

# Tri-Virtual Assistant Chatbot DEMO

Using NVIDIA Workbench AI and NVIDIA NIM

**Mentor :**

**Prof. Manjul Krishna Gupta**

**TEAM MEMBERS :**

1) Renu Bojja:1RVU22CSE129

2) Karmishtha P:1RVU22CSE077


# NVIDIA LaunchPad

Provides free access to enterprise NVIDIA hardware and software through an internet browser. Users can experience the power of AI with end-to-end solutions through guided hands-on labs or as a development sandbox. Test, prototype, and deploy your own applications and models against the latest and greatest that NVIDIA has to offer.

NVIDIA Entitlement Certificate - Ref

External

Inbox x



noreply@nvidia.com

to me

Fri, Nov 8, 3:27 PM (9 days ago)

☆

↶

⋮

Welcome to NVIDIA AI Enterprise! Your evaluation license has been approved.

**Run Generative AI Models Anywhere with NVIDIA NIM**

You can now deploy NIMs on your own infrastructure. Head over to the [API catalog](#) to generate an API key and download your model's NIM.

**Get Enterprise Support for NVIDIA software**

Attached to this email is your NVIDIA AI Enterprise entitlement certificate. This includes the PAK ID you need to file support tickets with [NVIDIA Enterprise Support](#).

**Experience all of NVIDIA AI Enterprise**

The entitlement also includes details for accessing exclusive NVIDIA AI Enterprise software, including production branches, vGPU and more.


The following is your order information:

| Entitlement Type | NVIDIA Sales Order | NVIDIA Delivery Number |
|------------------|--------------------|------------------------|
| EVAL-NVAIE       | NA                 | 163868                 |

**Questions?**

Need help? Please contact [us](#).

Thank you!



NVIDIA

NVIDIA Corporation

2788 San Tomas Expressway

SANTA CLARA CA 95051

USA

NOTICE

HOW TO USE THIS CERTIFICATE

Please refer to your [NVIDIA AI Enterprise Quick Start Guide](#) for information on how to get started, including additional instructions on how to register for your entitlement.

**Sales Type: EVAL**


1.

After you have successfully registered, you will receive an email from NVIDIA application hub to access NVIDIA Licensing Portal, NVIDIA Enterprise Support and NVIDIA NGC.

2.

In order to access your entitlements, please [login](#) and bookmark this site for future access/reference.

Rights and restrictions on the use, transfer and copying of the Software are set forth in NVIDIA's End User License Agreement.



NVIDIA

NVIDIA Corporation

2788 San Tomas Expressway

SANTA CLARA CA 95051

USA

NVIDIA® Entitlement Certificate

This certificate serves as evidence that NVIDIA has entitled you for the following product(s).

End Customer (74ec42a1-1fce-46c2-bb00-694f64d62a95)

none

NVIDIA Delivery

Entitlement Date

Entitlement Type

NVIDIA Sales Order

163868

08 NOV 2024

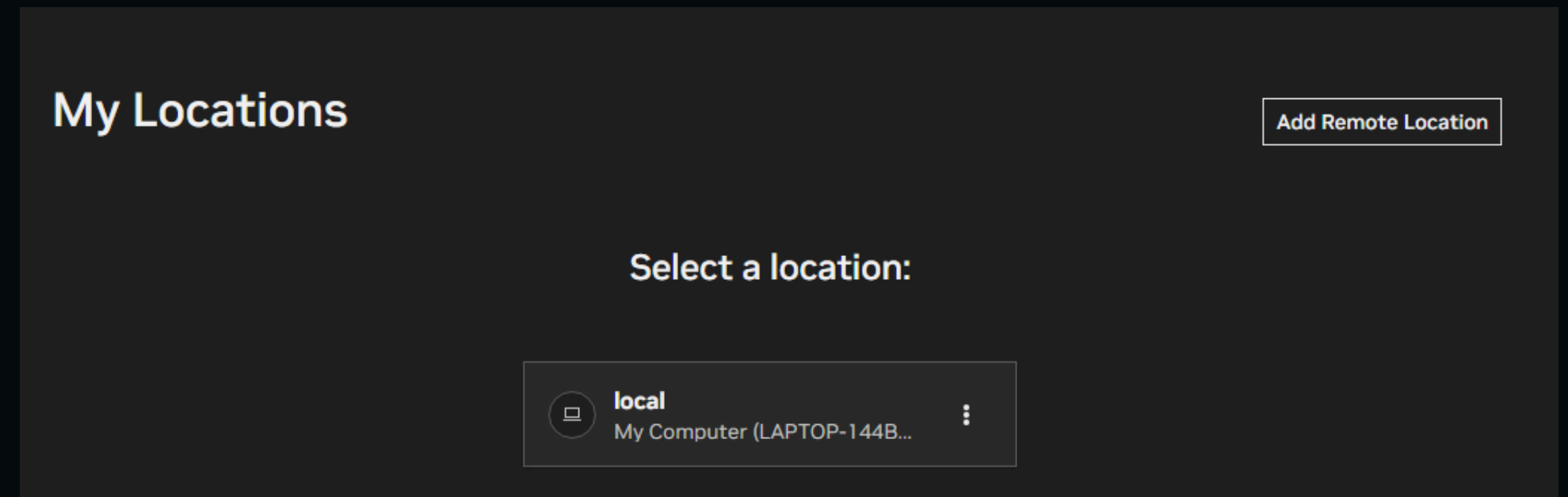
Eval - NVAIE

NA

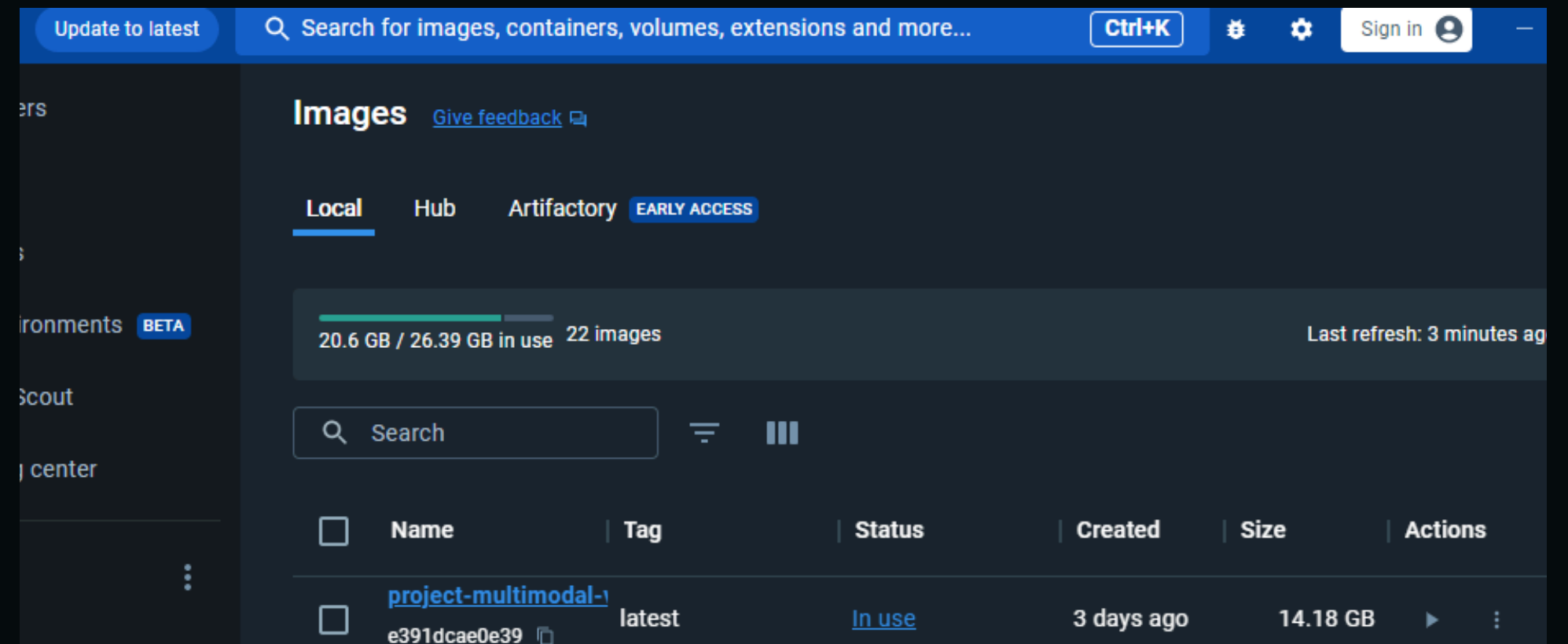
| No | Entitlement Description   | Quantity | Sales Type | Term | Start Date  | End Date    |
|----|---|----------|------------|------|-------------|-------------|
| 1  | NVIDIA AI Enterprise Evaluation<br>PAK ID: ofegqpwugo-77ojnlfod2-splwfpkx6t | 6 EA     | Initial    | N/A  | 08 NOV 2024 | 06 FEB 2025 |

DOWNLOAD LINK

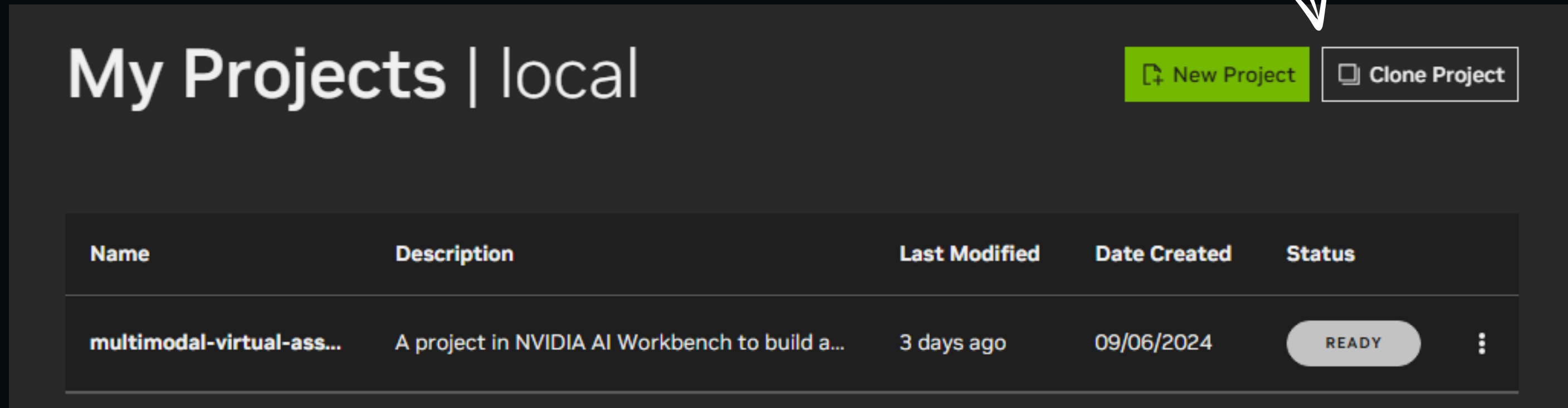
Select local location.



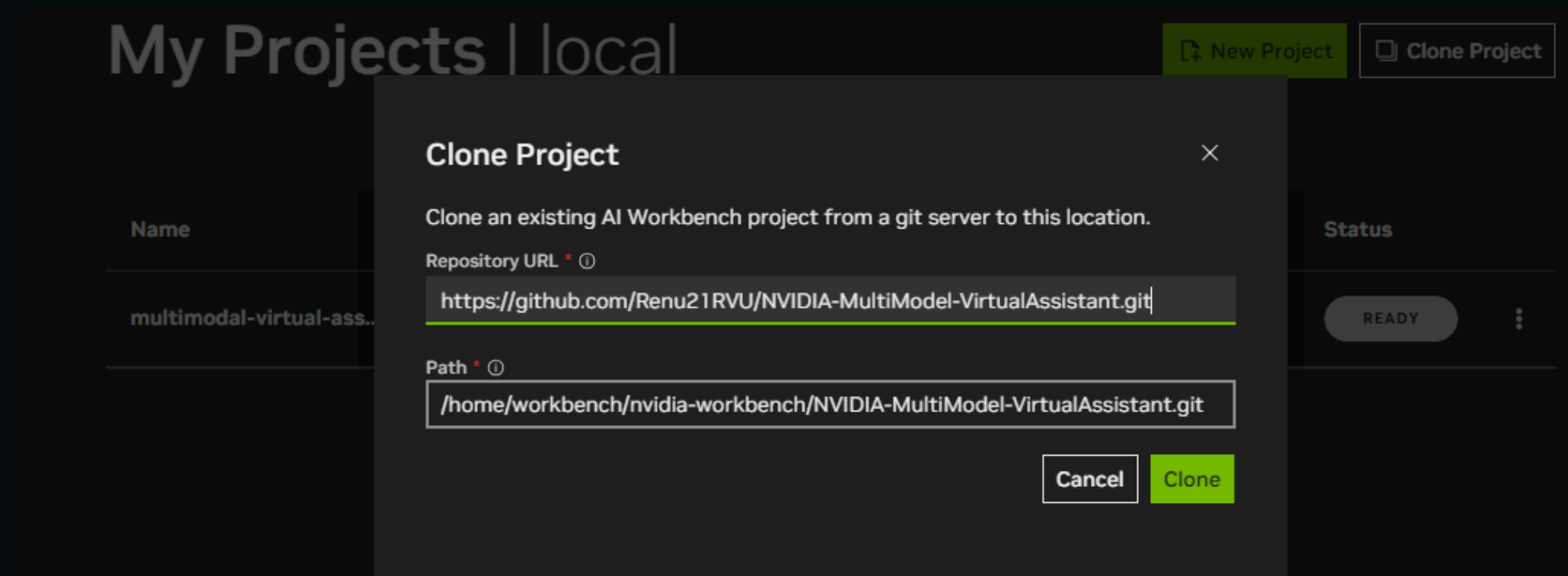
Docker opens automatically. Don't close it.



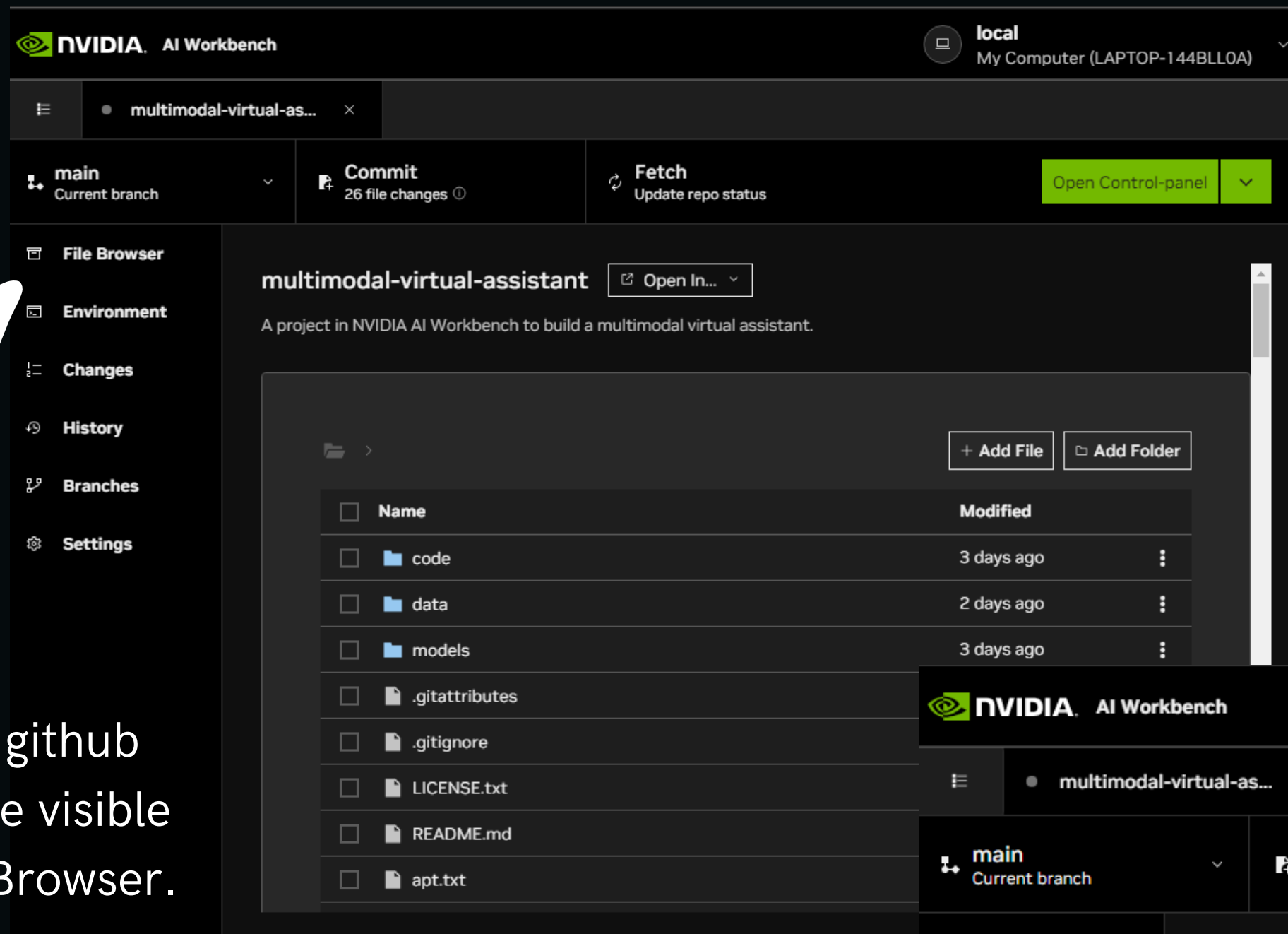
If you have your project on github click on Clone Project else you can click on New Project and start coding/upload your code files.



- Add your Repository URL in the space provided.
- Path is automatically created.
- Click on Clone.

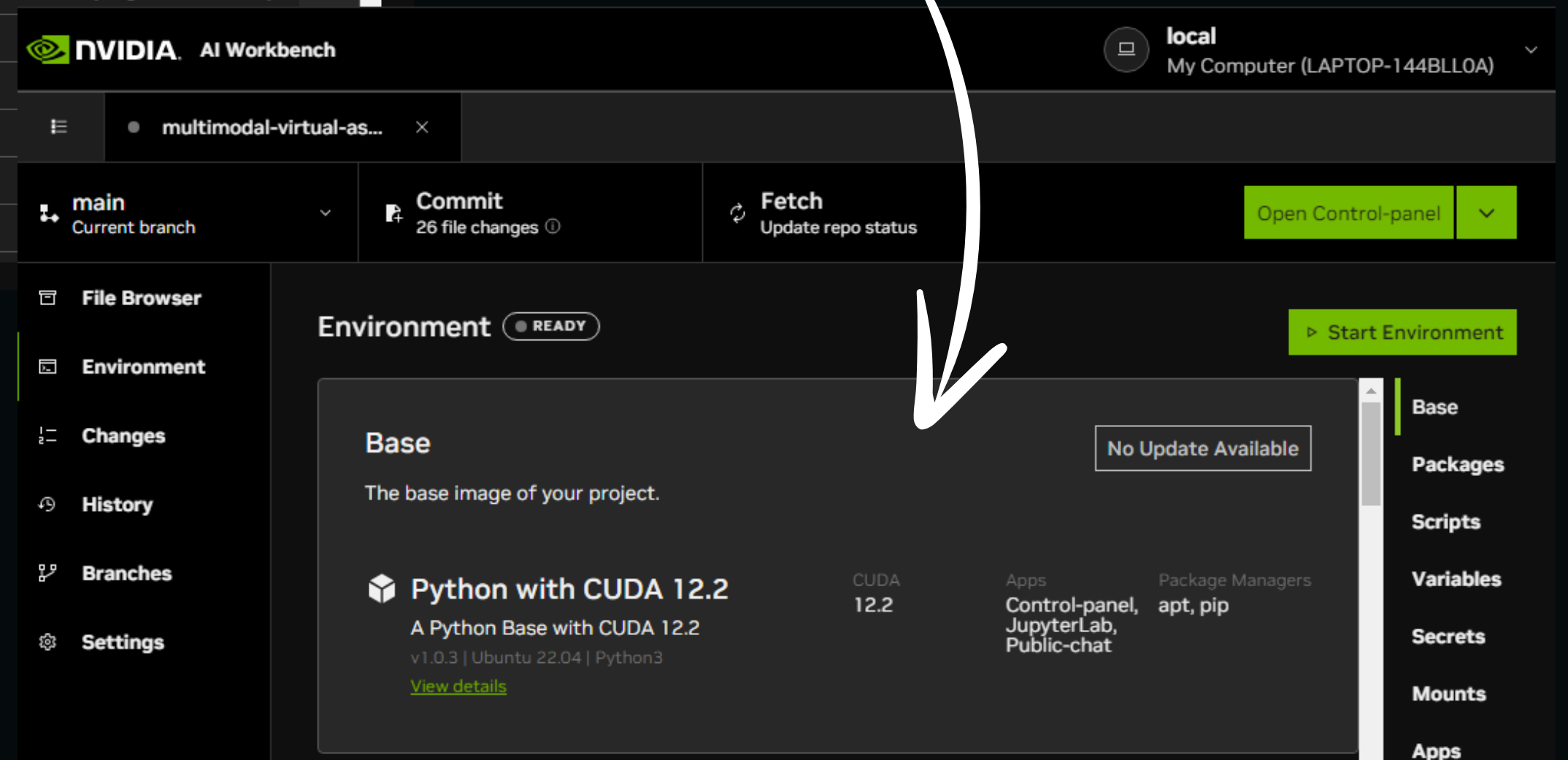


GITHUB : <https://github.com/Renu21RVU/NVIDIA-MultiModel-VirtualAssistant.git>

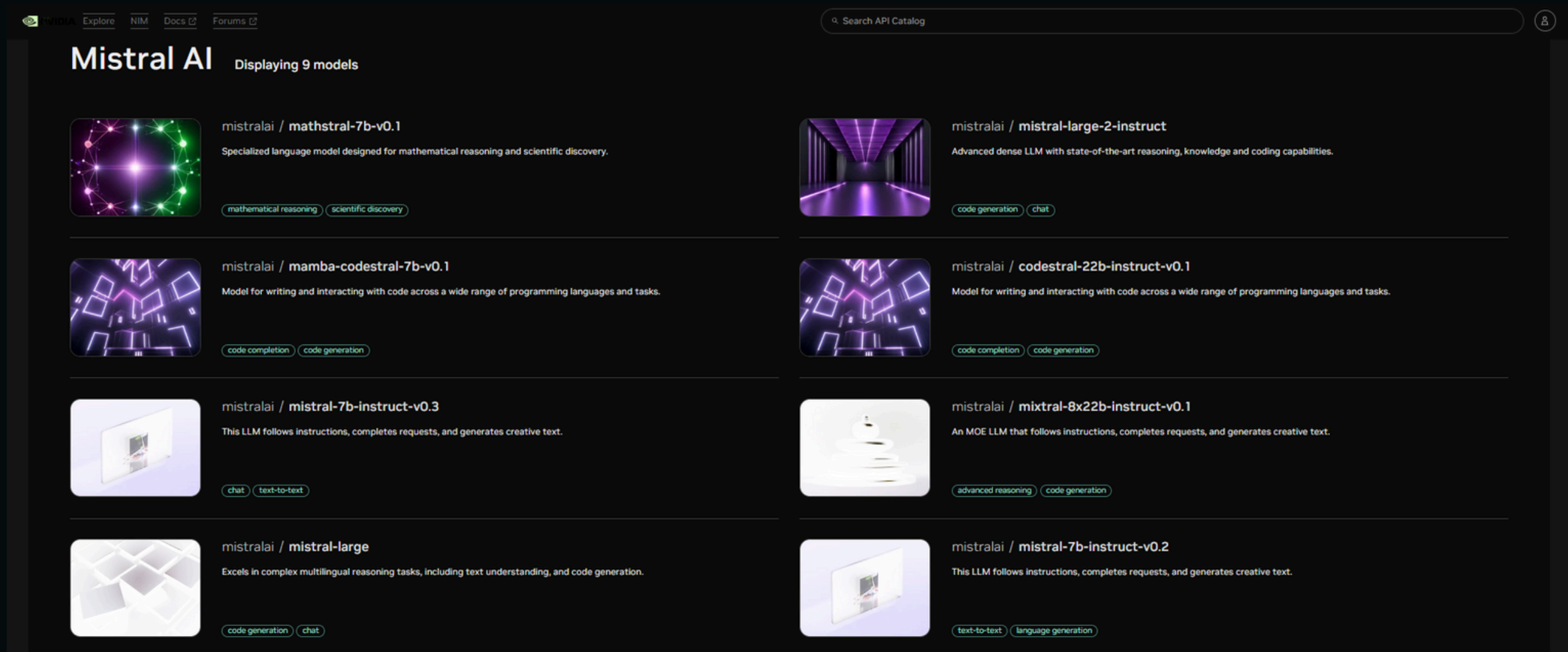


Your github files are visible in File Browser.

NVIDIA AI Workbench provides a default container with CUDA 12.2, Python 3.10, and JupyterLab installed. This container can be used as a starting point for creating a new project.

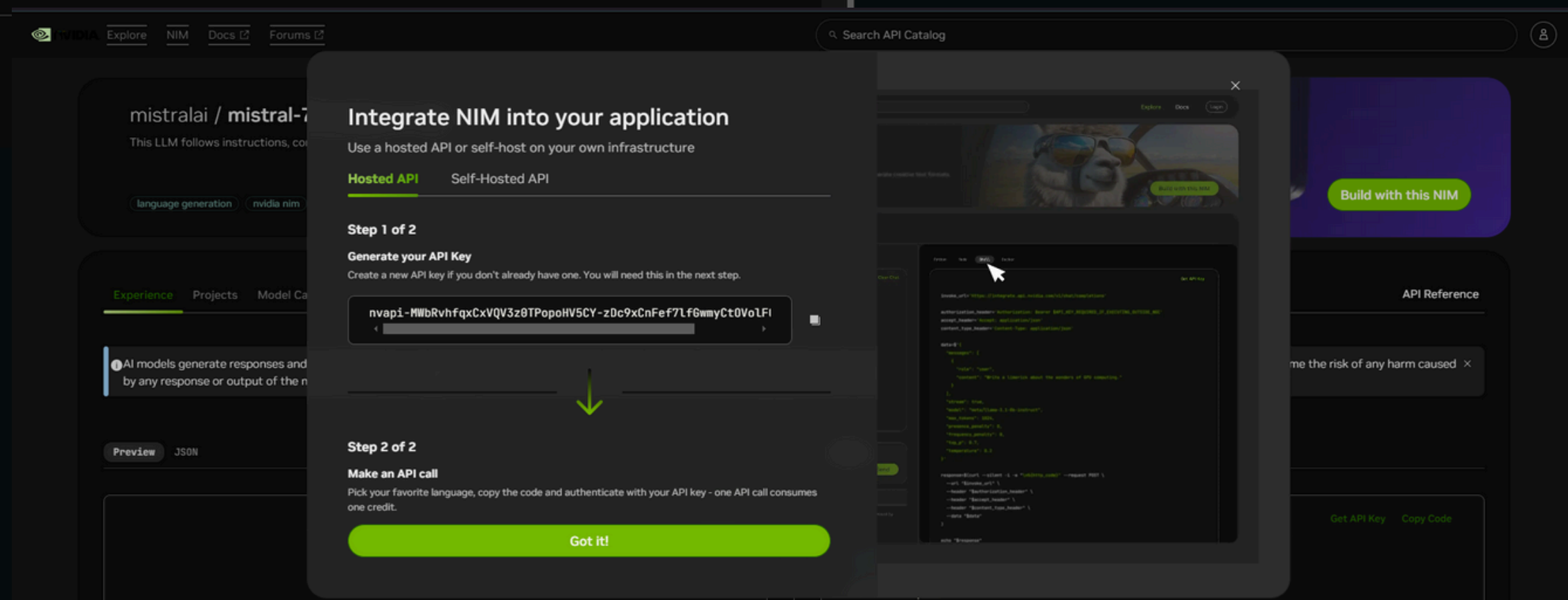


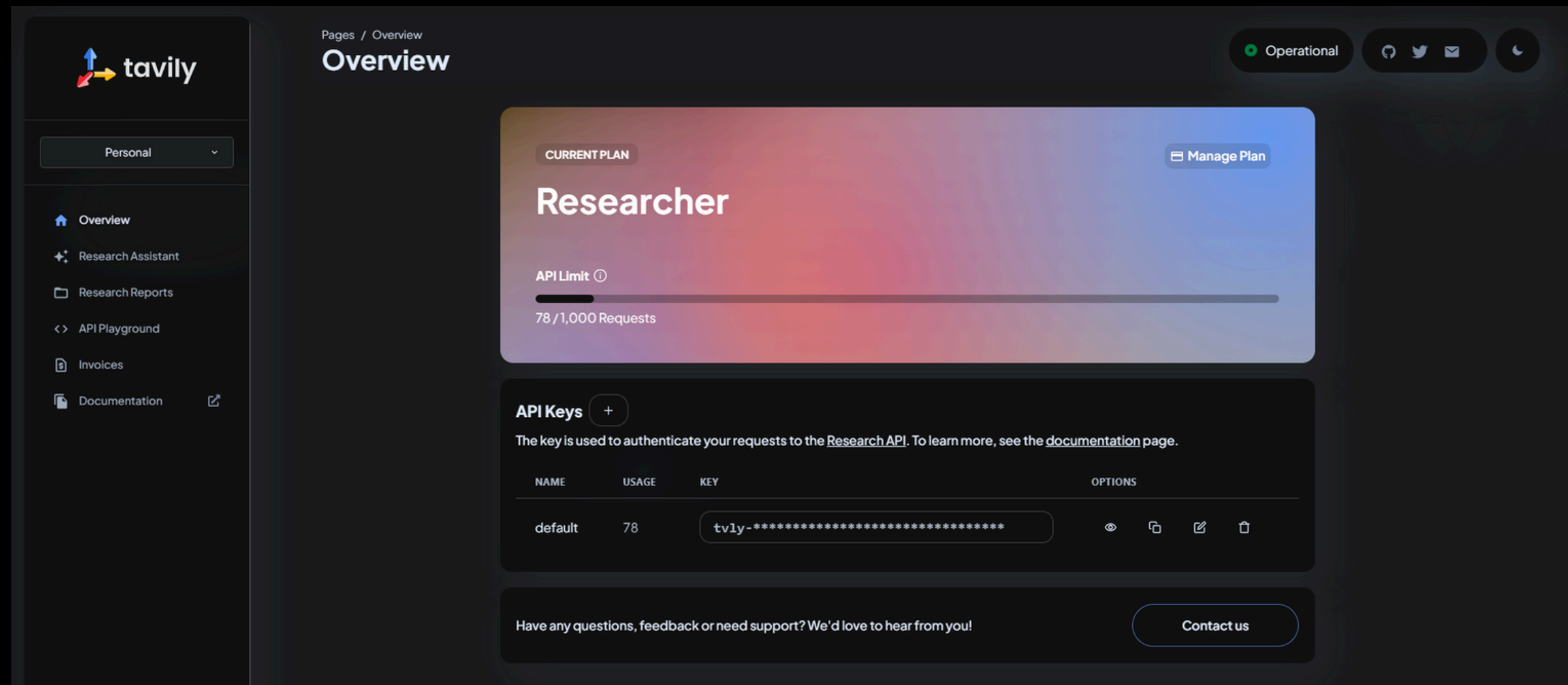




Mistral-7B-Instruct is a language model that can follow instructions, complete requests, and generate creative text formats. It is an instruct version of the Mistral-7B-v0.2 generative text model fine-tuned using a variety of publicly available conversation datasets.

NVIDIA\_API\_KEY :  
<https://build.nvidia.com/mistralai/mistral-7b-instruct-v2>





## Create an Account in Tavily:

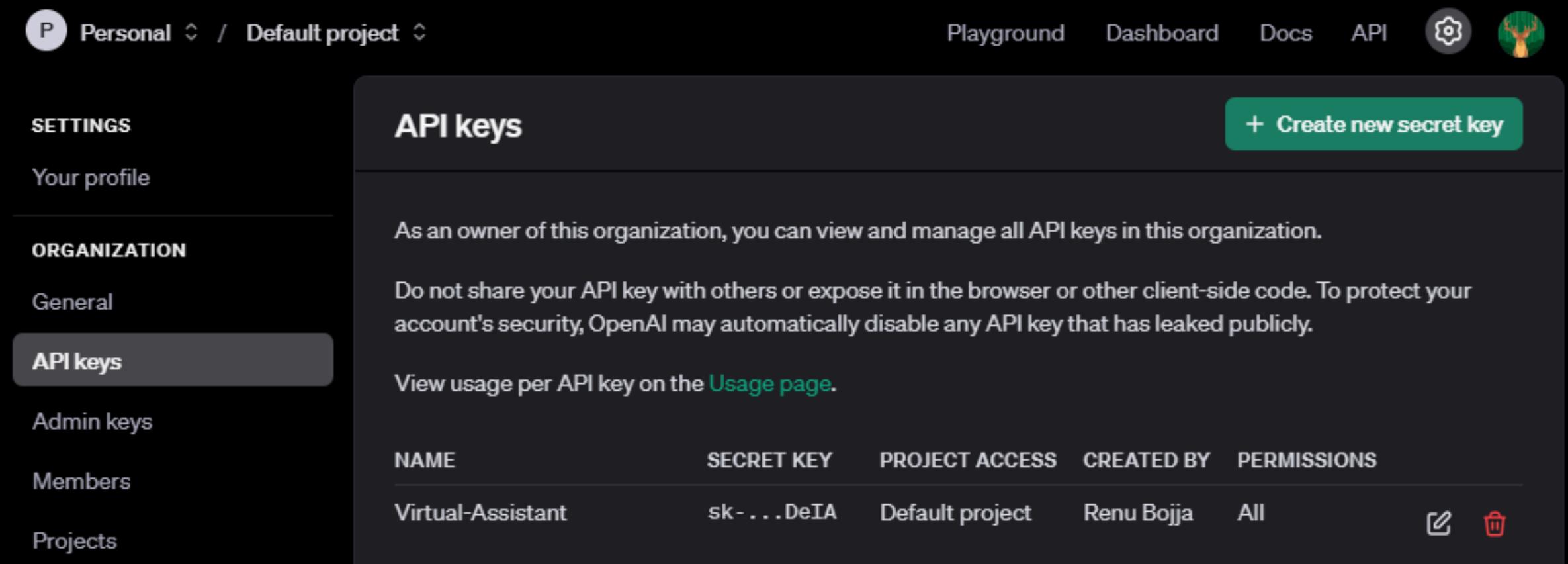
[https://app.tavily.com/home?code=wLiopT\\_LIUD\\_EsE5VOOHZri5V6DV80B1hqM5eHiFZlIFP&state=eyJy.ZXR1cm5Ubyl6li9ob21lIn0](https://app.tavily.com/home?code=wLiopT_LIUD_EsE5VOOHZri5V6DV80B1hqM5eHiFZlIFP&state=eyJy.ZXR1cm5Ubyl6li9ob21lIn0)

- Click on API Keys tab on the left
- Create a new secret key
- Each key has around 1000 times approx 100 days

## Why Tavily?

- Acts like a research assistant
- easily integrate search functionalities into their applications.
- searching, scraping, filtering and extracting the most relevant information from online sources. All in a single API call!

OPENAI\_API\_KEY : <https://platform.openai.com/settings/organization/api-key>



File Browser

Environment

Changes

History

Branches

Settings

Environment ● READY

Start Environment

Secrets

Environment variables for configuring sensitive information, like an API key. The name and description are stored and synced, but values are only stored locally. A Secret's value must be set every time the project is cloned to a new Location.

Add

| Name           | Value    | Description           |
|----------------|----------|-----------------------|
| NVIDIA_API_KEY | *****... | NVIDIA API Key        |
| TAVILY_API_KEY | *****... | Tavily Search API Key |
| OPENAI_API_KEY | *****... | OpenAI API Key        |

Base  
Packages  
Scripts  
Variables  
Secrets  
Mounts  
Apps  
Compose  
Hardware

**Come back to your Workbench and insert the relevant keys**

- Step 1 : Side bar go to environment
- Step 2: Insert the keys in relevant keys
- Step 3 :Start Environment



Click on Start Environment and then click on Open Control-panel  
Browser : <http://localhost:10000/projects/multimodal-virtual-assistant/applications/control-panel/>

localhost:10000/projects/multimodal-virtual-assistant/applications/control-panel/

# NVIDIA AI Workbench Virtual Assistant

Welcome

Hide All Settings

Welcome to your virtual assistant! I can help with all kinds of questions related to NVIDIA AI Workbench. Use me to troubleshoot, get started with, or to simply learn more about AI Workbench.

Press the button below to initialize the application. Happy chatting!

Initialize Virtual Assistant

Enter text and press ENTER

Clear history

What OS versions are supported by AI Workbench?

How do I get started with AI Workbench?

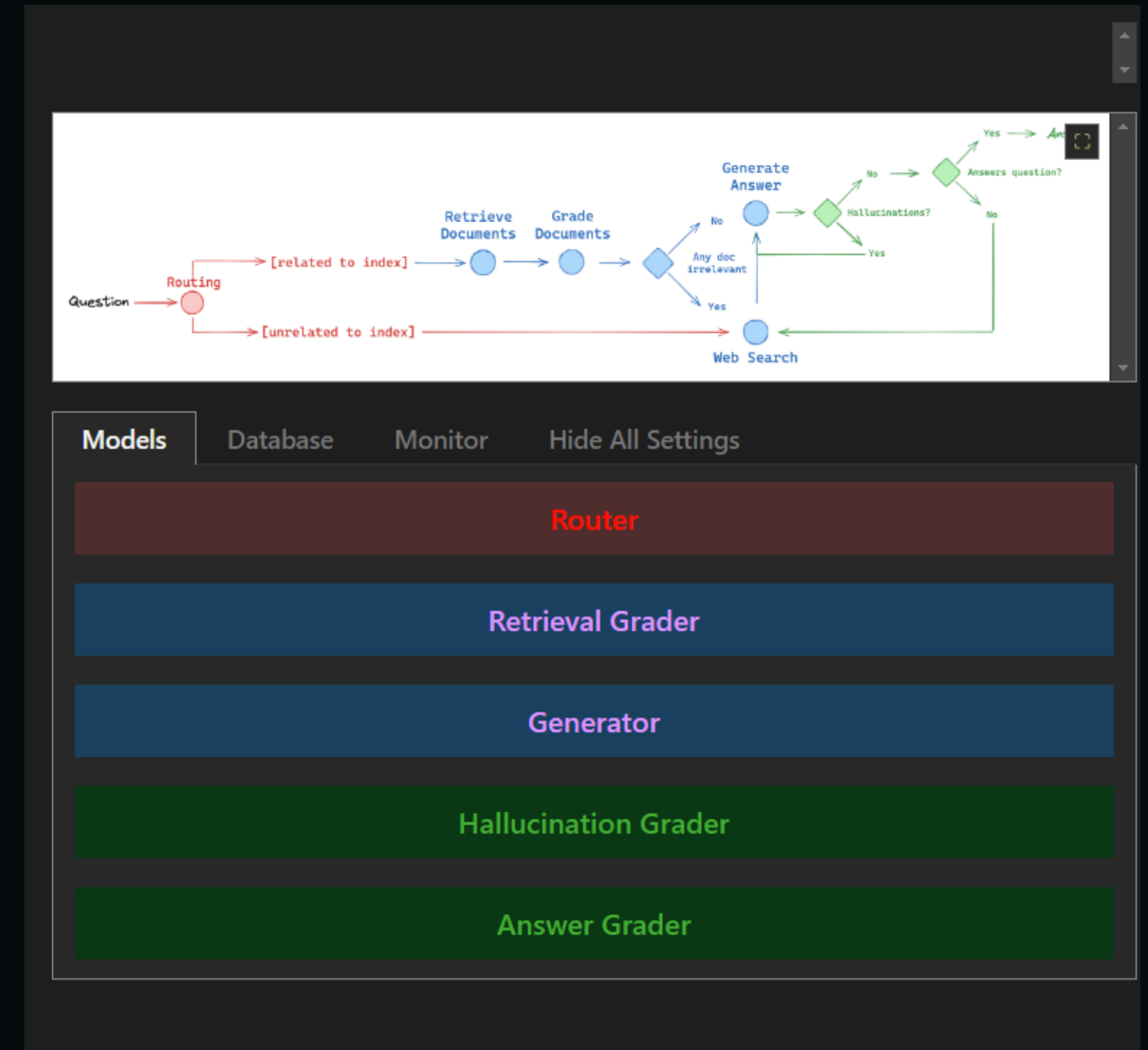
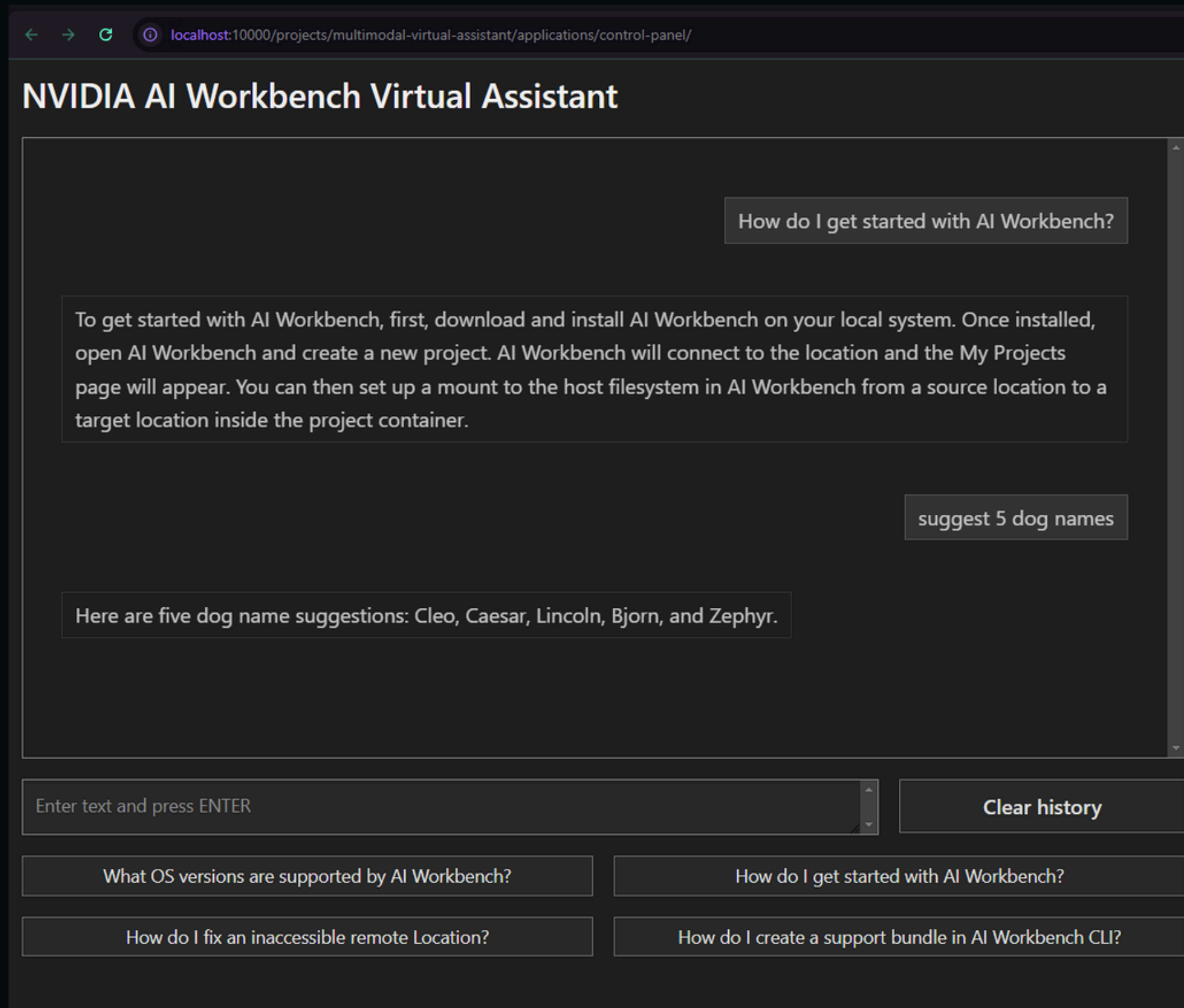
How do I fix an inaccessible remote Location?

How do I create a support bundle in AI Workbench CLI?

```
graph LR; Question --> Routing((Routing)); Routing -- "[related to index]" --> Retrieve((Retrieve Documents)); Routing -- "[unrelated to index]" --> WebSearch((Web Search)); Retrieve --> Grade((Grade Documents)); Grade --> AnyDocIrrelevant{Any doc irrelevant?}; AnyDocIrrelevant -- No --> GenerateAnswer((Generate Answer)); AnyDocIrrelevant -- Yes --> WebSearch; GenerateAnswer --> Hallucinations{Hallucinations?}; Hallucinations -- No --> AnswersQuestion{Answers question?}; Hallucinations -- Yes --> GenerateAnswer; AnswersQuestion -- Yes --> Answer[Answer]; AnswersQuestion -- No --> GenerateAnswer;
```

# A visual representation after the website is build

- Left side presents you the chatbot where you can insert the queries
- Right side shows the Rag pipeline on which this is been build



Question → Routing → [related to index] → Retrieve Documents → Grade Documents → Any doc irrelevant? → No → Generate Answer → Hallucinations? → No → Answers question? → Yes → Answer

Question → Routing → [unrelated to index] → Web Search → Generate Answer → Hallucinations? → Yes → Web Search

Models Database Monitor Hide All Settings

Upload webpages, pdfs, images, and videos to the vector store.  
**Note:** Clearing docs will empty the database!

Webpages File Upload

Webpage URLs

YouTube URLs

Here you can insert a you tube URL or a website URL to receive information and ask queries

As this is a multimodal chatbot you can also upload pdf ,image ,videos here

Question → Routing → [related to index] → Retrieve Documents → Grade Documents → Any doc irrelevant? → No → Generate Answer → Hallucinations? → No → Answers question? → Yes → Answer

Question → Routing → [unrelated to index] → Web Search → Generate Answer → Hallucinations? → Yes → Web Search

Models Database Monitor Hide All Settings

Upload webpages, pdfs, images, and videos to the vector store.  
**Note:** Clearing docs will empty the database!

Webpages File Upload

PDF Upload

Image Upload

Video Upload

Clear Docs