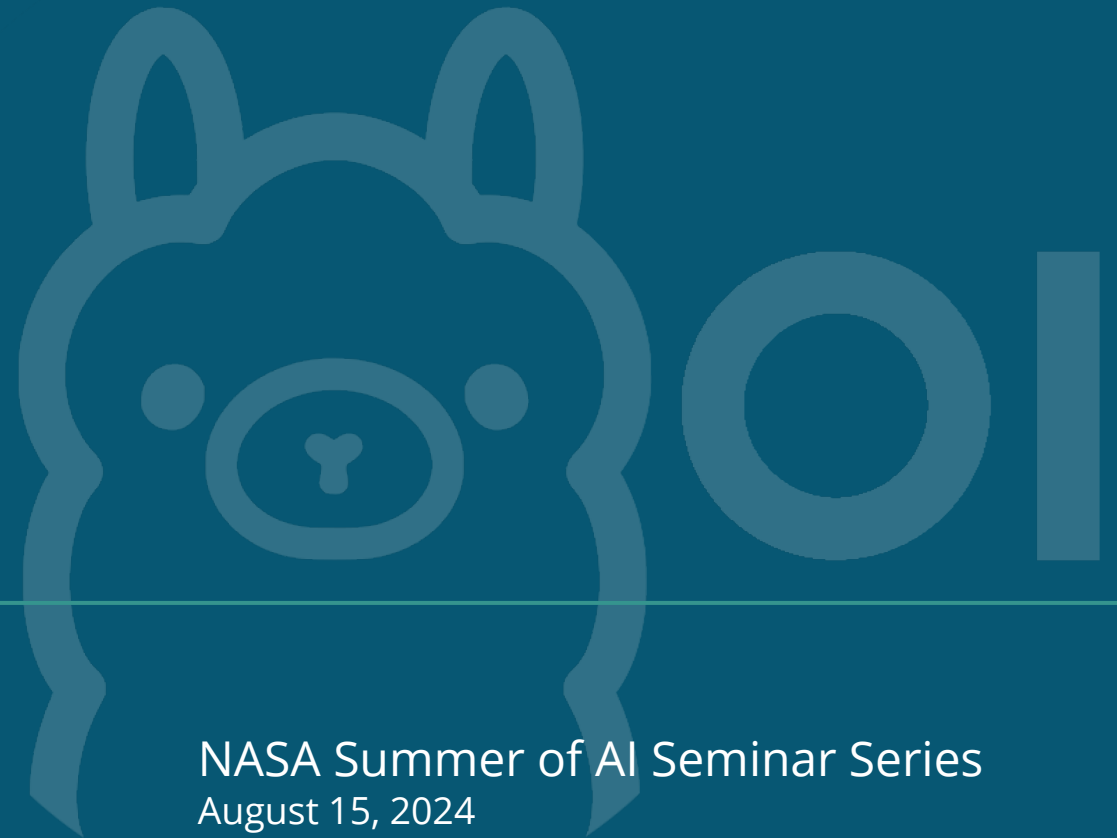


# GETTING STARTED WITH LOCAL LARGE LANGUAGE MODELS (LLMs) AND OPEN WEBUI: A STEP-BY-STEP GUIDE

Alexander Schepelmann, Ph.D.  
NASA Glenn Research Center



NASA Summer of AI Seminar Series  
August 15, 2024

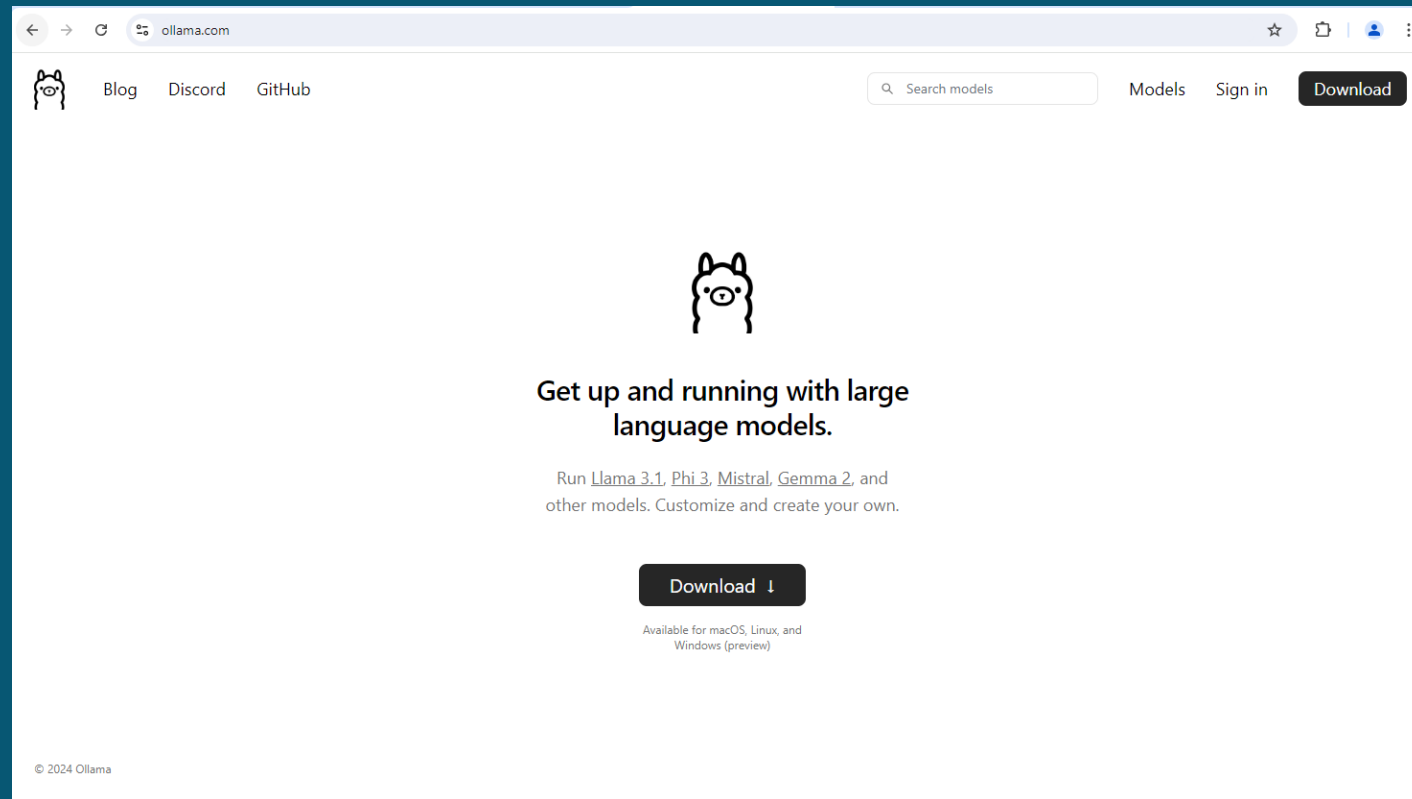


# LARGE LANGUAGE MODELS ARE NOT JUST CHATGPT!

- ChatGPT is the most popular LLM tool since launching in 2022.
- ChatGPT uses GPT series of LLMs, but there are many more!
- LLMs also don't need internet access and can run on your own PC.
- Private LLMs enable unlimited queries and more security.



# OLLAMA IS A TOOL FOR RUNNING LOCAL LLMs



*Ollama download screen at ollama.com.*



# ACCESS OLLAMA VIA THE COMMAND PROMPT

- Run *ollama*: `C:\Users\aschepel>ollama run llama2`

- Chat:

```
>>> Hello llama2! Please tell me about yourself!  
  
👉 Of course, friend! *adjusts sunglasses* I'm just an AI, but I'll do my best to  
give you a fun and interesting profile. 👉  
  
Well, I was trained on a massive dataset of text from the internet, so I have access  
to a vast amount of knowledge and information. *nods* But don't worry, I won't bore  
you with too much technical stuff! 👉 | ♂👉
```

- Exit *ollama*: `>>> /bye`

# OI

## OPEN WEBUI: OLLAMA WITH CHATGPT-LIKE INTERFACE

- Open WebUI can be installed via Docker (a containerization software)
- After Docker is installed, run the following command:

```
docker run -d -p 3000:8080 --gpus all --add-host=host.docker.internal:host-gateway -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:cuda
```

*\*This command is long and gross. Please DON'T TYPE THIS YOURSELF!  
Just head to <https://docs.openwebui.com/> and copy the command from there!*

- Simply access Open WebUI from your browser at *localhost:3000*



# INTERACT WITH YOUR DOCUMENTS, LOCALLY!

The screenshot shows a web browser window at localhost:3000. The interface is for a chat application using llama2:latest. A user has uploaded a document titled "2022\_optitracktm\_schepel...". The chat history shows a user message: "Please summarize the document for me." The AI response is as follows:

**OI llama2:latest**  
Sure! The document appears to be a technical report detailing the characterization of an infrared optical motion tracking system in NASA's Simulated Lunar Operations (SLOPE) Laboratory. Here is a summary of the main points:

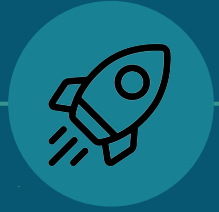
- The motion tracking system was calibrated using an active calibration wand and OptiTrack's Motive software package before each session.
- The calibration resulted in a mean overall wand error that quantifies the difference between the measured and known values.
- The document presents violin plots with interquartile ranges of mean cell error for "level-driving" scenarios on Days 1-3, indicating the variability of the motion tracking system's performance during these days.
- The table below the violin plots shows the mean cell error for each day, along with the minimum and maximum values observed in each cell.

Overall, the document provides information on the calibration and performance of an infrared optical motion tracking system in a laboratory setting, with a focus on the variability of its performance across different days.

1 2022\_optitrackTM\_schepelmann.pdf

The interface also shows a sidebar with "New Chat", "Workspace", and "Search" options. The user's name "alex" is visible in the bottom left corner.

*Llama2 summarizing a user-uploaded document in Open WebUI.*



## WHAT IF YOU HAVE NASA-SPECIFIC QUESTIONS?

- Uploading all relevant NASA documents is infeasible.
- Check out NASA's cloud-based Mission Cloud Platform (MCP)!

**THANK YOU!**

