Natural Language Discovery of NSF NCAR Scientific Data

Sofia Borukhovich

Mentors: Nathan Hook, Eric Nienhouse, Jason Cunning

NCAR CIT Credentials required

BACKGROUND

The NSF NCAR Research Data Archive (RDA) supports atmospheric and Earth system research spanning topics such as meteorology, atmospheric composition, and oceanographic observations, and model outputs.

Despite its scale and value, discovering relevant datasets can be challenging. Current tools often require users to understand and apply domain-specific keywords or metadata filters, which can pose a barrier to new or interdisciplinary users.

15.7 PB of data **49.6 million files**

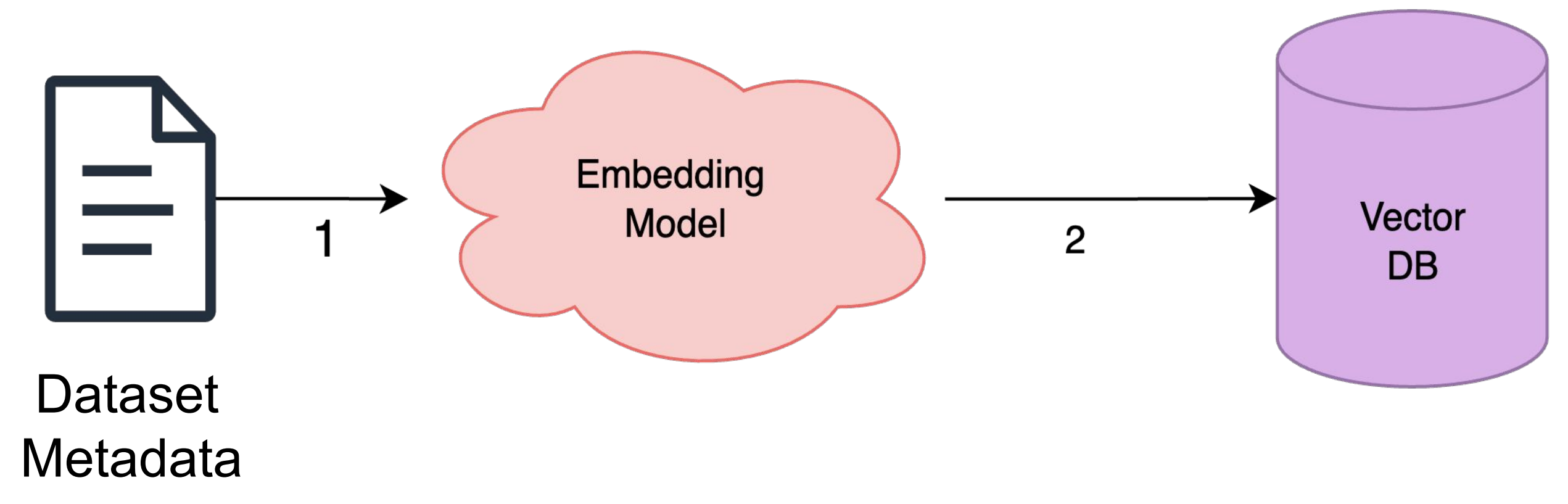
OBJECTIVES

Explore the use of natural language search allowing users to ask questions in plain English and receive dataset suggestions powered by language models.

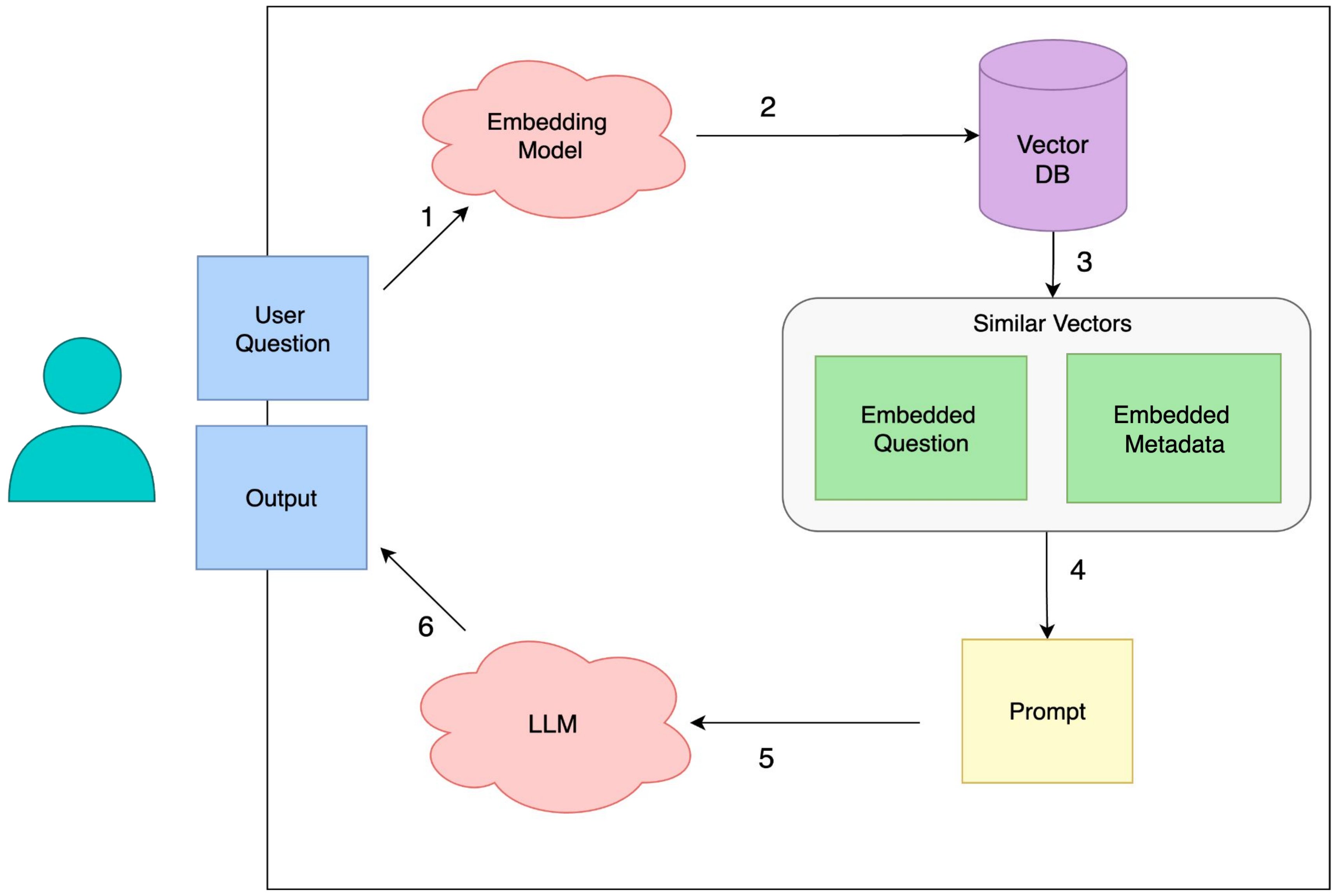
Goals:

- Reduce the technical barrier for novice users
- Empower more users to use RDA for research
- Lay the foundation for future improvements

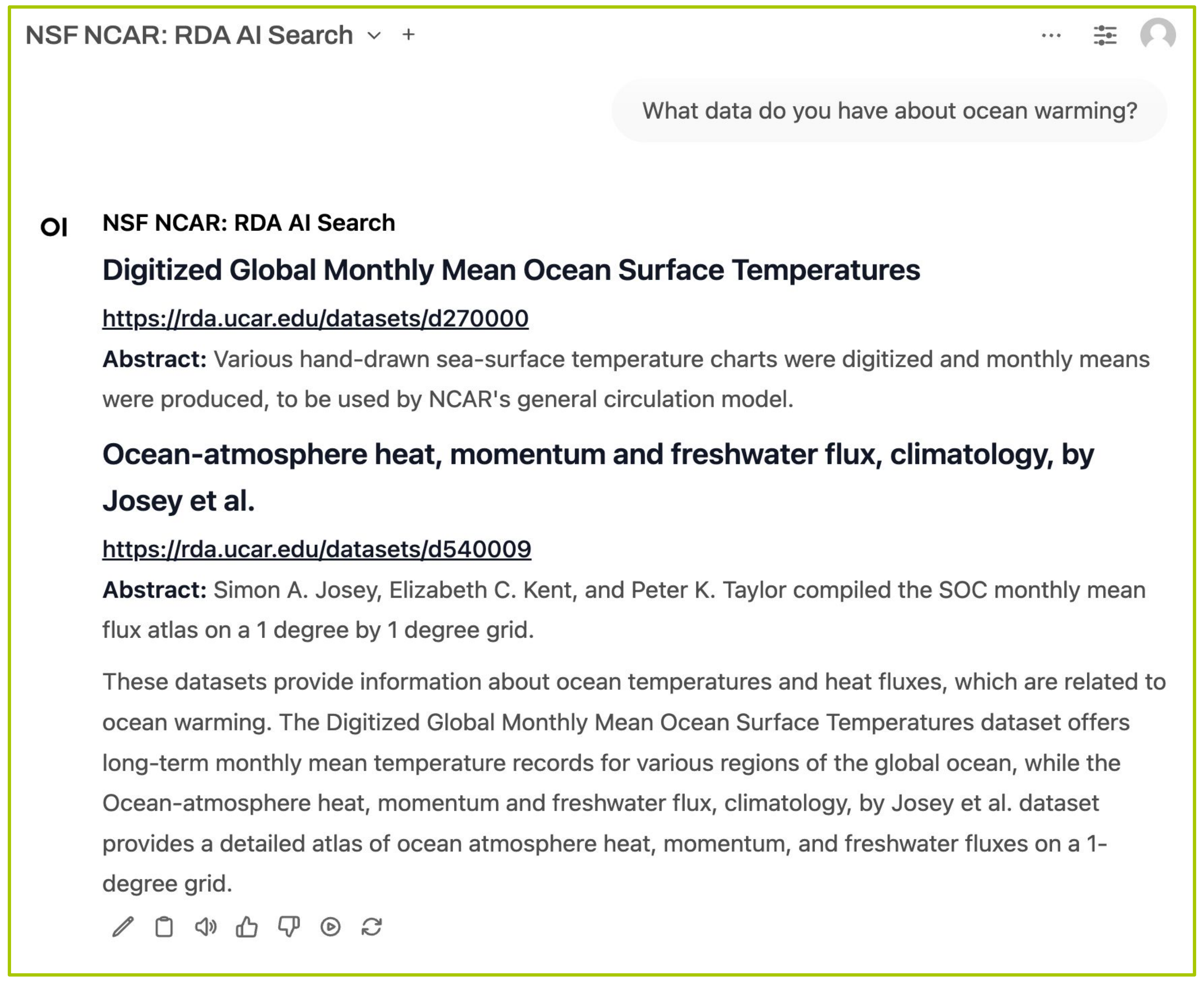
INITIALIZE WORKFLOW



QUESTION WORKFLOW



RESULTS



What data do you have about ocean warming?

NSF NCAR: RDA AI Search

Digitized Global Monthly Mean Ocean Surface Temperatures

https://rda.ucar.edu/datasets/d270000

Abstract: Various hand-drawn sea-surface temperature charts were digitized and monthly means were produced, to be used by NCAR's general circulation model.

Ocean-atmosphere heat, momentum and freshwater flux, climatology, by Josey et al.

https://rda.ucar.edu/datasets/d540009

Abstract: Simon A. Josey, Elizabeth C. Kent, and Peter K. Taylor compiled the SOC monthly mean flux atlas on a 1 degree by 1 degree grid.

These datasets provide information about ocean temperatures and heat fluxes, which are related to ocean warming. The Digitized Global Monthly Mean Ocean Surface Temperatures dataset offers long-term monthly mean temperature records for various regions of the global ocean, while the Ocean-atmosphere heat, momentum and freshwater flux, climatology, by Josey et al. dataset provides a detailed atlas of ocean atmosphere heat, momentum, and freshwater fluxes on a 1-degree grid.

FUTURE WORK

Expand Search Capabilities

- Include more metadata in embeddings:
  - Keywords
  - Location
  - Time range
- Expand beyond RDA

Learn from User Feedback

- Categorize feedback
  - "How do I download data?"
  - "How do I contact someone from the RDA?"
- Use AI agents to perform these actions

Refine AI

- Experiment with:
  - LLM Models
  - Embedding Models
  - Prompts
- Explore chunking strategies

Evaluate Performance

- Compare facet-based vs. natural language search
- Measure: accuracy, relevance, usability
- Automate Data Updates
  - Webhooks

ACKNOWLEDGMENTS

A sincere thank you to my mentors Nathan Hook, Eric Nienhouse, and Jason Cunning for their guidance, support, and encouragement throughout this summer.

I would like to also thank NSF NCAR and CISL for making this research opportunity possible. Special thanks to the Administration team for organizing the program, arranging housing and travel, and coordinating such valuable experience. I am also grateful to the IT support and staff at NCAR for their dedication in ensuring a smooth and productive environment throughout the internship.

This material is based upon work supported by the U.S. National Science Foundation National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1755088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation.