Natural Language Processing in Environmental Health Science

Zlatan Feric, *Computer Engineering Graduate Student*, *Northeastern University*
Mouad Tiahi, *Northeastern University*
David Kaeli, *COE Distinguished Professor*, *Northeastern University*

Northeastern University
College of Engineering

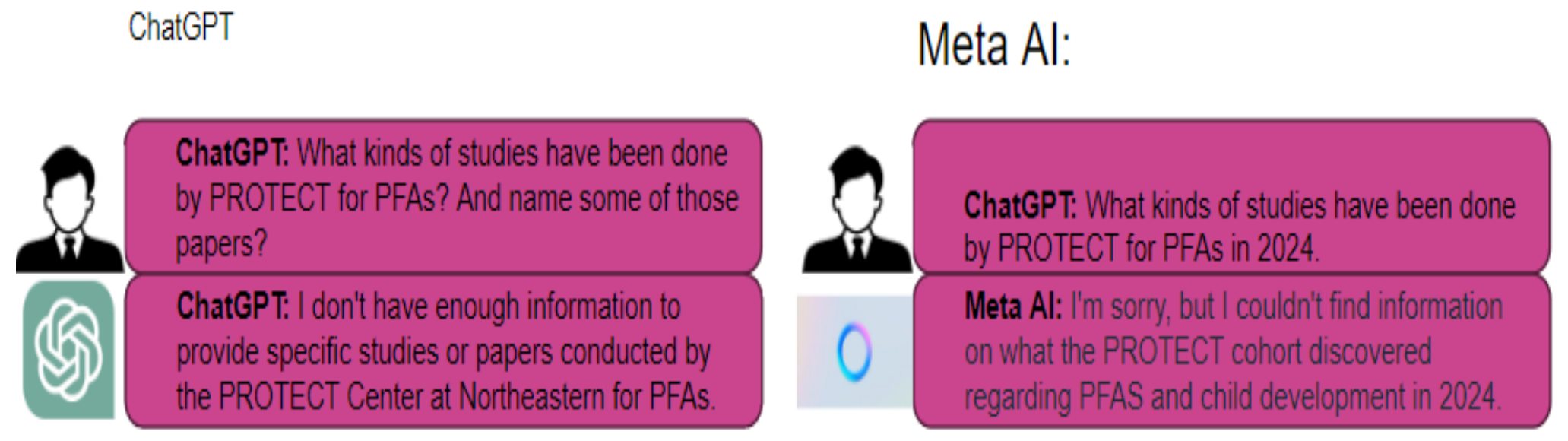
Northeastern University
Michael B. Silevitch and Claire J. Duggan Center for STEM Education

Abstract

- The PROTECT initiative at Northeastern University addresses environmental health issues, specifically the impact of contaminants on pregnancy and preterm birth in Puerto Rico, using a Large Language Model (LLM) chatbot to enhance community engagement and data collection.
- The chatbot, leveraging Natural Language Processing (NLP), interacts with community members to disseminate information about environmental risks, health advisories, and preventive measures while collecting qualitative data for researchers.
- The study indicates that integrating an LLM chatbot in environmental health initiatives significantly improves community awareness, engagement, and provides valuable insights for targeted public health interventions.
- **Tools Used:** Python, Jupyter Notebooks, Anaconda and PyTorch were utilized to design a Retrieval-Augmented Generation (RAG) model for PROTECT.

Motivation

Mainstream LLMs and GPTs have difficulty identifying specifics about the PROTECT research that has been done and with this project, we wanted to develop a new method to answer questions about the research and assist the 2000 researchers within PROTECT to have the capability for transmitting data between different researchers effectively and reliably



ChatGPT

ChatGPT: What kinds of studies have been done by PROTECT for PFAs? And name some of those papers?

ChatGPT: I don't have enough information to provide specific studies or papers conducted by the PROTECT Center at Northeastern for PFAs.

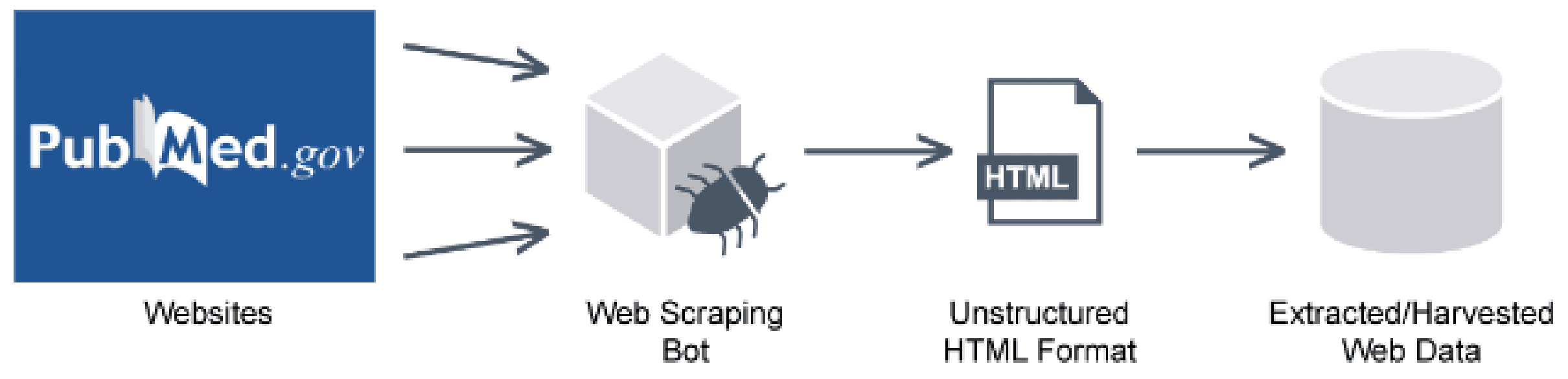
Meta AI:

ChatGPT: What kinds of studies have been done by PROTECT for PFAs in 2024.

Meta AI: I'm sorry, but I couldn't find information on what the PROTECT cohort discovered regarding PFAS and child development in 2024.

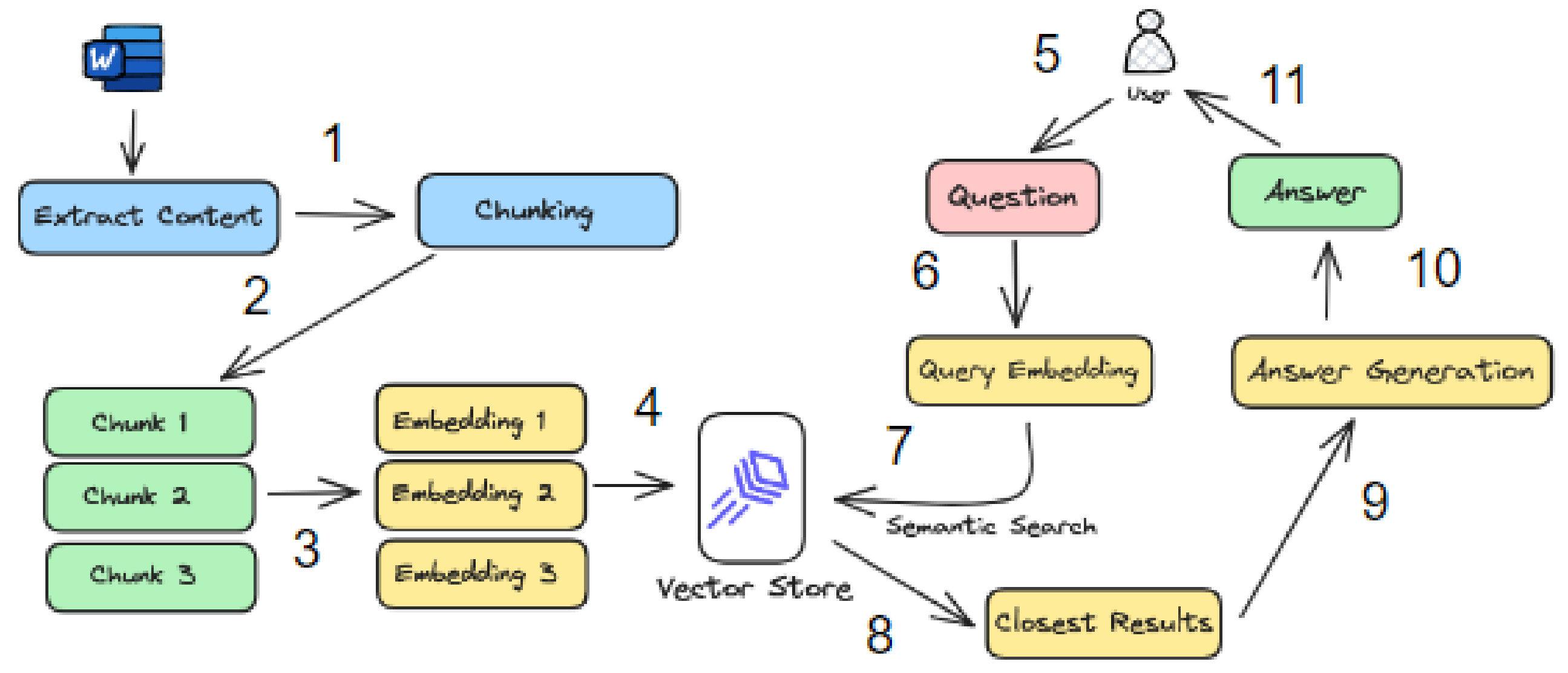
Methodology

- **Web Scraping and Data Preparation:** Developed Python scripts to scrape relevant environmental health data from PubMed PROTECT research. The collected data was then cleaned and organized using Pandas and stored in CSV format for further processing.



- **Embedding Comparisons:** Utilized transformer-based models to generate text embeddings of the collected data. These embeddings were compared to assess similarities and relevance to the PROTECT initiative's focus areas.
- **RAG Model Development:** Designed a Retrieval-Augmented Generation (RAG) model using PyTorch and Jupyter Notebooks. This model integrates the retrieval of relevant information with a generative component to provide accurate and contextually appropriate responses to user queries.

- Extract data and chunk it to prepare for model (1)
- Generated **embeddings (vector representations)** of queries and research papers (3)
- Implemented **retrieval component** which uses the user query (5) vector t retrieve relevant documents and pass it to the LLM. (9)
- **Passed** the user query and relevant papers to LLM (9) for interpretation and answer generation to arrive final answer (11)



- **Processing and Integration:** Implemented simultaneous execution of the generative and retrieval components of the RAG model to ensure real-time response generation. Text data was chunked into manageable segments to optimize the processing and retrieval efficiency, allowing the chatbot to handle large datasets effectively and provide precise and relevant answers.

Results

Question	ChatGPT Response	RAG Model Response
What kinds of studies have been done by PROTECT for PFAs? And name some of those papers.	I don't have enough information to provide specific studies or papers conducted by the PROTECT Center at Northeastern for PFAs.	PROTECT has conducted studies on per- and poly-fluoroalkyl substances (PFAS) in relation to various health outcomes. Some of the research papers published by PROTECT on PFAS include: 1. "Cross-sectional associations between prenatal maternal per- and poly-fluoroalkyl substances and bioactive lipids in three Environmental influences on Child Health Outcomes (ECHO) cohorts" by Himal Suthar et al., published on November 7, 2023. 2. "Racial and Ethnic Disparities in Phthalate Exposure and Preterm Birth: A Pooled Study of Sixteen U.S. Cohorts" by Irene Lafarga Previdi et al., published in December 2023. These studies contribute to the understanding of the health effects associated with PFAS exposure during pregnancy and highlight disparities in exposure levels among different racial and ethnic groups.
When did the PROTECT initiative start?	I'm not certain about the exact start date of the PROTECT initiative.	PROTECT (Puerto Rico Testsite for Exploring Contamination Threats) started on December 1, 2010.
Who is Akram Alshawabkeh?	Akram Alshawabkeh is a professor at Northeastern University known for his work in environmental engineering and remediation technologies. He has been involved with the PROTECT Center, which focuses on studying environmental contamination and its health effects, particularly in Puerto Rico.	Akram Alshawabkeh is a professor and researcher who has been involved in various studies related to environmental health and community engagement in emergencies, particularly in Puerto Rico. He has collaborated with other researchers on projects focusing on topics such as phthalate exposure and preterm birth, as well as trichloroethylene metabolites and their effects on energy metabolites and amino acids in human placental trophoblast models.

- RAG model responses were more in depth, thought out, specific and offered more insight than models such as ChatGPT, MetaAI and Gemini
- Processing speed matched these models and outperformed in some cases.
- RAG model accuracy can be improved upon by altering how much chunked data is processed at once and by fine tuning a custom LLM rather than relying on an API to utilize for text data.

Next Steps

- Develop UI within PROTECT website for easy, widespread accessibility to the model.
- Integrate agents for various search paradigms (websites, databases) and tools (search, functions, APIs, datasets) to expand the intelligence and dexterity of the model.
- Verify results with PROTECT experts and continue testing with various complex queries.

NSF REU-Site: Research Experience for Undergraduates

This work was supported by the National Science Foundation
Grant #2150417