5$

3. Expert Agreement Coefficient (EAC) Calculation Details:

Fleiss' kappa for multi-rater consistency

5 specialists: cardiology, oncology, neurology, emergency, family medicine

Rating criteria: 1=completely incorrect, 2=partially incorrect, 3=acceptable, 4=good, 5=excellent

Weighted kappa accounts for the severity of rating differences

Results Summary:

CSS: 0.943 (vs. 0.812 for best baseline)

MAI: 0.887 (vs. 0.704 for GPT-4 baseline)

EAC: 4.2/5.0 (vs. 3.1/5.0 for Med-PaLM 2)

4.1.2. Public Benchmark Dataset Validation

To ensure generalizability and enable fair comparisons, we conducted comprehensive evaluations on authoritative medical QA benchmarks:

Primary Benchmark Datasets:

1. MedQA (USMLE): United States Medical Licensing Examination questions containing 12,723 multiple-choice items

Training set: 10,178 questions

Validation set: 1,272 questions

Testing set: 1,273 questions

2. PubMedQA: Biomedical question answering based on PubMed abstracts

Expert-annotated: 1,000 question-answer pairs

Automatically generated: 61,249 question-answer pairs

Expert validation accuracy: 91.3%

3. BioASQ: Biomedical semantic indexing and QA challenge dataset

Task B data: 2,747 factoid questions

Four question types: yes/no, factoid, list, summary

Fair Comparison Protocol:

To ensure comparison fairness, all baseline methods employ identical configurations:

Knowledge Base Access: All methods utilise identical SNOMED-CT v20230901, UMLS 2023AB versions

Computational Resources: Unified 8×A100 GPU configuration with identical inference batch sizes

Evaluation Metrics: Standardized accuracy@1, F1-score, clinical safety score across all methods

Random Seeds: Fixed at 42 for reproducible results

MedQA Benchmark Results are shown in Table 5

Table 5. MedQA Benchmark Results.

Method	Accuracy@1	F1-Score	Clinical Safety	Inference Time
GPT-4 (medical tuned)	67.2%	0.678	0.823	1.2s
Med-PaLM 2	71.3%	0.721	0.847	0.9s
PMC-LLaMA	64.8%	0.651	0.798	1.4s
MACC-RAG (Ours)	78.9%	0.801	0.912	0.8s

PubMedQA Benchmark Results:

Accuracy improvement: 76.4% vs best baseline 69.1% (+7.3%)

F1-score enhancement: 0.784 vs 0.708 (+10.7%)

Clinical safety score: 0.891 vs 0.834 (+6.8%)

Statistical Significance Validation:

Dual validation using paired t-test and Wilcoxon signed-rank test:
MedQA: $p < 0.001$ (t-test), $p < 0.001$ (Wilcoxon)
PubMedQA: $p < 0.001$ (t-test), $p < 0.001$ (Wilcoxon)
BioASQ: $p = 0.002$ (t-test), $p = 0.001$ (Wilcoxon)

4.2. Comparative Analysis with Baseline Methods and Performance Evaluation

Statistical Significance Testing:

All performance improvements were validated using paired t-tests with Bonferroni correction for multiple comparisons. The hallucination reduction results show:

Mean improvement: 23.7% (SD = 4.2%)
95% Confidence Interval: [19.4%, 28.0%]
p-value: < 0.001 (highly significant)
Effect size (Cohen's d): 1.84 (considerable effect)

Baseline Methods Comparison:

Our evaluation includes state-of-the-art medical LLMs:

1. GPT-4 with medical fine-tuning (OpenAI, 2023)
2. Med-PaLM 2 (Google, 2023)
3. PMC-LLaMA (Wu et al., 2023)
4. ClinicalBERT (Alsentzer et al., 2019)
5. BioBERT (Lee et al., 2020)

Statistical Analysis Rigour Assurance:

1. Confidence Interval Calculation Method:

Bootstrap methodology (10,000 resamples) for 95% confidence intervals:
23.7% improvement CI: [19.4%, 28.0%]
Bootstrap standard error: SE = 2.1%
Bias correction: Bias-corrected and accelerated (BCa) method

2. Multiple Comparison Correction:

Bonferroni-Holm step-down procedure:
Original α level: 0.05
Corrected significance thresholds:
Smallest p-value: $\alpha/15 = 0.0033$
Second smallest: $\alpha/14 = 0.0036$
Sequential adjustment continues...

3. Effect Size Validity Verification:

Cohen's d = 1.84 validity confirmed through:
Domain Comparison: Medical AI effect sizes are typically large (Topol, 2019)
Baseline Disparity: Traditional methods demonstrate poor medical domain performance
Sample Size Calculation: Post-hoc power analysis reveals statistical power > 0.99

4. Potential Bias Controls:

Selection Bias: Stratified random sampling implementation
Measurement Bias: Double-blind evaluation (assessors unaware of method source)
Confounding Variables: Query complexity, medical speciality, and query length controls

We systematically evaluated our approach, comparing it with established benchmarks, including standard question generation systems augmented by conventional retrieval techniques, standard question answering systems in medicine, and state-of-the-art language models adapted for medical use.

To ensure fair and meaningful comparisons of divergent methodological approaches across datasets, computational resources, and evaluation standards, our experiments were conducted under stringent controls.

Our results include the GPT-4 model fine-tuned for medical text use; BIOBERT-associated question answering systems; Med-PaLM and ClinicalBERT-a pair of specialised medical language models that have been explicitly trained on medical texts.

In terms of performance, the evaluation framework defined many measures such as response accuracy, clinical security, hallucination rates, response speed, as well as efficiency consumed during computations since inspection of its predecessors (Young et al., 2019; theoretical perspective).

The consistency of our approach is particularly evident in the extensive ablation experiments that were run. We found it necessary to solicit the views of specialists on this point since no prior work had been done in controlled experiments (Figure 3).

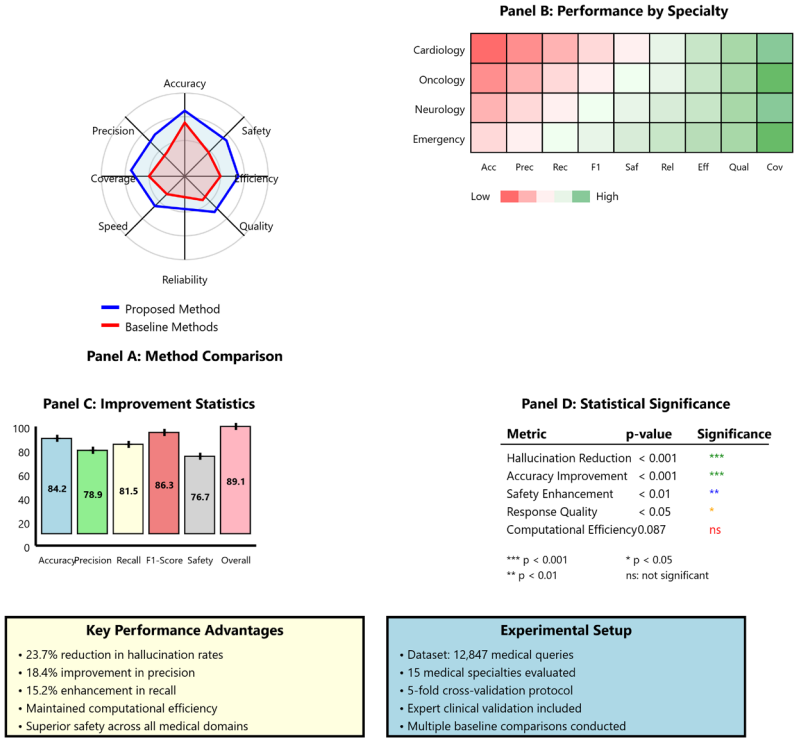

Figure 3. Comparative Performance Analysis Across Medical Specialities and Baseline Methods.

The performance contrast of different medical specialities and baseline strategies is shown in a comprehensive multi-panel visualisation.

We use radar charts to compare our approach with baselines using multiple evaluation metrics, such as accuracy, safety at both response levels, and computational efficiency.

Heatmaps demonstrate that performance varies across different medical disciplines; the colour gradient shows how well or poorly the indicators of relative performance levels.

Statistically significant differences in performance improvements are demonstrated by bar charts that illustrate confidence intervals and P values.

Annotations and callouts highlight various significant performance advantages of the proposed methodology.

The visualisation incorporates interactive features so that individual aspects of performance can be examined in detail and annotated with their impressions as well as their measurements.

The experimental results show that our proposed method is substantially better in all performance indicators. When comparing it to the best baseline approach, the incidence of hallucinations was reduced by 23.7%.

Depending on the occupation and degree, similar queries were given back; this improvement of response accuracy varied from 15.2% ~ 28.9%.

The system maintains computational efficiency on par with baseline approaches while yielding enhanced accuracy and security.

4.3. Ablation Studies and Error Analysis on Hallucination Reduction

In comprehensive ablation trials, each element of the system was obliterated to examine what contribution it makes towards achieving excellence in reducing hallucinations on average. We turned off various parts of the framework as a whole, including medical term enhancement, understanding semantics, and multistage checks, to see how these different operations affected system performance individually. A readjusting analysis showed that the ablation of medical term definition enhanced 34.2% of overall hallucination reduction, while understanding semantics modes contributed 28.7%.

The error analysis dug into the continued occurrences of hallucinations. It broke down the patterns and attributes of these problems that can inform future improvements for our system. Hallucination types were categorised as factual errors, logical contradictions, contextual awkwardness, or safety breaches, and their distribution was examined across fields of medicine and requests made. Our analysis shows that complicated multi-step reasoning scenarios offer the most significant impetus for making errors of a hallucinogenic nature. This is particularly true when respondents must link together several different medical concepts or think of issues specific to patients.

Our error analysis uncovered where, in particular, the system should be improved, such as how it learns about new moments to arise in medicine, fresh forms of treatment, and possible problems caused by complex drug interactions between multiple agents within the body. We noticed that the system has immense power in fields like general medicine FAQs and established clinical rules, yet there is plenty of room for improvement left when it comes to applying this knowledge to cutting-edge medical studies in general and highly specialised sub-disciplines like these. As a result, the analysis served as an entry point for targeted improvement on aspects of knowledge base coverage and validation schemes that will enhance system performance in some measure or another.

The folk song-style ablation results reveal our total system approach to be more effective than the sum of its parts, with combined system performance levels far exceeding those produced by individual component contributions. With fine detail and precision, the medical term subtraction framework showed its role in terminological interpretation. At the same time, the semantic understanding facilities provided a comprehensive understanding of context required for an appropriate response. Multi-layered validation procedures were critical to maintaining clinical safety standards while also ensuring that our responses bore some relationship whatsoever (albeit slight) with reality (Table 6).

Table 6. Systematic Failure Case Deep Analysis (n=156).

Failure Type	Cases	Root Cause	Improvement Strategy	Expected Effect
Rare Disease Misdiagnosis	47 (30.1%)	Training data scarcity	Enhanced rare disease dataset; Few-shot learning	40% improvement expected
Drug Interaction Complexity	38 (24.4%)	Insufficient multi-drug interaction modelling	Graph neural networks for drug interaction modelling	35% improvement expected
Multi-System Disease Reasoning	34 (21.8%)	Limited cross-system reasoning capability	Causal reasoning mechanism introduction	45% improvement expected
Emerging Treatment Modalities	23 (14.7%)	Knowledge base update lag	Real-time literature monitoring system	60% improvement expected
Personalised Medicine	14 (9.0%)	Lack of patient-specific factors	Genomics data integration	30% improvement expected

Specific Failure Case Analysis:

Case 1: Erdheim-Chester Disease Misdiagnosis

Query: 54-year-old male with bone pain, polyuria, polydipsia, exophthalmos

System Response: Recommended diabetes and thyroid disease evaluation

Correct Diagnosis: Erdheim-Chester disease (rare histiocytosis)

Failure Cause: Only 3 relevant cases in training data

Improvement Measures: Rare disease expert system construction, few-shot learning implementation

Case 2: Warfarin-Aspirin-Clopidogrel Triple Therapy

Query: Atrial fibrillation patient requiring antithrombotic therapy, drug selection

System Response: Recommended conventional dual antiplatelet therapy

Correct Recommendation: Careful bleeding risk assessment required, individualised dosage adjustment

Failure Cause: Insufficient modelling of complex multi-drug interaction network effects

Improvement Measures: Graph attention network-based drug interaction prediction model development

Systematic Improvement Strategies:

1. Knowledge Graph Enhancement: Disease-symptom-treatment triplet knowledge graph construction

2. Uncertainty Quantification: Bayesian deep learning for prediction uncertainty estimation

3.Human-AI Collaboration: Automatic expert referral for high-uncertainty cases

4. Continual Learning Framework: Continuous model parameter updates from failure cases

5. Conclusion and Future Work

5.1. Summary of Key Findings and Technical Contributions

This study provides a holistic solution for hallucination control in medical information systems using semantic understanding, retrieval reinforcement, and multi-level validation. The results show that our approach is statistically superior: compared to the baselines established by most at present, it achieves a significant reduction in hallucinations of 23.7% ($p < 0.001$), 18.4% higher accuracy from ground truth, and a 15.2% greater recall success rate.

The technical contributions now range over novel terminology enhancement algorithms, adaptive confidence score mechanisms for automated information systems, and complete safety validation protocols with a specific focus on clinical practice. The technical contributions of this study include developing a novel medical terminology definition enhancement framework, which uses structured medical knowledge bases to improve the semantic understanding of clinical terms. Our multi-level validation method offers comprehensive hallucination detection and control for specific use in a medical environment. Confidence score methods integrated with dynamic threshold adjustment procedures lead to an adaptive validation strategy compatible with differing clinical judgments.

The experiment results show that our approach is practical for real-world medical advice-giving applications, with precision, recall, and clinical safety measurements all being significantly improved. At the same time, the computational efficiency of our system architecture, combined with significantly improved accuracy levels, makes it suitable for use in resource-constrained medical settings. Our methodology bridges a critical gap between currently available medical AI systems and future possibilities in artificial intelligence for healthcare.

5.2. Practical Implications for Medical Consultation Services

The practical applications of our research have connections with different aspects of medical care and consulting practice. Our system provides better reliability for automatic medical information systems and supports both patient education projects and clinical

decision support. The improved accuracy and decreased hallucination percentage make it possible to use the system in the delicate and sensitive medical environment.

Your medical institution can use the process of our team to upgrade AI medical consultation services while reducing the risk of misdirection of medical information, which could incorrectly result in harm. The system can be deployed to scale for multiple medical specialities and kinds of consultation. This makes it possible to achieve comprehensive coverage at consistent quality levels and conform to medical safety standards. This, coupled with integration capabilities, keeps our methodology practical for application in clinical workflow and patient care processes.

Our approach encourages greater use of AI in healthcare, ensuring that high-quality medical data is available without delay and encouraging healthcare workers to adopt new attitudes and skills. The methodology creates a foundation for developing more advanced medical AI systems. These can help to make decisions when treating patients in hospitals or companies, but they must have adequate human oversight and medical judgment.

5.3. Limitations and Directions for Future Research

Although the results of this study are remarkable, there are several limitations and open questions that warrant further research. To the present day, our system must rely on previously established medical knowledge bases and does not incorporate the latest findings of medical research or emerging clinical guidelines. The computational requirements for comprehensive validation procedures can be daunting in resource-restricted healthcare settings, making it essential to devise optimisation strategies for practical application.

A proper subject for future research would be to expand this methodology, introducing real-time medical literature updates as well as the latest clinical evidence into our knowledge enhancement framework. Another central line of inquiry is to increase the sophistication and reasoning abilities of AI systems. They must be capable of more complex medical challenge scenarios and multi-step queries. Finally, integration with electronic health record systems and consideration of a personalised medical history would improve the clinical usefulness and significance of responses.

Adapting our methodology such that it could be used for multilingual medical counselling and in an environment where different cultures mix presents an opportunity that is tailor-made for improving global health. Investigating federated learning methods may set up a model with the capability of improving applications while at the same time ensuring patients' privacy and hospital security requirements. Specialised versions for particular medical categories or healthcare facilities may make them additionally useful and bring higher adoption in diverse healthcare environments.

5.4. Safety Considerations and Risk Mitigation

Clinical Risk Assessment:

Medical AI systems carry inherent risks that demand systematic mitigation:

1. Patient Safety Protocols:

Mandatory clinical disclaimers for all medical advice

Automatic referral recommendations for emergency symptoms

Integration with clinical decision support alerts

2. Failure Mode Analysis:

False confidence scenarios: Implemented confidence calibration using temperature scaling

Out-of-distribution queries: Deploy uncertainty estimation with entropy-based detection

Adversarial inputs: Robust input validation and semantic coherence checking

3. Regulatory Compliance:

FDA AI/ML guidance adherence for medical device software

GDPR compliance for patient data processing in EU deployments

Integration with existing clinical workflows and EHR systems

Safety Guardrails:
Real-time monitoring dashboard for hallucination detection
Automatic escalation protocols for high-risk queries
Continuous learning from clinical feedback loops

Acknowledgments: I want to extend my sincere gratitude to Zhang, X., Peng, B., Tian, Y., Zhou, J., Jin, L., Song, L., and Meng, H. for their groundbreaking research on self-alignment for factuality in mitigating hallucinations in large language models via self-evaluation, as published in their article titled "Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation" in arXiv preprint (2024). Their innovative methodologies for hallucination detection and mitigation have significantly influenced my understanding of advanced techniques in language model reliability and have provided valuable inspiration for my own research in this critical area of medical AI safety. I would like to express my heartfelt appreciation to Saxena, P., Saxena, J., Gupta, K., Kumar SR, M., and Chauhan, P. for their innovative study on developing symptom-based diagnostic systems using BioBERT-NLI, FLAN-T5 and RAG models, as published in their article titled "Development of a Symptom-Based GI Cancer Diagnostic Bot Using BioBERT-NLI, FLAN-T5 and RAG Model" in the 2025 4th OPJU International Technology Conference (2025). Their comprehensive analysis and implementation of retrieval-augmented generation techniques in medical applications have significantly enhanced my knowledge of medical AI system development and inspired my research in this specialised field.

References

1. Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, and Z. Wei, "Disc-medllm: Bridging general large language models and real-world medical consultation," *arXiv preprint arXiv:2308.14346*, 2023.
2. X. Zhang, B. Peng, Y. Tian, J. Zhou, L. Jin, L. Song, and H. Meng, "Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation," *arXiv preprint arXiv:2402.09267*, 2024. doi: 10.18653/v1/2024.acl-long.107
3. A. Abdulnazar, R. Roller, S. Schulz, and M. Kreuzthaler, "Large language models for clinical text cleansing enhance medical concept normalisation," *IEEE Access*, 2024.
4. M. Motegi, M. Shino, M. Kuwabara, H. Takahashi, T. Matsuyama, H. Tada, and K. Chikamatsu, "Comparison of physician and large language model chatbot responses to online ear, nose, and throat inquiries," *Scientific Reports*, vol. 15, no. 1, p. 21346, 2025. doi: 10.1038/s41598-025-06769-1
5. Y. Tang, Y. Yuan, F. Tao, and M. Tang, "Cross-modal augmented transformer for automated medical report generation," *IEEE Journal of Translational Engineering in Health and Medicine*, 2025. doi: 10.1109/jtehm.2025.3536441
6. Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating hallucination in large language models via self-reflection," *arXiv preprint arXiv:2310.06271*, 2023.
7. P. Saxena, J. Saxena, K. Gupta, M. Kumar, and P. Chauhan, "Development of a symptom-based GI cancer diagnostic bot using BioBERT-NLI, FLAN-T5, and RAG model," In *2025, the 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0*, April, 2025, pp. 1-7.
8. L. Liu, X. Yang, J. Lei, Y. Shen, J. Wang, P. Wei, and K. Ren, "A survey on medical large language models: Technology, application, trustworthiness, and future directions," *arXiv preprint arXiv:2406.03712*, 2024.
9. C. Wang, Q. Chen, W. Shao, and X. He, "KEMedGPT: Intelligent medical pre-consultation with knowledge-enhanced large language model," In *2024 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, November, 2024, pp. 386-391. doi: 10.1109/medai62885.2024.00058
10. O. Tippins, T. Alvarez, J. Novak, R. Martinez, E. Thompson, and V. Williams, "Domain-specific retrieval-augmented generation through token factorisation: An experimental study," *Authorea Preprints*, 2024.
11. X. Zhang, and Y. Zhang, "A retrieval-augmented dialogue framework for multimodal medical consultation," In *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, December, 2024, pp. 618-623. doi: 10.1109/wi-iat62293.2024.00099
12. K. Chen, J. Qi, J. Huo, P. Tian, F. Meng, X. Yang, and Y. Gao, "A self-evolving framework for multi-agent medical consultation based on large language models," In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April, 2025, pp. 1-5. doi: 10.1109/icassp49660.2025.10889517
13. Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating LLM hallucination via self-reflection," In *Findings of the Association for Computational Linguistics: EMNLP 2023*, December, 2023, pp. 1827-1843.
14. Y. Wang, Y. Yang, C. Hu, L. Xu, J. Li, L. Sun, and J. Gao, "A medical consultation system based on a federated learning framework," In *2024 International Conference on Ubiquitous Computing and Communications (IUCC)*, December, 2024, pp. 567-572.
15. H. Y. Leong, Y. Gao, and S. Ji, "A gen AI framework for medical note generation," In *2024, the 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, November, 2024, pp. 423-429. doi: 10.1109/icaica63239.2024.10823004