

## Abstract

Large Language Models (LLMs) are powerful tools for processing and extracting information. The PROTECT Center has been studying preterm birth rates in Puerto Rico for over a decade, publishing more than 300 research papers and collecting raw data on over 2000 women. Papers written by PROTECT researchers utilize information from the dataset, but does not cite which variables they used in the data dictionary. Our project employs LLMs to link the papers and the data dictionary variables, thereby enhancing reproducible research. Our baseline approach involved giving each paper to GPT-4 with the data dictionary and asking for the variables, but this led to context window overflow, hallucinations, and irrelevant results. To address the issues, we developed a customized Retrieval Augmented Generation (RAG) pipeline that uses an LLM to identify variables in the papers without the data dictionary and matches the LLM's output to the dictionary variables using cosine similarity on the embeddings. We used the LLM-as-a-judge method to evaluate the second step of our preprocessing.

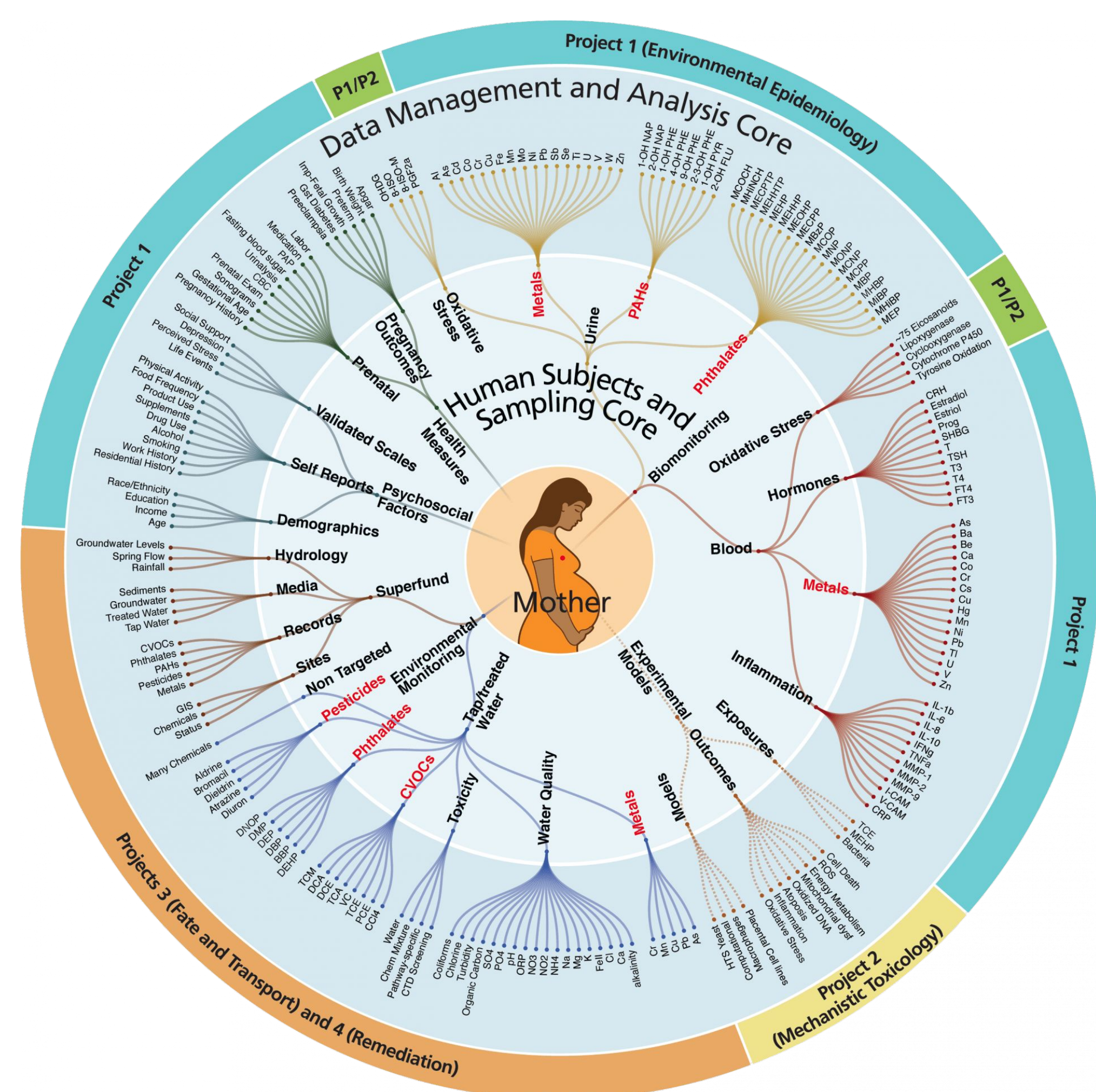
## Background

### Large Language Models (LLMs)

- Designed to predict the next word of a text
- Trained on large amounts of text
- Type of machine learning model and use neural networks
- Useful in a variety of fields, especially for processing text data

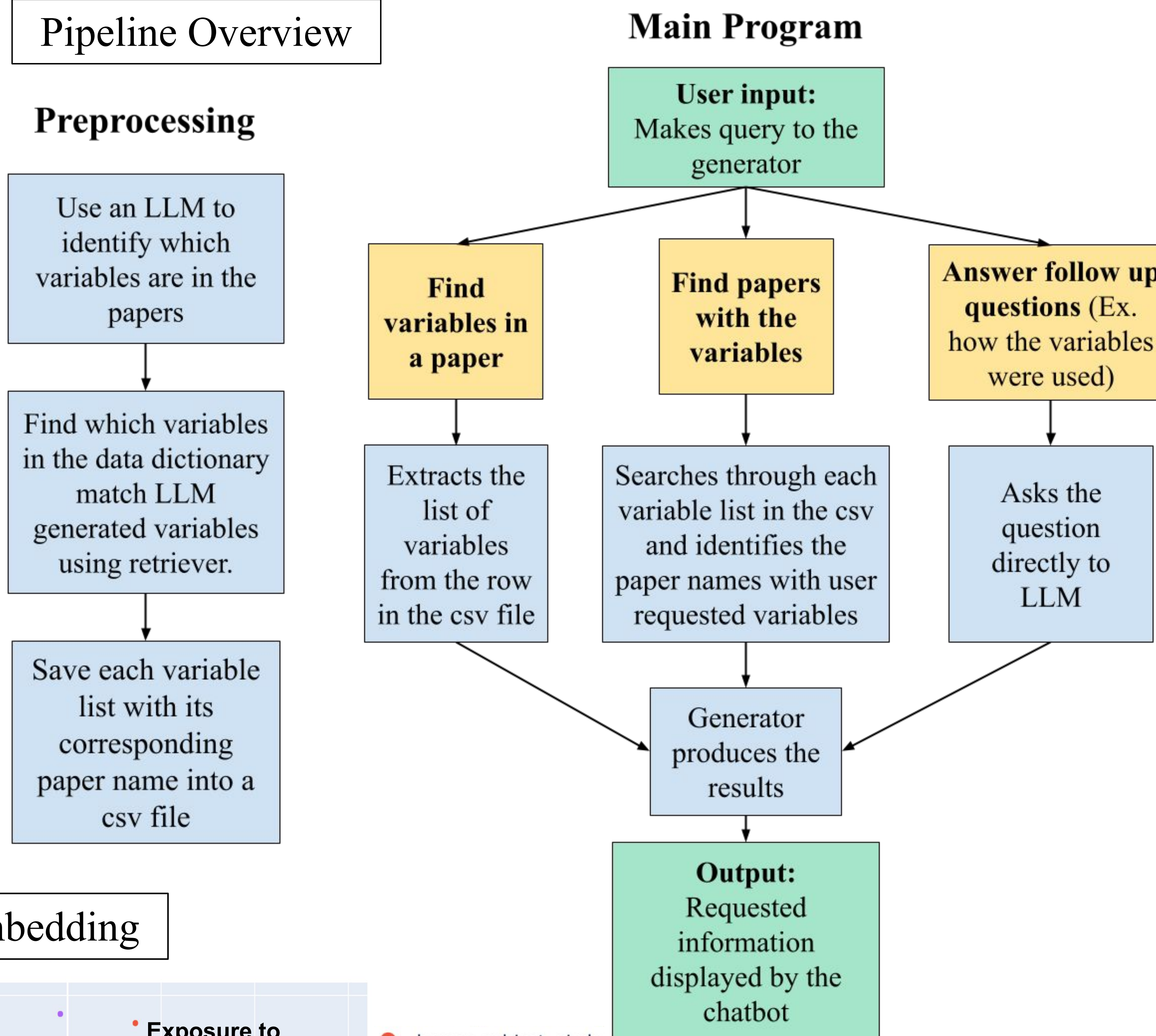
### PROTECT Data Dictionary

- Contains metadata about data collected by PROTECT
  - Variable names, descriptions, options, units, etc.



Data Dictionary Visualization

## Methods



### Vector Embedding



- Vector Embedding
  - LLMs can't do math with words
  - Use embedding model to capture semantic meaning of text in high-dimensional vectors
  - Embedded dictionary variable names and descriptions

### Cosine Similarity

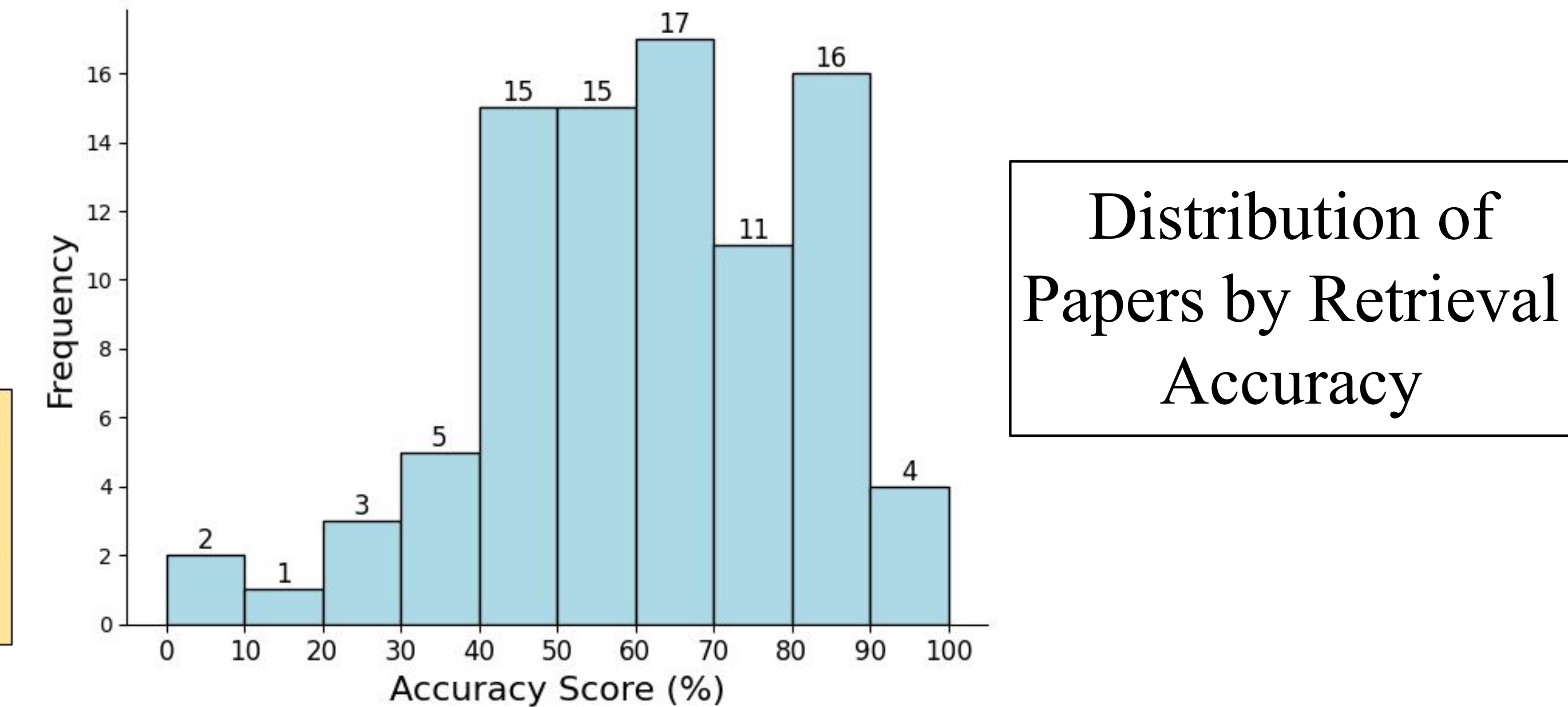
$$A \cdot B = |A||B|\cos(\theta)$$

$$A \cdot B = \sum_{i=1}^n A_i B_i$$

$$\cos(\theta) = \frac{\sum_{i=1}^n A_i B_i}{|A||B|}$$

- Cosine Similarity
  - Calculate the angle between two vectors of the same number of dimensions
  - Compared vector embeddings of dictionary descriptions to variables produced in first preprocessing step

## Results



- Evaluated the retrieval component using LLM-as-a-judge
  - Average accuracy of 71.58%

## Conclusions and Next Steps

### Successes

- Biological Data
- Human subject data
  - Correctly identifies weight, height and other information from specific visits

### Areas for improvement

- Environmental data
  - LLM provides too general data
- Human subject data
  - Fails to distinguish related terms
- Runtime

### Next Steps

- Improve on initial prompt to LLM with in-context learning
- Integrate with existing PROTECT RAG Chat
- Improve evaluation

## References

1. Hugging Face. (n.d.). *Understanding NLP and LLMs* (Chapter 1, Section 1). In *The Hugging Face LLM Course*. Retrieved July 22, 2025, from <https://huggingface.co/learn/llm-course/chapter1/1>
2. Jurafsky, D., & Martin, J. H. (2025, January 12). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed., online manuscript). Stanford University. <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>
3. Sentence-Transformers team. (2023). *multi-qa-mpnet-base-dot-v1* [Pre-trained sentence embedding model]. Hugging Face. <https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1>
4. Tiahi, M., Kumar, M., Feric, Z., & Kaeli, D. (2024). *Natural language processing in environmental health research*. Proceedings of the 2024 IEEE MIT Undergraduate Research Technology Conference (URTC). <https://doi.org/10.1109/URTC65039.2024.10937629>

## Acknowledgements

### NUCAR Lab

David Kaeli  
 Zlatan Feric  
 Mouad Tiahi

### Center for STEM Education

Claire Duggan, Executive Director  
 Jennifer Love, Associate Director  
 D'mitra Mukasa, Victoria Berry, Ahmed Othman, YSP Coordinators  
 Nicolas Fuchs, Program Manager  
 Mary Howley, Administrative Officer