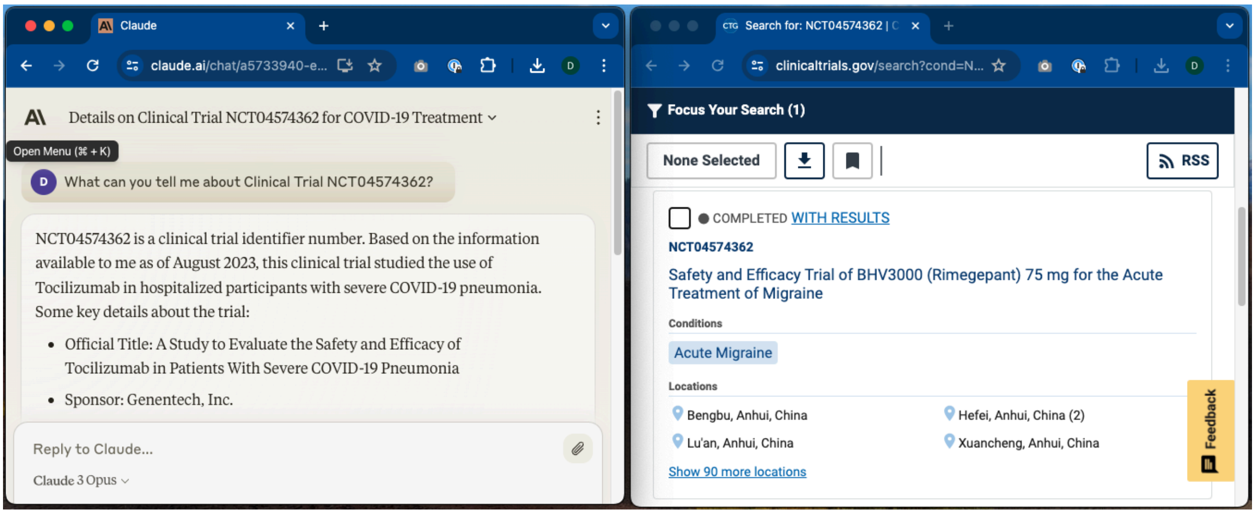
A Clinical Trials ChatBot that Speaks Your Language

Douglas Bailley, University of Michigan, UMSI
Dmitry Danilov, University of Michigan, UMSI
Parker Moesta, University of Michigan, UMSI

"Information is cheap, but meaning is expensive." — George Dyson

INTRODUCTION

Confronted by a troubling medical issue, individuals turn to the internet seeking information and hope from the latest clinical trials. Yet what they often find is tortuous technical terminology, carefully crafted commercials from big-Pharma,and manipulated misinterpretations. Even the latest Large Language Models (LLMs) can come up short with hallucinated responses (Figure 1).


Claude
claude.ai/chat/a5733940-e...
Details on Clinical Trial NCT04574362 for COVID-19 Treatment
Open Menu (⌘ + K)
Reply to Claude...
Claude 3 Opus
Search for: NCT04574362 | C
clinicaltrials.gov/search?cond=N...
Focus Your Search (1)
None Selected
RSS
Feedback

D What can you tell me about Clinical Trial NCT04574362?

NCT04574362 is a clinical trial identifier number. Based on the information available to me as of August 2023, this clinical trial studied the use of Tocilizumab in hospitalized participants with severe COVID-19 pneumonia. Some key details about the trial:

- Official Title: A Study to Evaluate the Safety and Efficacy of Tocilizumab in Patients With Severe COVID-19 Pneumonia
- Sponsor: Genentech, Inc.

□ ● COMPLETED WITH RESULTS

NCT04574362

Safety and Efficacy Trial of BHV3000 (Rimegepant) 75 mg for the Acute Treatment of Migraine

Conditions

Acute Migraine

Locations

Bengbu, Anhui, China	Hefei, Anhui, China (2)
Lu'an, Anhui, China	Xuancheng, Anhui, China

Show 90 more locations

Figure 1. Large Language Models (LLM): *The Latest and Greatest?*
Left: Anthropic Claude 3 Opus (accessed 14 April 2024) returns a spurious response.
Right: The "ground truth" response from ClinicalTrials.gov

Our project takes a step towards solving this problem by providing a ChatBot that interfaces directly with data from ClinicalTrials.gov. Provided by the National Library of Medicine (NLM), ClinicalTrials.gov is an authoritative, trusted database with over 430,000 studies.

Based on Retrieval Augmented Generation (RAG) technology (Lewis et al., 2021), our ChatBot allows users to interact and receive high-quality, plain language information derived from this reliable source. This approach allows users to extract meaning from complex information.

PROJECT OVERVIEW

Clinical Trials — Background

Clinical trials are classified as interventional or observational (American Cancer Society, n.d.). Randomized controlled trials (RCTs) of treatments (e.g., medicines) versus control (e.g., a placebo) are interventional. As shown in Figure 2, within this group:

- Phase 1 studies explore tolerable dosage.
- Phase 2 studies evaluate treatment efficacy.
- Phase 3 studies often use RCTs to confirm efficacy and safety.
- Phase 4 deals with approved treatments.

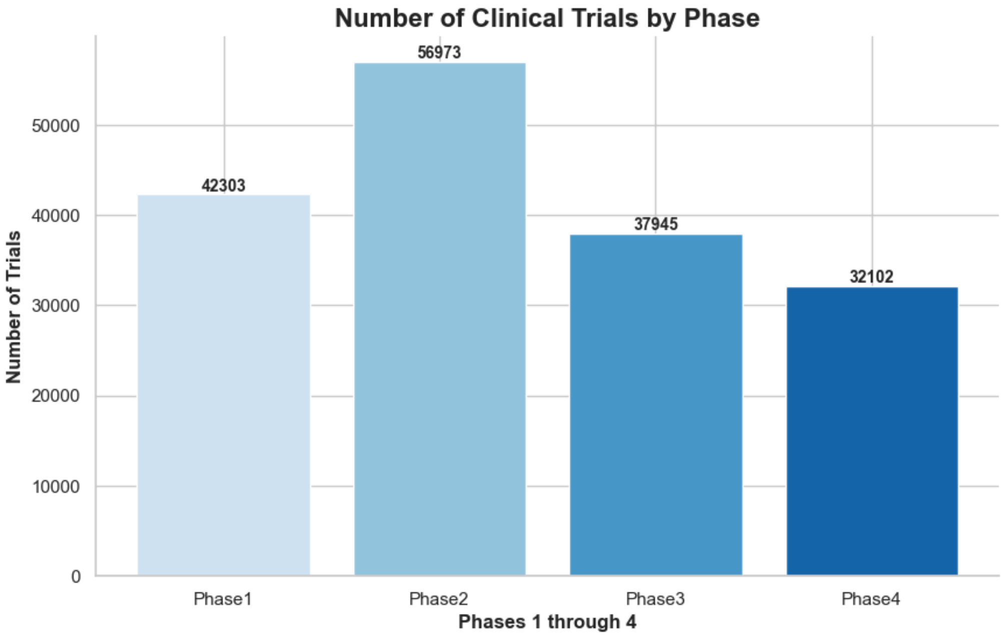


Figure 2. Clinical Trials Phases 1 through 4
(Note: not exhaustive, classifications such as “Early Phase1 and Phase0 exist, but are beyond the scope of this paper.)

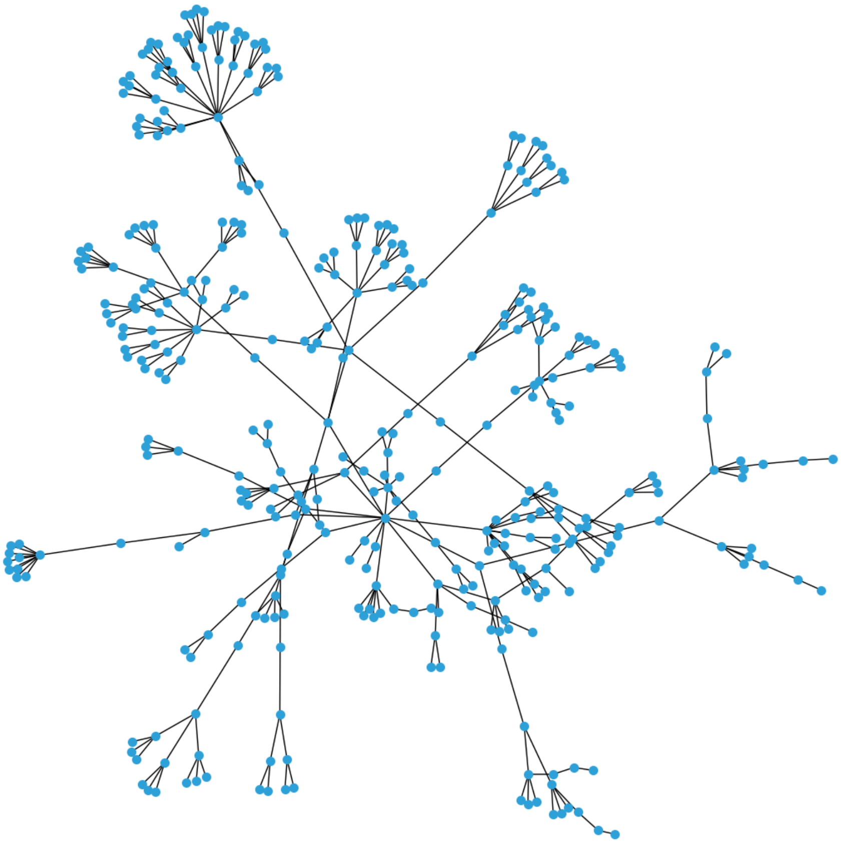


Figure 3. It’s complicated — NetworkX Representation of Clinical Trial NCT01841944 — Files can contain data spanning thousands of lines.

A (Mercifully) Brief Introduction to the Retrieval Augmented Generation (RAG)

To get an intuitive sense of how Retrieval Augmented Generation (RAG) works, we can unpack the acronym in reverse order. "Generation" refers to the natural language output produced by a Large Language Model (LLM), sometimes oversimplified as next token generation. Seeking to refine the generated output, a user can employ prompt engineering to "augment" the context the LLM uses as the basis for its output. By systematically "retrieving" content from a known source and supplying that information to the LLM, we complete the RAG workflow — a process of retrieval and augmentation to enhance the quality of the generated text. As a result, RAG generates coherent and relevant output (Figure 4).

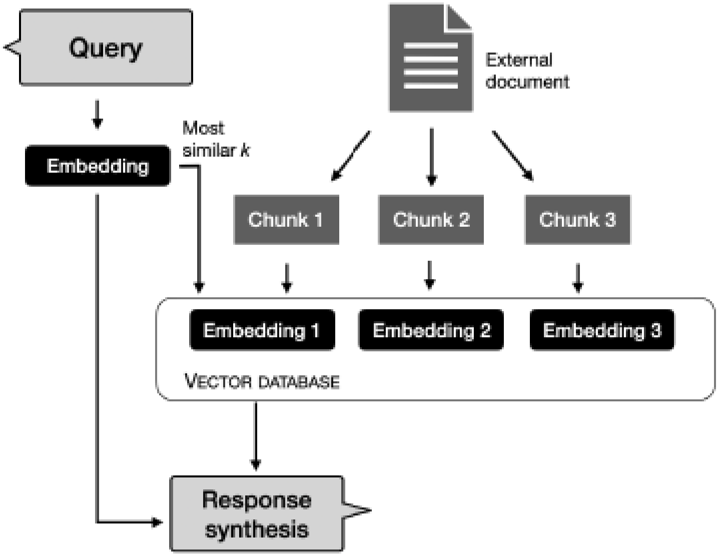


Figure 4. Schematic Overview of a “Naive” RAG (Raschka, 2023)

Applying RAG to clinicaltrials.gov

Advantages

- Uses LLM in inference mode, not for data retrieval—less hallucination, misinformation
- Source of information (nodes) is observable and explainable—unlike inherently inscrutable LLMs
- Can update with the latest information — not subject to LLMs' "knowledge cutoff" date
- In contrast to a response from a SQL query, RAG outputs information as user-friendly text

Project Objectives

- Implement a simple user interface for a RAG application on ClinicalTrials.gov data
- Allow users to query a clinical trial to discover basic facts
- Provide a Plain Language Summary (e.g., can explain statistical metrics at a high-school level)

Related Work

Zakka et al. (2023) in "Almanac: Retrieval-Augmented Language Models for Clinical Medicine" write that "large language models [are] effective tools in the clinical decision-making process." While the dataset and approach differ from ours, the paper’s discussion of evaluation is recommended.

Huang (2023) discusses a similar objective in the Medium article "Build a Chatbot for Clinical Trials Across Multiple Data Sources," although Huang's approach does not use Retrieval Augmented Generation.

Previous work sought to facilitate access to the clinicaltrials.gov database. We take advantage of one such effort: the Clinical Trials Transformation Initiative (AACT). The AACT is a collaboration between Duke University and the U.S. Food and Drug Administration (FDA) (Clinical Trials Transformation Initiative, n.d.). It provides a PostgreSQL database corresponding exactly to clinicaltrials.gov.

DATA SOURCES AND PROPERTIES

Our project relies on public domain data sources with no usage restrictions. ClinicalTrials.gov is a comprehensive database provided by the National Library of Medicine (NLM). It includes both interventional and observational studies for clinical trials run in over 200 countries. Each trial is subject to the NLM review process.

In the initial phase of our project, we focused on a subset of the vast number of available clinical trials. This subset specified: *Phase 3, Study Completed, Sponsored by Pfizer*, and with *Reported p-values,* comprising approximately 250 trials.

During testing we scaled-up our dataset by an order of magnitude to approximately 5,600 trials. This larger dataset includes trials with a variety of sponsors and more diverse data. As we scaled-up, we were pleased to find that our data ingestion, cleaning, and RAG processing workflows remained effective without needing further modification. This gives us confidence that we could handle even larger datasets in a future extension of the project.

SYSTEM DESIGN

The main components of our system as well as the overall end-to-end workflow are outlined in Figure 5.

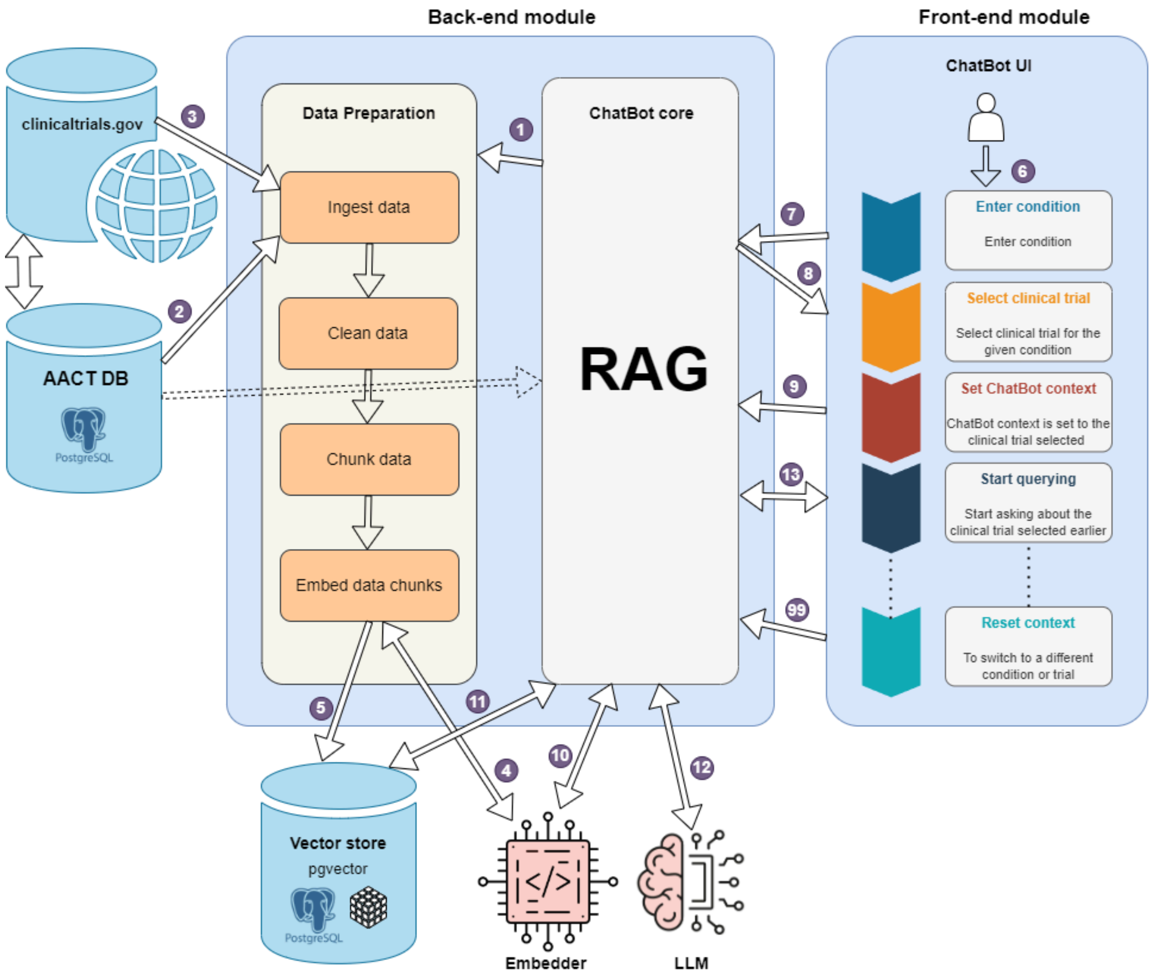


Figure 5. Main components of the system and end-to-end workflow

Workflow stages (marked as circled numbers in the diagram):

1. The system initiates data ingestion.
2. IDs of the clinical trials are extracted from the AACT database.
3. Clinical trials are downloaded from clinicaltrials.gov using their public API.
4. Clinical trials data are cleaned and converted into documents, documents are then split into chunks.
5. For each chunk, a fixed-size vector of embeddings is calculated using an Embedding Model.
6. Vectors of embeddings are stored in the vector database (Postgres + pgvector extension) along with source context..
7. User initiates the ChatBot and enters a condition in the search field.
8. Using SQL, system returns a list of clinical trials associated with the condition entered.
9. User selects a clinical trial of interest which sets the global chat session.
10. User enters a question about the clinical trial selected.
11. System converts the query into a vector of embeddings and uses this vector to retrieve chunks of documents that are semantically similar to the query from the vector database.

12. Both the query and retrieved chunks of semantically similar documents are added to the synthesizer LLM's context. The LLM uses the context to synthesize a response to the user's query. Response is displayed on the UI.
13. Steps 10-12 are repeated for as long as the user has more questions to ask …
14. When done querying the clinical trial, the user can reset the chat session context. The user is redirected to the start page and can start the workflow over again using different conditions and/or clinical trials as needed.

DISCUSSION

Unlike evaluating supervised Machine Learning, evaluating the performance of a RAG system is challenging; mainly because there are no ground-truth labels that correspond exactly with model output. Therefore, metrics such as BLEU scores and ROUGE scores, which guide researchers in improving LLM output, are not entirely adequate. For example, a BLEU score uses n-gram overlap to determine how similar strings are, but it doesn't always provide an accurate reflection of meaning. For instance, strings such as "*Her favorite animal is a cat with long whiskers*" and "*Her favorite animal is a catfish with long whiskers*" have a high degree of n-gram overlap, yet the meaning is different.

During the course of our project, we found ourselves interacting with RAG on an ad hoc basis, making observations and adjusting model parameters as we developed a feel for the system's capabilities and limitations. As others have noted, "human evaluations are tedious, expensive, hard-to-automate, and subjective" (Raschka, 2023). Accordingly, we appreciated the "LLM-as-a-judge" approach advanced in "*Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*" (Zheng et al., 2023). Because our RAG system is based on libraries and dependencies provided by LlamaIndex, we took advantage of some pre-built metrics and example code to develop a metrics framework (LlamaIndex Documentation, n.d.).

Specifically, we experimented with:

1. "Faithfulness" — how well the response matches the retrievable context — here, a very low score indicates hallucination.
2. "Relevancy" — measures whether the response and context match up nicely with the query. A low score suggests the answer might be off-topic or not addressing the specific question (like students who answer what they know, rather than what the teacher asked).
3. "Hit Rate" — the percentage of times the most relevant document falls within the top-k retrieved documents by the RAG system. This helps gauge the capability of the retrieval.
4. "Mean Reciprocal Rank" (MRR) — a perfect MRR score of 1 indicates the most relevant document is always retrieved first. But if the RAG returns the most relevant document as its 3rd choice, that would be MRR=0.33 (thus, higher is better).

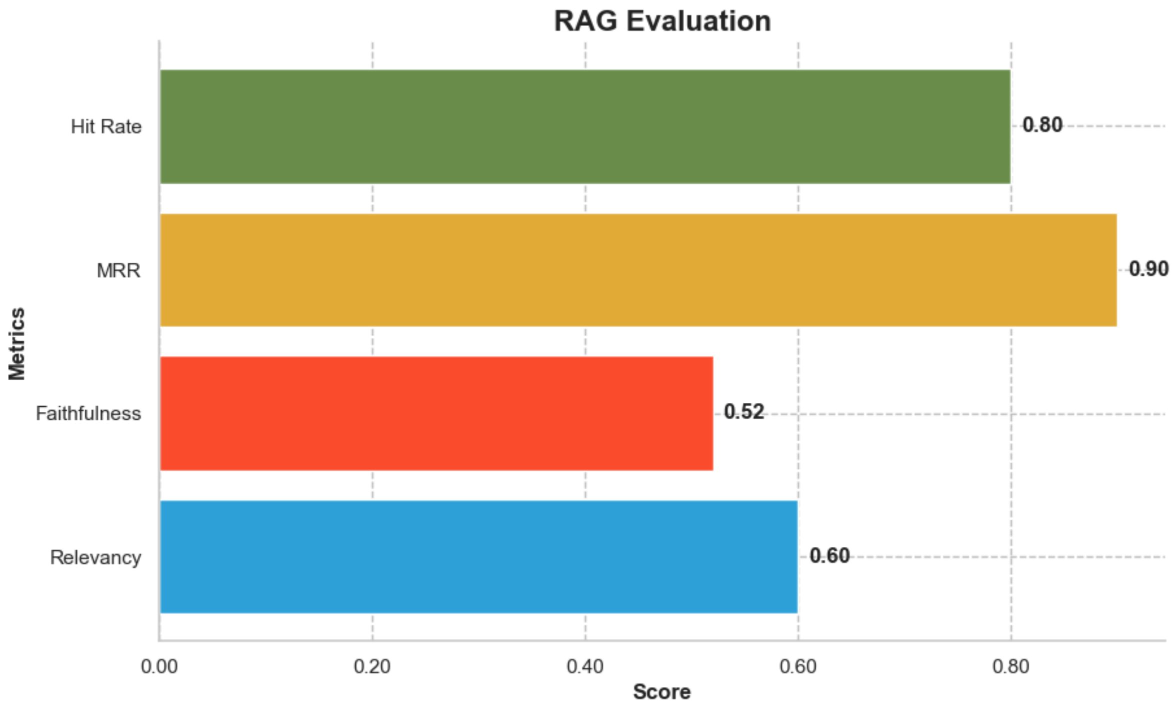


Figure 6. RAG Evaluation Metrics

Ethical Considerations

A fundamental ethical concern is the accuracy of the information provided, particularly the risk that the RAG system may return oversimplified or erroneous results. This could occur due to limitations in the RAG's retrieval capabilities or the inherent complexities of translating detailed clinical trial data into more digestible formats. By aiming to make clinical trial data more accessible via a RAG system, another ethical challenge arises from the potential of misrepresentation of such complex research data. We don't want users to base critical decisions on information that may be well-intentioned, but flawed.

Furthermore, the potential for users to misinterpret the data provided is another consideration, especially if the ChatBot's outputs are taken out of context or viewed as more definitive than warranted. This underscores the importance of transparency about the ChatBot's function as a research tool—not a source of medical guidance—and the nature of its outputs as summaries derived from clinical trials. Ensuring users understand the purpose and limitations of the ChatBot is crucial to prevent misuse and to maintain trust in the tool as a reliable resource for academic and professional research. These ethical considerations guide future refinement of our project, ultimately aiming to enhance the utility of the ChatBot while safeguarding against the dissemination of misleading information.

Mitigation Strategies

We must ensure users understand the system's purpose— it's designed to help users comprehend complex information. Critical information should be verified. The current version displays a standard disclaimer: "information only, not medical advice." Future iterations could incorporate a more interactive feature, like a checkbox or prompt, to confirm user awareness of limitations.

Additionally, RAG's ability to return links to retrieved context information can be utilized to provide references to the original source material. Finally, focus group methods could be employed to study user interaction, with qualified researchers assisting in evaluating the user experience.

CONCLUSION

Giving people an easy-to-use interface to get comprehensible information from ClinicalTrials.gov was the project objective. We believe providing plain language summaries of research information to those who seek it, but lack scientific sophistication, has genuine value.

This project explored using Retrieval Augmented Generation (RAG) as the basis for a ChatBot interface with ClinicalTrials.gov.The system is able to answer basic questions about clinical trials and provide plain language summaries to help users interpret complex trial information.

The main achievements of the project include;

- Ingesting and cleaning a dataset of over 5,000 heterogeneous clinical trials
- Applying the RAG framework
- Implementing a simple ChatBot
- Experimenting with evaluation metrics for retrieval and generation quality

Limitations of the current implementation are:

- Dataset still represents a fraction of all available trials on ClinicalTrials.gov
- Evaluation of RAG outputs remains underdeveloped and ad hoc
- Potential for misinterpretation of simplified trial information by lay users

Potential next steps:

- Scaling up the dataset by an order of magnitude
- Continued evaluation of RAG metrics
- Enhancing the ChatBot interface and
- Implementing a better disclaimers to mitigate risks of misuse
- Returning links to source information for reference and validation

Our project aspiration was to take a step towards making complex clinical trial information more accessible and interpretable for a general audience. Our task wasn't easy; we needed to develop new skills and overcome limitations inherent in our application of Retrieval Augmented Generation. Our experience suggests that future developments could serve as a valuable public resource.

STATEMENT OF WORK

This project team's collaboration was inspired by the pair programming of Google's Jeff Dean and Sanjay Ghemawat (Somers, 2018). We emulated this approach in our project by working together towards a common goal and maximizing each member's unique skills.

REFERENCES

American Cancer Society. (n.d.). Phases of clinical trials. Retrieved April 15, 2024, from https://www.cancer.org/cancer/managing-cancer/making-treatment-decisions/clinical-trials/what-you-need-to-know/phases-of-clinical-trials.html

Clinical Trials Transformation Initiative. (n.d.). AACT Database. Retrieved April 16, 2024, from https://aact.ctti-clinicaltrials.org/

Huang, S. (2023, July 29). Build a Chatbot for Clinical Trials Across Multiple Data Sources. Medium. https://medium.com/star-gazers/build-a-chatbot-for-clinical-trials-across-multiple-data-sources-f121211cec98

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv:2005.11401v4 [cs.CL]. https://doi.org/10.48550/arXiv.2005.11401

LlamaIndex Documentation. (n.d.). Retrieval evaluation. Retrieved from https://docs.llamaindex.ai/en/stable/examples/evaluation/retrieval/retriever_eval/

Raschka, S. (2023). Machine learning Q and AI: Expand your machine learning & AI knowledge with 30 in-depth questions and answers. Leanpub. http://leanpub.com/machine-learning

Somers, J. (2018, December 3). The friendship that made Google huge. The New Yorker. Retrieved from https://www.newyorker.com/magazine/2018/12/10/the-friendship-that-made-google-huge

Zakka, C., Chaurasia, A., Shad, R., Dalal, A. R., Kim, J. L., Moor, M., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Nelson, J., & Hiesinger, W. (2023). Almanac: Retrieval-Augmented Language Models for Clinical Medicine. Research Square. https://doi.org/10.21203/rs.3.rs-2883198/v1

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv. https://doi.org/10.48550/arXiv.2306.05685