Rare Disease RAG Chatbot

A Retrieval-Augmented Generation System for Evidence-Based Answers on Rare Diseases

Shlomi Jakubowicz – June 2025 (BIU DS18)

Why This Project

- **Rare diseases affect millions**, yet expert knowledge is limited and scattered across sources.
- **Clinicians struggle to find reliable, evidence-based answers quickly**; traditional LLMs often hallucinate or lack biomedical grounding.
- **Goal:** Build a trustworthy AI assistant for 20[A] rare diseases using structured (ORDO, mim2gene) and unstructured (PubMed) data.

- **Demo:**
  https://rare-disease-rag-app-etaz9a8nphysbuuhxlyj85.streamlit.app/

- [A] Diseases list: Cystic Fibrosis, Huntington's Disease, Duchenne Muscular Dystrophy, Spinal Muscular Atrophy, Hemophilia A, Hemophilia B, Gaucher Disease, Pompe Disease, Neurofibromatosis Type 1, Prader-Willi Syndrome, Angelman Syndrome, Rett Syndrome, Fragile X Syndrome, Phenylketonuria, Alpha-1 Antitrypsin Deficiency, Marfan Syndrome, Ehlers-Danlos Syndrome (Hypermobile Type), Sickle Cell Anemia, Thalassemia Major, Crigler-Najjar Syndrome Type 1

Data Sources and Tools

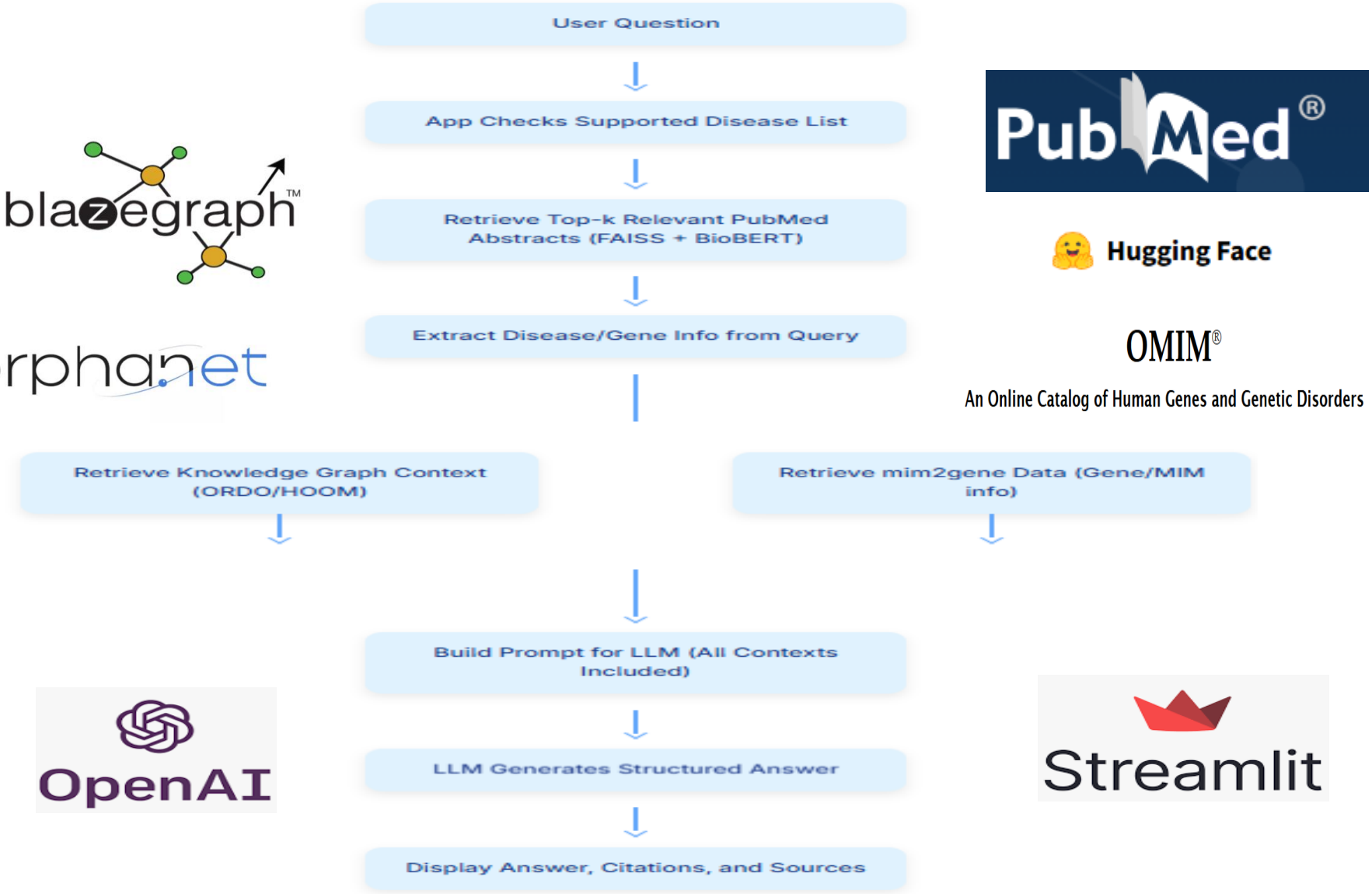
Data

- **PubMed Abstracts:** ~20,000 abstracts covering 20 rare diseases.
- **ORDO via Blazegraph:** Structured rare disease data from the *Orphanet Rare Disease Ontology*, queried locally using SPARQL.
- **mim2gene.txt:** Gene-disease mappings from OMIM for gene-level context.

Tools

- **BioBERT + FAISS:** Embedding model and vector search engine for semantic retrieval.
- **GPT-4:** Generative models for answer synthesis.
- **Streamlit :** Frameworks for building the user interface.

System Architecture



Example Use Cases LLM vs RAG Comparison

Question: What are the common genetic mutations associated with Cystic Fibrosis, what are their clinical manifestations, and what are the latest gene-editing therapies under investigation for CF?"

Feature/Aspect	Copilot's Response (General LLM)	Your RAG System's Response	Advantage
Overall Scope	Comprehensive general overview	Direct, concise, and focused on "latest" and specifics	Both
Mutation Details	Mentions key mutations, categorizes by class	Clearly identifies CFTR, highlights F508del, mentions others	RAG
Gene-Editing Therapies	Lists types (CRISPR, Base/Prime, mRNA, Gene Replacement)	Specific mentions (CRISPR, Prime, In Utero), explicit challenges	RAG
Source/Evidence	General, numbered citations (e.g., 1, 2, 3)	**Crucially, Direct PMIDs with titles**	RAG
Verifiability	Less directly verifiable	**Highly verifiable**, allows direct lookup of sources	RAG
Factual Accuracy	Generally accurate, but potential for generality/hallucination for niche details	**High accuracy**, grounded in retrieved evidence	RAG
Freshness of Info	Stated "as of 2025," but less explicit verification of latest research	Reflects "latest investigations" via recent PMIDs	Both
Actionability	Provides good information for understanding	**Actionable for researchers/clinicians** to explore primary literature	RAG

LLM: excels at generating fluent, broad-ranging text from its vast training data, including diverse news and public information.
RAG: system's strength is providing highly accurate, verifiable, and up-to-date answers by grounding its responses in specific, evidence-based retrieved data.

Challenges, Summary & Next Steps

Challenges:

- Dependency hell ☺ – Tackled.
- Original idea was to connect online to Orpahnet endpoint and query (not done due API availability & restrictions.
- Result hallucinations (e.g. PMID : 123456) – Tackled.

Achievements:

- Built a RAG system for rare disease Q&A
- Integrated structured (ORDO, mim2gene) and unstructured (PubMed) biomedical data
- Demonstrated use cases with improved evidence-backed answers

Next Steps:

- Expand coverage to more rare diseases
- Enhance retrieval accuracy and relevance
- Pursue additional data (e.g., clinical trials data, drugs DB)
    - Improve retrieval accuracy
    - Clinical validation and deployment

Thank you!

I appreciate your time.