https://github.com/gordonwatts/snowmass-chat

Retrieval Augmented Generation for Particle Physics: A Case Study with the Snowmass White Papers and Reports

Benjamin Galewsky
Gordon Watts

Why an LLM for Snowmass?

- There are **642 papers** including all the final reports **1000's of pages**
- Work over greater than **2 years** by many **hundreds of US and world physicists!**

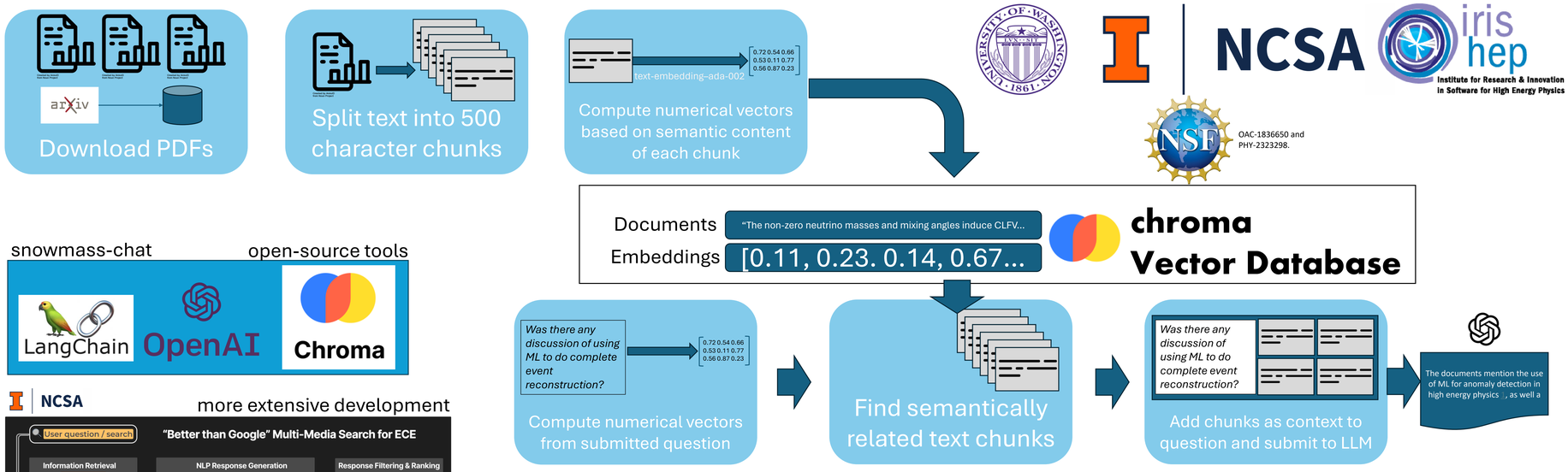
Can Machine Learning and the recently trained Large Language Models (LLM) help **make this corpus of work accessible**?

What is RAG?

- ***Retrieval Augmented Generation***
- A rudimentary **search engine combined with a LLM** to summarize the search results
    1. **Text is extracted** from downloaded PDF's
    2. **Text is split** into convenient sizes
    3. ML techniques are used to **assign a distance vector** to each chunk of text
    4. The vector **distance between the chunks of text and any question** are used to select the ~6 closest chunks
    5. The LLM is asked to **answer the question using the 6 closest chunks**.
- RAG is popular because it is ***easy*** to implement.
- Open-Source Libraries are available that take care of most of the work (e.g. langchain).
- Make it possible to evaluate various configurations and experiment
- We had to good fortune to link with Kasten Day at NCSA who had developed a university app prior to the open-source tools being available.

What Matters?

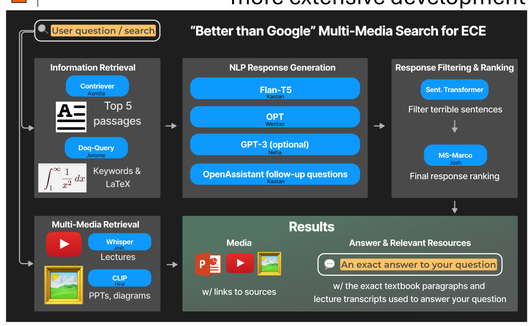
- **The LLM**: difference between OpenAI's gpt-3.5-turbo and gpt-4-turbo-preview is huge in getting a coherent answer
- **A modern semantic embedding model** is crucial in finding the most relevant text chunks. OpenAI's new models made almost as large a difference in the quality of the answer as the LLM.
- Chunk size, number of chunks, etc. seemed to have much smaller effects
- There are many variations we did not explore (summarizing, prompt compression, etc.). As LLM's get better these aren't as necessary.



OAC-1836650 and PHY-2323298.

snowmass-chat open-source tools

NCSA more extensive development



Successful?

- Gets the job done
- UIUC has references, which are crucial
- But good embedding and GPT-4 are quite good!
- Very affordable for personal use (less than $10 was spent on this project)

BUT

- NCSA work taught us: **we should not write this ourselves**
- Many commercial services do exactly what we want
- Anything with some flexibility will track industry best practices
- **Leaving us to physics and distributed computing** and **use this**!

Ethics

- The arXiv has restrictions
    - Do not re-serve documents
    - Rate limits for downloading papers and meta data
    - Do encourage anything which provides access into their database
- Some members of our community are uncomfortable with the use of LLM's
    - Ethical training data, power consumption, implications for the job market, hallucinations, etc.
- Costs are not zero
    - OpenAI charges for the use of their LLM's and embedding services
    - Note the cost is small for this project!
- We have made the source code public
- We have not made the project accessible

What does the MATHUSLA Detector Do?

Where are the Results?

- Using some fraction of the papers for more than personal use **may** violate the arXiv license.
- What licenses are there?
    - The arXiv has two main types licenses.
    - Create Commons license – includes the ability to "remix" the results – and would be just fine with this technique: "… distribute, remix, adapt… "
    - The standard arXiv license seems to say no: "…limits re-use of any type from other entities or individuals." Question: Is storing it on your own computer (Mac or Windows which have built in search) allowable?
    - Note that metadata (which includes authors and abstracts) is all under creative commons.
- What was the explicit violation?
    - Extracting the full text from the PDF might violate the "re-use of any type".
- However, the arXiv pages are ambiguous – they explicitly talk of bulk download of full text for scanning and building indices. Which is step one of what is done in this project.

- **How did we stumble on this**?
    - Some members of the Snowmass community wrote to us to let us know they "did not authorize their papers to be used as data"
    - Discussion on the Snowmass Slack #general channel showed that is was more than one person that was uncomfortable with this use of papers they were an author on.
    - Without understanding the arXiv licenses, copyright fair use, and the time before the conference, we decided present the results of sample searches.
    - Most important lessons are in the **What Matters** section to the left.
- What is next?
    - We are contacting the arXiv maintainers to understand what they think of as "ok" use of the full text.
    - **ATLAS and CMS publish their papers under a CC license**. This is a fall back for the paper version of this ACAT poster.