A Multidimensional Evaluation Framework for Healthcare RAG Conversational Agents

Jose Escarraman-Reyes[1 [0009−0009−4938−4416]], Andrea Morales-Garzón[1 [0000−0002−3458−0694]], M. Dolores Ruiz[1 [0000−0003−1077−3173]], and Maria J. Martin-Bautista[1 [0000−0002−6973−477X]]

Dept. of Computer Science and Artificial Intelligence, University of Granada (Spain)
jose.escarraman@ugr.es, {amoralesg, mdruiz, mbautis}@decsai.ugr.es

Abstract. Retrieval-augmented generation (RAG) has improved the factual grounding capabilities of healthcare conversational agents by integrating external medical knowledge into response generation. However, evaluating these systems remains challenging, as traditional Natural Language Processing (NLP) metrics do not adequately capture retrieval quality, clinical safety, factual grounding, or usability. This paper reviews current evaluation approaches for healthcare conversational agents, with emphasis on RAG-based systems, and synthesizes the main evaluation dimensions, metrics, and methodological practices reported in recent literature. Based on these findings, we propose a multidimensional evaluation framework integrating retrieval quality, factual grounding, clinical safety, usability, and evaluation transparency within a unified workflow combining automated and clinician-centered assessment. The proposed framework aims to contribute toward more standardized, reproducible, and clinically meaningful evaluation practices for healthcare conversational AI.

Keywords: conversational agents · clinical validation · LLM · retrieval-augmented generation · evaluation metrics

1 Introduction

The adoption of Large Language Models (LLMs) based conversational agents in healthcare has increased rapidly in recent years [1]. Although systems based on models such as GPT-4 have shown promising clinical capabilities [15,8], establishing reliable and reproducible evaluation methodologies for healthcare chatbots remains a major challenge.

Recent healthcare deployments increasingly rely on RAG architectures [17], which combine LLMs with external medical knowledge bases to improve factual grounding and reduce hallucinations. By grounding responses in curated clinical sources, RAG systems can improve traceability, safety, and clinical relevance [20]. However, they also introduce additional evaluation challenges related to retrieval quality, grounding fidelity, faithfulness, and error propagation across retrieval and generation stages [17,20].

Recent frameworks and reporting guidelines such as CHART [6] and HAICEF [13] have advanced standardisation efforts for healthcare chatbot evaluation, covering dimensions including safety, trustworthiness, and usability. In parallel, automated frameworks such as RAGAS (Retrieval-Augmented Generation Assessment Score) assess retrieval quality and grounding fidelity in scalable ways [10]. However, existing approaches either focus on general chatbot evaluation or on isolated RAG-specific metrics, without providing a unified framework that explicitly addresses the retrieval–generation interaction characteristic of healthcare RAG systems.

The objectives of this paper are as follows:

- To provide a structured review of current strategies for evaluating and validating LLM-based conversational agents in clinical settings, with particular emphasis on RAG architectures.
- To identify and synthesize the main evaluation dimensions, metrics, and methodological practices reported in the literature.
- To propose a practical, RAG-oriented evaluation framework that explicitly captures the interaction between retrieval quality, response grounding, and clinical safety.
- To illustrate the application of this framework through a small empirical use case based on representative clinical queries.

Building on these objectives, this work makes three main contributions: (1) it consolidates and organizes a fragmented body of literature into a coherent view of evaluation practices for RAG-based clinical conversational agents; (2) it introduces a structured evaluation framework that operationalizes key dimensions and standardizes the process; (3) it shows the practical applicability of the proposed framework through an empirical illustration, providing initial evidence of how different configurations behave under clinically relevant evaluation criteria.

2 Lessons from Existing Evaluation Approaches

This section synthesizes the main findings identified across the reviewed literature on healthcare conversational agent evaluation, with particular emphasis on RAG systems. Specifically, we examine the core evaluation dimensions proposed in existing frameworks, summarize the most commonly adopted metrics and methodological strategies, and review representative empirical studies illustrating how these approaches are applied in practice.

To identify the most relevant work, we conducted a focused literature review combining searches on the main biomedical and technical databases, followed by title/abstract screening, full-text review, and backward/forward citation chaining. The resulting set of studies was then qualitatively synthesized to extract recurring evaluation dimensions, metrics, and methodological patterns.

2.1 Core Evaluation Dimensions for healthcare chatbots

Recent literature increasingly recognizes that evaluation of healthcare conversational agents cannot rely solely on generic NLP benchmarks or isolated performance indicators. Instead, existing frameworks converge toward a multidimensional perspective combining technical performance, clinical safety, transparency, usability, and operational considerations. Although these frameworks differ in scope and methodological depth, they consistently identify several core dimensions as essential for clinical deployment.

One of the most emphasized dimensions is *clinical safety and factual reliability*. Frameworks proposed by Ding et al. [9], Hua et al. [13], and Asgari et al. [3] stress the importance of minimizing hallucinations, omissions, misleading recommendations, and other clinically harmful outputs. In healthcare settings, evaluation must therefore consider not only linguistic quality, but also the potential clinical consequences of incorrect or weakly grounded responses. This concern is particularly relevant in RAG-based systems, where retrieval deficiencies may propagate into factually incorrect but plausible responses.

Another recurring dimension involves *trustworthiness, transparency, and explainability*. Existing frameworks emphasize reporting model characteristics, documenting limitations, and enabling traceability between generated outputs and supporting evidence. For example, CHART highlights transparency in data sources, evaluation procedures, and harm reporting [6], while HAICEF incorporates broader notions of fairness, accountability, and operational trustworthiness [13]. In RAG architectures, these concerns naturally extend to retrieval quality, grounding fidelity, and evidence attribution.

A further dimension concerns *usability and user-centered performance*. Clinically acceptable systems must not only be accurate, but also understandable, contextually appropriate, and aligned with the needs of clinicians and patients. Existing frameworks therefore include readability, empathy, communication clarity, perceived helpfulness, and user trust [9,13].

Finally, several frameworks stress the importance of *operational and methodological rigor*, including structured validation procedures, deployment-stage evaluation, and transparent reporting practices [9,6]. Despite progress toward standardisation, existing approaches remain fragmented and often insufficiently adapted to the retrieval-generation interaction characteristic of RAG systems.

2.2 Evaluation Strategies and Metrics

The reviewed literature consistently recommends combining automated and evaluation strategies centered in human for healthcare conversational agents [2]. Automated metrics provide scalable and reproducible assessment, whereas human evaluation remains essential for identifying clinically meaningful errors and safety risks. Consequently, recent work increasingly advocates hybrid evaluation pipelines integrating automated assessment with expert adjudication, particularly in high-risk clinical settings.

Automated evaluation commonly targets linguistic quality, semantic relevance, factual consistency, and retrieval performance. Traditional metrics such as BLEU, ROUGE, and BERTScore remain widely used, although their limitations for assessing clinical correctness and safety are well recognized [2,20]. More clinically oriented approaches prioritize factuality and grounding metrics, including faithfulness checks, semantic similarity, FactScore, and readability indices for patient-facing systems [19,2].

RAG-based systems introduce additional evaluation requirements related to retrieval quality and grounding fidelity [20]. Frameworks such as RAGAS evaluate *faithfulness*, *answer relevancy*, *context precision*, and *context recall*, often complemented by Recall@k, MRR, nDCG, entailment-based checks, and LLM-assisted evaluators [10]. These metrics help distinguish failures caused by retrieval deficiencies, hallucinations, or incomplete synthesis, although automated evaluators may require complementary expert validation in medical contexts [10,20].

Human expert evaluation remains central for assessing clinical safety, appropriateness, trustworthiness, and empathy through structured clinician review, Likert-scale ratings, and safety-oriented assessment of hallucinations, contraindications, omission severity, and escalation appropriateness [2,4,3]. To improve reproducibility, several studies recommend annotator calibration, adjudication workflows, and inter-rater agreement measures such as Cohens κ and Krippendorffs α [23,21,14,4]. Overall, for RAG-based healthcare systems, the literature suggests that robust evaluation of healthcare RAG systems requires multidimensional and risk-dependent assessment combining automated retrieval-aware diagnostics with clinician-centered validation [2,9].

2.3 Evaluation Practices and Use Cases

This subsection illustrates how the evaluation dimensions and metrics discussed in Sections 2.1 and 2.2 are operationalized in concrete healthcare use cases. The reviewed studies generally adopt hybrid evaluation pipelines that combine automated, RAG-oriented metrics with expert human review, reflecting the need to assess technical performance, factual grounding, and clinical safety.

Several works provide representative examples of this hybrid approach. Baur et al. [5] presented a mixed-methods evaluation of a RAG-based chatbot for orthopedic and trauma surgery patient education, combining automated RAG-specific metrics such as answer relevance, contextual precision, and faithfulness with Likert-scale assessments carried out by clinicians, medical students, and lay users. Similarly, in [12], the authors evaluated a postoperative-instruction virtual assistant using large-scale simulated queries together with blinded surgeon review, measuring retrieval precision, completeness, faithfulness, and hallucination-related safety outcomes. These studies show how retrieval-aware metrics and human validation can be integrated in practical clinical evaluation workflows.

Other comparative studies highlight the value of such evaluation protocols when different system types are contrasted. In [22], a RAG-based chatbot outperformed a non-augmented GPT-4 baseline in medical accuracy and agreement

with evidence while maintaining comparable linguistic quality. Khabaz et al. [16] compared general-purpose, RAG-enhanced, and healthcare-specialized conversational systems for thyroid-eye-disease question answering, finding that RAG-based systems achieved higher clinical accuracy, whereas general-purpose models showed advantages in empathy and readability. Taken together, these findings suggest that evaluation should be aligned with the intended clinical context and the priorities of the target users.

Beyond these comparative analyses, several studies have explored specialized evaluation strategies for clinically relevant dimensions. Choo et al. [7] proposed an AI-assisted evaluation framework in which simulated patient agents interact with healthcare chatbots while an evaluator model automatically assesses conversational quality, obtaining strong agreement with human ratings. Other work has focused on high-risk and patient-centered dimensions such as hallucination severity [3] and empathetic communication [18], extending evaluation beyond generic correctness toward more clinically meaningful assessment.

Overall, the reviewed studies reveal recurring dimensions, metrics, methodological practices, and RAG-specific challenges that can be synthesized into a unified view, summarized in Table 1. These patterns directly informed the design considerations and structure of the evaluation framework proposed in the following section.

Table 1. Synthesis of core evaluation dimensions, representative metrics, and RAG-specific challenges identified in the reviewed literature.

Evaluation Dimension	**Identified Metrics / Methods**	**Human Evaluation Required**	**Challenges of Healthcare RAG Chatbots**	**Refs**
Safety and clinical risk	Hallucination rate, harm grading, contraindication checks, escalation appropriateness, expert rubric-based review	High	Error propagation from retrieval to generation, clinically harmful hallucinations, unsafe omissions	[3,13,9]
Factuality and grounding	Faithfulness, FactScore, NLI/entailment checks, claim verification, attribution analysis	Moderate to High	Unsupported claims despite retrieved evidence, grounding mismatch, selective omission of qualifiers	[10,19,20]
Retrieval quality	Recall@k, contextual recall, contextual precision, MRR, nDCG, retrieval relevance	Low to Moderate	Clinically irrelevant retrievals, outdated evidence, semantic mismatch between query and retrieved passages	[10,17]
Answer relevance and usefulness	Answer relevancy, completeness scores, utility judgments, expert usefulness ratings	Moderate	Incomplete answers despite correct retrieval, loss of clinically actionable information	[10,2]
Usability and trustworthiness	Likert-scale ratings, readability indices, user satisfaction questionnaires, trust scores	High	Trust calibration between perceived fluency and factual correctness	[2,13]
Empathy and patient-centered communication	EMRank, comparative ranking, empathy rubrics, human conversational ratings	High	Trade-offs between clinical precision and empathetic communication quality	[18,16]
Transparency and reporting quality	Documentation of data sources, model configuration, harm reporting, evaluation traceability	Moderate	Limited reproducibility and incomplete reporting of retrieval pipelines	[6]
Evaluation process quality	Inter-rater agreement (κ, α), adjudication workflows, evaluator calibration	High	Low reproducibility of expert evaluations and variability across annotators	[23,21,14,4]

3 Proposed Framework

The literature reviewed in Section 2 suggests that evaluation of healthcare conversational agents, particularly RAG systems, requires multidimensional assessment beyond conventional NLP benchmarks. Existing studies consistently emphasize the need to jointly evaluate retrieval quality, factual grounding, safety,

usability, transparency, and human-centered performance [13,9]. However, current approaches remain fragmented, often focusing on isolated dimensions or insufficiently addressing the interaction between retrieval and generation stages in RAG architectures.

Based on these findings, the proposed framework was designed according to four principles:

1. Evaluation should be *retrieval-aware*, distinguishing retrieval failures from generation failures.
2. Evaluation should be *risk-oriented*, prioritizing clinically relevant safety aspects such as hallucinations, omissions, and contraindications.
3. Evaluation should combine *automated and human-centered assessment*, balancing scalability with expert validation.
4. The framework should remain *practical and modular*, adaptable to different healthcare contexts and system architectures.

Accordingly, the framework organizes evaluation into complementary dimensions covering retrieval performance, grounding, clinical safety, usability, and evaluation process quality, aiming to support more consistent, transparent, and clinically meaningful assessment of healthcare RAG systems.

3.1 Framework dimensions

Building on the recurring dimensions, methodological practices, and specific RAG challenges identified in Section 2, we propose a structured evaluation framework for healthcare RAG conversational agents. The framework supports multidimensional assessment across both retrieval and generation stages while combining automated evaluation with targeted expert validation.

The framework organizes evaluation into five interconnected dimensions: (D1) retrieval quality, (D2) factual grounding and faithfulness, (D3) clinical safety and risk assessment, (D4) usability and communication quality, and (D5) evaluation process quality and transparency. These dimensions are evaluated through a hybrid pipeline integrating retrieval-aware automated metrics, review by clinicians, and methodological quality-control procedures. Figure 1 illustrates the overall structure of the framework and the interaction between its main components.

The framework adopts a staged perspective in which retrieval and generation are evaluated separately before integrating higher-level clinical and user-centered assessment. This distinction is particularly important in RAG systems, where failures may originate either from retrieval deficiencies (e.g., missing or irrelevant evidence) or from generation problems such as hallucinations, unsupported synthesis, or omission of clinically important qualifiers.

In D1, *retrieval quality* evaluates whether the system retrieves clinically relevant and sufficiently comprehensive evidence for a given query. Representative metrics include Recall@k, contextual precision, contextual recall, MRR, and nDCG. D2 focuses on *factual grounding and faithfulness*, assessing whether generated responses are adequately supported by retrieved evidence through metrics

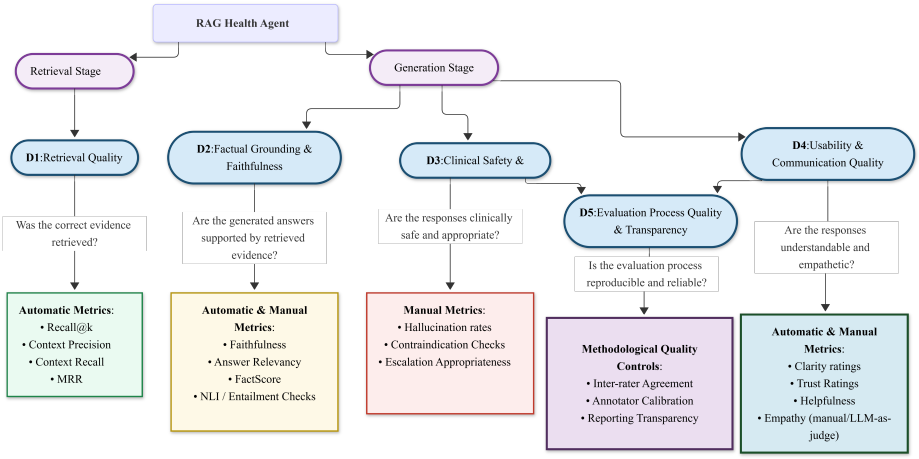


Fig. 1. Proposed multidimensional evaluation framework.

such as faithfulness, answer relevancy, entailment-based grounding checks, and FactScore-style claim verification.

D3 addresses *clinical safety and risk*. Since healthcare conversational agents operate in safety-critical settings, evaluation must explicitly assess hallucinations, contraindications, omission severity, escalation appropriateness, and potential patient harm, primarily through structured clinician review and expert adjudication. D4 evaluates *usability and communication quality*, including empathy, readability, clarity, trustworthiness, and perceived usefulness, combining automated readability metrics with human-centered assessments, e.g., Likert-scale ratings or comparative ranking methods.

Finally, the framework incorporates a transversal dimension (D5) related to *evaluation quality and transparency*. This includes annotator calibration, inter-rater agreement analysis, documentation of retrieval pipelines and model configurations, and explicit reporting of evaluation procedures to improve reproducibility and methodological rigor. Table 2 summarizes the proposed framework, including representative metrics, evaluation types, and common failure modes.

3.2 Recommended Evaluation Workflow and Reporting Practices

To operationalize the framework, we recommend a lightweight multi-stage evaluation workflow combining automated RAG metrics with clinician-centered validation. The objective is not only to estimate performance, but also to identify whether failures stem from retrieval, grounding, safety, or communication issues.

Step 1: Benchmark construction. Define a representative query set covering routine questions, high-risk clinical scenarios, contraindications, ambiguous prompts, and edge cases. Ground-truth references and supporting evidence should be curated/validated by domain experts whenever possible.

Step 2: Automated retrieval and grounding evaluation. Evaluate retrieval quality using metrics such as Recall@k, Context Precision, Context Recall, MRR, and nDCG. Generated responses should then be assessed using Faithfulness, Answer Relevancy, FactScore, and semantic grounding checks.

Step 3: Clinician-centered validation. A subset of high risk or low confidence outputs should undergo structured expert review assessing hallucinations, omission severity, contraindications, and escalation appropriateness. Multiple reviewers and adjudication procedures are recommended for safety-critical applications. After confirming that responses are clinically safe, usability and communication quality should also be assessed through dimensions such as clarity, readability, trustworthiness, and empathy, e.g., using Likert-scale ratings.

Step 4: Reporting and reproducibility. Studies should report retrieval corpora, model versions, prompts, evaluator profiles, scoring rubrics, and edge-case testing procedures, following reproducibility-oriented recommendations such as CHART [6]. In addition to individual metrics, reporting dimension-level composite scores can support interpretable comparison between conversational systems.

3.3 Empirical use case

To illustrate the practical applicability of the proposed framework, we present a small-scale use case based on *KneeChat*, a patient-oriented RAG-based conversational agent for orthopedic support [11]. A subset of 50 representative patient-oriented orthopedic queries was selected from the original study, covering common postoperative and rehabilitation-related scenarios such as recovery guidance, symptom interpretation, mobility recommendations, medication-related concerns, and general patient education. The objective of this example is not benchmark-scale validation, but rather to demonstrate how the proposed workflow can be operationalized in practice through dimension-oriented reporting. Table 3 presents an illustrative example of how the proposed framework can be used to structure multidimensional evaluation reporting for healthcare RAG-based conversational agents.

Table 2. Framework operationalization for healthcare RAG conversational agents.

Evaluation Dimension	Objective	Metrics / Methods	Evaluation Type	Main Failure Types
Retrieval Quality	Assess whether the system retrieves relevant and sufficiently comprehensive evidence	Recall@k, Context Precision, Context Recall, Mean Reciprocal Rank (MRR), normalized Discounted Cumulative Gain (nDCG)	Automatic	Irrelevant retrievals, missing evidence, outdated sources, semantic mismatch between query and retrieved passages
Factual Grounding and Faithfulness	Evaluate whether generated responses are supported by retrieved evidence	Faithfulness, Answer Relevancy, FactScore, entailment/NLI checks, attribution analysis	Automatic / Hybrid	Unsupported claims, hallucinations, grounding mismatch, omission of clinically relevant qualifiers
Clinical Safety and Risk	Identify clinically harmful or unsafe outputs	Clinician rubric-based review, hallucination assessment, contraindication checks, harm severity grading, escalation appropriateness	Manual (Clinician-centered)	Unsafe recommendations, dangerous omissions, clinically harmful misinformation
Usability and Communication Quality	Assess clarity, trustworthiness, readability, empathy, and usefulness	Readability indices, Likert-scale ratings, trust scores, empathy rubrics, EMRank, user satisfaction surveys	Hybrid	Low readability, poor empathy, misleading fluency, non-actionable responses
Evaluation Process Quality and Transparency	Improve reproducibility, consistency, and methodological rigor	Inter-rater agreement (κ, α), adjudication workflows, annotator calibration, reporting transparency	Manual / Methodological	Evaluator inconsistency, poor reproducibility, incomplete reporting of retrieval and evaluation procedures

Table 3. Illustrative application of the proposed evaluation framework to the *KneeChat* RAG-based conversational agent. Composite dimension scores correspond to the mean of normalized metrics associated with evaluation dimensions.

Evaluation Dimension	**Metrics**	**Observed Values**	**Composite Dimension Score**	**Interpretation**
Retrieval Quality	Context Precision Context Recall	0.972 0.901	**0.93**	Retriever consistently identified clinically relevant and sufficiently comprehensive evidence.
Factual Grounding and Faithfulness	Faithfulness Answer Relevancy FactScore Semantic Similarity	0.883 0.919 0.901 0.947	**0.913**	Generated responses remained strongly supported by retrieved evidence with limited hallucination risk.
Clinical Safety and Risk	Clinician accuracy rating Escalation appropriateness Hallucination assessment	0.94 0.92 0.90	**0.930**	Expert review suggested high clinical appropriateness and low frequency of unsafe outputs.
Usability and Communication Quality	Empathy rating Clarity rating Language adequacy	4.76 / 5 4.58 / 5 100%	**0.942**	Responses were perceived as clear, understandable, and empathetic for patient-oriented communication.
Evaluation Process Quality and Transparency	Inter-rater agreement (κ) Annotator calibration Edge-case testing Protocol documentation	0.82 Completed Included Reported	-	Evaluation process showed strong methodological consistency and reproducibility-oriented reporting.

Evaluation combined automated retrieval-aware metrics with clinician assessment. Retrieval and grounding dimensions were evaluated using metrics such as Context Precision, Context Recall, Faithfulness, and Answer Relevancy, while clinical safety and communication quality were assessed by healthcare professionals using structured review criteria and Likert-scale ratings. To improve interpretability and facilitate comparison across systems, individual metrics were normalized into dimension-level composite scores ranging from 0 to 1, where higher values indicate better performance. Composite scores were computed as the mean of normalized metrics associated with each evaluation dimension.

This example shows how the proposed framework enables interpretable evaluation of healthcare RAG-based conversational systems. Instead of relying on isolated metrics, the framework supports structured evaluation across retrieval performance, factual grounding, clinical safety, usability, and quality, enabling clearer identification of system strengths, limitations, and deployment readiness.

4 Discussion

Across the reviewed literature, a consistent finding emerges: evaluation of healthcare conversational agents cannot rely on isolated natural language generation metrics or general-purpose chatbot benchmarks alone. Instead, robust assessment requires multidimensional evaluation strategies integrating retrieval quality, factual grounding, clinical safety, usability, and methodological transparency [9,13,2].

A key observation from the reviewed studies is that RAG introduces evaluation requirements beyond those of conventional conversational AI systems. In RAG-based architectures, clinically unsafe responses may arise from failures in evidence retrieval or from unsupported synthesis during generation. Consequently, recent frameworks increasingly advocate decomposing evaluation into

retrieval-oriented and generation-oriented components [10,20]. Metrics such as Context Precision, Context Recall, Faithfulness, and Answer Relevancy provide useful signals for diagnosing retrieval and grounding failures, while clinician-centered assessment remains essential for identifying harmful hallucinations, unsafe omissions, and contextually inappropriate recommendations [3,4].

Another important finding is the growing recognition that evaluation must extend beyond factual correctness toward broader patient-centered and operational dimensions. Several studies emphasized the importance of empathy, readability, trustworthiness, and communication quality, particularly in patient-facing systems [18,16]. Frameworks such as CHART and HAICEF highlighted the importance of reproducibility, transparent reporting, and methodological rigor [6,13]. This supports the need for evaluation-process quality and transparency as transversal concerns in healthcare RAG assessment.

5 Conclusion

Healthcare conversational agents, particularly those based on RAG, require evaluation strategies beyond traditional NLP benchmarks. This paper reviewed current evaluation approaches for healthcare conversational systems and synthesized the main evaluation dimensions, metrics, and methodological practices reported in the literature. The review shows that effective evaluation increasingly relies on hybrid strategies combining automated retrieval-aware metrics with clinician-centered validation [9,13,10].

Based on these findings, we proposed a multidimensional evaluation framework for healthcare RAG-based conversational agents. The framework organizes evaluation around retrieval, grounding, safety, usability, and transparency, providing a structured way to distinguish retrieval failures from generation failures.

An empirical use case illustrated how the framework can support staged evaluation and dimension-level reporting. Overall, this work contributes a retrieval-aware, clinician-centered, and reporting-oriented perspective for evaluating healthcare RAG conversational systems.

Future work should focus on validating the framework across diverse clinical domains and developing weighting strategies for composite dimension-level scores. The framework also remains dependent on the quality and calibration of automated evaluators, which may require adaptation for specialized medical language.

Acknowledgments. The research reported in this paper was supported by the FederaTrans project:
Grant PID2024-158373OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. It is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309 and CUSTOMAI project, grant agreement no. 101226029).

References

1. Healthcare Chatbots Market Size, Share & Industry Analysis, By Component, By Deployment, By Application, and Regional Forecast (2024), https://www.fortunebusinessinsights.com/healthcare-chatbots-market-114375
2. Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin, B., Gevaert, O., Li, L.J., Jain, R., Rahmani, A.M.: Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. npj Digital Medicine **7**(1), 82 (2024). https://doi.org/10.1038/s41746-024-01074-z
3. Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J.A., Pimenta, D.: A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine **8**(1), 274 (2025). https://doi.org/10.1038/s41746-025-01670-7
4. Awasthi, R., Bhattad, A., Ramachandran, S.P., Mishra, S., Khanna, A.K., Cywinski, J.B., Maheshwari, K., Mahapatra, D., DiRosa, I., Cohen, A., Arshad, H., Atreja, A., Alshukaili, A., Vohra, A., Singh, N., Papay, F.A., Atreja, A., Kashyap, R., Mathur, P.: Human evaluation of large language models in healthcare: Gaps, challenges, and the need for standardization. npj Health Systems **2**(1), 40 (2025). https://doi.org/10.1038/s44401-025-00043-2
5. Baur, D., Ansorg, J., Heyde, C.E., Voelker, A.: Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study. JMIR AI **4**, e75262–e75262 (2025). https://doi.org/10.2196/75262
6. CHART Collaborative: Reporting guideline for chatbot health advice studies: The Chatbot Assessment Reporting Tool (CHART) statement. BMJ medicine **4**(1) (2025). https://doi.org/10.1136/bmjmed-2025-001632
7. Choo, S., Yoo, S., Endo, K., Truong, B., Son, M.H.: Advancing Clinical Chatbot Validation Using AI-Powered Evaluation With a New 3-Bot Evaluation System: Instrument Validation Study. JMIR Nursing **8** (2025). https://doi.org/10.2196/63058
8. Clark, M., Bailey, S.: Chatbots in Health Care: Connecting Patients to Information. Canadian Journal of Health Technologies **4**(1) (2024). https://doi.org/10.51731/cjht.2024.818
9. Ding, H., Simmich, J., Vaezipour, A., Andrews, N., Russell, T.: Evaluation framework for conversational agents with artificial intelligence in health interventions: A systematic scoping review. Journal of the American Medical Informatics Association : JAMIA **31**(3), 746–761 (2023). https://doi.org/10.1093/jamia/ocad222
10. Es, S., James, J., Espinosa Anke, L., Schockaert, S.: RAGAS: Automated Evaluation of Retrieval Augmented Generation. In: Aletras, N., De Clercq, O. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. pp. 150–158. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.eacl-demo.16
11. Escarraman-Reyes, J., Zapatero Herrera, M., Rivas-Ruiz, F., Morales-Garzón, A., Ruiz, M.D., Martin-Bautista, M.J.: Grounding retrieval-augmented LLMs in patient-derived qualitative data for surgical pathway support (2026), manuscript submitted for publication to IEEE Journal of Biomedical and Health Informatics
12. Haider, S.A., Prabha, S., Gomez Cabello, C.A., Genovese, A., Collaco, B., Wood, N., London, J., Bagaria, S., Tao, C., Forte, A.J.: The Development and Evaluation

of a Retrieval-Augmented Generation Large Language Model Virtual Assistant for Postoperative Instructions. Bioengineering **12**(11), 1219 (2025). https://doi.org/10.3390/bioengineering12111219
13. Hua, Y., Xia, W., Bates, D., Hartstein, G.L., Kim, H.T., Li, M., Nelson, B.W., Stromeyer Iv, C., King, D., Suh, J., Zhou, L., Torous, J.: Standardizing and Scaffolding Health Care AI-Chatbot Evaluation: Systematic Review. JMIR AI **4** (2025). https://doi.org/10.2196/69006
14. Jin, Y., Chandra, M., Verma, G., Hu, Y., De Choudhury, M., Kumar, S.: Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. In: Proceedings of the ACM Web Conference 2024. pp. 2627–2638. ACM (2024). https://doi.org/10.1145/3589334.3645643
15. Katz, U., Cohen, E., Shachar, E., Somer, J., Fink, A., Morse, E., Shreiber, B., Wolf, I.: GPT versus Resident Physicians A Benchmark Based on Official Board Scores. NEJM AI **1**(5) (2024). https://doi.org/10.1056/AIdbp2300192
16. Khabaz, K., Parekh, Z., Saeed, S., Krakauer, M., Silas, M.R., Farooq, A.V., Shah, H.: Incorporating information retrieval into AI chatbots for patient education on thyroid eye disease. International Journal of Medical Informatics **207**, 106213 (2026). https://doi.org/10.1016/j.ijmedinf.2025.106213
17. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
18. Luo, M., Warren, C.J., Cheng, L., Abdul-Muhsin, H.M., Banerjee, I.: Assessing Empathy in Large Language Models with Real-World Physician-Patient Interactions (2024). https://doi.org/10.48550/arXiv.2405.16402
19. Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12076–12100. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.741
20. Neha, F., Bhati, D., Shukla, D.K.: Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. AI **6**(9), 226 (2025). https://doi.org/10.3390/ai6090226
21. Sallam, M., Barakat, M., Sallam, M.: Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models. Cureus (2023). https://doi.org/10.7759/cureus.49373
22. Steybe, D., Poxleitner, P., Aljohani, S., Herlofson, B.B., Nicolatou-Galitis, O., Patel, V., Fedele, S., Kwon, T.G., Fusco, V., Pichardo, S.E.C., Obermeier, K.T., Otto, S., Rau, A., Russe, M.F.: Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw. Journal of Cranio-Maxillofacial Surgery **53**(4), 355–360 (2025). https://doi.org/10.1016/j.jcms.2024.12.009
23. Tam, T.Y.C., Sivarajkumar, S., Kapoor, S., Stolyar, A.V., Polanska, K., McCarthy, K.R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G.E., Sun, C., Peng, Y., Wang, Y.: A framework for human evaluation of large language models in healthcare derived from literature review. npj Digital Medicine **7**(1), 258 (2024). https://doi.org/10.1038/s41746-024-01258-7