

Frostbyte: An AI Assistant for Science Gateways Discovery

Joshua Harrell

Elizabeth City State University

Aguiluchos2004@gmail.com

 <https://github.com/LaJoshua>

Coreen Mullen

Elizabeth City State University

coreenmullen.8052@gmail.com

 <https://github.com/coreen-mullen>

Lisha Ramon

City College of New York

lisharamontn@gmail.com

 <https://github.com/LishaRamon>

Luz Rivera

Elizabeth City State University

lzrivera24@gmail.com

 <https://github.com/Lzrivera>

Dickson Acheampong

Howard University

adomacheampong dickson@gmail.com

 <https://github.com/dickson202>

Seth Mack

Morehouse College

sethamack89@gmail.com

 <https://github.com/SethMack611>

Je'aime H. Powell

Elizabeth City State University

jeaimhep@gmail.com

 <https://github.com/jeaimhep>

Abstract

This research project introduces an AI-powered virtual assistant for the Science Gateways Catalog (<https://sciencegateways.org/resources/browse>) named "*Frostbyte*", aimed at aiding researchers, educators, and students in navigating and leveraging science gateway tools. Utilizing natural language processing (NLP) and machine learning (ML), the assistant employs tools such as Beautiful Soup, and Hugging Face models to deliver features like catalog recommendations, simulation setup guidance, ethical research support, and direct resource links. Targeting faculty, students, data curators, and developers, it prioritizes accessibility for the Science Gateways community. The methodology integrates site scraping and API connectivity for real-time data access, developed collaboratively during the SGX3 Codeathon. Key outcomes include enhanced usability and inclusivity, simplifying access to complex scientific resources.

Tools and Application Interfaces

The chatbot is designed to provide a user-friendly interface for exploring the Science Gateways Catalog, allowing users to search for relevant resources using natural language queries. The key features include:

Natural Language Queries:

The chatbot understands user queries and maps them to structured search parameters (e.g., name, abstract, tags, category). Users can ask questions in different formats such as:

- *'Find all gateways related to climate change.'*
- *'Show me resources tagged with 'biosecurity.'*
- *'What science gateways are available for genomics research?'*

Search Assistance & Filtering:

The chatbot allows users to refine searches based on multiple criteria, including:

- Name of the science gateway
 - Keywords in the abstract
 - Publication date
 - Tags associated with the resource
 - Category classification
- This ensures users can efficiently locate the most relevant information.

User-Friendly Responses:

The chatbot returns results in a structured format, displaying key details such as the name, abstract, category, and relevant links. It also supports pagination for large datasets and provides suggestions for refining searches.

Backend Implementation:

1. A JSON dataset containing science gateways is loaded into the Flask API.
2. The API allows filtering and searching using query parameters such as name, abstract, tags, and category.
3. The chatbot interacts with the API by sending user queries as search parameters.

Front-End Design & User Interaction:

1. The chatbot is accessible via a web interface where users input their queries.
2. The chatbot dynamically processes input and sends requests to the API.
3. Responses are formatted in a clear, structured manner for easy readability.

API Connection Workflow:

1. User submits a query (e.g., *'Find science gateways for climate research.'*).
2. The chatbot processes the request and converts it into API parameters.
3. The Flask API searches the dataset for matching records.
4. The results are formatted and displayed to the user.

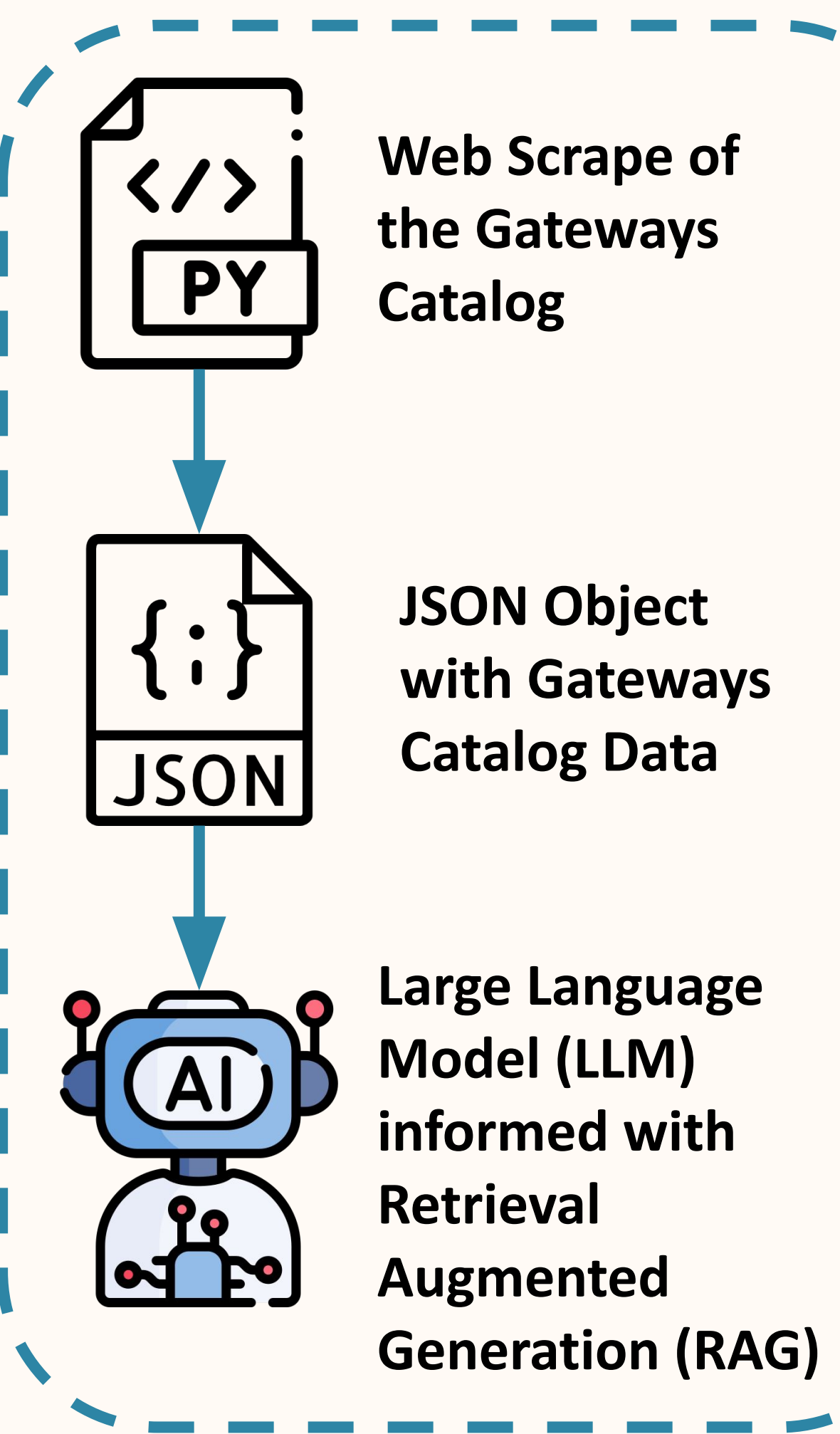


Results

SGX3

GATEWAY CATALOG:

<https://sciencegateways.org/resources/browse>



Type your message OR Say 'voice' to speak: Show me a list of science gateways related to polar science

Science Gateway AIs: Here are some science gateways related to polar science that you might find interesting:

1. Arctic Data Center
Category: Science Gateways
Abstract: The Arctic Data Center serves as a repository for data generated from research funded by the National Science Foundation's Arctic Sciences Section. The gateway provides data discovery, access, and user support for Arctic research data.
Website: Arctic Data Center (<https://www.arcticdata.io/>)
2. Antarctic Glaciological Data Center (AGDC)
Category: Science Gateways
Abstract: The Antarctic Glaciological Data Center (AGDC) is a gateway that provides access to data related to Antarctic glaciology, including ice sheet mass balance, glacier dynamics, and ice core records. The gateway aims to facilitate research and collaboration in Antarctic science.
Website: Antarctic Glaciological Data Center (AGDC) (<https://www2.umaine.edu/antarctic-data/>)
3. PolarHub: Polar Earth Observing Network Gateway
Category: Science Gateways
Abstract: PolarHub is a gateway that serves as a portal

Type your message OR Say 'voice' to speak:

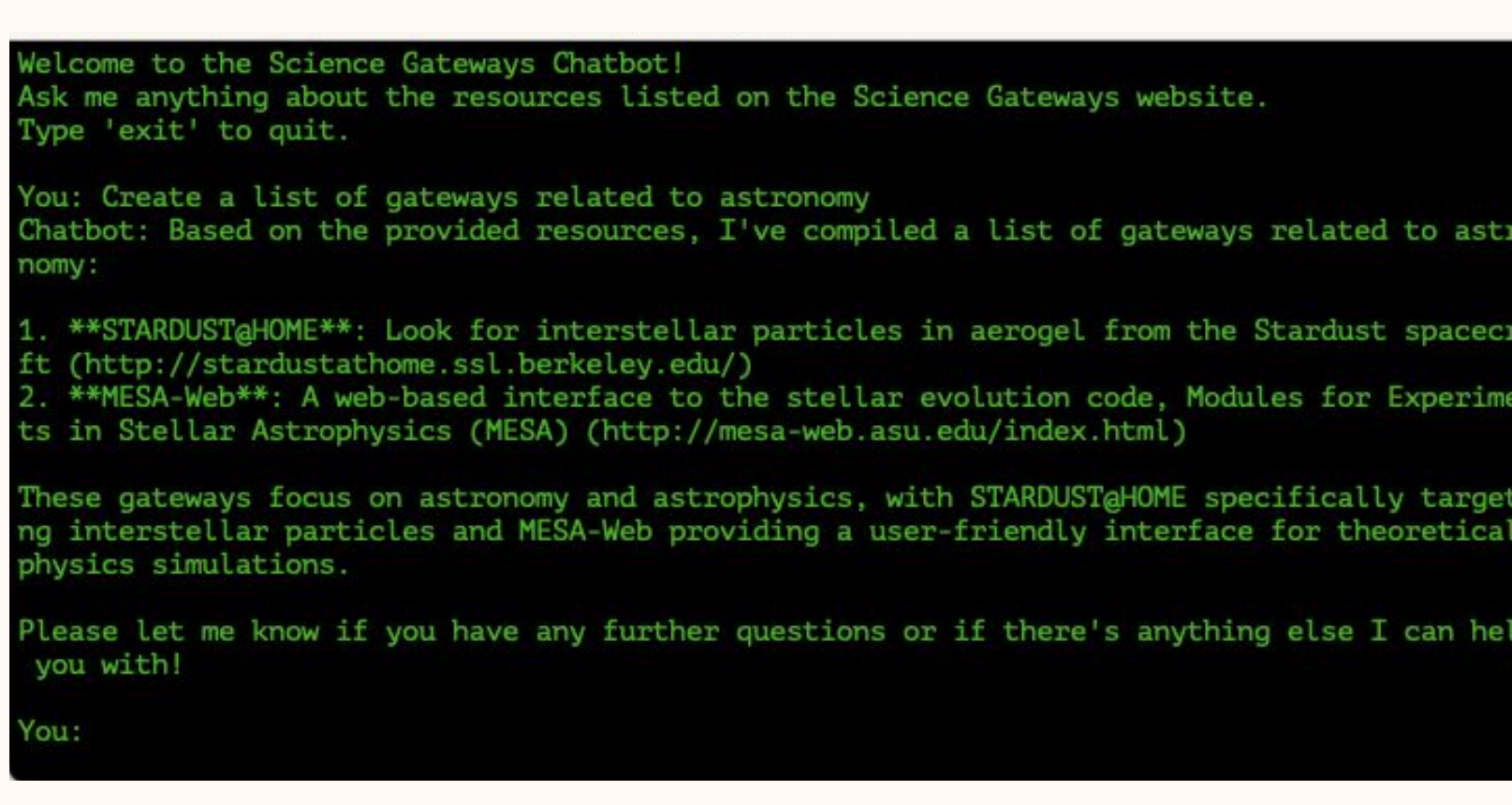
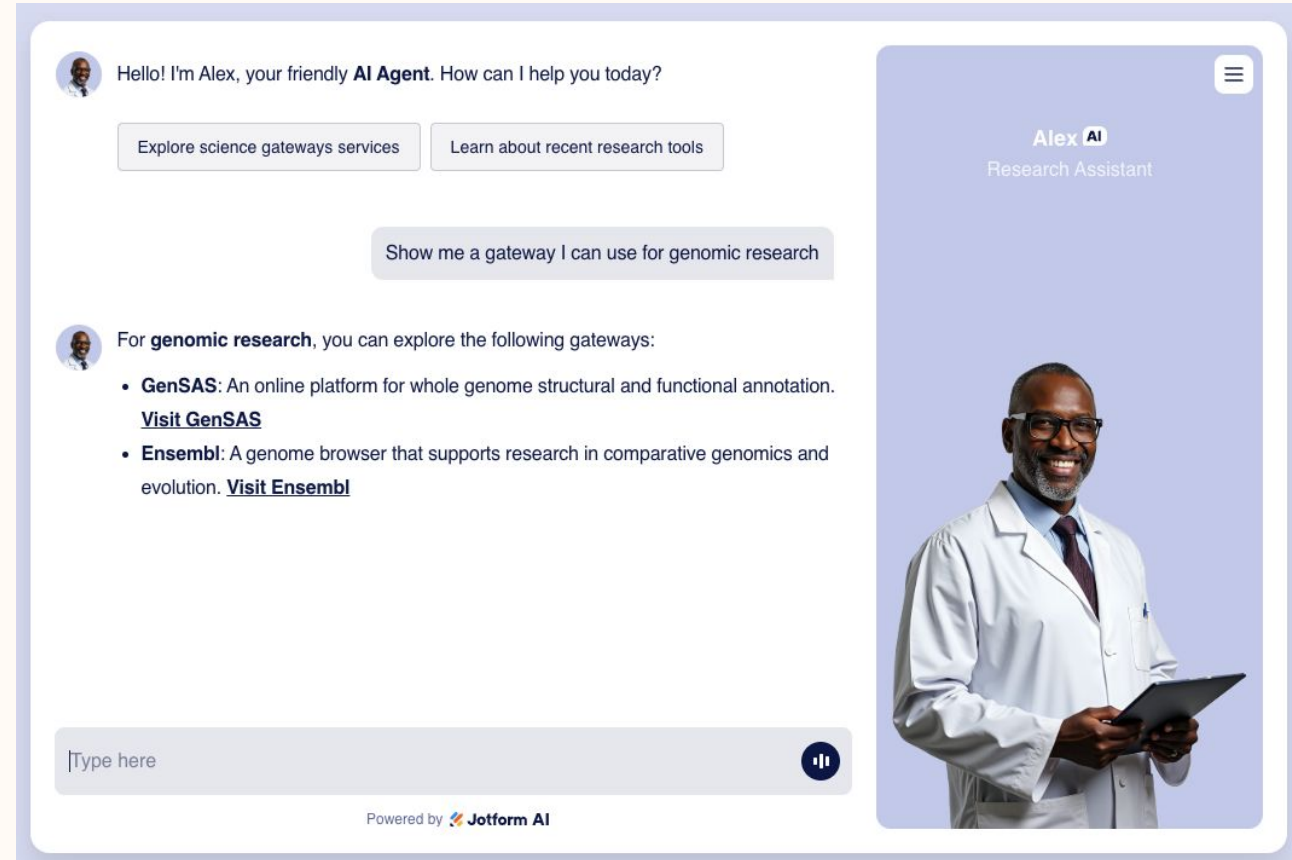
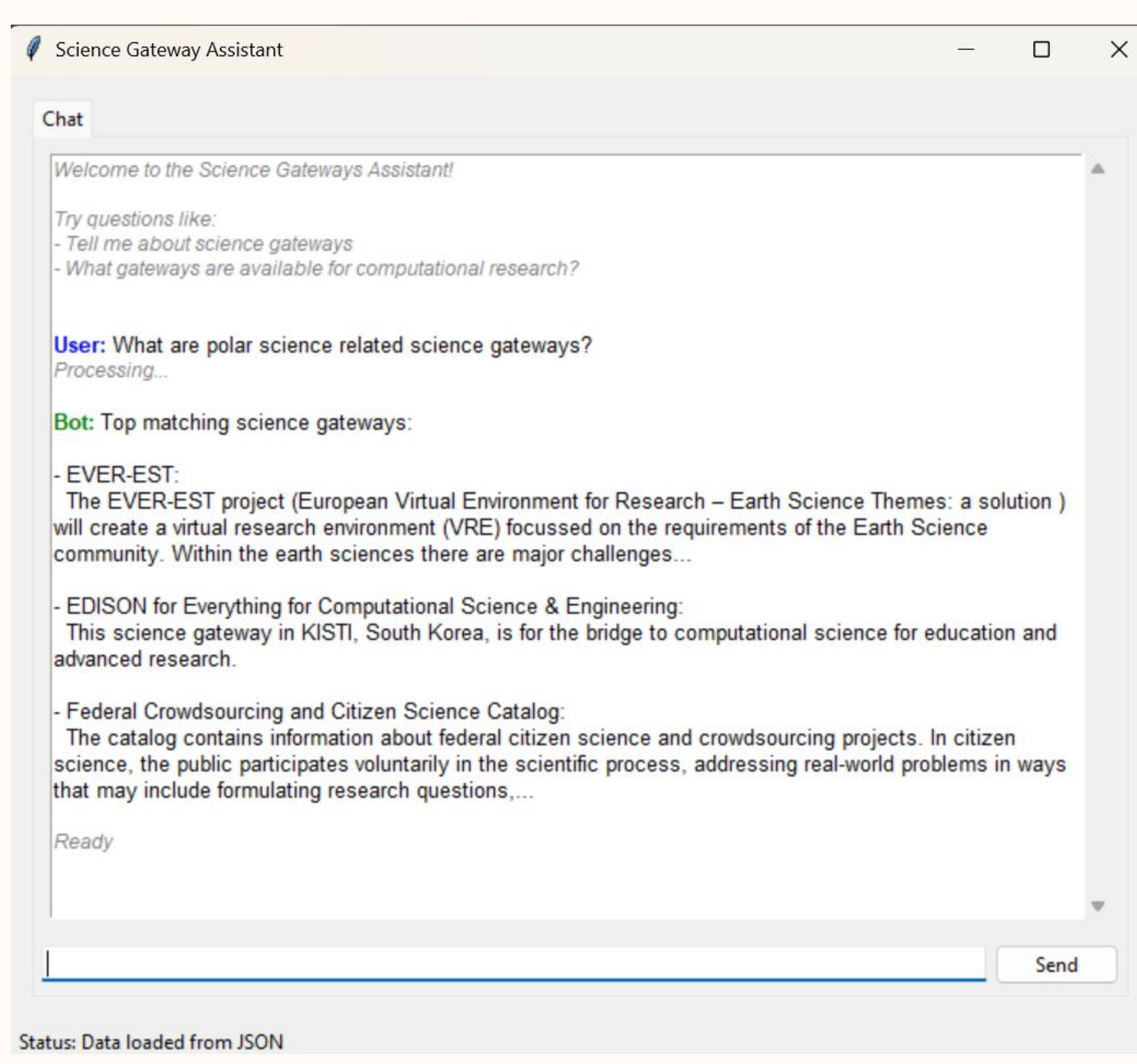
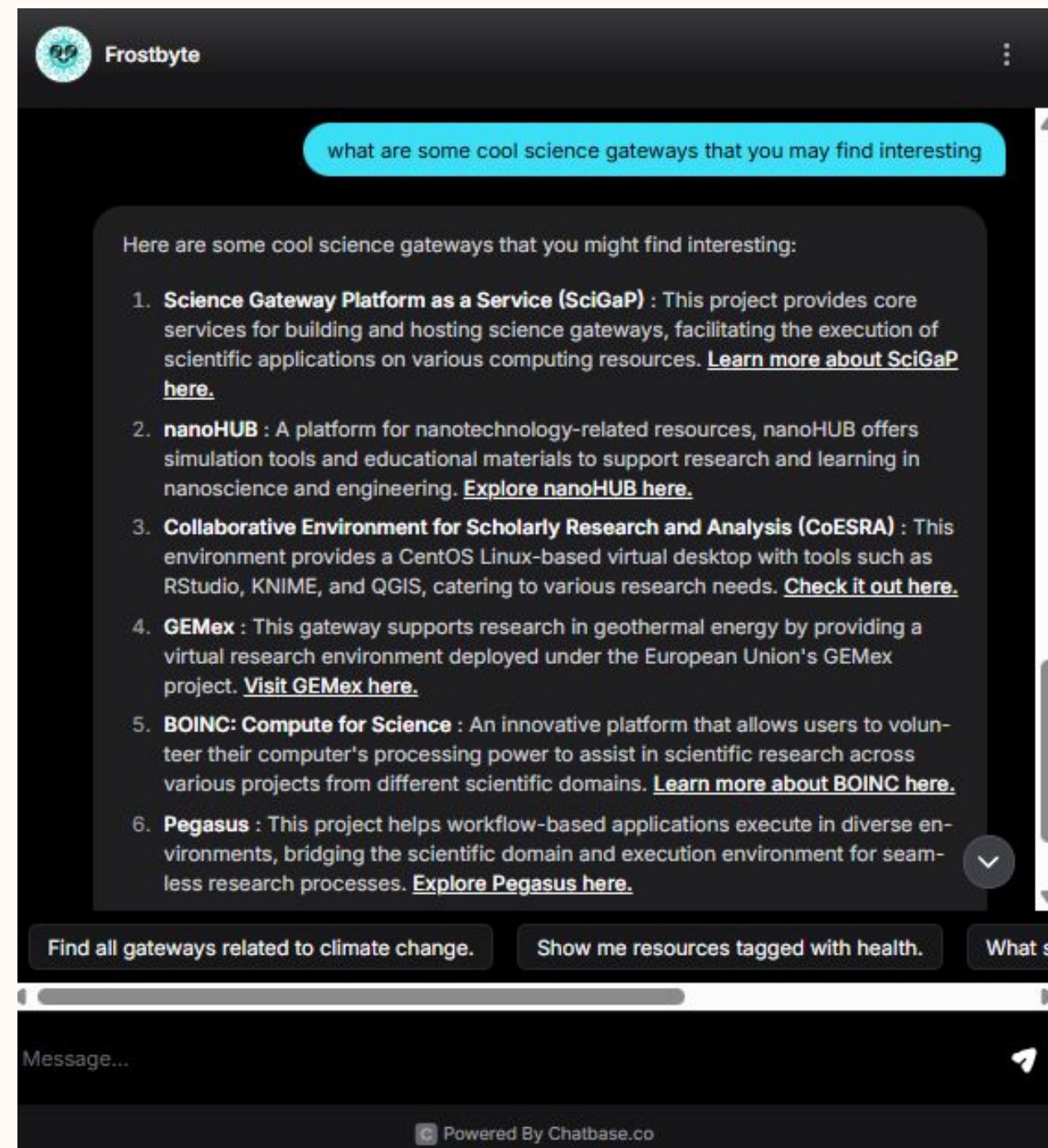
Welcome to the Science Gateways Chatbot!

Commands: 'list', 'describe [gateway name]', 'search [keyword]', 'quit'

What would you like to do? list

Here are all the science gateways:

- Custos Admin Gateway
- EDGE-Bioinformatics-Gateway
- CyberWater



Introduction

The Science Gateways Catalog serves as a vital resource for researchers, educators, and industry professionals, offering access to a diverse range of computational tools and applications. However, navigating this extensive catalog can be overwhelming due to information overload, complex navigation, and limited personalized recommendations. Users often struggle to identify the most relevant tools for their research, leading to inefficiencies in discovery and utilization.

To address these challenges, we propose Frostbyte an AI-powered Science Gateway Assistant, an intelligent chatbot designed to improve access to scientific resources, assist with computational tasks, and foster collaboration. By leveraging natural language processing (NLP) and Retrieval-Augmented Generation (RAG), this assistant provides a user-friendly interface for discovering relevant science gateway tools, automating metadata generation, and offering tailored research recommendations.

A key aspect of this assistant is its ability to dynamically retrieve and update information from the Science Gateways Catalog using web scraping techniques and RAG AI frameworks. Python-based libraries like Beautiful Soup allow the assistant to extract structured data from the catalog, ensuring that users receive up-to-date information on available tools and resources. Additionally, Large-Language Models (LLMs) enhance the chatbot's ability to provide context-aware responses by retrieving relevant documents from the catalog before generating answers. This approach combines real-time information retrieval with the generative power of AI, improving the accuracy and reliability of recommendations.

This assistant offers several key features to enhance user experience. It enables smart tool discovery, allowing researchers to efficiently navigate the catalog. Automated metadata generation improves search accuracy, ensuring that users can quickly locate the most relevant resources. The assistant also provides research assistance by integrating scientific journals and datasets, offering deeper insights and recommendations. Additionally, it includes collaboration tools such as file sharing and research paper suggestions, fostering teamwork among researchers. A guided, interactive dashboard delivers real-time updates and enhances usability, making it easier to interact with science gateway tools.

The assistant's core functionalities include personalized recommendations, helping users find relevant science gateway tools based on their research topics. It also facilitates navigation and tool discovery through natural language queries, making the catalog more intuitive. By offering concise summaries, key findings, and comparisons between related resources, the assistant streamlines information retrieval. Furthermore, automated metadata generation enhances search efficiency, while collaborative features, such as research paper suggestions and dataset recommendations, create a more interconnected research environment.

Objective

The goal of this project is to reduce barriers to information retrieval, improve research productivity, and enhance collaboration within the research community. By integrating web scraping and RAG models, the assistant ensures that users have access to the most recent and relevant scientific resources. This AI-powered solution makes the Science Gateways Catalog a more accessible, efficient, and valuable tool for researchers, educators, students, and industry professionals worldwide.

Methodology

Overview of the LLM Used

For the development of the AI-powered Science Gateway Assistant, a Large Language Model (LLM) is employed to handle natural language understanding and generation. The core model for this project is based on Hugging Face's transformers and pre-trained models such as BERT or GPT, optimized for scientific content and user interaction. The model is capable of understanding complex queries related to science gateway tools and providing context-aware, conversational responses. Custom fine-tuning is applied to tailor the model's responses to specific domain knowledge about science gateways, research tools, and relevant datasets.

Development Process

- **Data Collection:**
The data for fine-tuning the LLM is collected from multiple sources including the Science Gateways Catalog, which provides detailed information about various scientific tools, simulations, and research datasets. Additional data is gathered from reputable scientific articles, datasets like Kaggle and NIH Data Commons, and other publicly available resources.
- **Preprocessing:**
The collected data is preprocessed to remove irrelevant information, ensuring it is clean and consistent for training. This involves text normalization (removing special characters, handling case sensitivity), tokenization (splitting text into smaller, manageable parts), and the extraction of key entities related to science gateways (e.g., tool names, research topics, datasets).
- **Model Fine-Tuning:**
Fine-tuning is performed using specific scientific datasets to adapt the pre-trained model to the nuances of scientific language and terminology used in the context of science gateways. This process involves training the model on domain-specific questions and answers, allowing it to handle queries such as "What is the best tool for data visualization?" or "How can I set up a simulation on a science gateway platform?"
- **Integration into a Chatbot Interface:**
The fine-tuned LLM is integrated into a user-friendly chatbot interface. This interface is designed to allow users to interact with the assistant in real-time, asking queries about research tools, simulations, datasets, and more. The chatbot provides context-sensitive responses to guide users through complex processes, suggest relevant resources, and answer common scientific questions.

Tools and Technologies Employed

- **Python:** The primary programming language for development, used for data processing, model fine-tuning, and backend integration.
- **Hugging Face:** Utilized for accessing pre-trained transformer models (such as GPT and BERT) and for fine-tuning the LLM on scientific content.
- **Beautiful Soup:** A Python library used for web scraping to gather data from various scientific repositories and portals like Science Gateways and Kaggle.

Source Code Repository: <https://github.com/HackHPC/sgx3codeathontraining25>



Conclusions

- *Frostbyte*, an AI-powered assistant for the Science Gateways Catalog, successfully meets its objectives of reducing barriers to information retrieval, boosting research productivity, and enhancing collaboration.

- Leveraging NLP and ML with tools like Hugging Face, Beautiful Soup, and a Flask-based API, it delivers smart tool discovery, automated metadata generation, and user-friendly responses, significantly improving usability and accessibility for researchers, educators, and students—especially from underrepresented groups.

- Its broader implications extend beyond the catalog, offering potential applications in other science gateways and educational platforms, fostering inclusive scientific exploration worldwide.

- *We gratefully acknowledge the support of ADMI, the SGX3 Codeathon, our advisors: Je'aime Powell and Linda Hayden, and funding bodies for empowering this innovation through workforce development initiatives.*

- Presented at the 2025 ADMI Symposium, Frostbyte exemplifies the transformative power of AI in scientific discovery.

Future Work

Building upon the current base model of our chatbot, we will be implementing several key enhancements to improve its functionality and user experience. One of the primary updates will be the **implementation of multilingual support**, allowing the chatbot to communicate effectively with users in various languages. This is especially important due to the diversity of users of Science Gateways. Additionally, we will be incorporating **accessibility options** such as voice-to-text capabilities and customizable text sizes to ensure that the chatbot is usable by individuals with diverse needs. Another crucial step in our development process will be the **integration of the chatbot into the Science Gateways website**, expanding its reach and utility. Finally, we plan to add a **feedback feature**, enabling users to provide input on their interactions with the chatbot, which will be invaluable for ongoing improvements and refinements to the system.