Computers and Artificial Intelligence；Vol.3, No.1; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n1.4152
Published by: Zeus Press

Design and Evaluation of an Intelligent Agent-Based Home Healthcare System for Clinical Decision Support

Zhanyan Zhu[1], Yifan Peng[2], Leyi Xu[3], Weitong Hu[4,*], Jiale Feng[5], Yuqing Liu[6], Ke Zhou[7] and Yang Zhang[8]

[1]*School of Beijing-Dublin International Education, Beijing University of Technology, Beijing 100071, China*

[2]*School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*

[3]*School of Management, Xi'an Jiaotong University, Xi'an 710049, China*

[4]*School of Business English, Shandong University of Finance and Economics, Quanzhou 362200, China*

[5]*School of Law and Economics, Wuhan University of Science and Technology, Wuhan 430000, China*

[6]*School of International Business, Dongbei University of Finance and Economics, DaLian 116025, China*

[7]*School of Life Sciences, Zhuhai College of Science and Technology, Chengdu 610047, China*

[8]*School of Business, Beijing Language and Culture University, Beijing 100080, China*

**Corresponding author: Weitong Hu*

Abstract

The rapid development of large language models (LLMs) has stimulated growing interest in medical intelligent agents for clinical decision support. However, existing systems often suffer from limited grounding in authoritative medical knowledge, potential safety risks, and a tendency to generate definitive diagnostic conclusions without sufficient clinical context. In this work, we present the design of a medical intelligent agent aimed at supporting clinical decision-making through evidence-grounded information retrieval and safety-aware interaction. The proposed system focuses on two primary functions: (i) providing drug usage guidance, dosage information, and food–drug interaction warnings based on authoritative medical knowledge sources, and (ii) retrieving relevant clinical guidelines in response to patient-reported symptoms to assist clinicians with differential diagnostic considerations rather than definitive diagnoses. To mitigate safety risks, the agent is explicitly constrained to avoid diagnostic claims and instead emphasizes guideline-based recommendations and referral suggestions when appropriate. The agent integrates structured medical knowledge retrieval with natural language interaction, enabling users to obtain context-aware, interpretable and clinically relevant responses. By grounding outputs in curated medical references and enforcing non-diagnostic constraints, the system aims to reduce hallucinations and enhance reliability in medical consultations. This work highlights the potential of retrieval-augmented medical intelligent agents as supportive tools for clinical decision support, medical education, and patient-facing health information services, while underscoring the importance of safety, transparency, and scope limitation in medical AI deployment.

Keywords

medical intelligent agent, clinical decision support, retrieval-augmented generation, medication counseling, clinical guideline retrieval, differential diagnosis, medical AI safety

1. Introduction

Introduction Large language models (LLMs) have accelerated the development of conversational systems for health information access and clinical support. However, translating these models into reliable medical assistants remains challenging. In real-world use, users frequently ask medication-related questions that require precise, context-dependent answers—such as dosing instructions, drug–drug interactions, drug–food contraindications, and cautions for special populations (e.g., pregnancy, pediatrics, renal/hepatic impairment). Unconstrained LLM responses may exhibit unsupported claims, omission of contraindications, and overconfident recommendations, which can introduce safety risks in home healthcare settings where clinicians are not always present. These risks motivate medical agents that prioritize evidence traceability, task-appropriate scope, and safety-by-design interaction policies rather than open-ended "free-chat" medical advice.

Clinical practice guidelines (CPGs) and authoritative drug information sources provide systematically curated, evidenceinformed recommendations intended to standardize care and reduce preventable errors. Grounding an agent's outputs in such sources offers a principled pathway to improve reliability and interpretability. Yet, simply attaching generic retrieval-augmented generation (RAG) to an LLM does not guarantee safety. Guideline documents are structurally complex (e.g., recommendation statements, contraindication lists, dosage tables, and special-population sections), and naive chunking or similarity-based top-k retrieval may return irrelevant fragments or miss critical constraints. Moreover, user queries often omit essential context (age, comorbidities, concurrent medications), and the agent must decide when information is insufficient for a safe response. In high-risk scenarios, an agent should not attempt to "fill in" missing clinical details with guesses; instead, it should provide structured escalation guidance and point users back to authoritative sources or clinicians.

In this work, we present a guideline-first, safety-constrained intelligent agent for home healthcare support, focusing on two practical and high-impact use cases. First, the agent supports medication-oriented assistance, including dosage-related information and drug–drug/drug–food contraindication checks grounded in authoritative evidence. Second, it provides symptom-oriented decision support by retrieving relevant guideline recommendations and presenting supportive, non-diagnostic guidance—such as differential considerations, self-care boundaries, and when-to-seek-care prompts—without replacing clinicians. Importantly, the system is explicitly constrained to avoid autonomous diagnosis or treatment decisions. When queries are high risk, out of scope, or cannot be supported by retrieved evidence, the agent refuses to provide medical directives and instead returns a structured refusal with escalation recommendations.

To operationalize this design, we develop a CPG-tailored RAG pipeline with domain-specific knowledge engineering. We preprocess and segment multiple authoritative documents into clinically meaningful units aligned with guideline structure (e.g., indications, contraindications, dosing instructions, special populations, monitoring, warning statements, and interaction sections) and build searchable indexes to support reliable retrieval. At inference time, the system performs (i) query parsing and task classification (e.g., dosing vs. interaction vs. symptom support), (ii) safety triage to detect red-flag or high-risk requests, (iii) task-aware constrained retrieval over the guideline knowledge base, and (iv) structured response generation in which key claims are traceable to the retrieved evidence. This design aims to reduce hallucinated medical statements and improve interpretability by making the evidence basis explicit and auditable.

The knowledge base in the current prototype is constructed from four authoritative, publicly accessible sources (document titles, versions, and publication dates to be finalized in the appendix: WHO Model List of Essential Medicines, WHO International Classification of Diseases (ICD), a canonical traditional Chinese materia medica compendium, and public references covering symptoms of common baseline diseases . We evaluate the proposed system on a curated benchmark of approximately 400 queries spanning medication dosing, drug–drug/drug–food contraindications, special-population cautions, and symptom-support scenarios ([dataset construction details]). We compare against baseline LLM and generic RAG configurations ([baseline models and settings]) and report metrics emphasizing correctness, evidence-support consistency (citation alignment), hallucination rate, and safety compliance under high-risk prompts ([metrics definition]). In summary, this paper makes the following contributions:

- **Guideline-first, safety-constrained agent for medication-related support.** We present a medical conversational agent that prioritizes authoritative guideline grounding for medication counseling, interaction checking, and symptom support, while explicitly avoiding autonomous diagnosis or treatment directives and triggering conservative responses under high-risk or out-of-scope requests.
- **Task-aware RAG pipeline with evidence traceability.** We develop a structured retrieval-augmented generation pipeline that performs task routing and slot extraction, hybrid retrieval with query rewriting and reranking, and evidence-conditioned generation with explicit claim-to-evidence linkage to improve faithfulness and auditability.
- **Structured output schema and controlled regeneration.** We introduce a structured JSON output schema covering task type, medical entities, recommendations, safety signals (e.g., contraindications, interaction alerts, red flags), and citations; the system performs schema validation and triggers controlled regeneration when required fields are missing or inconsistent.
- **Safety review stage for policy compliance and escalation.** We implement an additional reviewer stage to detect unsafe or unsupported outputs (e.g., prohibited medical directives, missing escalation in red-flag contexts, or claims without evidence) and to enforce rewriting or refusal/escalation policies when safe correction is not feasible.
- **Reliability-oriented evaluation protocol.** We propose an evaluation protocol emphasizing correctness, evidence-support consistency, hallucination rate, and safety compliance, and report results on approximately 400 benchmark queries reflecting medication, interaction/contraindication, and high-risk scenarios.

2. Related Work

Healthcare conversational agents have been studied extensively for their potential to improve access to health information and support patient engagement. Laranjo et al. [1] provided an early systematic review of unconstrained natural-language conversational agents used for health-related purposes, summarizing prevalent application categories (e.g., education, monitoring, and behavioral support) and common evaluation practices. Importantly, they also highlighted persistent limitations in clinical validity, safety assessment, and integration into real-world workflows, which remain central concerns for deploying medical dialogue systems in practice. Beyond application-oriented studies, the design of language and interaction strategies has been recognized as a key determinant of user understanding, trust, and adherence in health communication. For example, Shan et al. [2] systematically reviewed language use in conversational agent–based health communication and identified recurring linguistic strategies (e.g., strategic wording, human-like conversational framing) as well as failure modes where inconsistent or ambiguous responses undermine effectiveness and safety. These findings motivate explicit attention to safety-aware communication and controllable generation behaviors in medical conversational agents.

More recently, large language models (LLMs) have been introduced into medical conversational systems, enabling more flexible interaction and stronger open-ended reasoning. Representative efforts such as HuatuoGPT [3] demonstrate the feasibility of domain adaptation for medical consultation via medical corpora and alignment techniques. However, many LLM-based medical agents are still primarily optimized for general consultation and may lack explicit mechanisms tailored to medication-centric guidance (e.g., drug–drug/drug– food contraindications) and safety-constrained interaction policies. As a result, these systems can produce overly confident or implicitly diagnostic outputs, raising concerns about reliability in patient-facing settings.

To address factuality and safety, a growing body of work explores retrieval-augmented generation (RAG) for clinical medicine, where external evidence (e.g., guidelines) is retrieved and used to ground generation. For instance, Almanac [4] incorporates retrieval from curated clinical resources and reports improved clinician-rated factuality and safety compared with non-retrieval baselines. Meanwhile, recent benchmark-driven studies (e.g., MIRAGE/MedRAG) further investigate how corpus choice, retriever design, and reranking affect medical RAG performance and failure modes [5]. These findings suggest that end-to-end workflow design (retrieval, evidence selection, and output structuring) is crucial for robust deployment.

In summary, prior work establishes strong foundations for healthcare conversational agents and LLM-based medical systems, but important gaps remain for patient-oriented, medication-focused, and safety-governed workflows. In contrast, our work targets an end-to-end medical intelligent agent that integrates symptom-driven guideline retrieval with medication usage guidance and dietary contraindication support, while enforcing non-diagnostic constraints, structured outputs, and conservative escalation behaviors to improve practical usability and patient safety. Retrieval-augmented generation (RAG) has recently been explored to improve factuality and safety in clinical language generation by grounding responses in external evidence such as guidelines and curated medical resources. Almanac reports improved clinician-rated factuality and safety compared with non-retrieval baselines, suggesting that guideline-grounded generation can be beneficial when evidence is carefully retrieved and integrated [4]. Beyond system proposals, benchmark-driven analyses further investigate how corpus choice, retriever design, and reranking strategies affect medical RAG performance and failure modes, indicating that end-to-end component choices can substantially influence reliability [5].

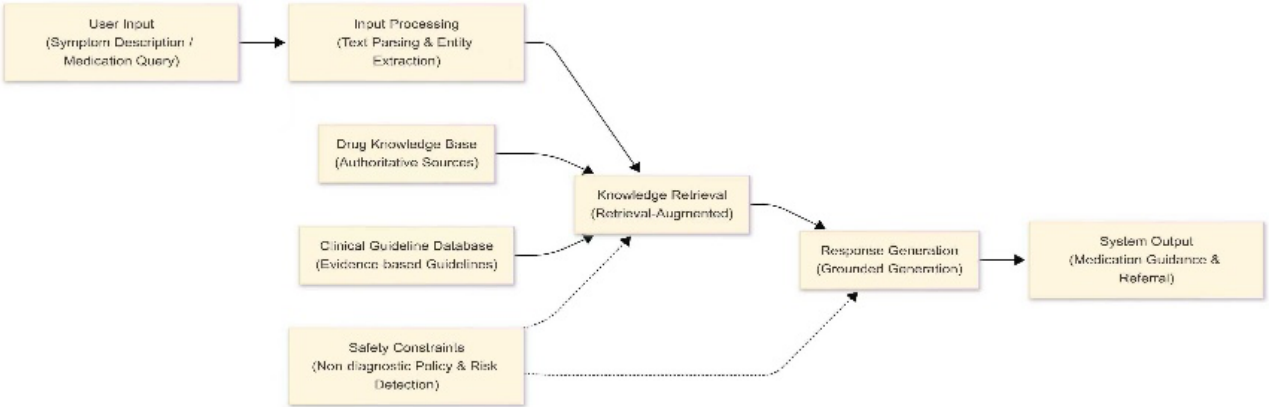
In parallel, agentic and multi-stage reasoning workflows have been studied as a way to improve complex medical reasoning. Role-based multi-agent collaboration has been shown to enhance zero-shot medical reasoning via deliberation among complementary roles, providing a useful perspective on staged orchestration beyond single-pass prompting [6]. More generally, ReAct-style agentic workflows with explicit tool use have been explored in clinical task settings, motivating our use of modular orchestration and tool-augmented reasoning for controllability and robustness [7].

Multimodal clinical systems increasingly combine text with additional modalities and align external medical knowledge, reinforcing the relevance of multimodal perception coupled with evidence-grounded reasoning for practical clinical decision support [8].

3. Methodology Analysis

3.1 System Overview

Figure 1: Overall architecture of the proposed medical intelligent agent

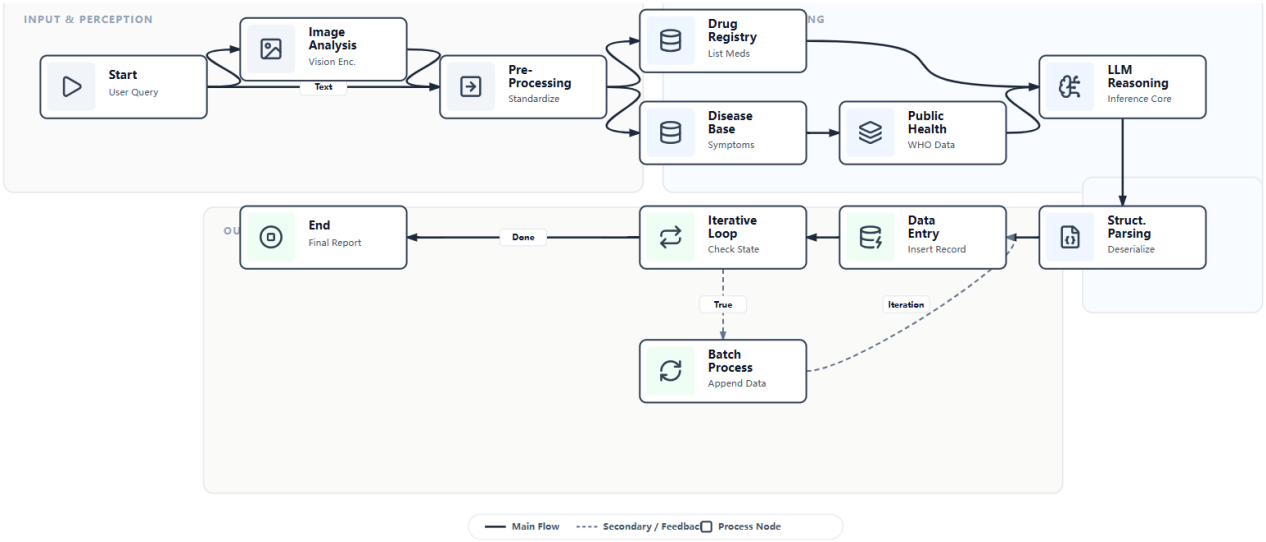


- **Knowledge sources.** The current knowledge base is constructed from four document collections: (1) *Compendium of Materia Medica (Bencao Gangmu)*; (2) *WHO-registered common medicines list*; (3) *WHO-registered diseases*; and (4) *Common basic diseases and their symptoms*. For reproducibility, all experiments are conducted on a fixed snapshot of these sources.
- **Model configuration.** Response generation and policy validation are implemented using Doubao-family large language models (LLMs) with deep reasoning enabled in our deployment. Decoding parameters (e.g., temperature and nucleus sampling) are fixed during evaluation.

3.2 Workflow Orchestration

The agent is implemented as an orchestrated workflow consisting of modular stages with explicit inputs and outputs. The workflow proceeds as follows.

Figure 2: Orchestrated workflow of the proposed medical intelligent agent



Step 1: Multimodal input understanding. The system accepts user text and optionally images (e.g., medication packaging or screenshots containing medication names). A multimodal understanding component extracts salient entities and key spans from images and merges them with the user text to form a normalized query representation q'.

Step 2: Task routing and slot extraction. The normalized query q' is parsed into structured slots such as medication entities, symptoms, and contextual cues (e.g., duration or severity markers). The system routes the query into task types including *medication counseling*, *interaction/contraindication*, *symptom support*, and *high-risk triage*. Task routing combines lightweight rules with LLM-based classification to handle colloquial and incomplete user inputs.

Step 3: Safety triage and scope control. Before retrieval and generation, the system checks whether the request is high risk or out of scope. When triggered (e.g., emergency red-flag patterns or insufficient context for safe guidance), the system avoids producing medical directives and instead returns structured escalation guidance (seek urgent care / consult clinicians or pharmacists).

Step 4: Knowledge base retrieval. If the request passes triage, the system retrieves evidence from the knowledge base using the platform's hybrid retrieval strategy. For each request, the retriever returns up to the top-20 evidence units ranked by relevance. To improve robustness, we optionally rewrite the query to better match the indexing space and apply result re-ranking using relevance and quality signals. Items below a configurable minimum-match threshold are filtered out. Retrieval is subject to access control so that only user-owned or developer-authorized documents are searchable.

Step 5: Constrained generation with evidence binding. A Doubao-based generator produces an answer conditioned on the normalized query and the retrieved evidence. The prompt enforces two constraints: (i) key claims must be supported by retrieved evidence; and (ii) the output follows a structured template with citations that link claims to evidence identifiers. If evidence is insufficient, the system defaults to conservative behaviors, including asking clarifying questions, issuing a limited informational response, or refusing with escalation guidance.

Step 6: Structured parsing. The generated output is parsed into a structured JSON record to support consistent rendering and auditing. Instead of free-form text, the record explicitly encodes the task category, medical entities, recommendations, safety signals (e.g., contraindications, interaction alerts, red flags, and when-to-seek-care guidance), and the corresponding citations. If parsing fails or required information is missing, the system triggers controlled regeneration under stricter constraints.

Step 7: Reviewer validation (Doubao). To reduce unsafe or unsupported outputs, an additional reviewer stage implemented with Doubao evaluates the structured response for: (i) prohibited content (diagnosis or treatment directives), (ii) missing escalation in red-flag contexts, and (iii) evidence alignment (claims without citations). When violations are detected, the system either rewrites the response under stricter constraints or returns a refusal/escalation response when safe correction is not feasible.

To improve auditability, our system enforces claim-level evidence traceability, requiring that generated statements be supported by retrieved sources and that citations be verifiable at the level of specific claims rather than only at the document level [9]. Prior studies also show that even when LLMs provide citations or URLs, a non-trivial portion of statements may remain unsupported by the referenced sources, motivating automated source verification and conservative generation policies in safety-critical medical settings [10].

3.3 Knowledge Base Construction and Indexing

Each document collection is preprocessed to remove formatting artifacts and preserve semantically meaningful structure (headings, bullet points, and tables where applicable).

Chunking and segmentation. Document segmentation follows the platform-default chunking configuration (*auto segmentation*), which produces retrievable units aligned with the underlying knowledge-base ingestion pipeline. Unless otherwise specified, we use the default segmentation parameters provided by the platform and keep the configuration fixed across all experiments for reproducibility.

Vector indexing. Each chunk is embedded into a dense vector representation using the platform-provided embedding configuration and stored in a vector index. At query time, the normalized query q' is embedded and used to retrieve relevant chunks. Retrieval is performed with maximum recall $k = 20$ and a minimum match threshold θ, both fixed during evaluation.

3.4 Evidence-Traceable Output and Schema

To ensure consistent presentation, auditability, and evaluation, the system emits a structured JSON response with explicit evidence traceability. A simplified schema is shown below (the full schema is reported in the appendix).

Listing 1: Structured output schema (example).

```
{
"task_type " : "medication | interaction |
symptom_support | refusal ", "drug_name " : [ " ... " ],
"symptom " : [ " ... " ],
"recommendations " : [ " ... " ],
"contraindications " : [ " ... " ],
"interaction_alerts " : [{ "pair " : [ "A " , "B " ], "risk " : " ... " , "action " : " ... " }],
"red_flags " : [ " ... " ],
"when_to_seek_care " : [ " ... " ],
"citations " : [{ "source " : "WHO-registered common
medicines list " , "chunk_id " : "S2-C015 " }] }
```

3.5 Implementation Notes

The workflow is deployed in a platform that supports multimodal inputs, vector-based knowledge retrieval, and multi-stage LLM calls. We enable the platform settings for hybrid retrieval, query rewriting,

and result reranking, and restrict retrieval to user-owned or developer-authorized documents. Unless otherwise specified, all configurations are kept fixed for evaluation to ensure reproducibility.

4. Experiment and Evaluation

4.1 Experimental Setup

We situate our safety-oriented evaluation within recent medical safety bench marking efforts that operationalize clinically meaningful unsafe behaviors (e.g., harmful recommendations, missing escalation in red-flag contexts) and quantify policy compliance at scale [11].

This section describes the experimental setup used to evaluate the proposed medical intelligent agent. We design scenario-based evaluations to assess the system's ability to provide accurate medication counseling and retrieve relevant clinical guidelines under safety constraints.

Specifically, we construct a set of test scenarios covering common medication inquiries and symptom-based guideline retrieval tasks. Each scenario consists of a user query and a corresponding reference answer derived from authoritative medical sources. The system responses are generated using the same configuration as in deployment, without manual intervention.

To ensure reproducibility, all experiments are conducted using a fixed system configuration, including the same retrieval strategy, knowledge sources, and response generation parameters.

4.2 Evaluation Metrics

We evaluate the proposed system from the perspectives of accuracy, safety, and clinical relevance.

For medication counseling tasks, we assess whether the generated responses provide correct drug usage guidance, dosage information, and food–drug interaction warnings consistent with authoritative references.

For clinical guideline retrieval tasks, we evaluate the relevance of the retrieved guidelines with respect to patient-reported symptoms, focusing on their usefulness for differential diagnostic considerations rather than definitive diagnoses.

In addition, we analyze safety-related behaviors, including the absence of diagnostic claims and the appropriateness of referral suggestions in potentially high-risk scenarios. All evaluations are performed through expert review or reference-based comparison to ensure reliability.

5. Conclusion

This study takes AI agents as the central object of investigation. We systematize the foundational concepts and technical implications of AI agents, examine the feasibility of constructing an end-to-end agent system from a practical implementation standpoint, and analyze the design rationale, structural characteristics, and application-oriented properties of representative core architectures. We further focus on retrieval-augmented generation (RAG), investigating its use in agent systems and its internal operational mechanisms, and summarizing the key procedures and essential alignment points for integrating RAG with agent architecture in real deployments.

5.1 Contributions

Our investigation indicates that RAG can effectively address recurrent practical issues in AI agents, including information staleness, insufficient knowledge coverage, and limited response accuracy. Moreover, the results underscore that a well-founded system architecture is pivotal for ensuring the stable and efficient operation of AI agents, and that the degree of alignment between an agent's technical architecture and its target application scenarios directly determines its practical effectiveness. Overall, our findings support the practical relevance of AI agents in academic research, industrial practice, and related domains, and delineate key design considerations and technical requirements for building RAG-based agent systems, thereby offering implementation-oriented references for subsequent development, optimization, and real-world adoption.

5.2 Limitations

This work primarily provides a conceptual and architectural investigation, together with feasibility-oriented verification, rather than an exhaustive evaluation across a wide range of professional domains and business scenarios. In addition, while the integration principles between RAG and agent architecture are summarized, the practical effectiveness of specific architectural choices may vary with scenario characteristics and domain requirements, calling for further empirical testing and iterative refinement in real deployments.

5.3 Future work

Future research will prioritize in-depth architectural optimization for RAG-based agent systems and improve overall technical performance through approaches such as core algorithm refinement and architectural recon-figuration. Concurrently, we will evaluate AI agents in more specialized domains and concrete business scenarios, collect practical deployment data and user feedback, and iteratively adjust system design according to domain-specific requirements and scenario characteristics. These efforts are expected to further enhance the adaptability, practicality, and operational efficiency of agent technology across diverse scenarios, and to promote deeper integration with both academic research and industrial practice.

6. Discussion and Limitations

6.1 Lessons from Representative Failure Cases

Prior experiences with medical AI systems suggest that clinical risk frequently emerges from a coupled failure of (i) evidence currency, (ii) workflow alignment, and (iii) governance mechanisms, rather than from model errors alone.

6.1.1 Case 1: IBM Watson for Oncology

Public investigations have reported that Watson for Oncology produced incorrect or unsafe treatment suggestions in certain scenarios, including recommendations inconsistent with prevailing clinical practice and insufficiently flagged safety constraints (e.g., drug conflicts). Such failures highlight two recurring pitfalls: reliance on limited or weakly structured training evidence without adequate real-world validation, and insufficient alignment with clinicians' decision-making processes. In addition, IBM later divested major Watson Health assets, reflecting the difficulty of translating ambitious clinical AI positioning into sustainable, validated deployment [12], [13].

6.1.2 Case 2: Rare Diseases and Pediatric Diagnostic Contexts

Rare diseases remain a high-risk context for AI-assisted decision support due to atypical symptom patterns, sparse samples, and heterogeneous clinical presentations, often contributing to delayed or incorrect diagnosis in real-world pathways [14]. In pediatric pneumonia-related contexts, recent evidence also cautions that general-purpose AI tools can show inconsistent and unreliable diagnostic performance, under-scoring the risks of unsupervised or over-trusted AI outputs in high-stakes pediatric settings [15]. high-stakes pediatric settings [15], [16].

6.1.3 Implications for System Framing

These cases motivate a conservative and responsibility-preserving framing: medical AI should be deployed as *decision support* rather than diagnostic substitution, with explicit scope control, auditable accountability, and workflow-integrated human verification.

6.2 System Design Choices and Their Practical Implications

In response to the above lessons, our system is designed as a guideline-grounded decision-support agent for family users. It adopts retrieval-augmented generation (RAG) over authoritative clinical guidelines, explicitly displays supporting evidence excerpts alongside responses, and maintains usage logs and version

management for traceability and pos thoc review. These design choices aim to improve evidence currency, reduce unsupported generation, and support accountability in real-world use.

6.3 Limitations

Despite these safeguards, several limitations constrain the system's applicability and the strength of its safety assurances.

Despite its promise, RAG does not universally improve medical answering quality: evidence retrieval and selection can become failure points, and inappropriate evidence can mislead generation or create a false sense of support. Recent analyses therefore caution that standard RAG pipelines may yield inconsistent gains, reinforcing our emphasis on safety constraints, structured auditing, and conservative fallback behaviors when evidence is weak or ambiguous [17].

6.3.1 Scope Constrained to Low-acuity, High-frequency Needs

The agent is primarily oriented toward common minor illnesses, medication guidance, and safe-medication education for families. This positioning improves relevance to everyday household health needs, but it also limits applicability to complex, high-acuity, or atypical presentations (e.g., severe comorbidities, rapidly deteriorating symptoms, and rare diseases), where guideline generalities may be insufficient and specialist assessment is indispensable.

6.3.2 Decision Support Cannot Replace Clinical Diagnosis

Although the agent is guideline-grounded and evidence-supported, it is not a diagnostic authority. Its outputs should be interpreted as informational guidance rather than definitive diagnosis or treatment. Consequently, the system cannot be considered safe for autonomous decision-making; it requires escalation to professional care when symptoms are severe, uncertain, or outside the guideline scope.

6.3.3 Residual Gaps and Variability in Guideline Coverage

RAG-based grounding reduces hallucination risk but does not eliminate limitations stemming from incomplete or uneven guideline coverage. Some conditions, populations, or region-specific practices may be under-represented in the guideline corpus, which can lead to lower-quality support in edge cases. Moreover, guideline updates and regional differences necessitate continuous corpus maintenance and explicit disclosure of coverage boundaries.

6.3.4 Dependence on User-provided Information in Consumer Settings

Family users may provide incomplete, ambiguous, or inaccurate symptom descriptions and medication histories. Such input uncertainty can degrade retrieval quality and downstream responses, especially for medication guidance where contraindications depend on detailed context (e.g., allergies, comorbidities, concurrent drugs). While evidence excerpts improve transparency, they do not fully compensate for missing or incorrect inputs.

6.3.5 Over-reliance Risk Persists Despite Evidence Display

Even with displayed citations and supporting excerpts, users may over-trust fluent responses. This risk is amplified by high cognitive load, low health literacy, or strong demand for certainty. Therefore, safety relies not only on technical grounding but also on user-facing interaction design (e.g., prominent uncertainty cues, red-flag symptom prompts, and escalation recommendations).

6.3.6 Traceability does not Equal Clinical Validation

Logging and version management support accountability and debugging, but they do not substitute for rigorous clinical validation. The system's real-world impact on user behavior, safety outcomes, and healthcare utilization cannot be inferred from architectural features alone and requires prospective evaluation.

6.4 Future Work

Future work will focus on two directions. First, we will strengthen technical and governance mechanisms by continuously updating the guideline corpus, expanding coverage across populations and regions, and

conducting periodic audits of retrieval behavior and safety-related failure modes. Second, we will pursue linkage with hospital systems and clinical partners to enable supervised integration pathways (e.g., referral and escalation workflows, clinician-reviewed feedback loops), thereby supporting more reliable handoff from family-facing decision support to professional care. These efforts will facilitate a translational trajectory in which evidence-grounded consumer decision support evolves toward clinically integrated infrastructure under appropriate oversight and evaluation.

7. Appendix and Ethical Considerations

The deployment of medical decision-support agents in clinical practice raises ethical challenges that extend beyond model accuracy, encompassing accountability, equity, human factors, and information governance. In high-stakes settings, ethical compliance should be operationalized as a set of auditable controls spanning system design, deployment, and post-deployment monitoring. This section identifies key risks and proposes concrete mitigation measures to support the responsible use of guideline-grounded decision-support agents.

7.1 Ambiguous Liability and Accountability

Although decision-support agents are intended to function as information providers, incorrect or misleading guidance may contribute to adverse outcomes and trigger disputes over responsibility among hospitals, technology vendors, and clinicians. To mitigate this risk, responsibility boundaries should be explicitly defined in governance and contractual instruments. Product documentation and clinical use protocols should state that agent outputs serve as guideline references and do not replace independent clinical judgment, with the licensed clinician retaining responsibility for patient care decisions.

Beyond statements of intent, accountability requires traceability. Institutions should maintain comprehensive lifecycle logs, including query context, retrieved evidence (e.g., guideline passages), model outputs, timestamps, and user interactions (e.g., acceptance, modification, override). Such logging supports incident investigation, post hoc review, and quality improvement, and it also enables clearer attribution of failure modes (e.g., knowledge base gaps, retrieval errors, or misuse). Where applicable, escalation workflows should be established for high-risk recommendations, enabling supervision or second review before action.

7.2 Algorithmic Fairness and Bias

Fairness concerns arise primarily from (i) incomplete or uneven coverage of clinical guidelines (e.g., rare diseases, special populations, or region-specific practice patterns) and (ii) bias introduced through semantic retrieval and ranking mechanisms. These limitations can lead to differential quality of decision support across patient groups or clinical scenarios, undermining the principle of equitable healthcare access.

Mitigation should proceed along two complementary fronts. First, the guideline corpus should be continuously expanded and curated to improve representativeness across diseases, geographies, and demographic groups, with explicit documentation of coverage boundaries and known gaps. Second, periodic fairness audits should be conducted to detect systematic disparities in outputs. Audits may stratify performance by clinically relevant attributes (e.g., age groups, sex, comorbidity profiles, and region-specific contexts) and evaluate whether the agent's retrieval and responses disproportionately disadvantage specific populations. Findings should feed into iterative updates of retrieval policies, ranking objectives, and corpus curation, accompanied by versioning to ensure that changes are trackable and reviewable.

7.3 Over-Reliance and Human-Factors Risk

A central ethical risk in clinical decision support is clinicians' over-reliance on structured agent responses, which may reduce critical engagement with primary evidence and erode evidence-based reasoning. This risk is particularly salient for early-career clinicians and in high-throughput environments where cognitive load is high.

To preserve professional autonomy and analytical rigor, the agent should be positioned and designed as a supportive tool rather than an authority. Training and usage guidelines should require clinicians to verify key

conclusions against primary guideline sources and to integrate patient-specific factors (e.g., comorbidities, preferences, and contextual constraints) into final decisions. Interface and workflow design can further discourage passive reliance by emphasizing cited evidence, presenting uncertainty or scope limitations, and providing prompts that encourage verification and clinical reflection. Educational programs should reinforce complementary use—agent assistance plus clinical judgment—rather than substitution.

7.4 Data Privacy and Security

Decision-support agents often process sensitive health information, creating risks related to privacy breaches, unauthorized access, and secondary use beyond clinical intent. Responsible deployment therefore requires strong information governance and security controls. Data minimization principles should be applied (collecting only what is necessary for clinical support), and secure handling should be ensured through access control, encryption in transit and at rest, and strict retention policies. Where feasible, de-identification or pseudonymization should be adopted for logs and analytics while preserving sufficient context for safety review. Institutions should also implement role-based permissions, audit trails for access, and incident-response procedures aligned with applicable regulations and hospital policies.

7.5 Transparency, Explainability, and Scope Control

To support safe adoption, the agent's outputs should be transparent with respect to their evidence basis and applicability. In guideline-grounded systems, transparency can be operationalized by displaying the retrieved guideline sources, the relevant excerpts, and the version/date of the referenced guideline. Clear scope statements should be provided, including what the system can and cannot support, and under what circumstances clinicians must disregard or escalate recommendations (e.g., when guideline coverage is absent, evidence is outdated, or patient context is atypical). Such scope control reduces misinterpretation, supports informed reliance, and strengthens accountability.

7.6 Post-Deployment Monitoring and Continuous Improvement

Ethical assurance is not a one-time deliverable; it requires continuous monitoring. Institutions should implement post-deployment surveillance to detect error patterns, drift in guideline currency, and emerging fairness issues. Mechanisms for user feedback and incident reporting should be built into clinical workflows, enabling rapid escalation and iterative improvement. Updates to the guideline corpus, retrieval components, and response policies should be version-controlled and accompanied by change logs and periodic re-validation.

These measures jointly support a safety culture that treats decision-support agents as continuously governed clinical infrastructure.

References

[1] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, "Conversational agents in healthcare: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248– 1258, 2018.

[2] R. Shan, X. Ding, L. Chen, A. B. Kocaballi, L. Laranjo, and E. Coiera, "Language use in conversational agent-based health communication: Systematic review," *Journal of Medical Internet Research*, vol. 24, no. 7, p. e37403, 2022.

[3] H. Zhang, X. Li, Y. Wang, Y. Li, Y. Zhang, Y. Shen, R. Zhang, Z. Liu, and M. Sun, "Huatuogpt: Towards taming language model to be a doctor," 2023.

[4] C. Zakka, A. Chaurasia, R. Shad, A. R. Dalal, J. L. Kim, M. Moor, K. Alexander, E. Ashley, J. Boyd, K. Boyd, K. Hirsch, C. Langlotz, J. Nelson, and W. Hiesinger, "Almanac: Retrieval-augmented language models for clinical medicine," 2023.

[5] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval- augmented generation for medicine," 2024.

[6] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein, "Medagents: Large language models as collaborators for zero-shot medical reasoning," 2023.

[7] L. Yue, S. Xing, J. Chen, and T. Fu, "Clinicalagent: Clinical trial multi- agent system with large language model-based reasoning," 2024.

[8] Y. Zhu, C. Ren, S. Xie, S. Liu, H. Ji, Z. Wang, T. Sun, L. He, Z. Li, X. Zhu, and C. Pan, "Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models," 2024.

[9] K. Wu, E. Wu, A. Casasola, A. Zhang, K. Wei, T. Nguyen, S. Riantawan, P. Shi, D. E. Ho, and J. Zou, "How well do llms cite relevant medical references? an evaluation framework and analyses," 2024.

[10] K. Wu, E. Wu, K. Wei, A. Zhang, A. Casasola, T. Nguyen, S. Riantawan, P. Shi, D. Ho, and J. Zou, "An automated framework for assessing how well llms cite relevant medical references," *Nature Communications*, vol. 16, p. 3615, 2025.

[11] T. Han, A. Kumar, C. Agarwal, and H. Lakkaraju, "Medsafetybench: Evaluating and improving the medical safety of large language models," 2024.

[12] C. Ross, "Ibm's watson recommended 'unsafe and in- correct' cancer treatments, internal documents show," STAT News, Jul. 2018, accessed: 2026-02-13. [Online]. Avail- able: https://www.statnews.com/2018/07/25/ibm-watson-recommended- unsafe-incorrect-treatments/

[13] IBM, "Francisco partners to acquire ibm's healthcare data and analytics assets," IBM Newsroom, Jan. 2022, accessed: 2026-02-13. [Online]. Available: https://newsroom.ibm.com/2022-01-21-Francisco-Partners-to-Acquire-IBMs-Healthcare-Data-and-Analytics-Assets

[14] EURORDIS-Rare Diseases Europe, "The diagnosis odyssey of people living with a rare disease," Rare Barometer survey report, May 2024, accessed: 2026-02-13. [Online]. Available: https://www.eurordis.org/publications/rb-diagnosis-odyssey/

[15] J. Gillette, M. Lu, and T. F. Heston, "Large language models perform at chance level in the diagnosis of pediatric pneumonia using chest radiographs," *Cureus*, vol. 17, no. 9, p. e92596, Sep. 2025.

[16] UW Medicine Newsroom, "Ai fails to reliably detect pediatric pneumonia on x-ray," UW Medicine - Newsroom, Dec. 2025, accessed: 2026-02-13. [Online]. Available: https://newsroom.uw.edu/blog/ai-fails- to-reliably-detect-pediatric-pneumonia-on-x-ray

[17] H. Kim, J. Sohn, A. Gilson, N. Cochran-Caggiano, S. Applebaum, H. Jin, S. Park, Y. Park, J. Park, S. Choi, B. A. H. Contreras, T. Huang, J. Yun, E. F. Wei, R. Jiang, L. Colucci, E. Lai, A. Dave, T. Guo, M. B. Singer, Y. Koo, R. A. Adelman, J. Zou, A. Taylor, A. Cohan, H. Xu, and Q. Chen, "Rethinking retrieval-augmented generation for medicine: A large-scale, systematic expert evaluation and practical insights," 2025.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights