MedAgent: Trustworthy Personalized Consumer Health Search via Multi-Agent Retrieval and Verification

Keshav Trikha[1], Blazej Madrzyk[2,*], Ashleyn Castelino[2], and Aram Bahrini[2]
**madrzyk2@illinois.edu*

[1]*Siebel School of Computing and Data Science*
Grainger College of Engineering
University of Illinois at Urbana-Champaign
Champaign, IL, USA

[2]*Department of Business Administration*
Gies College of Business
University of Illinois at Urbana-Champaign
Champaign, IL, USA

***Abstract*—Access to reliable consumer health information is increasingly challenging as more people seek medical guidance online. Prior studies show that a majority of adults seek health information online, while the quality of search results remains inconsistent. However, large language models can generate fluent answers that may contain hallucinations or reflect bias, which can be especially harmful for users with limited health literacy. A key challenge is the mismatch between how users express health queries in natural language and how medical knowledge is structured in authoritative sources. Although modern search interfaces increasingly present AI-generated summaries, their underlying retrieval pipelines often remain insufficiently aligned with biomedical ontologies and rarely incorporate patient-specific context, thereby limiting grounding and personalization. We present MedAgent, an agentic consumer health information retrieval system built around three design principles. First, MedAgent uses a coordinated multi-agent architecture that decomposes health search into four distinct agents: intent interpretation, retrieval planning, evidence synthesis, and verification. Second, it applies a tiered source strategy that distinguishes authoritative medical literature from community health experiences, allowing the system to weight evidence differently and express uncertainty when sources conflict. Third, it constructs personal health knowledge graphs from a user's medical history to support context-aware query interpretation and more relevant retrieval. We evaluate MedAgent on 200 sampled examples from OpenAI's HealthBench, a benchmark based on 5,000 physician-authored rubrics graded by GPT–4o. MedAgent achieves an overall score of 0.468 compared with 0.320 for a GPT–4o direct-answer baseline. These results indicate that structured agentic retrieval and synthesis can improve benchmark performance relative to direct-answer large language model baselines. MedAgent provides a practical framework for grounded, personalized consumer health information retrieval.**

***Index Terms*—Multi-agent systems, Consumer health information retrieval, Health literacy, Personal health knowledge graphs**

I. Introduction

As of 2022, more than half (58.5%) of U.S. adults used the internet to look for health or medical information, and this behavior increasingly shapes healthcare decisions [1]. However, the quality of available information remains inconsistent, with a recent meta-analysis reporting a 47% prevalence of health misinformation among older adults [2]. These challenges are compounded by widespread health literacy gaps, as approximately 88% of U.S. adults have less than proficient health literacy [3]. Low health literacy is associated with an estimated $106 to $238 billion in annual healthcare costs, driven largely by misinterpreted information, delayed care, and incorrect treatment management [4].

Recent advances in large language models (LLMs) have improved the fluency of generated health responses, but introduce new risks. Studies show that AI-generated health summaries introduce new forms of bias and misinformation [5], and adversarial hallucination rates of 50–82% have been observed across leading LLMs in clinical decision support settings [6]. Beyond accuracy, existing systems often do not adapt responses to users' individual health context, instead producing largely similar responses regardless of whether the user is a clinician, a patient managing a chronic condition, or a caregiver with limited medical background [7].

A further challenge is the structural discrepancy between how consumers express health queries and how medical knowledge is organized in authoritative sources [8]. Existing systems rely on keyword matching or generic embeddings that do not exploit biomedical structure or patient-specific context, limiting both retrieval quality and response reliability [7]. Retrieval-augmented generation (RAG) has shown promise in addressing these limitations, improving LLM accuracy from 73.4% to 80.0% when external knowledge is integrated [9]. However, most deployments remain non-personalized, with limited ability to reason over heterogeneous evidence sources or to explicitly model uncertainty in generated responses.

Prior work shows that agentic LLM architectures can improve accuracy, with a 2025 systematic review reporting gains ranging from 14–17 percentage points for multi-agent systems to 53 percentage points for single-agent tool-augmented agents over LLM baselines [10]. To address these limitations in the

consumer health domain, we present *MedAgent*. The main contributions of this work are summarized as follows:

- We design a multi-agent architecture that decomposes consumer health search into specialized, interpretable stages.
- We introduce a hybrid retrieval engine that combines dense semantic and sparse lexical retrieval via Reciprocal Rank Fusion (RRF), guided by a structured knowledge graph spanning 20 consumer health categories.
- We propose a tier-aware verification mechanism that prioritizes authoritative sources, calibrates response disclaimers across four confidence thresholds, and surfaces uncertainty flags in synthesized outputs.
- We evaluate MedAgent using OpenAI's HealthBench for response-quality benchmarking, together with analyses of personalization effectiveness and confidence calibration.

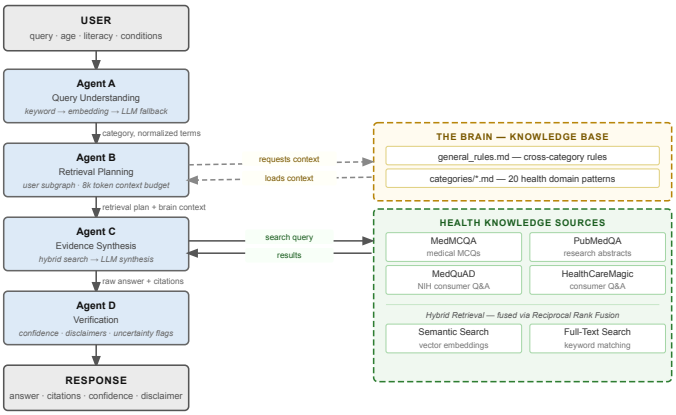

Fig. 1. Proposed end-to-end MedAgent system architecture. Additional implementation details are available in the public repository [19].

II. Related Work

A. Healthcare RAG and Biomedical Retrieval

RAG has become an important approach for grounding LLM responses in healthcare, with a 2025 systematic review reporting consistent improvements across clinical and consumer-facing tasks when external knowledge is incorporated [9]. A scoping review of 67 studies further identifies text-based RAG, knowledge graph-enhanced RAG, and agentic RAG as three prominent architectural categories, while benchmarks such as MIRAGE evaluate medical RAG pipelines along dimensions of retrieval quality and answer accuracy [7], [11]. Hybrid retrieval methods that combine dense embedding-based search with sparse keyword-based retrieval have been widely used to improve robustness in biomedical search. Reciprocal rank fusion (RRF) has proven effective for combining heterogeneous retrieval signals in biomedical settings [12], and domain-specific embedding models such as MedCPT have demonstrated strong zero-shot performance on biomedical information retrieval tasks [13]. However, most such systems remain oriented toward general biomedical or research search and do not incorporate patient-specific context or consumer-oriented interpretation of health queries.

B. Multi-Agent LLM Systems

Multi-agent LLM systems have emerged as a promising paradigm for handling complex reasoning tasks by decomposing them into coordinated, specialized components. Prior work shows that such task decomposition improves performance on multi-step reasoning and medical decision-making tasks, particularly when problems require sequential inference or integration of heterogeneous information sources [10]. In healthcare settings, these multi-agent systems have been used for tasks such as diagnosis, clinical decision support, and medical report generation [14].

However, existing multi-agent frameworks are largely designed for clinical or general-purpose reasoning and are not tailored to the unique practical requirements of consumer health search. In particular, they often lack mechanisms for communicating uncertainty, prioritizing sources based on authority and reliability, and adapting outputs to varying levels of health literacy.

C. Personalization and Consumer Health QA

Personalization in health information systems has been explored through recommender systems and patient-specific modeling, often using demographic or clinical data to tailor recommendations [15], [16]. A survey of consumer health question answering systems identifies key challenges including query ambiguity, source heterogeneity, and the need for lay-language responses [8]. Biomedical knowledge graphs such as PrimeKG [17] and SPOKE [18] have been used to enrich retrieval with structured relational knowledge, but integration with personalized LLM pipelines remains limited. These gaps motivate a framework that integrates personalization, structured medical knowledge, and LLM-based reasoning for consumer health search.

III. System Architecture

MedAgent processes each user query through a structured four-agent pipeline designed to improve retrieval quality, personalization, and response reliability. Each agent performs a distinct function and communicates via a shared `QueryContext` object that is passed sequentially through the pipeline. This design enables modular reasoning while preserving coherence across all stages. The pipeline is instantiated once at application startup and invoked for each query, with shared components such as the base knowledge graph and hybrid searcher constructed once and reused across queries. Fig. 1 illustrates the end-to-end MedAgent architecture and the flow of information across the four-agent pipeline.

A. The Brain: Structured Medical Knowledge Base

A central component of MedAgent is the "brain", a structured knowledge base that encodes medical knowledge across 20 consumer-facing health categories. Each category is represented as a markdown file that captures several types of structured information. This includes a terminology map

(consumer-to-medical translations), source priority overrides, common query patterns, category-specific retrieval and verification rules, known pitfalls, and explicit link metadata connecting related categories. A separate rules file defines system-wide policies that apply to every query, including the four-tier source trust hierarchy, confidence thresholds, multi-source requirements, and personalization baselines.

The brain serves two purposes. First, it grounds Agent A's terminology matching by providing a structured vocabulary index that maps lay language to clinical terms. Second, it provides Agent B with a token-budgeted context block that guides Agent C's synthesis prompt, ensuring that generated responses are shaped by domain-specific knowledge and source priorities rather than by generic LLM behavior alone.

B. Agent A: Query Understanding Agent

Agent A classifies the user's natural language input into one of 20 brain category files and normalizes consumer language into medically meaningful terminology. Classification proceeds through a three-stage cascade designed to balance efficiency and coverage, while maintaining high classification accuracy. In the first stage, the agent performs exact phrase and token-level matching against a terminology index parsed from all brain category files at startup. Full phrase substring matches are given priority, and individual token hits are used as a fallback within this stage, with ties resolved only when a single clear winner exists. If keyword matching fails to produce a confident result, the agent embeds the query using a small text embedding model and computes cosine similarity against prebuilt category centroids derived from the embeddings database. We use a similarity threshold to ensure that the proper category is selected. If the embedding step also fails, the agent falls back to using an LLM as a classifier, prompting it to return a single category option from the known list.

Following classification, Agent A performs Hypothetical Document Embedding (HyDE) query expansion [20]. It rewrites the consumer query into one to two sentences in the style of a clinical case presentation, seeded with medical terms from the brain's terminology map for the classified category. This bridges the vocabulary gap between consumer language and the clinical language in the embedding corpus, producing richer search terms for downstream retrieval.

C. Agent B: Retrieval Planning Agent

Agent B takes the classified query and the user profile produced by Agent A to construct a personalized retrieval plan. It begins by building a per-user overlay of the base knowledge graph through a five-stage sequence. First, the user's active medical conditions are anchored to corresponding subcategory nodes in the graph. Second, related nodes are expanded from those anchor points. Third, known comorbidity interactions are applied. For example, a user with both diabetes and chronic kidney disease will receive boosted relevance toward cardiovascular categories, reflecting clinically established relationships encoded in the brain. Fourth, demographic factors such as age range and biological sex are used to apply category-level adjustments. For instance, users over 50 years of age receive increased relevance toward cardiovascular and musculoskeletal categories. Fifth, active medications boost the medications and drug safety category, as well as any related condition categories. The resulting user subgraph encodes which areas of the knowledge graph are most relevant to the specific user submitting the query.

Agent B then merges this user subgraph with the primary query category to produce a `RetrievalPlan`, which specifies which database tables to prioritize, which related categories to include as supplementary context, and the effective weights to apply during post-retrieval reranking. Finally, Agent B assembles the brain context block that is passed to Agent C as part of the synthesis prompt. It loads the primary category file, any supplementary category files within a token budget, and the system-wide general rules.

D. Agent C: Evidence Synthesis Agent

Agent C executes hybrid search using the retrieval plan produced by Agent B and synthesizes a personalized, evidence-grounded response. The search draws from four structured medical databases: MedMCQA [21], a large collection of clinical multiple-choice questions and explanations; PubMedQA [22], a biomedical research question answering dataset; MedQuAD [23], a consumer-facing health QA dataset derived from NIH sources; and HealthCareMagic [24], a dataset of over 100,000 real-world consumer health conversations between patients and physicians. Semantic and lexical retrieval signals are fused via Reciprocal Rank Fusion, and results are subsequently reranked using the graph-derived category weights from the retrieval plan. This allows categories that are relevant to the user to surface even when they rank lower in the raw retrieval lists.

The synthesis prompt is tailored to the user's health literacy level across three tiers. These tiers correspond to simplified language targeting a low-literacy reading level, plain language with defined medical terms for moderate literacy, and full clinical terminology for high-literacy users. The model is instructed to identify red flags and emergency warning signs, describe evidence-supported treatment considerations, including dosing information when supported by retrieved sources, and highlight appropriate referral thresholds. It is further guided to provide mechanistic explanations rather than surface-level summaries. When a query lacks critical context that could materially affect the response, the agent first generates targeted clarifying questions before proceeding. Citations are extracted from retrieved documents and passed separately to Agent D, rather than embedded directly in the generated response.

E. Agent D: Verification Agent

Agent D evaluates the synthesized response for reliability and safety, operating solely on retrieved evidence and confidence signals available in the pipeline context. It first filters out low-relevance results using a minimum cosine-similarity threshold, thereby reducing noise in subsequent scoring.

Confidence is computed as a weighted combination of source quality and retrieval relevance, with additional weight assigned to corroborated evidence across multiple sources. Based on this score, responses are routed through four confidence tiers. High-confidence responses are returned with minimal modification, while lower-confidence levels trigger progressively more conservative messaging, ranging from light clinical reminders to general information cautions to limited evidence warnings. At intermediate levels, an epistemic hedge is introduced to reflect appropriate uncertainty in the response.

Two safety flags are applied when necessary. An uncertainty flag is raised for low-confidence responses, while a high-risk flag indicates cases requiring provider verification before action. A hard safety cap also limits confidence for responses supported by only a single source, reducing the risk of over-confident outputs in low-evidence settings. Finally, citations are deduplicated by source and record identifier and ranked by source quality and relevance before being returned to the user.

IV. Methods

This section describes the core methodological components of MedAgent: hybrid retrieval, personalized graph-based retrieval planning, and the confidence and safety verification framework.

A. Hybrid Retrieval

MedAgent retrieves candidate evidence from four structured biomedical datasets stored in a PostgreSQL database: *medmcqa_records*, containing approximately 182,000 clinical multiple-choice questions and explanations; *pubmedqa_records*, containing 1,000 biomedical research question-answering pairs; *medquad_records*, containing approximately 16,000 consumer-facing health QA pairs derived from NIH sources (excluding copyright-restricted MedlinePlus subsets); and *healthcaremagic_records*, containing approximately 112,000 real-world consumer-physician conversations. Together, these four tables provide complementary coverage across clinical, research, and consumer-oriented health information.

1) Semantic Retrieval: Semantic retrieval is performed using dense vector embeddings, which produce high-dimensional representations for both queries and documents. Document embeddings are stored in PostgreSQL via the `pgvector` extension, with IVFFlat indexing to support efficient approximate nearest neighbor search. Retrieval is performed using cosine distance, returning the top 20 most semantically similar records per source table for each query.

2) Lexical Retrieval: Lexical retrieval uses PostgreSQL's native full-text search infrastructure. Document fields are pre-processed into `tsvector` representations and indexed using Generalized Inverted (GIN) indexes to support high-throughput keyword lookup. At query time, the HyDE-expanded query is parsed and matched against indexed columns using `ts_rank` scoring. The top 20 results per source table are retrieved, providing complementary precision for exact-match entities such as medication names, dosage specifications, and diagnostic terminology that may not be well captured by similarity in embeddings alone.

3) Reciprocal Rank Fusion: The ranked lists produced by semantic and lexical retrieval are merged using Reciprocal Rank Fusion (RRF) [12], which combines heterogeneous retrieval signals without requiring score normalization across retrievers. We set $k = 60$ following standard practice [12], and assign a fallback rank to documents absent from one retrieval list so they still contribute a small positive signal rather than being discarded entirely. To provide sufficient candidate coverage before downstream reranking, the system overfetches by a factor of three, collecting three times the target number of final results prior to graph-guided reranking.

4) Graph-Guided Post-Fusion Reranking: Following RRF fusion, retrieved candidates are reranked using category weights derived from the personalized user subgraph. Documents associated with categories that are more relevant to the user's health profile receive a proportional score boost based on the effective graph weight, with the boost scaled by a factor of 0.3. In addition, documents belonging to categories designated as must-load in the retrieval plan receive a flat additive bonus of 0.10. This reranking step helps highly relevant categories surface in the final ranking even when they appear lower in the raw retrieval lists.

B. Personalized Retrieval via Knowledge Graphs

The base graph encodes 20 consumer health categories as nodes, with typed, directed edges classified as *strong*, *weak*, or *related*, reflecting the degree of clinical association between pairs of categories. Agent B constructs a user-specific subgraph overlay as described in the Agent B section, by adding profile-derived boosts to the base graph relationship weights and clipping the final value at 1.0 to prevent over-amplification when multiple risk factors overlap. The resulting subgraph is encoded into a `RetrievalPlan` specifying source table priorities, supplementary category inclusions, and per-category weights for post-fusion reranking.

C. Confidence and Safety Framework

1) Confidence Scoring: After filtering out documents below a minimum cosine similarity threshold, each source is assigned a quality tier that reflects its level of authority, ranging from peer-reviewed biomedical literature at the highest tier to unstructured consumer-physician conversations at the lowest. The overall confidence score is computed as a weighted combination of mean source quality and mean retrieval relevance, with an additional diversity bonus for corroboration across multiple independent source tables, and is then clipped to the $[0, 1]$ range.

2) Tiered Disclaimer and Uncertainty Propagation: The confidence score c routes the response through one of four output tiers. High confidence responses are returned with standard clinical context reminders. Intermediate scores trigger progressively more conservative messaging, ranging from a

light clinical reminder to a medium caution advisory. Responses that fall below the lowest threshold receive a limited evidence warning. An uncertainty flag is additionally raised for responses below a minimum confidence threshold.

V. Evaluation

A. Experimental Setup

We evaluate MedAgent using OpenAI's HealthBench [25], a benchmark designed to assess the quality of AI-generated health responses across clinically relevant dimensions. HealthBench evaluates responses along five axes, including accuracy, communication quality, context awareness, completeness, and instruction following. It also groups queries into seven thematic categories, including emergency referrals, hedging, health data tasks, and context seeking. We compare MedAgent against a GPT–4o baseline using $n = 200$ examples drawn from the benchmark. All evaluations use the same underlying model (GPT–4o) to help ensure that observed differences reflect the contribution of the MedAgent pipeline rather than model capability.

B. Evaluation Metrics

We evaluate system performance across three dimensions: HealthBench score, personalization effectiveness, and confidence calibration. HealthBench score measures overall response quality as a composite of the five axes described above, enabling direct comparison against the GPT–4o baseline across thematic categories. Personalization effectiveness assesses whether the system produces measurably different responses across user literacy profiles by measuring the Flesch-Kincaid grade level for the same queries run under low, medium, and high literacy settings. Confidence calibration evaluates Agent D by measuring the correlation between per-query confidence scores and per-query HealthBench rubric pass rates, assessing whether the verification mechanism tracks higher quality responses.

C. Results

1) HealthBench Comparison: Table I reports HealthBench scores for MedAgent and the GPT–4o baseline across seven thematic categories, along with an overall aggregate score. MedAgent outperforms the baseline on six of seven themes, with an overall score of 0.468 versus 0.320 (+46%). The largest gains are observed in context seeking (+71%), global health (+54%), and health data tasks (+50%), suggesting that retrieval grounding and user-specific context have the greatest impact on queries requiring situational awareness and information synthesis. Gains in hedging (+47%) and communication (+49%) reflect the contribution of the confidence and safety framework, which introduces calibrated uncertainty messaging that the baseline does not produce by default. A small decline is observed in complex responses (-1%), indicating that multi-step clinical reasoning remains an area for improvement.

Table II further breaks down MedAgent performance across the five HealthBench evaluation axes. Communication quality (0.734) and accuracy (0.580) are the strongest axes, while completeness (0.344) and context awareness (0.392) represent the weakest. These results suggest that the pipeline produces well-structured and accurate responses, while evidence coverage and user-specific contextualization remain opportunities for future improvement.

TABLE I
HealthBench scores by theme, MedAgent vs. GPT–4o baseline ($n = 200$).

Theme	GPT–4o	MedAgent	Δ
Overall	0.320	0.468	+46%
Communication	0.404	0.602	+49%
Emergency referrals	0.463	0.597	+29%
Hedging	0.354	0.520	+47%
Health data tasks	0.318	0.477	+50%
Context seeking	0.205	0.351	+71%
Complex responses	0.347	0.342	-1%
Global health	0.210	0.323	+54%

TABLE II
MedAgent HealthBench axis scores ($n = 200$).

Axis	Score
Accuracy	0.580
Communication quality	0.734
Context awareness	0.392
Completeness	0.344
Instruction following	0.540

2) Personalization Effectiveness: To validate health literacy adaptation, we evaluate whether responses generated under different literacy profiles exhibit measurable differences in readability. We compute the Flesch-Kincaid (FK) grade level for responses generated under low, medium, and high literacy settings across $n = 200$ queries.

The results show a clear monotonic increase in reading complexity across profiles (LOW: 11.5 ± 2.1, MEDIUM: 13.3 ± 1.9, HIGH: 14.5 ± 2.0), with a total spread of $\Delta = +3.0$ grade levels between the low and high profiles. This exceeds our target threshold of a 2 grade level difference, confirming that literacy-specific prompting meaningfully modulates response complexity.

Although the absolute FK score for the low-literacy profile (11.5) exceeds the intended 6th-grade target, this is consistent with a known floor effect in medical content. Clinical terminology (e.g., drug names, anatomical terms) inherently increases syllable counts, inflating FK scores even when responses are simplified. Despite this limitation, the monotonic ordering and consistent separation across profiles support the intended personalization behavior.

3) Confidence Calibration: To evaluate the reliability of Agent D's confidence scoring, we measure the correlation between per-query confidence scores and per-query HealthBench rubric pass rates. We observe a weak and statistically insignificant correlation between confidence and rubric performance (Pearson $r = 0.011$, Spearman $\rho = 0.038$, $p > 0.5$), indicating that the current confidence signal does not reliably predict answer quality.

This result reflects a structural limitation of the confidence formulation. Agent D computes confidence based on retrieval

quality (source tier weights and relevance scores), which measures the quality of the supporting evidence rather than the quality of the final generated answer. As a result, high-quality sources do not necessarily translate into complete or accurate responses, and vice versa. This finding highlights the need for answer-level confidence estimation methods that incorporate generation quality, not just retrieval signals, and motivates future work on integrated verification mechanisms.

D. Discussion

MedAgent performs best on queries that benefit from contextual grounding, uncertainty handling, and user-specific information. Its weaker performance on complex responses suggests that the current pipeline is more effective at evidence-guided synthesis than at multi-step clinical reasoning. The low completeness score further indicates that broader evidence coverage and deeper synthesis remain important limitations. Future work should therefore focus on deeper retrieval, stronger multi-step synthesis, and answer-level verification mechanisms that better align confidence with response quality.

Ultimately, this approach establishes a robust foundation for consumer health search, outperforming standalone LLMs in contextual grounding and uncertainty management. While current limitations in completeness and complex responses highlight the broader challenge of multi-step reasoning, our multi-agent architecture successfully mitigates immediate hallucination and sourcing risks. Future enhancements will focus on deeper retrieval, advanced synthesis, and structured evaluation frameworks, ultimately bridging the gap between complex medical knowledge and accessible, trustworthy consumer health search.

VI. Conclusion

We presented MedAgent, a personalized consumer health search framework designed to improve the reliability, relevance, and accessibility of medical information. By decomposing search into specialized agents for query understanding, retrieval planning, evidence synthesis, and verification, the system supports modular and interpretable processing. MedAgent combines hybrid retrieval, graph-based personalization, and tier-aware confidence control to better align retrieved evidence with user context and to communicate uncertainty more explicitly.

While the current system demonstrates the feasibility of integrating personalization and verification into health information retrieval, several limitations remain. Future work will focus on broader evaluation, user studies of health literacy adaptation, lower-latency deployment, and tighter integration with clinical-grade knowledge sources. Taken together, these directions position MedAgent as a promising foundation for continued progress in grounded, personalized consumer health information retrieval.

References

[1] X. Wang and R. A. Cohen, "Health information technology use among adults: United States, July–December 2022," NCHS Data Brief, no. 482, National Center for Health Statistics, Centers for Disease Control and Prevention, Oct. 2023.

[2] B. Hu, X. Liu, C. Lu, and X. Ju, "Prevalence and intervention strategies of health misinformation among older adults: A meta-analysis," *J. Health Psychol.*, vol. 30, no. 2, 2025.

[3] M. Kutner, E. Greenberg, Y. Jin, and C. Paulsen, "The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy," NCES 2006-483, National Center for Education Statistics, U.S. Department of Education, Washington, DC, Sep. 2006.

[4] J. A. Vernon, A. Trujillo, S. J. Rosenbaum, and B. DeBuono, "Low health literacy: Implications for national health policy," Department of Health Policy, School of Public Health and Health Services, The George Washington University, Washington, DC, 2007.

[5] C. Wardle, S. Urbani, and E. Wang, "Evolving health information-seeking behavior in the context of Google AI Overviews, ChatGPT, and Alexa: Interview study using the think-aloud protocol," *J. Med. Internet Res.*, vol. 27, e79961, 2025.

[6] M. Omar, V. Sorin, J. D. Collins, D. Reich, R. Freeman, N. Gavin *et al.*, "Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support," *Commun. Med.*, vol. 5, no. 330, 2025.

[7] Y. Miao, Y. Zhao, Y. Luo, H. Wang, and Y. Wu, "Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review," *J. Med. Internet Res.*, vol. 27, e80557, 2025.

[8] A. Welivita *et al.*, "A survey of consumer health question answering systems," *AI Magazine*, 2023.

[9] L. M. Amugongo *et al.*, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digit. Health*, vol. 4, no. 6, e0000877, 2025.

[10] A. Gorenshtein, M. Omar, B. S. Glicksberg, G. N. Nadkarni, and E. Klang, "AI agents in clinical medicine: A systematic review," *medRxiv*, 2025.

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

[12] J. Sun, S. Wu, X. Shen, C. Nugent, and H. Lu, "Subset selection based fusion for biomedical information retrieval tasks," *BMC Bioinformatics*, 2025.

[13] Q. Jin *et al.*, "MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval," *Bioinformatics*, vol. 39, no. 11, btad651, 2023.

[14] H. Yuan, "Agentic large language models for healthcare: Current progress and future opportunities," *Med. Adv.*, 2025.

[15] Y. H. Alfaifi, "Recommender systems applications: Data sources, features, and challenges," *Information*, vol. 15, no. 10, 660, 2024.

[16] D. Roy and M. Dutta, "A survey on personalized health recommender systems for diverse healthcare applications," in *Proc. 4th Int. Conf. Computing and Communication Systems (I3CS)*, 2023.

[17] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, 2023.

[18] J. S. Baranzini *et al.*, "The scalable precision medicine open knowledge engine (SPOKE): A massive knowledge graph of biomedical information," *Bioinformatics*, vol. 39, no. 2, 2023.

[19] K. Trikha, B. Madrzyk, and A. Castelino, "MedAgent," GitHub repository. [Online]. Available: https://github.com/ashcastelinocs124/MedAgent

[20] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," in *Proc. 61st Annu. Meeting Assoc. Computational Linguistics (ACL)*, 2023, pp. 1762–1777.

[21] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering," arXiv:2203.14371, Mar. 2022.

[22] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," arXiv:1909.06146, Sep. 2019.

[23] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinformatics*, vol. 20, no. 511, 2019.

[24] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge," arXiv:2303.14070, Mar. 2023.

[25] R. K. Arora, J. Wei, R. S. Hicks, P. Bowman, J. Quiñonero-Candela, F. Tsimpourlas *et al.*, "HealthBench: Evaluating large language models towards improved human health," arXiv:2505.08775, May 2025.