

CysticCare: An Evaluation Driven RAG AI Chatbot Targeted Towards Evidence Based Patient Education

¹Mehmet Seckin*, ²Yashwant Ponnaganti, ³Ankur Shah, ¹Resit Sendag, ¹Kunal Mankodiya

¹Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA

²Department of Computer Science, University of Maryland, College Park, MD, USA

³Division of Kidney Disease and Hypertension, Department of Medicine, Rhode Island Hospital, Providence, RI, USA

* Corresponding author: mseckin@uri.edu

Abstract. Autosomal Dominant Polycystic Kidney Disease (ADPKD) is a lifelong genetic and progressive condition. Patients with ADPKD need to understand a great deal of complex clinical and imaging information as their health conditions progressively decline. Clinical guidelines exist, but they are written for doctors and are hard for most patients to use on their own. General-purpose AI chatbots can render complex medical information in accessible language, but they are prone to hallucination and typically provide no verifiable provenance for their outputs, a critical limitation given the accuracy demands of clinical content. We built CysticCare to address this gap: a RAG system that answers questions about ADPKD using guidelines and peer-reviewed literature and shows its sources with every answer. The system routes each question to one of three reasoning agents, Standard RAG, Chain of Thought, or Stepback, depending on how complex the question is. A validation step checks each answer for relevance, correct sourcing, and safety before it reaches the user, regenerating the answer if a check fails. There was no established way to measure how well a system like this performs on real patient questions, so we also built MedRAGBench, a set of 286 questions spanning five types: standard factual, context-dependent, false premise, safety-critical, and unanswerable, across five clinical categories. Testing CysticCare on the full benchmark, its answers were semantically close to the reference answer in 98.3 percent of cases, with median similarity for each question score between 0.773 and 0.951 across the four answerable question types, and the system responded appropriately, answering when it should and declining when it should not, in 95.8 percent of cases. These results suggest that a disease-specific, multi-agent RAG system has potential to reliably support patient education while remaining transparent about what it does and does not know, a step toward AI tools that patients and clinicians can actually trust.

Keywords: retrieval augmented generation, RAG, patient education, medical chatbots, evidence grounding, ADPKD, benchmark, multi-agent systems

I. INTRODUCTION

Autosomal Dominant Polycystic Kidney Disease, or ADPKD, is one of the most common inherited kidney disorders and affects 12 million people worldwide [1]. It progresses over decades, with continuous cyst growth, gradual loss of kidney function, and effects outside the kidneys. Patients must make sense of imaging findings such as the Mayo Imaging Classification and Total Kidney Volume, laboratory results, genetic information about their PKD1 or PKD2 subtype, and treatment options such as tolvaptan [2]. Because ADPKD progresses differently in each patient and educational needs change over time, reliable ongoing access to education matters for self-management and shared decisions with a doctor.

Clinical guidelines for ADPKD are thorough [3] but written for specialists and hard for most patients to use. Many struggle

to interpret an imaging report or apply a general recommendation to their own diet, fluid intake, or activity level, and clinic visits rarely leave enough time, so patients search on their own through web sources of uneven quality [4], [5].

AI chatbots built on large language models (LLMs) are a natural fit for health education, explaining complex topics in plain language and holding a conversation [6]. The difficulty is that general-purpose models are not built for clinical reliability: they can hallucinate and usually do not show where an answer came from [7], [8]. RAG grounds a model's answers in an outside set of documents rather than only what the model knows; it pulls in real sources [9]. Even so, many healthcare RAG systems use a single retrieval step and give little visibility into how grounded, accurate, or safe their answers are [10], [11], [12], and studies of disease-specific chatbots tend to focus on whether patients like using them rather than how well the system can be evaluated [12].

There is also no established way to measure such a system on real patient questions. Existing medical RAG benchmarks tend to test general knowledge in multiple-choice format, focus on robustness rather than full system evaluation, or fail to reflect how patients phrase questions, and none combine a false assumption, a safety-sensitive topic, and a genuinely unanswerable question in one benchmark [13], [14], [15].

This paper presents CysticCare, a multi-agent RAG system built for patient education in ADPKD, and MedRAGBench, a benchmark built to evaluate systems like it. Our main contributions are as follows.

1) A multi-agent RAG architecture for ADPKD patient education. An agent selector routes each question to one of three reasoning agents (Standard RAG, Chain of Thought, or Stepback), so heavier reasoning is used only when needed. To our knowledge, this is among the first multi-agent RAG systems built specifically for patient education in a progressive kidney disease.

2) MedRAGBench, a disease-specific benchmark of 286 patient-centered questions that scores not just answer correctness but appropriate behavior on adversarial inputs. Its five types (standard factual, context-dependent, false-premise, safety-critical, and unanswerable) test whether the system resists false assumptions, handles sensitive topics carefully, and declines when evidence is lacking. To our knowledge, no prior disease-specific medical RAG benchmark scores correct abstention as an explicit outcome.

3) An evaluation approach pairing semantic similarity and response appropriateness over a controlled corpus where the benchmark and knowledge base share the same 80 publications. Because the documents are shared, performance differences

reflect retrieval and generation quality rather than coverage gaps, and a source-grounding check separates genuine retrieval from parametric memorization [12], [16].

II. BACKGROUND AND RELATED WORK

A. RAG for Healthcare Education

RAG has been applied across healthcare tasks including medical question answering, guideline summarization, and disease-specific education in oncology, cardiology, infectious disease, and oral health [9], [11], [16]. Retrieval tends to keep answers closer to the evidence and easier to trace than a standalone model. However, healthcare RAG systems vary widely in architecture, retrieval strategy, and source handling, making results hard to reproduce or compare [12], and most use a single retrieval pipeline with no adaptive reasoning, applying the same processing to every query.

A recent orthopedic RAG chatbot by Baur et al. [17] is the closest comparable system in domain-specific patient education and is representative of the simple, single-pass architecture common in current medical chatbots: one retrieval step feeds one generation call, with no query-dependent routing. It uses a mixed-methods evaluation with 30 human evaluators, but only 100 synthetic queries and RAGAS metrics without citation-level checking or adversarial types. Our work adds query-dependent routing, a substantially larger domain-specific benchmark, and a multi-metric evaluation.

B. Medical RAG Benchmarks

Several benchmarks exist for medical RAG. MIRAGE [18] offers 7,663 questions from medical examination datasets, but is limited to multiple-choice, general knowledge, not retrieval from a defined document set. MedRGB [19] tests robustness to noisy and adversarial retrieval contexts, focusing on model behavior under corrupted input rather than full-system performance with citations. RAGCare-QA [20] provides 420 multiple-choice questions across six specialties and three complexity levels, targeting general medical education rather than patient-centered questions with inline citations.

A common limitation is the absence of adversarial question types: they test recall, not judgment. Peng et al. [13] introduced a framework for unanswerable queries in RAG with six categories of unanswerability, and RefusalBench [14] evaluates selective refusal in grounded language models. A safety analysis in [15] showed that RAG systems are not inherently safer than standalone LLMs, motivating explicit safety evaluation. Still, no prior disease-specific benchmark combines false-premise, safety-critical, and unanswerable question types alongside standard factual evaluation. Table I compares MedRAGBench with these benchmarks.

TABLE I.
COMPARISON WITH EXISTING MEDICAL RAG BENCHMARKS

Benchmark	Key characteristics
MIRAGE [18]	7,663 multiple-choice questions from USMLE and MedMCQA; general medical scope; no adversarial types; no citation provenance.
MedRGB [19]	Tests robustness to noisy contexts; multiple-choice format; no disease-specific focus or citation evaluation.
RAGCare-QA [20]	420 multiple-choice questions across 6 specialties; no patient-centered framing; no inline citations.

Orthopedic RAG [17]	100 free-text questions (orthopedic); RAGAS metrics only; no adversarial types; no controlled corpus.
MedRAGBench (ours)	286 free-text questions (ADPKD); 5 types including adversarial; inline citations; controlled 80-paper corpus.

C. Evaluation Frameworks

RAGAS [21] evaluates faithfulness, answer relevancy, context precision, and context recall, but does not check citation accuracy or whether cited documents match. MedTrust-RAG [22] and MedRAGChecker [23] address claim-level evidence verification in biomedical QA, but neither offers a structured benchmark with a controlled corpus guaranteeing every reference answer is answerable from the indexed documents. Separating retrieval from memorization is a further difficulty: a model that already knows a fact can answer correctly even when retrieval contributes nothing, and few frameworks account for this. LLM-as-judge approaches [24] approximate expert-level medical reasoning evaluation at scale, motivating their use alongside similarity metrics.

D. ADPKD-Specific Systems

In nephrology, most AI work has targeted disease modeling, imaging analysis, and progression prediction; conversational education has received far less attention [25], and patient-facing kidney resources are typically static websites or general-purpose chatbots. Studies evaluating ChatGPT and GPT-4 against kidney guidelines [8], [26] test standalone LLMs on small question sets, 15 questions in [8], without a RAG architecture, benchmark, or controlled corpus, and a clinical decision support evaluation [27] focused on on-premises deployment rather than patient-facing education. CysticCare addresses these gaps with a complete RAG system, a disease-specific benchmark, a multi-agent architecture, and a multi-dimensional evaluation.

III. METHODOLOGY

A. CysticCare System Architecture

CysticCare is a cloud-hosted, three-layer RAG system. The presentation layer is a Flutter web app on Firebase Hosting, usable from any browser without an account, with a mandatory medical disclaimer, one-click example questions, and source citations shown with each answer. The application layer uses FastAPI (Python 3.11) on Google Cloud Run, exposing REST endpoints for request validation, query preprocessing, retrieval, and generation; all traffic runs over HTTPS, and each query gets a unique identifier so evidence, response, and citations are traceable. The retrieval and generation layer uses OpenAI's text-embedding-3-small for query embedding, ChromaDB with HNSW indexing [28] for the vector store, and GPT-4o for generation.

Fig. 1 below illustrates the query processing and retrieval pipeline mentioned above. A user question is normalized, expanded for common abbreviations, and converted into a 1536-dimensional embedding. The system then performs a cosine similarity search against the vector store and retrieves the top five most relevant passages with positive ranking scores, which are passed to GPT-4o along with their source metadata to generate a response with attached citations.

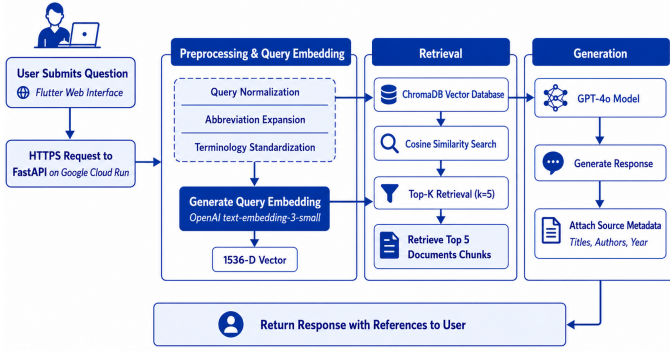


Fig. 1. Query processing and retrieval pipeline in CysticCare.

B. Multi-Agent Query Routing

A distinguishing feature is an adaptive agent selector that routes each query to one of three specialized reasoning agents based on the question, rather than applying the same processing to every query. The Standard RAG agent handles straightforward factual questions, for example “What is the PKD1 gene?”, retrieving the most relevant passages by cosine similarity and generating a direct, citation-backed response; it is optimized for speed and used when the question is focused and the evidence is unambiguous.

The Chain-of-Thought agent handles complex questions needing multi-step reasoning, for example “How do tolvaptan and ACE inhibitors interact in ADPKD management?”. It breaks the question into explicit reasoning steps, retrieves evidence for each, and builds a response makes its reasoning visible, which helps for medication and comparative treatment questions where a single retrieval pass could miss context.

The Stepback agent handles narrow questions that lack enough directly relevant material, for example “What is the role of mTOR signaling in cyst formation?”. It first reformulates the question more broadly, gathers wider background, then returns to the specific question with that fuller context, preventing a thin answer based on limited evidence.

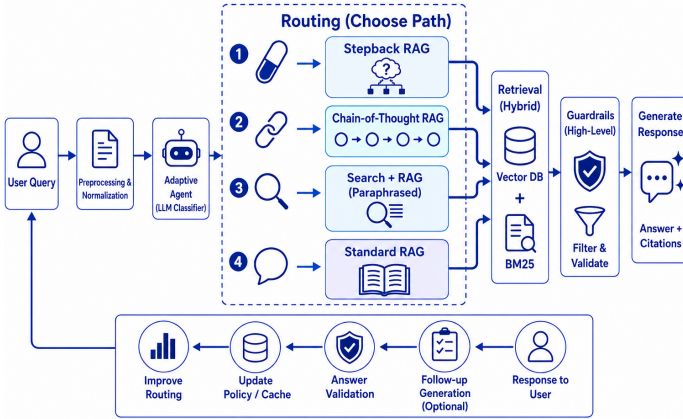


Fig. 2. Multi-agent routing architecture. The adaptive agent selector chooses a reasoning path, retrieves evidence through hybrid search, and validates the response before it reaches the user.

The agent selector looks at the complexity of the question, how domain-specific it is, and the relevance scores from an initial retrieval pass to decide which strategy to use, so that the more elaborate reasoning paths are only used when they are

actually needed. Fig. 2 shows the full routing process, including the retrieval and validation steps that follow agent selection.

C. Knowledge Base

The knowledge base is built from authoritative, publicly available sources relevant to ADPKD. Primary sources include peer-reviewed scientific publications and guideline-aligned educational materials from the PKD Foundation and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). After removing duplicates and filtering by language, more than 400 scientific articles in PDF format were retained. This full corpus is used for production deployment, and the smaller controlled subset is used for evaluation.

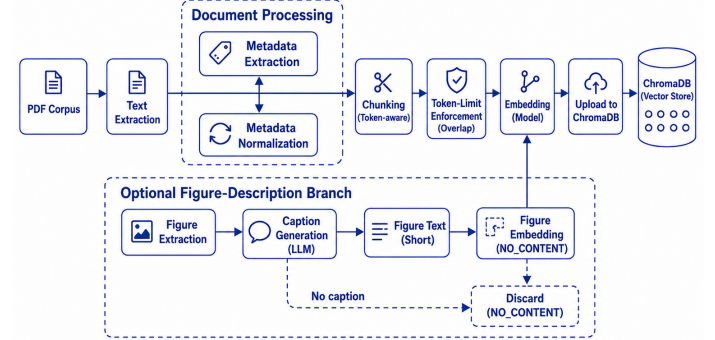


Fig. 3. Knowledge base ingestion pipeline.

Fig. 3 shows the ingestion pipeline: a hybrid approach combines pdfplumber for layout-aware text extraction and PyPDF2 for metadata. Text is segmented into overlapping 300-word chunks with 50-word overlap, yielding roughly 4,500 chunks. Each is embedded into a 1536-dimensional vector with text-embedding-3-small and stored in ChromaDB (HNSW, cosine similarity), with metadata (title, authors, year, source) for citation logging and evaluation.

D. MedRAGBench: Benchmark Construction

We built MedRAGBench, a benchmark dataset made of patient-centered, natural-language questions grounded in disease-specific peer-reviewed literature. The benchmark contains 286 questions across five types (Table II) and five clinical categories (Table III) (disease mechanisms, diagnosis and imaging, treatment and medication, diet and lifestyle, progression and prognosis), each assigned an easy, moderate, or hard difficulty by the synthesis required.

TABLE II
BENCHMARK QUESTION TYPES

Question type	What it is designed to test
Standard factual	Whether the system retrieves the right evidence and answers correctly for well-supported questions.
False-premise	Whether the system resists misleading assumptions instead of accepting and building on a flawed premise.
Unanswerable	Whether the system declines to answer when the sources do not adequately support an answer.
Context-dependent	Whether the system uses surrounding context correctly and does not over-generalize from a single passage.
Safety-critical	Whether the system avoids unsafe recommendations, especially about medication, and defers to clinical guidance.

TABLE III
QUESTION DISTRIBUTION ACROSS CATEGORIES AND QUESTION TYPES

Category	Standard factual	Context-dependent	False-premise	Safety-critical	Unanswerable	Total
Disease mechanisms	45 (15E/13M/17H)	8 (0E/1M/7H)	9 (1E/3M/5H)	0	2 (0E/1M/1H)	64
Diagnosis and imaging	38 (5E/27M/6H)	10 (0E/8M/2H)	3 (1E/1M/1H)	2 (0E/1M/1H)	5 (1E/2M/2H)	58
Treatment and medication	29 (2E/22M/5H)	6 (0E/3M/3H)	7 (1E/3M/3H)	18 (0E/10M/8H)	15 (0E/5M/10H)	75
Diet and lifestyle	17 (3E/12M/2H)	3 (1E/1M/1H)	3 (0E/3M/0H)	2 (0E/2M/0H)	6 (0E/6M/0H)	31
Progression and prognosis	33 (7E/23M/3H)	10 (0E/5M/5H)	5 (2E/3M/0H)	0	10 (0E/3M/7H)	58
Total	162	37	27	22	38	286

The parenthetical is the Easy/Medium/Hard split (E/M/H). Superscript a: total questions per category. Superscript b: total questions per question type.

TABLE IV
STRUCTURE OF EACH BENCHMARK DATASET ITEM

Field	Description
Question	Free-text, natural-language question as a patient would ask it.
Gold answer	One automatically assembled comprehensive reference answer per question.
Supporting evidence	The passages that support the gold answer.
Source	Which uploaded documents a correct system should retrieve.
Category	Topic area, such as disease mechanisms, medication, or lifestyle.
Difficulty	How demanding the question is, from standard factual to safety critical.
Expected behavior	What a reliable system should do, including when it should decline to answer.

Each benchmark item (see Table IV) records the question text, a reference answer with inline, passage-level citations such as [P1] and [P2] that link to specific passages, the supporting passages with cross-encoder reranking scores, source paper identifiers and metadata, the category and difficulty level, and the expected system behavior, which can be to answer, answer with synthesis, correct the premise, defer, or abstain.

Questions were generated with GPT-5 and Claude Opus 4.7 using prompts enforcing patient-centered, natural-language phrasing (as a patient, nurse, or clinician would ask, not referencing specific papers). To stay grounded yet general, the model saw sampled excerpts spanning all papers, and questions were produced in per-type batches with deduplication. Each was assigned a clinical category and difficulty level reflecting the degree of synthesis or specialized knowledge required.

The generated dataset passed through multi-stage automated quality filtering: near-duplicate questions were removed by embedding-similarity thresholding, off-topic questions were discarded, each question was checked for consistency with its assigned type (for example, a false-premise item must contain an actual false assumption), and questions supported by only a single passage or by low-scoring evidence were removed to ensure multi-source grounding. The final benchmark contains the 286 questions that survived all filtering stages.

The benchmark is built from a controlled 80-paper subset of the knowledge base, used specifically for evaluation. This controlled design guarantees that every reference answer is answerable from the indexed corpus. In that setting, a difference

in performance reflects retrieval accuracy and generation quality, not a gap in knowledge coverage. This addresses a limitation identified in prior RAG evaluations, where it is often unclear whether a failure comes from poor retrieval or from missing documents [12], [16].

E. Evaluation Framework

We evaluate CysticCare using two complementary metrics that together capture answer quality and system behavior. Neither is sufficient on its own: a fluent answer can still be inappropriate, and an appropriate refusal can look dissimilar to a substantive reference answer.

Semantic similarity: embedding cosine similarity between the reference answer and the RAG-generated answer, giving an overall measure of answer quality that is independent of citation accuracy.

Response appropriateness: whether the system does the right thing for each question type, giving a substantive, on-topic answer for the four answerable types and declining for Unanswerable questions. This is scored by an LLM relevance grader, with the scoring inverted for the Unanswerable type: a correct refusal produces no relevant answer to the question, so a relevance judgment of not relevant is counted as the appropriate outcome, and a substantive answer to an unanswerable question is counted as a failure.

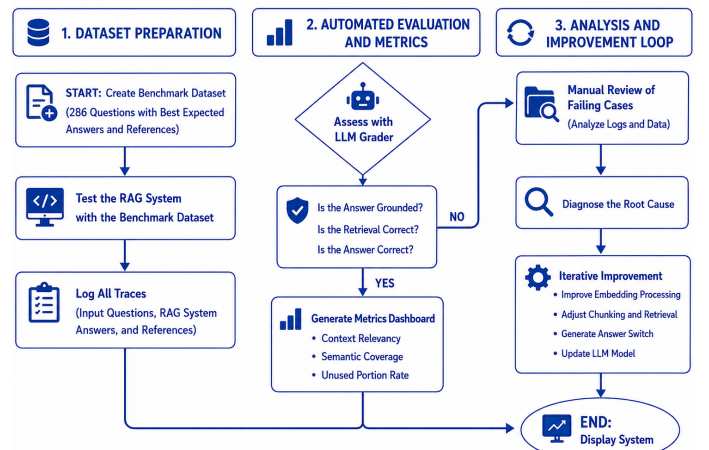


Fig. 4. End-to-end evaluation workflow.

The two metrics are computed independently, so agreement on a given item indicates a robust judgment rather than an artifact of one grader. Source-level provenance matters as much as answer correctness in medical systems [22], [23].

IV. RESULTS

A. Semantic Similarity to Reference Answers

Across all 286 questions, 281 (98.3 percent) produced answers judged semantically close to the reference. The five failures were four Standard factual questions where the response diverged from the reference, and one Unanswerable question where the system answered instead of declining. Fig. 5 shows this breakdown by question type.

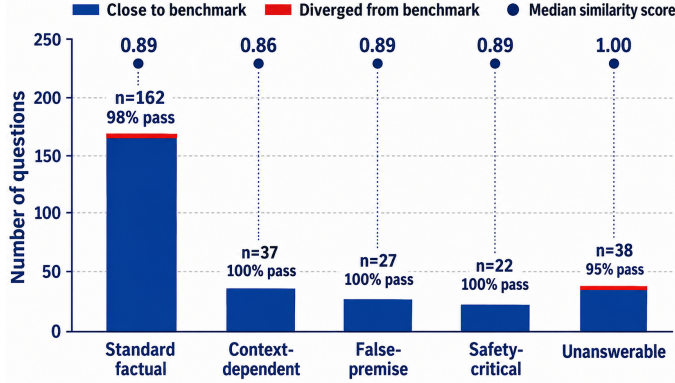


Fig. 5. Semantic similarity. Blue segments show answers judged close to the reference answer. Red segments show answers that diverged from it.

Average median similarity was consistent across the four answerable types: 0.882 (Standard factual), 0.864 (Context-dependent), 0.881 (False-premise), and 0.881 (Safety-critical), each with a narrow spread (Fig. 5), so answer quality does not drop for the harder adversarial types. The Unanswerable type shows a median of 1.000, driven by near word-for-word agreement between the system's refusal and the reference refusal on 37 of 38 items; this reflects textual agreement on declining rather than answer quality, and should be read alongside the behavior results below.

B. Response Appropriateness

The response-appropriateness grader checks whether a response matches the expected system behavior for its question type: giving a substantive, on-topic answer for the four answerable types, or declining for Unanswerable questions. Fig. 6 shows the resulting pass and fail counts by question type.

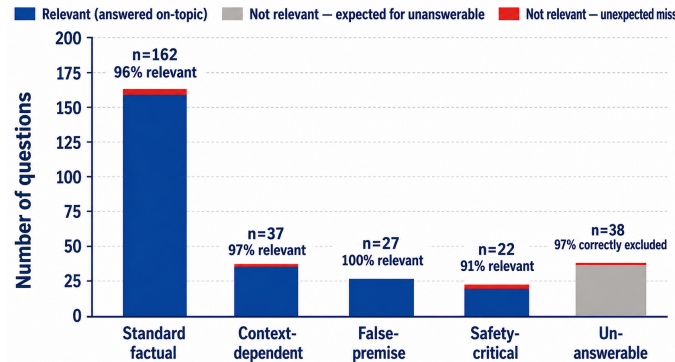


Fig. 6. Response appropriateness by question type. For Unanswerable questions, gray indicates a correct decline, the expected outcome, and red indicates an answer that should have been declined.

Overall, 274 of 286 responses (95.8 percent) showed appropriate behavior. Of the twelve inappropriate cases, eight

were Standard factual (screening timing, imaging modality choice, drug pipelines beyond tolvaptan), two Safety-critical (tolvaptan with other medications or fluid intake), one Context-dependent, and one the Unanswerable case below. All five semantic similarity failures were also flagged behavior-inappropriate, so the two independent metrics agree.

C. Case Analysis: The Unanswerable Failure

Of the 38 Unanswerable questions, 37 were correctly declined, typically through a scoped refusal that the system only addresses PKD-related questions. One, on the safety of creatine monohydrate supplementation in ADPKD, was answered directly, discussing creatine's effect on serum creatinine instead of declining. It fails both metrics, departing from the reference refusal and engaging with unsupported content. This shows why adversarial types matter: a system reliable on 279 of 286 items can still fail in a way standard factual questions alone would not reveal.

V. DISCUSSION

This study presents CysticCare, a multi-agent RAG system for ADPKD patient education, together with MedRAGBench, a disease-specific benchmark addressing several limitations of existing medical RAG evaluations.

The multi-agent architecture departs from the single-pipeline systems that dominate the literature. ADPKD questions range from simple lookups (“What is TKV?”) to multi-step reasoning, and routing to specialized agents handles this without applying the same overhead to every query, while a validation agent checks relevance, citation accuracy, and safety before the patient sees the response, a practical verify-then-generate approach [22].

MedRAGBench addresses the gap in Table I: to our knowledge, ours is the first to combine patient-centered questions, inline citations, and adversarial types in a disease-specific setting, with a controlled corpus so a performance difference reflects retrieval and generation quality rather than a coverage gap [12], [16].

The evaluation framework is designed so no single metric can hide a failure. Semantic similarity alone cannot distinguish a fabricated answer from a properly sourced one, which is why the response-appropriateness grader earns its place, as the creatine case shows: it scored acceptably on some dimensions yet still failed to abstain.

We also examined source-level provenance: on 93.5 percent of answerable questions the system retrieved at least one source paper used to build the reference answer, indicating responses are anchored in the intended evidence base rather than the model's prior knowledge. A full retrieval-recall analysis is left to future work.

Comparison with Prior Work

The closest published system, the orthopedic RAG chatbot of Baur et al. [17], reports strong RAGAS scores over an 899-document German corpus, but a direct numerical comparison is not meaningful given different corpora, question sets, and metric families, so we compare on design and evaluation scope (Table V). Two differences are central: where Baur et al. use a single static top-k step, CysticCare routes each query to one of three reasoning agents, the agentic RAG they name as future

work; and unlike Baur et al. or the benchmarks in Table I, MedRAGBench scores whether a system declines when the corpus cannot support an answer, through its false-premise, safety-critical, and unanswerable types. The creatine case in Case Study shows why: a system can score well on relevance and similarity yet still fail to abstain, a failure relevance-only evaluations such as RAGAS never surface.

TABLE V
SYSTEM-LEVEL COMPARISON WITH PRIOR PATIENT-EDUCATION RAG WORK

Dimension	Baur et al. [17]	CysticCare (ours)
Domain	Orthopedic / trauma surgery	ADPKD (nephrology)
Corpus	899 docs, German	80-paper controlled eval corpus
Architecture	Single static top-k RAG	Multi-agent routing + validation
Adversarial types	None	False-premise, safety-critical, unanswerable
Abstention scored	No	Yes (unanswerable items)
Eval metric	RAGAS (LLM-judge)	Semantic similarity + behavior grading
Eval set	100 synthetic queries	286 LLM-generated, filtered questions

Several limitations remain. The results reflect performance on the controlled 80-paper evaluation corpus, not the full 400-plus paper production knowledge base, and retrieval over the larger corpus has not yet been evaluated and may differ. A full retrieval-recall analysis over the production corpus, along with per-agent ablations isolating the contribution of each reasoning path, is in progress. The comparison with Baur et al. [17] is a design comparison rather than a head-to-head on identical metrics, since the two systems use different corpora and metric families. A direct comparison against flagship general-purpose models on MedRAGBench is planned to measure how much CysticCare's retrieval grounding and multi-agent routing improve over a strong standalone LLM.

VI. CONCLUSION

This paper presented CysticCare, a multi-agent RAG system for patient-centered ADPKD education, together with MedRAGBench, a 286-question benchmark spanning standard and adversarial question types and built from a controlled corpus that separates retrieval and generation quality from knowledge coverage.

On the full benchmark, CysticCare produced answers semantically close to the reference answer in 98.3 percent of cases and behaved appropriately, answering when supported and declining when not, in 95.8 percent of cases. The single adversarial failure observed underscores the value of adversarial question types in surfacing errors standard factual evaluation would miss. Full retrieval-recall analysis, clinician review, and a patient usability study remain as future work.

These results suggest CysticCare is a promising tool for disease-specific patient education, and the benchmark and evaluation framework are designed to transfer to other diseases as a template for trustworthy medical AI. All code is publicly available at <https://github.com/mhseckin/>.

REFERENCES

[1] A. B. Chapman, O. Devuyst, K.-U. Eckardt, et al., "Autosomal-dominant polycystic kidney disease (ADPKD): Executive summary from a KDIGO Controversies Conference," *Kidney Int.*, vol. 88, no. 1, pp. 17–27, 2015.

[2] V. E. Torres, C. Ahn, T. R. M. Barten, et al., "KDIGO 2025 clinical practice guideline for the evaluation, management, and treatment of ADPKD: Executive summary," *Kidney Int.*, vol. 107, no. 2, pp. 234–254, 2025.

[3] KDIGO ADPKD Work Group, "KDIGO 2025 clinical practice guideline for the evaluation, management, and treatment of ADPKD," *Kidney Int.*, vol. 107, no. 2S, pp. S1–S239, 2025.

[4] E. R. Lutz, K. L. Costello, M. Jo, et al., "A systematic evaluation of websites offering information on chronic kidney disease," *Nephrol. Nurs. J.*, vol. 41, no. 4, pp. 355–363, 2014.

[5] D. M. Taylor, S. Fraser, C. Dudley, et al., "Health literacy and patient outcomes in chronic kidney disease: A systematic review," *Nephrol. Dial. Transplant.*, vol. 33, no. 9, pp. 1545–1558, 2018.

[6] J. W. Ayers, A. Poliak, M. Dredze, et al., "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum," *JAMA Intern. Med.*, vol. 183, no. 6, pp. 589–596, 2023.

[7] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," *arXiv:2509.04664*, 2025.

[8] P. C. Acharya, R. Alba, P. Krisanapan, et al., "AI-driven patient education in chronic kidney disease: Evaluating chatbot responses against clinical guidelines," *Diseases*, vol. 12, no. 8, art. 185, 2024.

[9] Y. Gao, Y. Xiong, X. Gao, et al., "Retrieval-augmented generation for large language models: A survey," *arXiv:2312.10997*, 2023.

[10] F. Shi, X. Chen, K. Misra, et al., "Large language models can be easily distracted by irrelevant context," in *Proc. 40th Int. Conf. Machine Learning (ICML)*, PMLR 202, pp. 31210–31227, 2023.

[11] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nat. Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.

[12] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digit. Health*, vol. 4, no. 6, art. e0000877, 2025.

[13] X. Peng, P. K. Choubey, C. Xiong, and C.-S. Wu, "Unanswerability evaluation for retrieval augmented generation," in *Proc. 63rd Annual Meeting of the ACL (Vol. 1: Long Papers)*, 2025.

[14] A. Muhamed, L. F. R. Ribeiro, M. Dreyer, V. Smith, and M. T. Diab, "RefusalBench: Generative evaluation of selective refusal in grounded language models," *arXiv:2510.10390*, 2025.

[15] B. An, S. Zhang, and M. Dredze, "RAG LLMs are not safer: A safety analysis of retrieval-augmented generation for large language models," in *Proc. NAACL 2025 (Vol. 1: Long Papers)*, pp. 5444–5474, 2025.

[16] K. Hao and B. Wang, "Retrieval-augmented generation in healthcare: A narrative review of methods, contributions, and future directions," *Digit. Med.*, vol. 12, no. 1, art. e25-00015, 2026.

[17] D. Baur, J. Ansorg, C. E. Heyde, and A. Voelker, "Development and evaluation of a retrieval-augmented generation chatbot for orthopedic and trauma surgery patient education: Mixed-methods study," *JMIR AI*, vol. 4, art. e75262, 2025.

[18] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," in *Findings of the ACL 2024*, pp. 6233–6251, 2024.

[19] N. T. Ngo, C. V. Nguyen, F. Derroncourt, and T. H. Nguyen, "Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering," *arXiv:2411.09213*, 2024.

[20] J. Dobrev, I. Karasmanakis, F. Ivanisevic, et al., "RAGCare-QA: A benchmark dataset for evaluating retrieval-augmented generation pipelines in theoretical medical knowledge," *Data in Brief*, vol. 62, art. 112146, 2025.

[21] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in *Proc. 18th Conf. European Chapter of the ACL: System Demonstrations (EACL)*, pp. 150–158, 2024.

[22] Y. Ning, Y. Sun, L. Luo, Y. Wang, Y. Pan, and H. Lin, "MedTrust-RAG: Evidence verification and trust alignment for biomedical question answering," *arXiv:2510.14400*, 2025.

[23] Y. Ji, M. G. Kwak, H. Zhang, X. Wu, C. Li, and Y. Wang, "MedRAGChecker: Claim-level verification for biomedical retrieval-augmented generation," *arXiv:2601.06519*, 2026.

[24] S. Zhou, W. Xie, J. Li, et al., "Automating expert-level medical reasoning evaluation of large language models," *npj Digit. Med.*, vol. 9, no. 1, art. 34, 2025.

[25] O. A. Garcia Valencia, S. Suppadungsuk, C. Thongprayoon, et al., "Ethical implications of chatbot utilization in nephrology," *J. Pers. Med.*, vol. 13, no. 9, art. 1363, 2023.

[26] J. N. Shah, A. Ahmad, C. Thongprayoon, W. Cheungpasitporn, and P. C. Acharya, "The effectiveness of AI chatbots in patient education regarding intermittent fasting in CKD," in *Proc. ASN Kidney Week, Abstract*, 2024.

[27] K. Wolk, "Evaluating retrieval-augmented generation variants for clinical decision support: Hallucination mitigation and secure on-premises deployment," *Electronics*, vol. 14, no. 21, art. 4227, 2025.

[28] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, 2020.