

Using LLM Chat to Simplify User Interactions With GDEX Data



Anna Scribner,
Nathan Hook, Harsha Hampapura, Jason Cuning, Eric Nienhouse, Chia-Wei Hsu

National Science Foundation - National Center for Atmospheric Research and North Seattle College



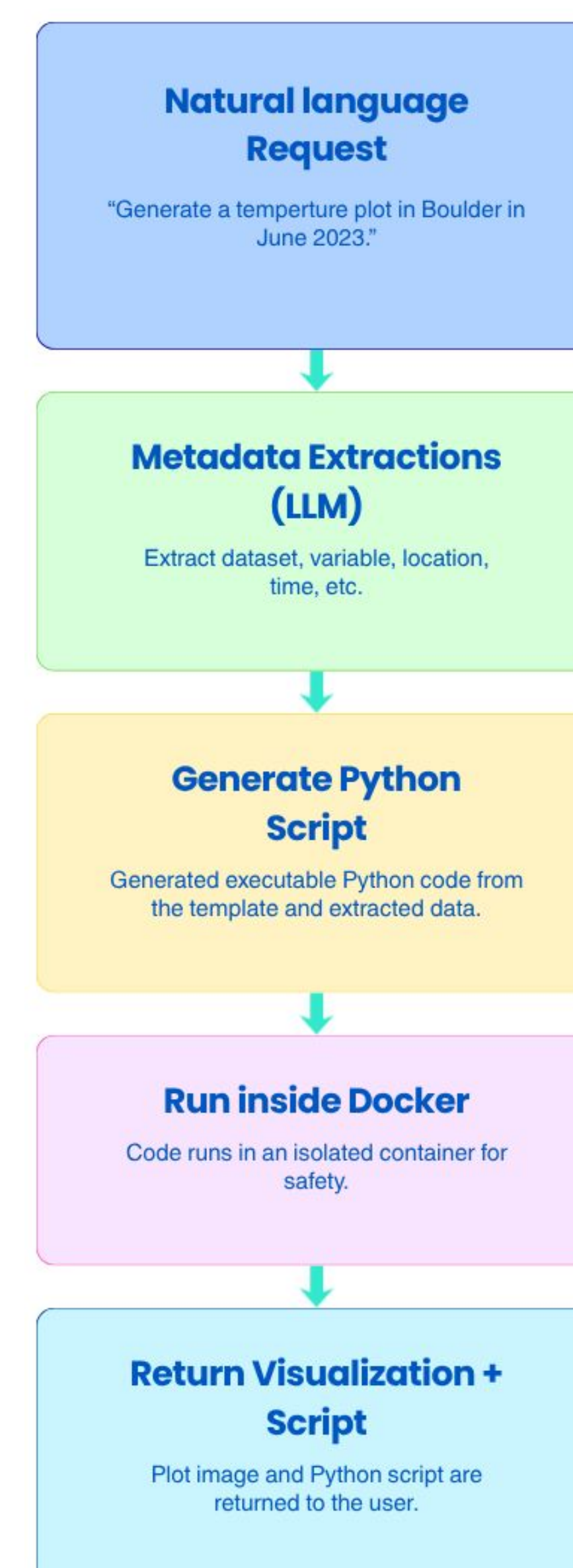
1. MOTIVATION

Accessing scientific datasets can be difficult, especially for people who are unfamiliar with the data or the tools required to analyze them. Users often need to understand the dataset's structure, variable names, formats, and programming libraries before they can explore their questions.

This project explores whether Large Language Models (LLMs) can reduce these barriers by allowing users to interact with scientific datasets using natural language.

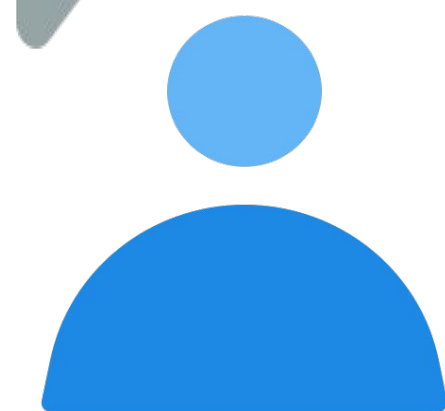
2. WHAT DID I DO?

Built a system that turns natural language requests into visualizations from GDEX hosted datasets.

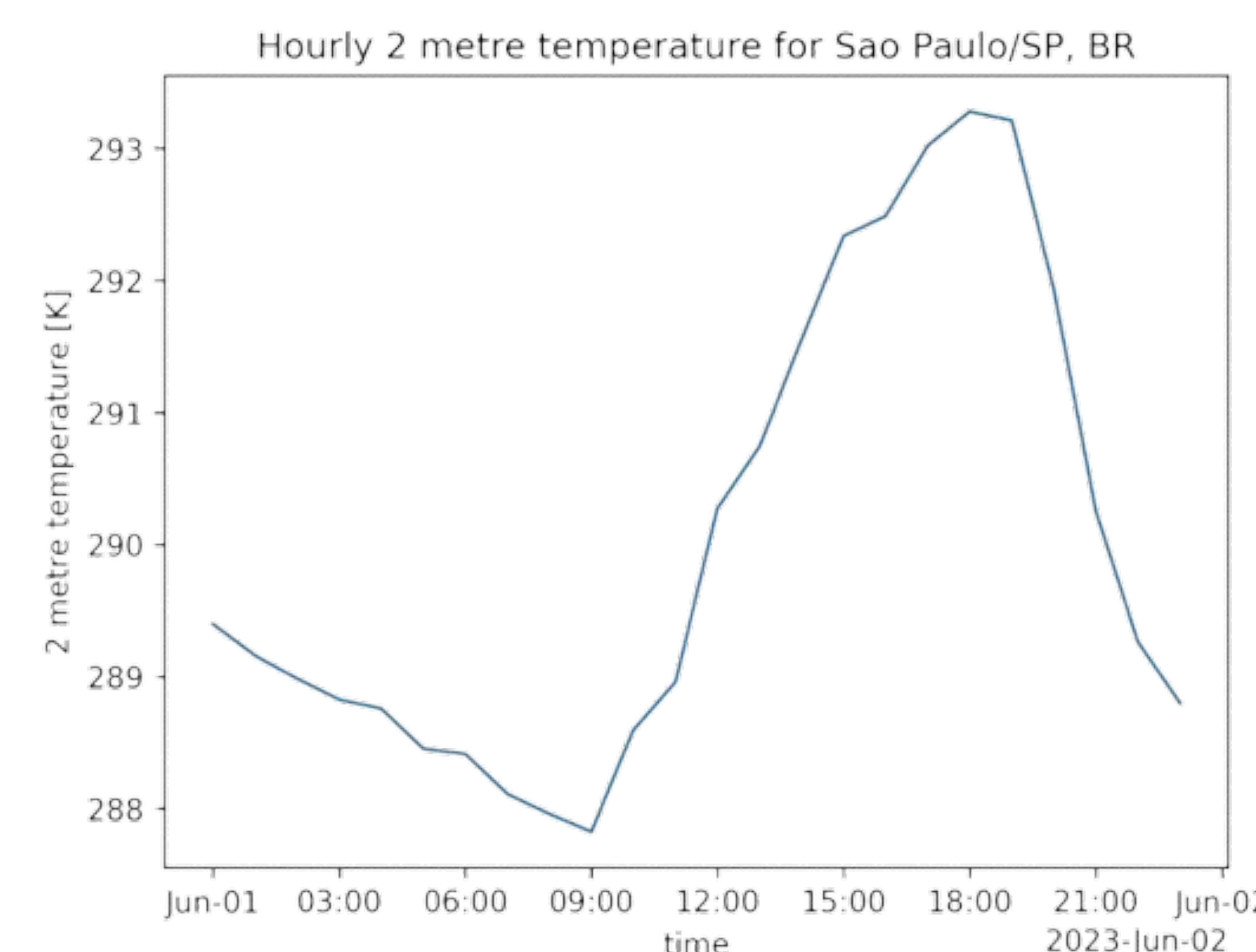
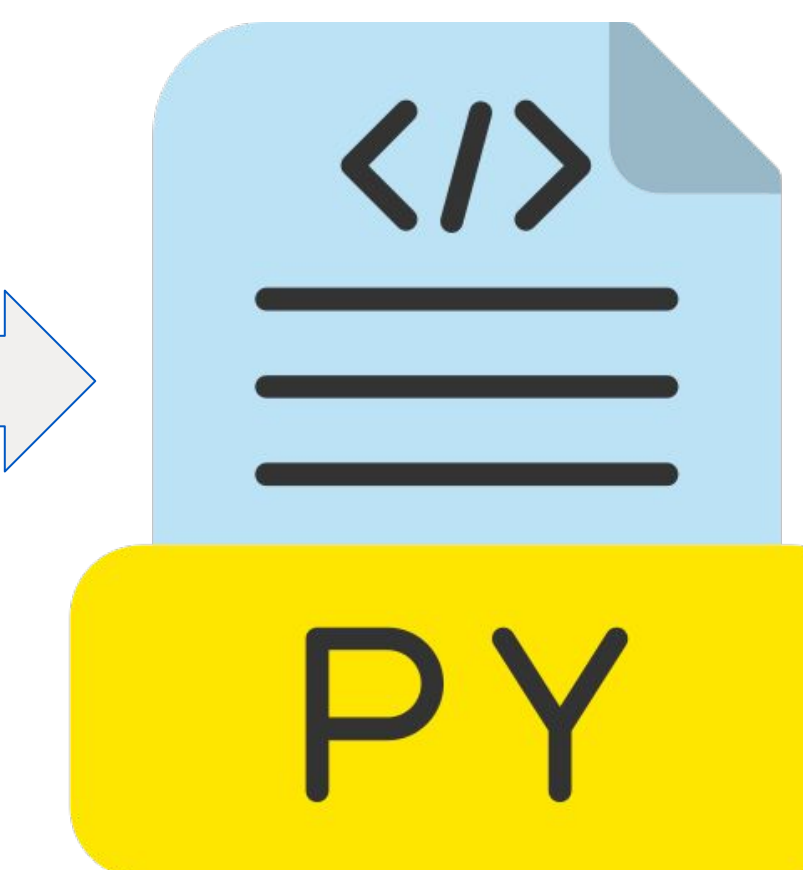


3. Example

Plot temperature in São Paulo for June 1 2023.



Dataset: *link*
Variable: VAR_2T
Location: São Paulo, BR
Time: 2023-06-01
Latitude: -23.5505
Longitude: -46.6333



4. WHY IT MATTERS?

Scientific datasets are becoming larger and more complex, making them difficult for new users to explore.

By allowing natural language interaction, this project attempts to lower the technical barriers to accessing scientific data and makes exploration more accessible to researchers, students, educators, and anyone curious about Earth system science.

5. Challenges

- LLM choosing incorrect dates
- Finding the correct prompts
- Templated vs Generated code
- Extracting metadata variables
- Error handling
- Recursive mount ambiguity (e.g. docker mounts inside of docker)

6. Future Work

- Continue improving the reliability of LLM generated code and metadata extraction
- Making the request conversational
- Include statistical analysis
- Real code generation
- Data validation

7. Acknowledgement

I would like to thank my mentors: Nathan Hook, Harsha Hampapura, Jason Cuning, Eric Nienhouse, and Chia-Wei Hsu.

I would also like to thank the SIParCS Administration Team, the CISL Team, the Sage Team, and NSF NCAR for their support throughout this project.

Finally, I would like to thank North Seattle College for supporting my academic journey.