Natural Language Discovery of NSF NCAR Scientific Data

Sofia Borukhovich
SIParCS Intern

July 29, 2025

Table of Contents

Introduction

Proposal

RAG

Workflow

Results

Future Work

Acknowledgments

Background

NSF NCAR & Research Data Archive
NSF NCAR produces and hosts vast amounts of scientific data across domains like atmospheric science, climate modeling, oceanography, and geoscience.

49.6 million files

The Challenge
Diversity and volume of data can make it challenging for users to locate what they need, especially without prior knowledge of terminology.

15.7 Petabytes

How can we help people find the right data, even if they're not familiar with NCAR's terminology or structure?

Proposal

Experiment
Natural-language search

Natural-language search allows

Search in plain English

Express complex ideas simply

Discover related data

Goal
Build a prototype that uses natural language search

RAG

Retrieval-augmented generation

- Advanced technique used in LLMs
- Enhances LLM capabilities
- Allows to retrieve relevant information

Why use it?

- Scalable and dynamic
- Relevant response

Metadata

Format

Keywords

Title

Publication date

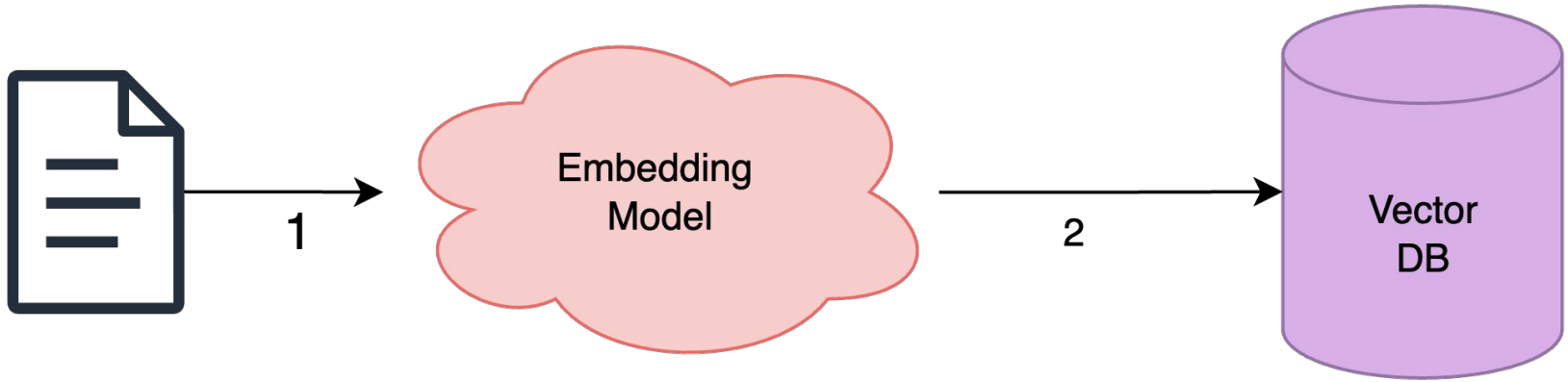
Abstract

Author

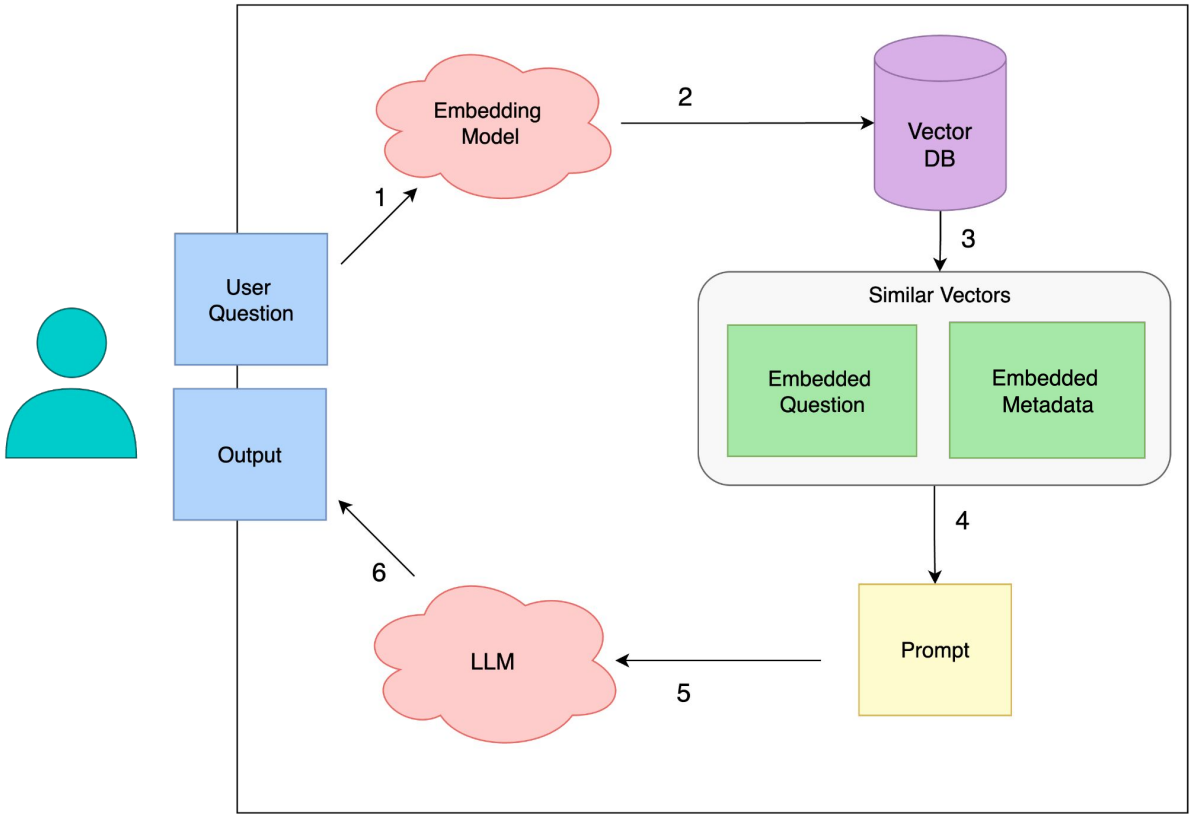
Temporal coverage

Geospatial data

Initialization Workflow



Question Workflow



Example Queries

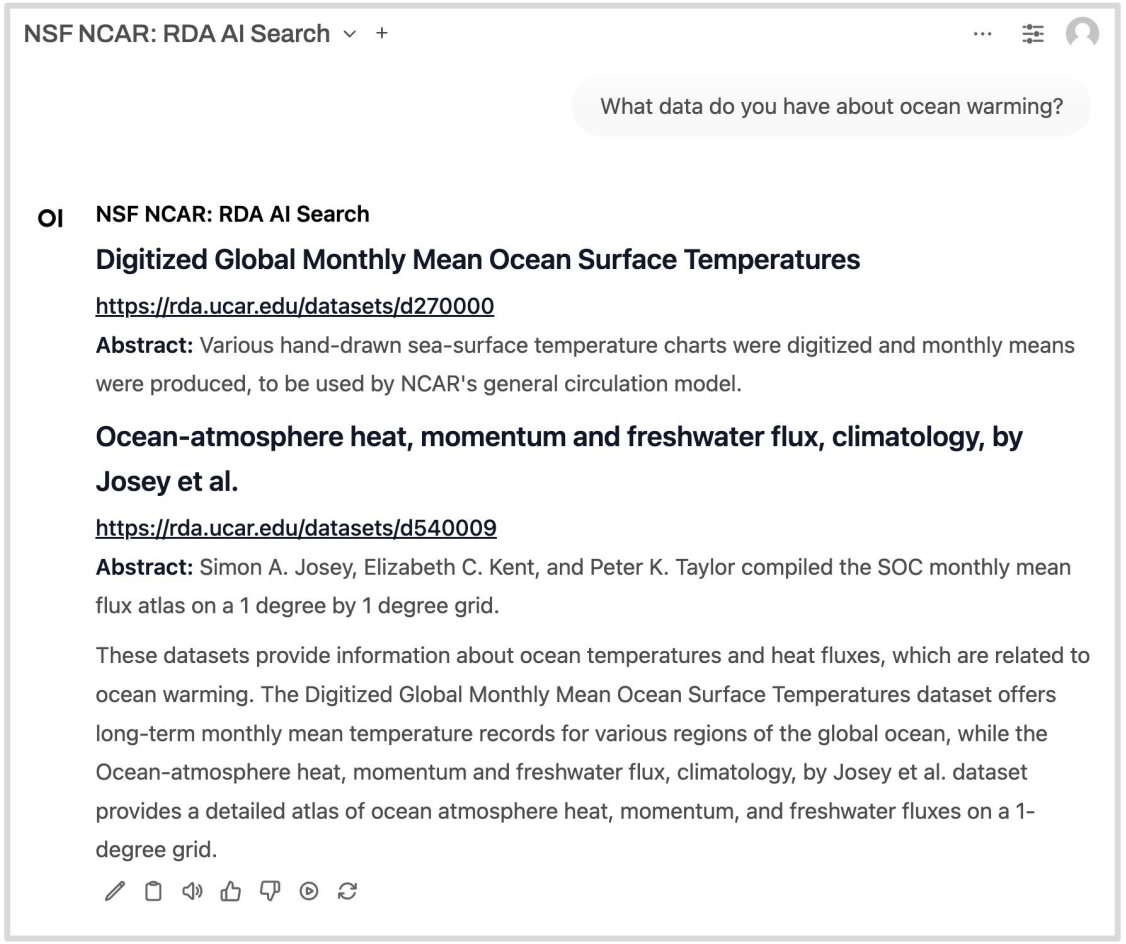
What long-term climate datasets do you have for North America?

Search for datasets related to soil science or agriculture.

I am a skier and I am interested in snowfall data.

What data do you have about ocean warming?

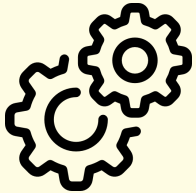
Example Queries & Results



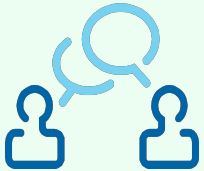
Future Work

Expand Search Capabilities

Refine AI



Learn from User Feedback



Evaluate Performance

Future Work

Expand Search Capabilities

- Add more metadata to embeddings:
  - Keywords
  - Geographic location
  - Time ranges
- Include other datasets beyond RDA

Refine AI

- Experiment with:
  - LLM Models
  - Embedding Models
- Fine-tune prompt
- Explore chunking strategies

Future Work

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

Evaluate and Optimize Performance

- Compare facet-based vs. natural language search
- Accuracy, relevance, usability
- Reduce manual maintenance
- Automate Data Updates
  - Webhooks

ACKNOWLEDGMENTS

NSF NCAR and CISL

Administration Team:
- Virginia Do, Jerry Cyccone, Jeff Weber,
- Jessica Wang, Sam Scalice

CIRRUS Team

Mentors
- Nathan Hook, Eric Nienhouse, Jason Cunning

CISL

Try It Yourself

NCAR CIT Credentials required

Disclaimer

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.