AI Verde MCP Server: Bridging Generative AI with CyVerse Data Resources

Illyoung Choi
iychoi@arizona.edu

Paul Sarando
psarando@cyverse.org

Edwin Skidmore
edwin@cyverse.org

Nirav Merchant
nirav@arizona.edu

June 19, 2025
iRODS User Group Meeting

Generative AI's Impact on Research & Education

- **Accelerated Research**
  - Quick information search within knowledge bases
  - Literature review acceleration
  - Hypothesis generation support
  - Data interpretation assistance
- **Enhanced Teaching & Learning**
  - Course material creation for educators
  - Assignment drafting assistance
  - Personalized Q&A and concept clarification
  - Content simplification for accessibility
  - Interactive learning experiences

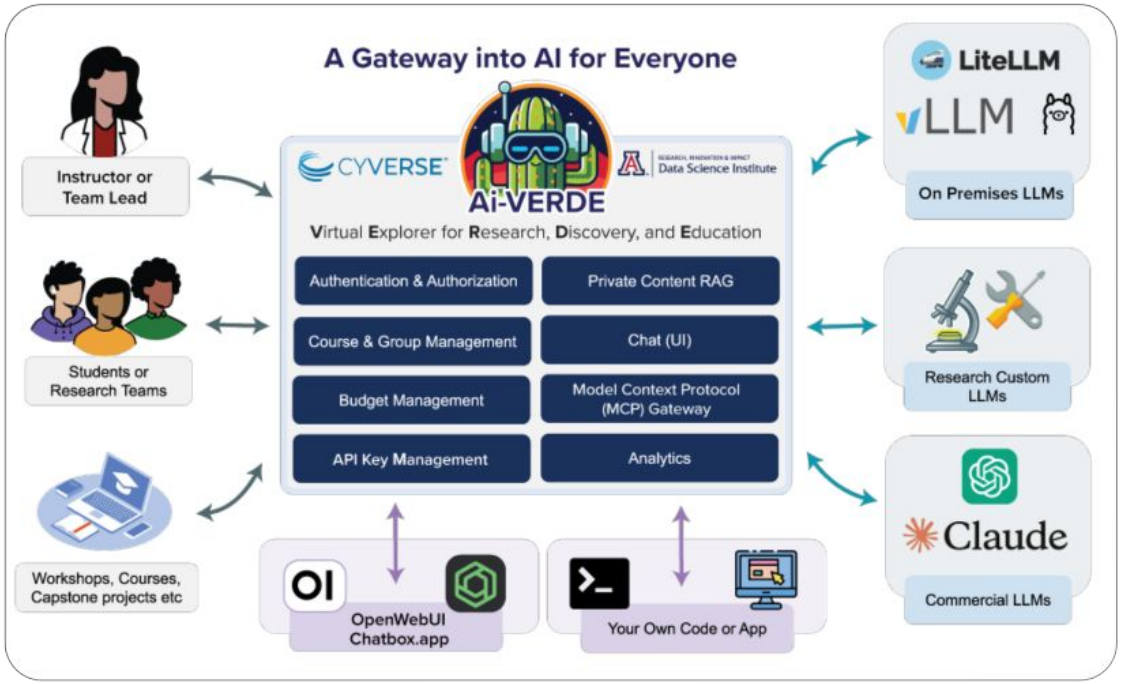
Limitations of Commercial Generative AIs

- Challenges in budget and group management for courses or research teams
- Limited access to diverse AI models
- Limited customization for specific domains or private data
- Data privacy and security risks

AI Verde

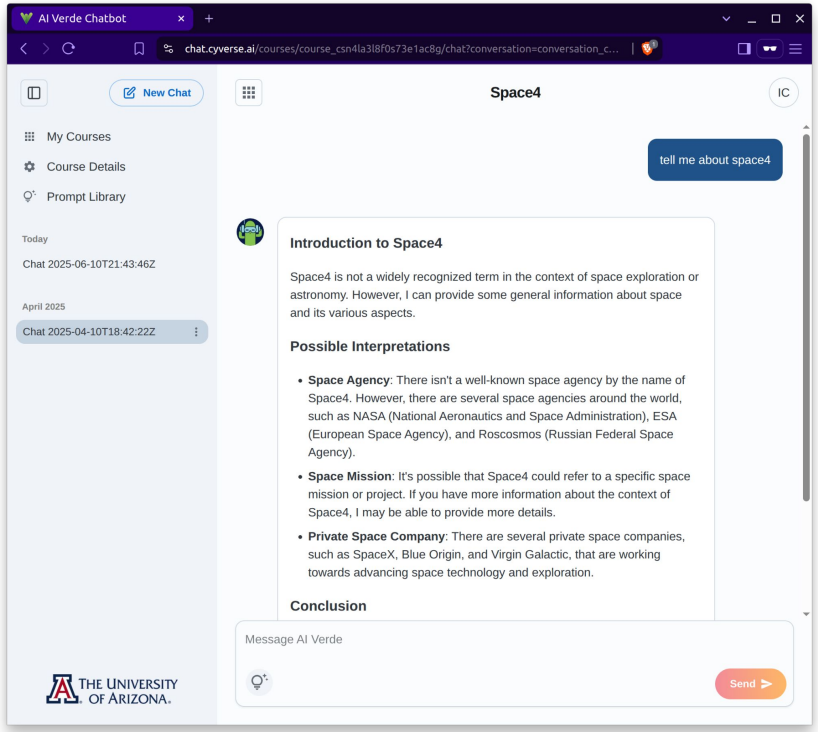
An integrated generative AI platform designed for research and education

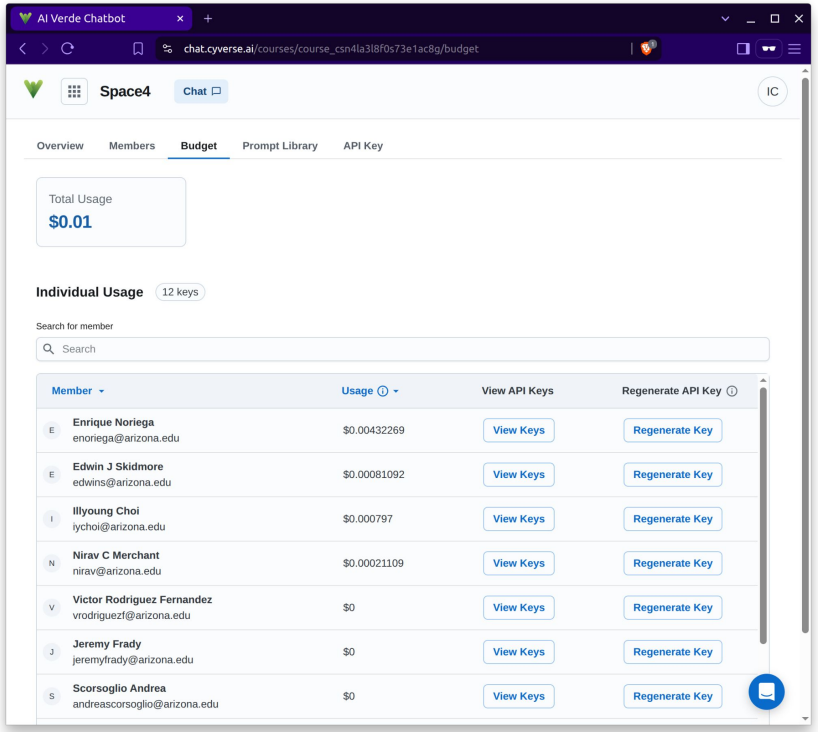
- Empowers research teams with AI on their private data
- Assists students and educators with AI, using customizable course materials
- Provides integrated group payments & budget management
- Integrates a wide range of LLMs, including commercial and open-source models
- Offers flexible access from diverse clients or directly in your code
- Leverages dedicated hardware: In-house & Jetstream2



(NSF's national infrastructure)

AI Verde Web Interface



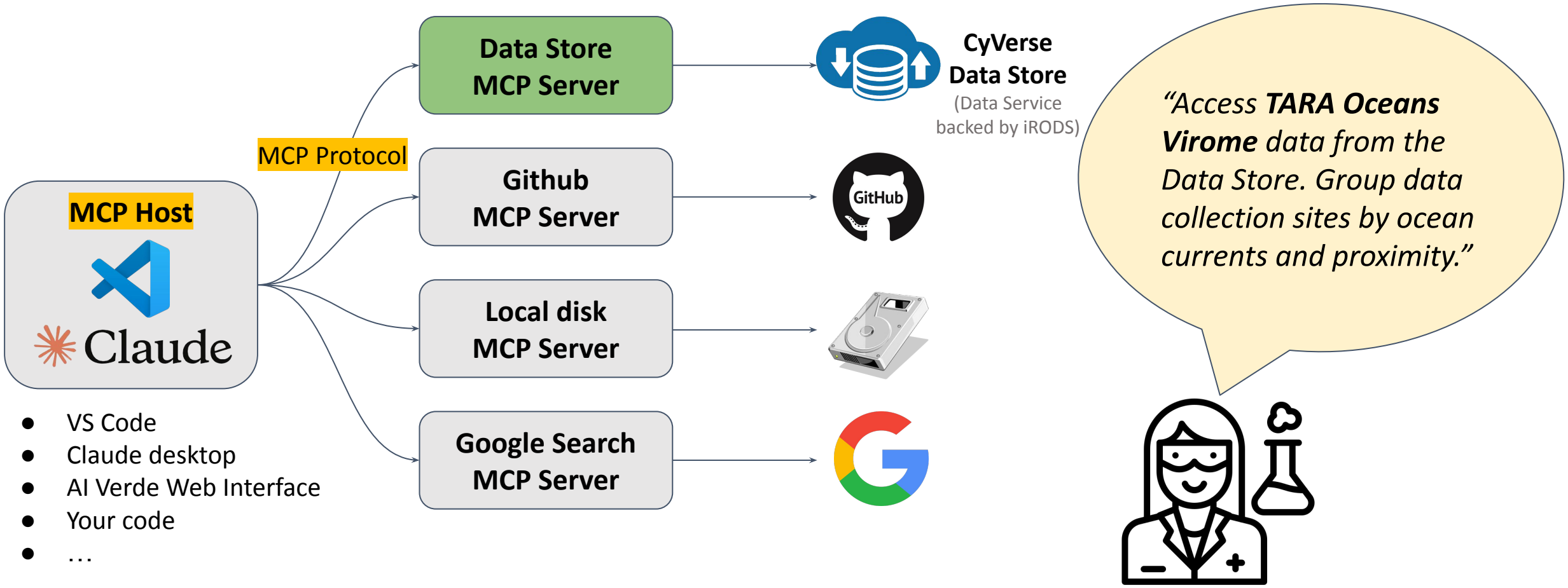


- Chatbot Interface
- Configurable Settings
  - Teams & Courses
  - Members
  - Budgeting
  - AI models
  - API Keys
  - External knowledge
  - …

Customizing LLMs with External Knowledge Bases

- **File Upload with RAG** (Retrieval Augmented Generation)
  - Limited number of files at a time
  - Assistance on specific documents
  - Content is read-only
  - E.g., Syllabi, lecture notes, research papers
- **Integration using MCP** (Model Context Protocol)
  - Direct access to external data services
  - Extended assistance with identifying and utilizing documents
  - Full data operations (e.g., file copy, modification)
  - E.g., Community research data, online real-time data

Model Context Protocol (MCP)



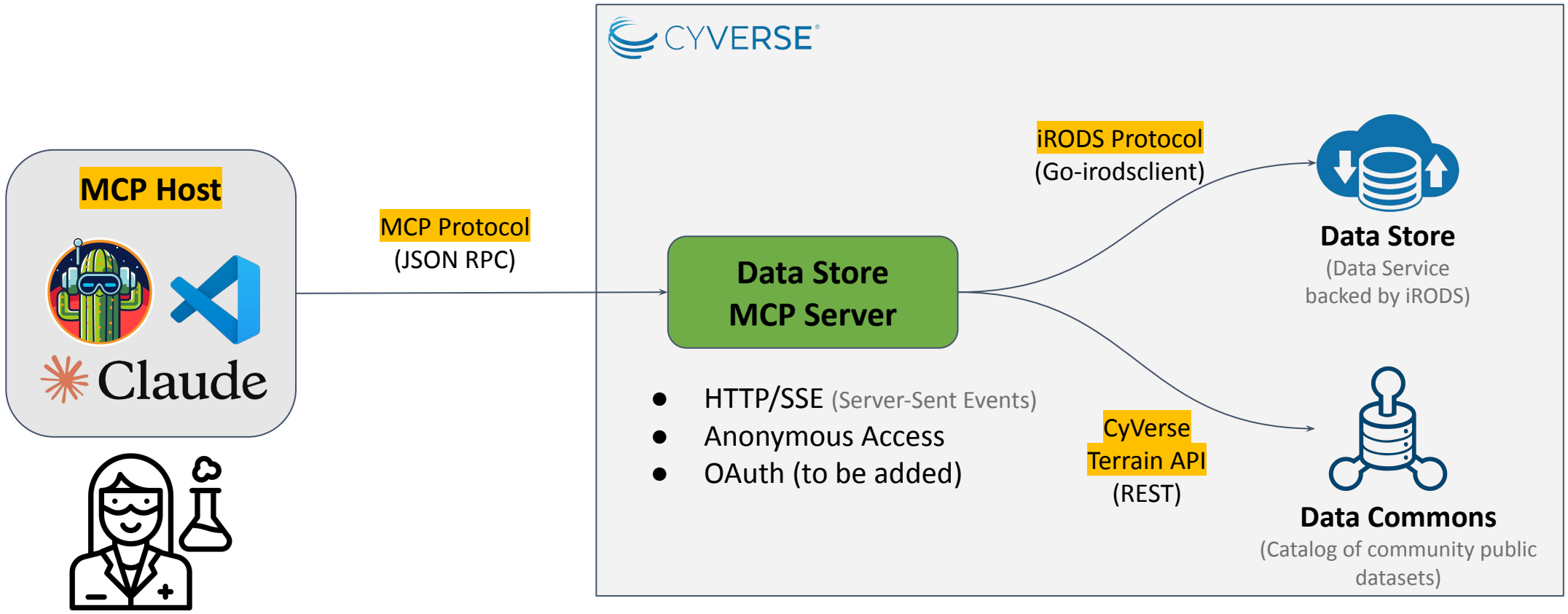
- VS Code
- Claude desktop
- AI Verde Web Interface
- Your code
- …

AI Verde MCP Servers in Development

- **Data Store MCP Server**
  - Provides general access to CyVerse Data Store (iRODS)
  - Provides access to CyVerse Data Commons (279 public datasets, 76 TB)
  - Currently closed-source (in research phase)

- **RAG MCP Server**
  - Provides access to indexing system for user-uploaded documents
  - Customizes answer based on user-uploaded data using Retrieval-augmented generation (RAG)

Data Store MCP Server Architecture



How the Data Store MCP Server Works

*Find supplementary material for "**The genome sequence of segmental allotetraploid peanut Arachis hypogaea**" in the data store*

MCP Host

Claude

Run
"ds_search_metadata"

API

Input
{"search_string": "Arachis hypogaea"}

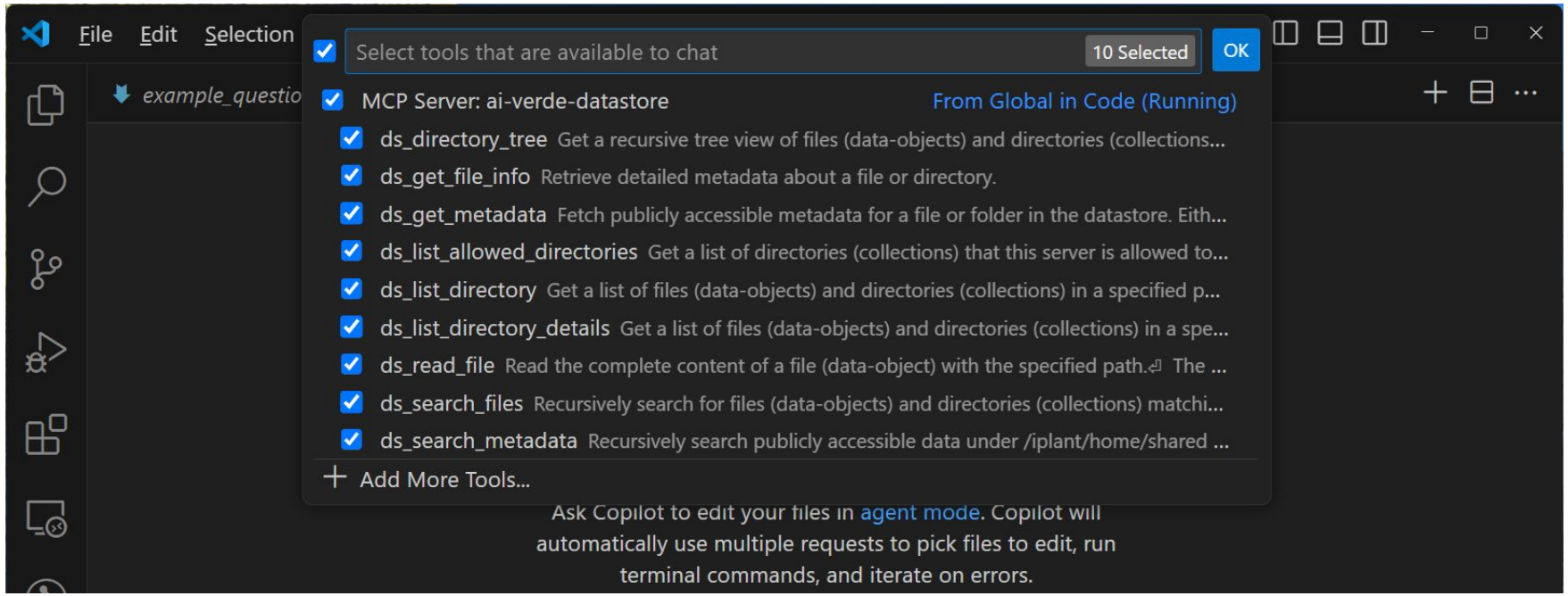
arguments

Data Store MCP Server


```
{
   "total": {
      "value": 1,
   },
   "hits": [
      {
         "_source": {
            "path": "/iplant/home/shared/commons_repo/
            "id": "2431259c-4f4a-11e9-a97e-000e1e0af2d
         },
      },
   ],
}
```


The supplementary material can be found in the Data Store at:
/iplant/home/shared/commons_repo/curated/Bertioli_Arachis_genome_supplement_TVDM_Mar2019

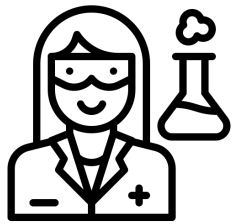
API Listing

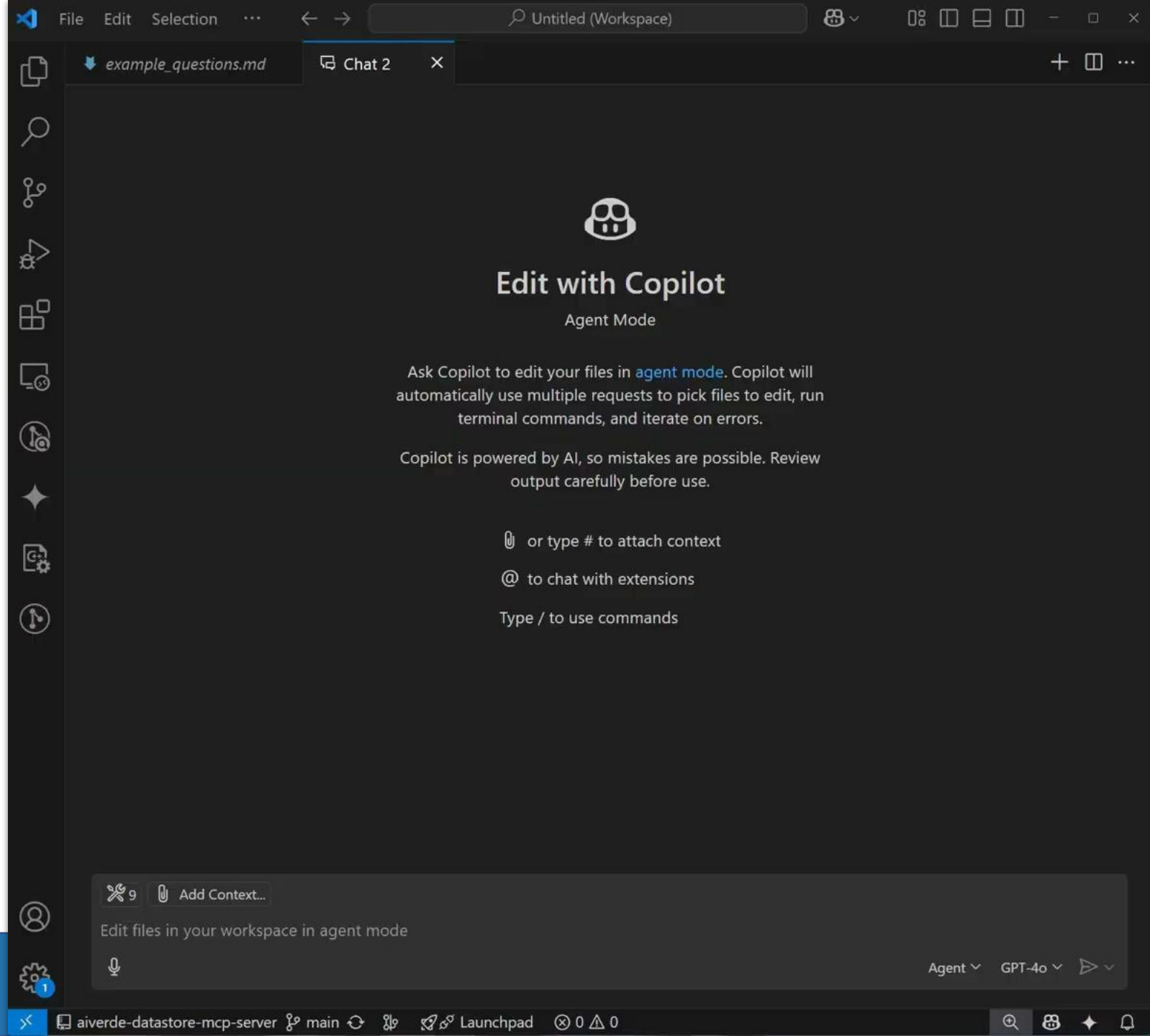

File
Edit
Selection
Select tools that are available to chat
10 Selected
OK
MCP Server: ai-verde-datastore
From Global in Code (Running)
ds_directory_tree Get a recursive tree view of files (data-objects) and directories (collections...
ds_get_file_info Retrieve detailed metadata about a file or directory.
ds_get_metadata Fetch publicly accessible metadata for a file or folder in the datastore. Eith...
ds_list_allowed_directories Get a list of directories (collections) that this server is allowed to...
ds_list_directory Get a list of files (data-objects) and directories (collections) in a specified p...
ds_list_directory_details Get a list of files (data-objects) and directories (collections) in a spe...
ds_read_file Read the complete content of a file (data-object) with the specified path. The ...
ds_search_files Recursively search for files (data-objects) and directories (collections) matchi...
ds_search_metadata Recursively search publicly accessible data under /iplant/home/shared ...
Add More Tools...
Ask Copilot to edit your files in agent mode. Copilot will
automatically use multiple requests to pick files to edit, run
terminal commands, and iterate on errors.

Demo #1

Listing a directory in the Data Store

*Show me files and directories in **/iplant/home/shared/terraref** directory in the Data Store and their **access permissions**.*



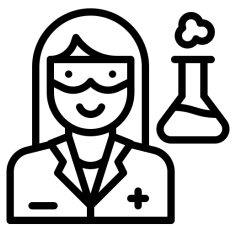


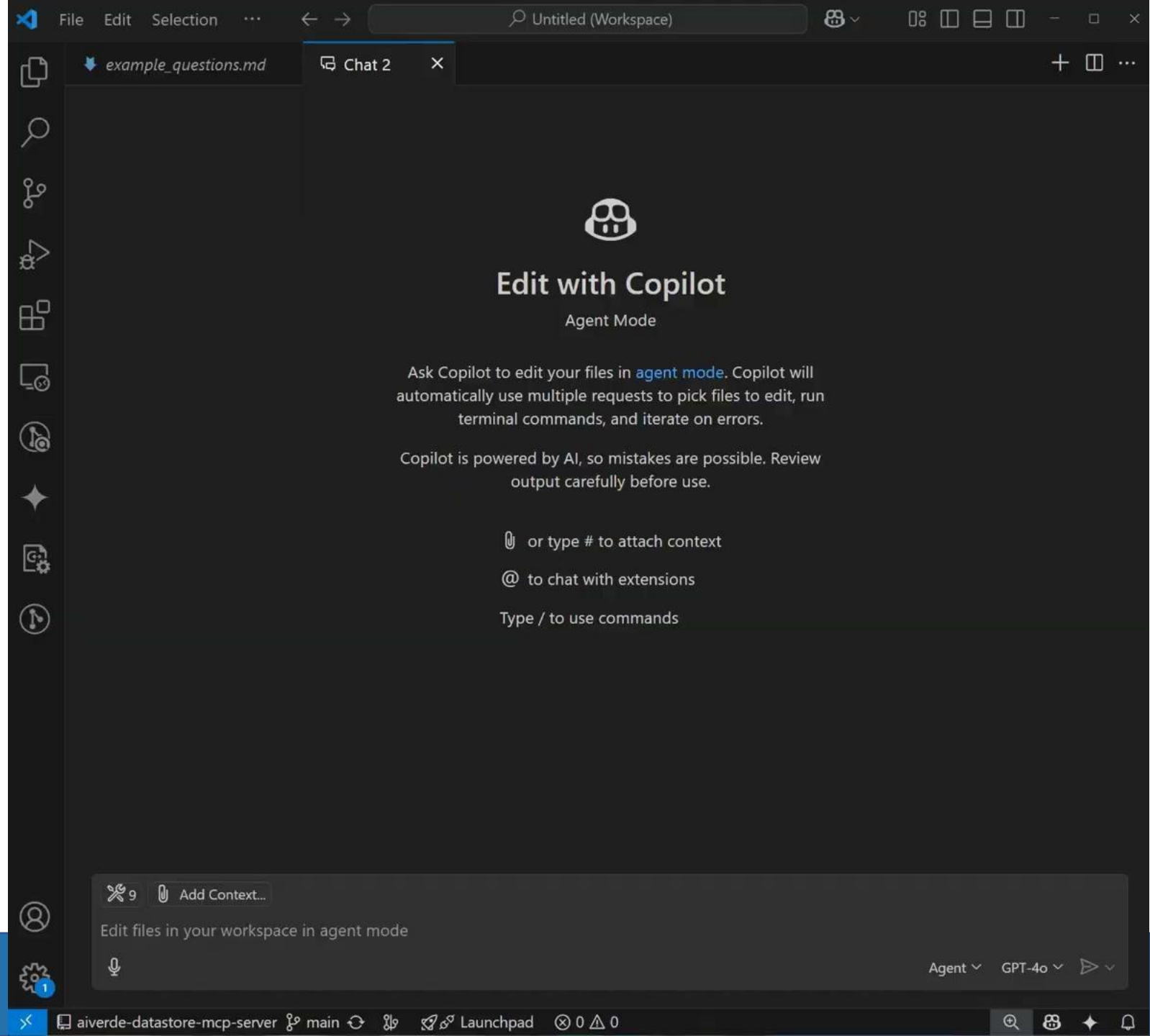
Demo #2

Searching a dataset in the Data Commons using metadata and summarizing ReadME file.

*Find a dataset for **wheat drought stress** in the Data Store and show the files in it. Show me the **details of the files**, such as **sizes, checksums** and **replica info**.*

*Summarize **ReadMe.txt** file.*





CYVERSE®

Future Work

- New Features
  - Implement file copy, upload, remove, and rename APIs
  - Implement OAuth authentication
  - Add caching for faster response

- Integration with AI Verde Web Interface
  - Allow instructors and users to add MCP servers to their teams & courses

Conclusion

AI Verde is an integrated generative AI platform

- Designed for research and education
- Supports private data, customization, integrated group payments and budgeting, diverse LLM models
- Currently free for evaluation purposes (Ask Nirav for access)

AI Verde Data Store MCP Server

- Provides access to Data Store and Data Commons, archiving large community public datasets (3.2 PB)
- AI Agents (MCP Hosts) automatically handle input and output for MCP Server APIs
- Simplifies access to Data Store and Data Commons
- Planned for open-source release soon

Questions?

Data Store MCP Server APIs

- Resources
  - Read-only file system access (listing and reading)
- Tools
  - Read-only access to community public datasets
  - Read-write access to home directory with authentication
  - List and search datasets by metadata (AVU)
  - Extensible for additional operations