Leveraging AI to improve diagnosis and treatment of rare diseases: a chat agent for equitable and accessible healthcare

Frank Hodges, Sundareswar Pullela, Jared Roach, and Stephen A. Ramsey

Oregon State University (contact-us@rarepath.ai)

Executive Summary

In this white paper, we describe Radiant (radiant.rarepath.ai), a free-to-use, artificial intelligence (AI)-driven chat agent designed to assist in the diagnosis, therapeutic target discovery, and management of rare diseases. By leveraging large-language models (LLMs) in combination with comprehensive databases extracted from scientific and medical literature, Radiant will provide users—patients, caregivers, and healthcare professionals—with timely, accurate, and evidence-based answers to their questions about rare diseases.

We have completed and released a prototype of Radiant that is geared toward two rare diseases, hypophosphatasia and Ehlers-Danlos syndrome, but a future version will be capable of handling the approximately 7,000 known rare diseases. The chat agent is designed to make healthcare more equitable and accessible, particularly for individuals affected by rare diseases. It will do so by offering insights into disease mechanisms, causal genes, and potential therapeutic approaches.

Through a transdisciplinary effort led by a team at Oregon State University, we continue to refine and validate Radiant through input from AI experts, healthcare providers, translational scientists, and a broad network of rare disease advocates. Our long-term goal is to significantly improve treatment and enable more patients to access effective treatment for rare diseases, improving outcomes and quality of life.

We invite you to take a brief survey to help us understand how an AI chat agent can assist healthcare providers and members of the rare disease community in retrieving useful, relevant, and up-to-date information about rare diseases, available here.

The impact and diagnostic challenges of rare diseases

Rare diseases[1] are a broad category of low-prevalence genetic disorders that reduce both quality of life and life expectancy, can cause debilitating symptoms, and have a significant public health burden. Globally, approximately 300 million people live with a rare disease, accounting for about 6% of the population [1]. The majority of rare disease patients lack access to effective treatments; 90% of rare diseases have no FDA-approved therapeutic agents available [2]. The FDA-approved therapies that are available most often target symptoms rather than underlying disease mechanisms.

Today, there is a lack of tools that are specifically designed to access current and scientifically accurate information about rare diseases, and this gap hinders timely diagnosis and accurate treatment. Overcoming this gap poses a significant challenge, in part because of the sheer number and diversity of conditions. Difficulties in accessing accurate, organized, relevant, and up-to-date information delay patient diagnoses and cause misdiagnoses. The average rare-disease patient experiences three misdiagnoses and consults with five doctors during a multi-year “diagnostic odyssey” before receiving a correct diagnosis. With the small number of specialists trained to diagnose and treat these diseases, and the uneven distribution of rare-disease specialists, patients typically must travel to access definitive care. The diagnostic odyssey worsens outcomes and increases the emotional and financial burden on patients and their families.

Why can’t we do better?

Established diagnostic criteria and treatment recommendations for many rare diseases lag behind the rapidly advancing scientific knowledge about the genes that cause rare diseases and the mechanisms by which they cause them. A clinician treating a rare disease patient must piece together the current state of the literature from searches using information tools such as PubMed, Medline Plus, ClinicalTrials.gov and using preprint archives like medRxiv; the information provided by these sources is not often curated for quality and not organized into a concise and actionable summary. Articles in expert-curated sources such as UpToDate provide well-organized answers for many common diseases, but may be out of date by months or even years, may not be accessible to patients and doctors, and do not address many rare diseases. On the other hand, current-generation AI chat agents such as ChatGPT and Perplexity AI both have the capability to summarize knowledge from multiple sources and augment their built-in knowledge with knowledge obtained via web searches, but they fare poorly at curating sources by quality and tend to favor secondary sources (e.g., press releases

[1] The World Health Organization defines a rare disease as one that affects fewer than 65 per 100,000 people.

and news pieces) rather than the primary biomedical literature and clinical trial registrations. Today, to access all relevant information for a rare-disease case, a clinician must search multiple sources and synthesize, curate, and organize the information themselves.

Radiant: a solution to address these gaps

As a first step to address these gaps, we are developing an AI-based knowledge exploration tool, called Radiant, that is designed for providing information about rare diseases. Our prototype Radiant tool is available to use online free-of-charge, at the web address radiant.rarepath.ai. Radiant provides a chat interface (Fig. 1) through which healthcare providers, patients, and researchers can ask questions in natural language (at this time, in English only) and receive summarized answers backed up by citations to primary sources such as peer-reviewed biomedical research articles and clinical trial registrations. The tool generates responses using a foundational large-language model whose built-in knowledge is augmented with text from biomedical research articles selected based on relevance to the user's query. Unlike general-purpose chat agents, Radiant will be fine-tuned to provide responses that are tailored to the unique challenges and scientific vocabulary of rare-disease diagnosis and management. Our team has designed Radiant with three main use-cases in mind: (1) help providers quickly and accurately diagnose rare diseases and explore what is known about how they occur; (2) help patients and caregivers research timely and beneficial treatment options within their healthcare and community setting; and (3) help researchers discover new treatments and therapeutic targets. Broad adoption of an AI-based approach that is dedicated to rare disease, like Radiant, has the potential to both decrease the time to diagnosis and reduce the frequency of misdiagnoses.

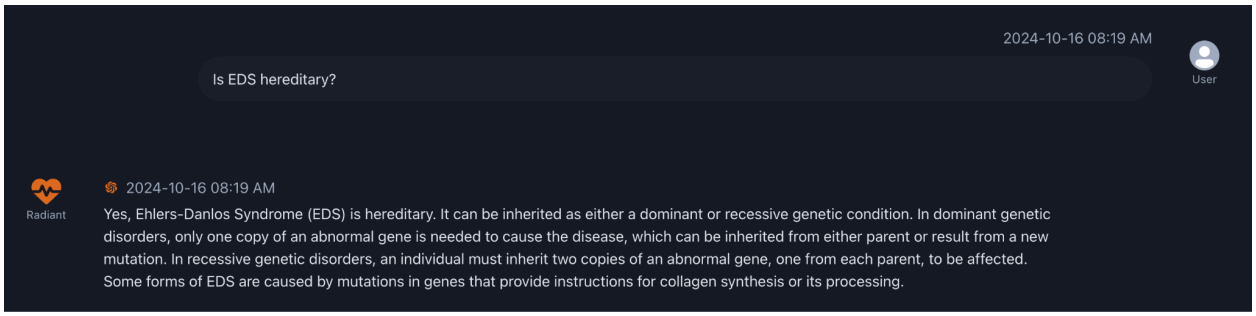

Figure 1: example usage of Radiant

Technological overview

Radiant incorporates some of the same AI technology that underpin general-purpose chat agents such as ChatGPT and Google Gemini, but in a manner specifically

designed for knowledge exploration and summarization for rare diseases. Collectively, rare diseases pose several challenges for an AI knowledge exploration tool: rapidly evolving knowledge due to advances in biomedicine and genomics; a very high prioritization on correctness of answers; and a very broad array of topics (namely, the 7,000 rare diseases and their associated genetic variants). These challenges motivated Radiant's design. In this section, we describe the main types of AI technology that are used by Radiant, including LLMs, fine-tuning, retrieval-augmented generation (RAG), and a vector database.

Large-language models

LLMs are sophisticated AI systems designed to comprehend and generate human language. They learn from extensive general-purpose datasets, consisting of millions of text examples, to recognize and predict patterns, structures, and relationships within language. Given an input query, the LLM generates output text by statistically predicting the most likely sequence of words based on its accumulated knowledge from the training text examples. This enables the model to produce coherent and contextually accurate replies across a wide range of topics. To further enhance the effectiveness of these models, fine-tuning is often employed.

Fine-tuning

Fine-tuning is a process that further trains an LLM on narrower, domain-specific text examples to tailor its capabilities to specific fields, vocabulary, or tasks. The process refines the LLM's ability to generate more accurate and relevant responses in specialized contexts by garnering associations between domain specific terminology, patterns, and expectations. This additional training makes the model more effective in its target area and more reliable in real-world applications where precision and relevance are critical.

Vector database

A vector is a numerical representation of data, created by an embedding function that transforms text, images, or other types of data into high-dimensional arrays of numbers (Fig. 2). These vectors capture the semantic meaning of the data in a way that computers can understand. A vector database is a specialized database designed to efficiently store, index, and retrieve these vectors. This allows for rapid similarity searches, where the system can quickly find and rank the most relevant data points based on their proximity in the vector space.

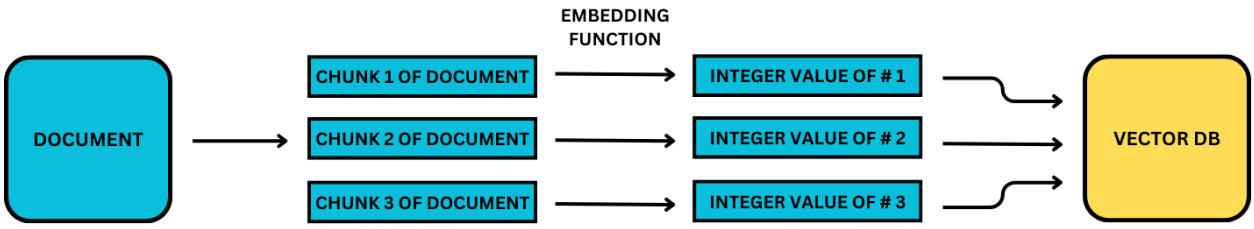

Figure 2: flowchart of the embedding process to create a vector database (DB)

Retrieval-augmented generation (RAG)

RAG systems enhance the performance of LLMs by referencing one or more vector databases before generating a response. The process begins with the user's question, which is first converted into an embedding—a mathematical representation—using the same embedding function that was applied to the documents stored in the vector database. This ensures consistency in how data is represented. The query's vector is then used to search the vector database for similar vectors, representing relevant information. When values meet a certain similarity threshold, they are retrieved as pertinent information. This retrieved information is then decoded into human-readable text and sent to the LLM along with the user's original query (Fig. 3). By providing the LLM with up-to-date and contextually relevant information, RAG systems improve the accuracy and relevance of the generated responses.

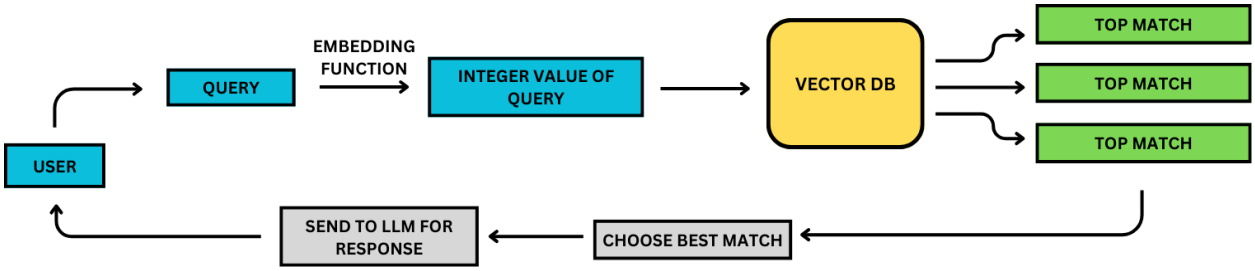

Figure 3: flowchart of RAG pipeline

Current limitations of AI

Although state-of-the-art AI technology is both powerful and useful, it is important to recognize the inherent limitations of using AI, which persist despite AI researchers' extensive efforts to minimize the limitations. These limitations include, but are not limited to, the possibility of generating inaccurate or misleading information (hallucinations), reliance on potentially outdated training data, overdependence on the data used during training, and the necessity for well-constructed inputs to achieve accurate results.

Our solution and how it minimizes issues with AI

To address the challenges posed by current AI technologies, Radiant uses a fine-tuned LLM with a comprehensive vector database and retrieval-augmented generation (RAG)

pipelines, delivering responses that are not only contextually appropriate but also accurate and grounded in real-time data. Radiant's approach involves a preparation step and then four steps at the time a user makes a query:

Preparation step. Radiant builds a vector database from carefully chosen literature specific to the rare disease in question, augmented as appropriate by high-quality literature of general medical and clinical relevance. The process of selecting domain-specific literature for each of the 7,000 rare diseases can be fully automated, and/or augmented with human curation for specific diseases.

Step 1. When a user submits a query, it first undergoes an augmentation process, where additional queries similar to the original are generated. This step ensures that the system considers the underlying intent of the question, not just its exact wording, thereby improving the quality of the information retrieved.

Step 2. Radiant searches through our extensive vector database using both the original query and the augmented queries to find several relevant document fragments. These text fragments are retrieved from the database and reranked by an algorithm that assesses their relevance to the original user query, with the top three most relevant fragments selected for further processing.

Step 3. The selected document fragments are passed to the LLM. Unlike traditional systems that rely solely on an LLM's built-in knowledge, our approach ensures that the LLM generates responses strictly using the context provided by the document fragments from Radiant's vector database. This method provides two key benefits. First, it ensures that Radiant's responses are maximally derived from curated and traceable sources. Second, it reduces the risk of LLM hallucinations.

Step 4. Finally, the generated responses are presented to the user, complete with clickable references to the original source documents. This allows users to explore the topic in greater depth, ensuring transparency and trust in the information provided.

What's next for Radiant

Our ongoing research and development efforts are focused on optimizing the performance of our AI chat agent and to maximize the accuracy and trustworthiness of its responses. We are refining our algorithms to enhance efficiency and accuracy in understanding medical queries and retrieving the most relevant information from our document database. To ensure our AI is informed by the most current and diverse data, we are actively expanding our data sources, including the latest research publications, clinical trial results, and case studies, while maintaining accuracy through continuous data validation and updates. Additionally, we are working to improve the transparency of the AI's decision-making process by enhancing the traceability of its outputs.

We are continuously improving Radiant's technology stack with the primary goal of expanding the agent's capabilities to cover all known rare diseases by consistently updating and integrating new data as medical knowledge advances, ensuring it remains a reliable resource for healthcare providers managing complex cases. Integrating patient-specific data through electronic health records (EHRs) could potentially enable a next-generation chat agent to access comprehensive information such as medical history, current medications, and previous diagnoses, providing more tailored and accurate recommendations. Additionally, by training the AI model to remember key details across interactions, the system can "learn" from previous conversations, recalling prior symptoms and treatments to build a more informed understanding of a patient's health and enhance the relevance of its responses over time. In a future iteration, Radiant will also utilize comprehensive structured databases of expert-curated disease information in the form of "knowledge graphs", to augment its text document-based capabilities.

As we improve Radiant and broaden its repertoire of rare diseases, our team is reaching out to the healthcare and rare-disease patient communities via surveys, dialog, and demonstrations. Our aim is to grow our network of rare-disease community contacts and early beta-testers of Radiant and use their feedback to guide future development and outreach efforts. We plan to extend the Radiant user interface to enable the user to easily provide feedback on the quality and transparency of answers provided by Radiant. We believe that systematizing the acquisition of structured feedback from beta-testers is essential to objectively measuring Radiant's performance and measuring the performance gains as we make improvements to the system.

Team

The Radiant tool is being developed by a multidisciplinary team that is unified by the belief that AI tools can be a force for good in the lives of rare-disease patients. The core development team includes Frank Hodges, a graduate student at Oregon State University (OSU) in AI who has a background in healthcare; Sundareswar Pullela, a graduate student in computer science at OSU who has expertise in biomedical knowledge graphs; Akshay Mulgund, a graduate student in computer science at OSU who has worked in health data analytics; Derrick Higgins, a postbaccalaureate student in computer science at OSU; Jared Roach, MD PhD, who has expertise in rare-disease genetics and genomics; and Stephen Ramsey, an OSU associate professor in Biomedical Sciences and in Computer Science whose research program is focused on applications of AI in biomedicine. To maximize our technology development and outreach efforts' broad benefits for the rare-disease community and to contribute to emerging standards for the ethical and inclusive use of AI in biomedicine, our team's work on AI for rare disease will be guided by a Community and Ethics Advisory Board

that will include members from the research community (AI and biomedicine), rare-disease healthcare providers, and rare-disease advocates.

Call to action

The Radiant team recognizes that collaboration will be key to driving meaningful advancements in the diagnosis and treatment of rare diseases. We are eager to partner with other rare-disease drug discovery efforts, AI experts, healthcare providers specializing in rare diseases, rare-disease patients, and rare-disease advocates. With the benefit of broad collaboration, we can refine and expand our AI-driven tools to ensure they provide maximum benefit in a manner that is inclusive and accessible to all. If you are interested in finding out more about Radiant or about our efforts on AI for rare disease, please contact our team at contact-us@rarepath.ai.

Conclusion

The Radiant tool is a timely advance in using AI to provide up-to-date knowledge about rare diseases. By integrating fine-tuned LLMs with retrieval-augmented generation (RAG) using current research articles, Radiant will deliver more accurate, relevant, and up-to-date information—information that is aimed at enhancing clinical decision-making and empowering patients.

As we continue to expand Radiant's capabilities and our outreach efforts our focus remains on (1) ensuring that Radiant evolves alongside advances in technology and biomedical knowledge and (2) delivering maximum benefit to the broader rare-disease community. With ongoing updates to our data sources and the incorporation of patient-provided information, Radiant will become an even more powerful tool, offering personalized, data-driven support that adapts to the needs of each individual. By revolutionizing the way rare diseases are diagnosed and managed, AI tools have the potential to significantly improve patient outcomes and quality of life.

References

1. Fehr and F. Prütz, "Rare diseases: a challenge for medicine and public health," *Journal of Health Monitoring*, vol. 8, no. 4, pp. 3-6, 2023. Available: PMC10790412.

2. Rare Diseases International, "Living with a rare disease," [Online]. Available: Living With a Rare Disease. [Accessed: Aug. 15, 2024].

3. Summit Health, "Challenges in treating rare diseases," [Online]. Available: Challenges of Treating Rare Disease. [Accessed: Aug. 15, 2024].

4. T. Willmen, L. Willmen, A. Pankow, S. Ronicke, H. Gabriel, and A. D. Wagner, "Rare diseases: why is a rapid referral to an expert center so important?" *BMC Health Services Research*, vol. 23, 2023. Available: PMC10463573.

5. Z. Zhang, "Diagnosing rare diseases and mental well-being: a family's story," *Orphanet J Rare Dis.*, vol. 18, p. 45, 2023. Available: PMC9990187.

6. National Organization for Rare Disorders, *30-year survey report: Barriers and challenges*, [Online]. Available: Barriers and challenges. [Accessed: Aug. 15, 2024].

7. "HIPAA and GDPR: Data Compliance with Anonymization," ShareMedix, [Online]. Available: HIPAA. [Accessed: 16-Aug-2024].

8. "Tackling Data Privacy and Regulatory Compliance in AI-Driven Drug Discovery," Medneed, 30-Nov-2023. [Online]. Available: data privacy and-regulatory compliance. [Accessed: 16-Aug-2024].

9. "Large Language Models (LLMs) in Healthcare," PubMed, [Online]. Available: DOI: 10.7326/M23-2772. [Accessed: 16-Aug-2024].

10. "HIPAA Encryption Requirements," *HIPAA Journal*, [Online]. Available: encryption requirements for HIPAA. [Accessed: 16-Aug-2024].

11. "Understanding Rare Disease: Undiagnosed Diseases," NORD, [Online]. Available: NORD. [Accessed: 16-Aug-2024].

12. "Regulations, Compliance, and Data Security in Digital Health," Bene Studio, [Online]. Available: Digital health data security. [Accessed: 16-Aug-2024].

13. "Large Language Models in Healthcare: A Review," NCBI, [Online]. Available : PMC8932585. [Accessed: 16-Aug-2024].

14. "How Symptoms of Rare Diseases Can Mimic Common Conditions," TGen, [Online]. Available: common symptoms mirror rare symptoms. [Accessed: 16-Aug-2024].