AI IN ACTION

Building Your Essential AI Toolkit

Building a Personal LLM Agent

Yvonne Leung, PhD

Assistant Teaching Professor

College of Professional Studies

Northeastern University

Agenda or Table of Contents

1 Introduction to Open-Source Models (5 mins)

2 Building a Personal Local Agent (20 mins)

3 How to interact with it with examples (15 mins)

4 Q&A (20 mins)

AI IN ACTION
Building Your Essential AI Toolkit

Teaching Assistants

DHRUVI MODI

Computer Engineer | Junior Consultant

DhruvI is a master's student at Northeastern University. Dhruv is a computer engineer by training. She has expertise in the RAG based system evaluation using the RAGAS framework.

DHRUV PATEL

Data Science Intern | Junior Consultant

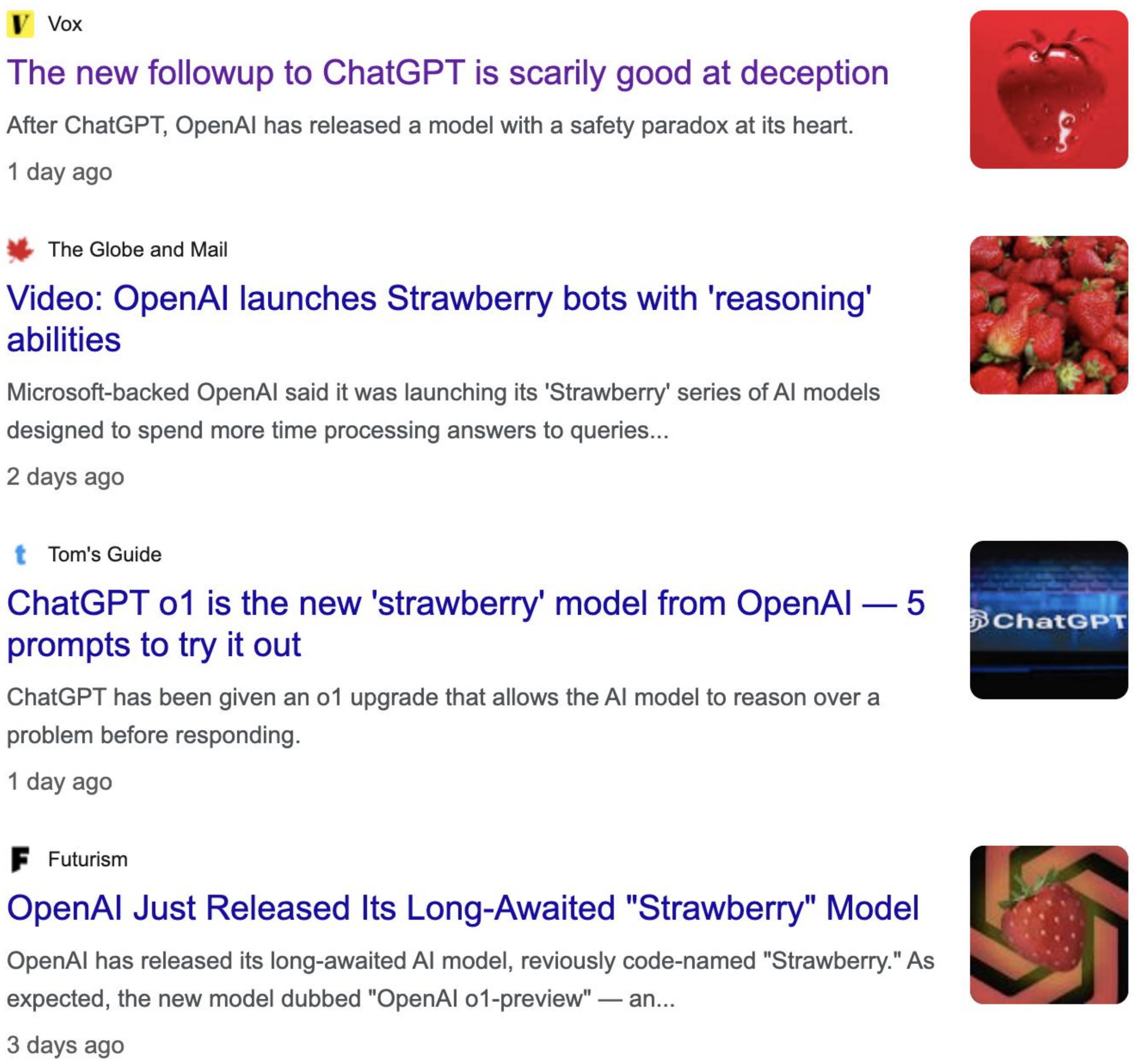
Dhruv is a master's student at Northeastern University. Dhruv is currently pursuing an internship at Ilcontent in the capacity of a data scientist.

TAPASWI SATYAPANTHI

Computer Engineer | Consultant

A Master's Student at Northeastern University, Tapaswi is a computer engineer by profession. He has expertise in the development of intelligent agents and RAGs using the Langchain Ecosystem and with various LLMs.

AGI is nearer!?



Vox

The new followup to ChatGPT is scarily good at deception

After ChatGPT, OpenAI has released a model with a safety paradox at its heart.

1 day ago

The Globe and Mail

Video: OpenAI launches Strawberry bots with 'reasoning' abilities

Microsoft-backed OpenAI said it was launching its 'Strawberry' series of AI models designed to spend more time processing answers to queries...

2 days ago

Tom's Guide

ChatGPT o1 is the new 'strawberry' model from OpenAI — 5 prompts to try it out

ChatGPT has been given an o1 upgrade that allows the AI model to reason over a problem before responding.

1 day ago

Futurism

OpenAI Just Released Its Long-Awaited "Strawberry" Model

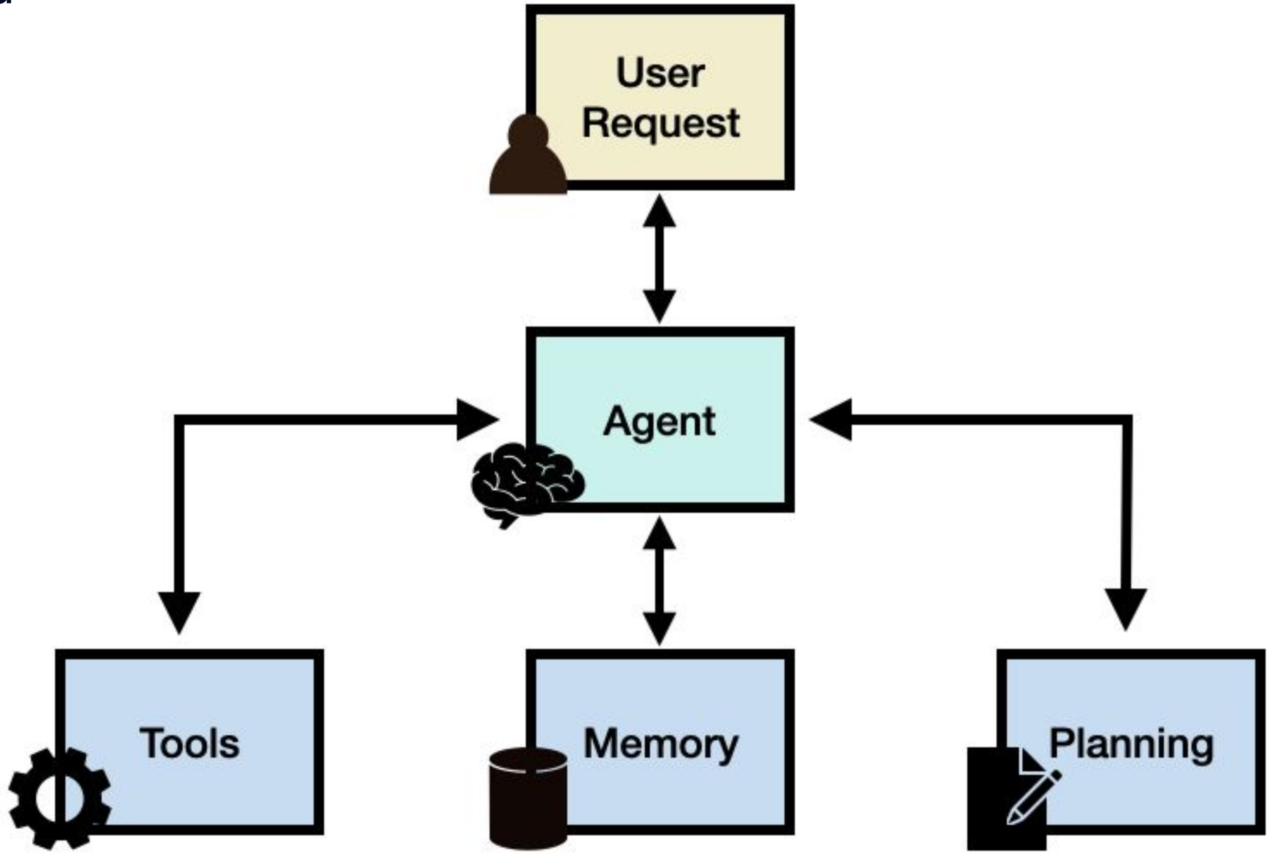
OpenAI has released its long-awaited AI model, reviously code-named "Strawberry." As expected, the new model dubbed "OpenAI o1-preview" — an...

3 days ago

What is an agent?

An advanced AI systems designed for creating complex text that needs sequential reasoning

- User Request - a user question or request.
- Agent/Brain - the agent core acting as coordinator.
- Planning - assists the agent in planning future actions.



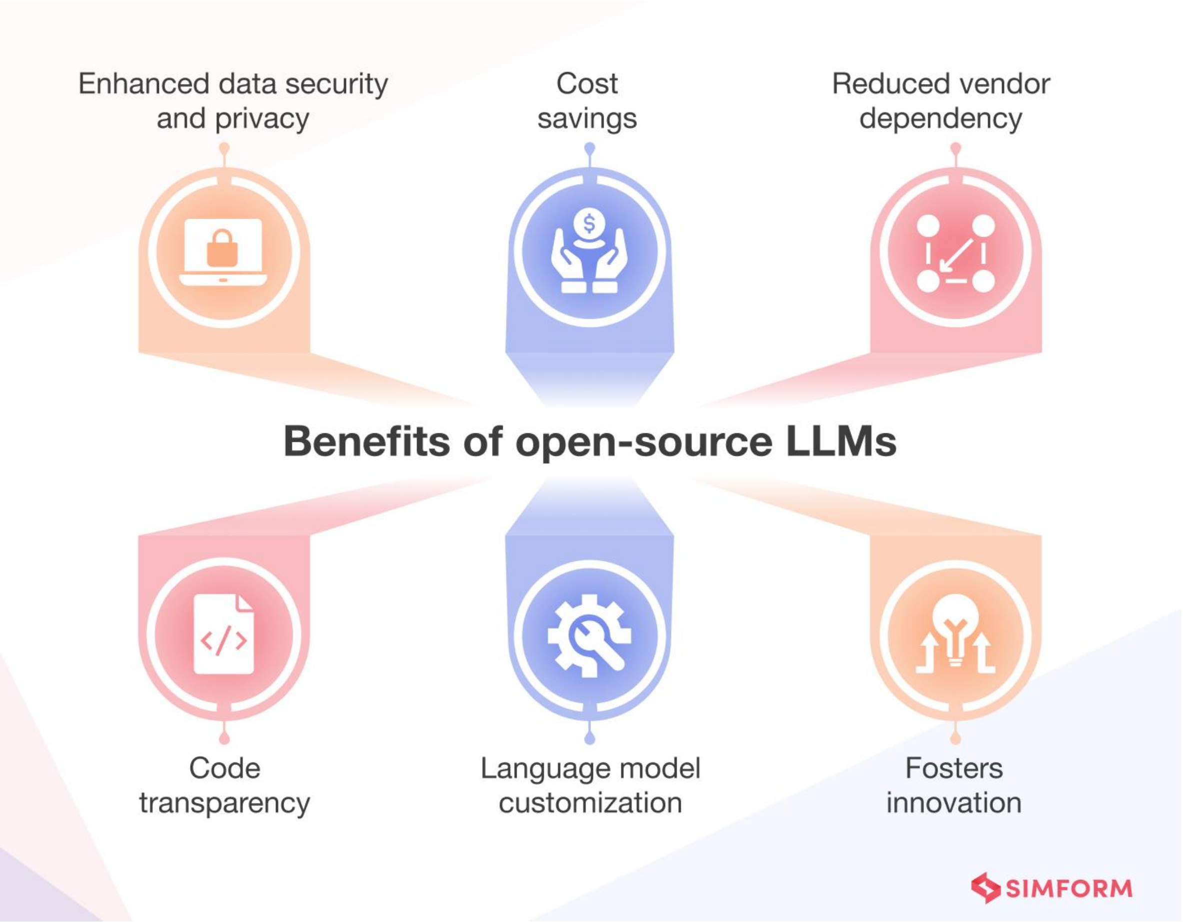
scripts calling

Keep track of action taken

Create action steps

An Overview of Open-Source Models

- Llama 3.1 by Meta (Best)
- Phi 3 by Microsoft (Very Good)



Enhanced data security and privacy

Cost savings

Reduced vendor dependency

Benefits of open-source LLMs

Code transparency

Language model customization

Fosters innovation

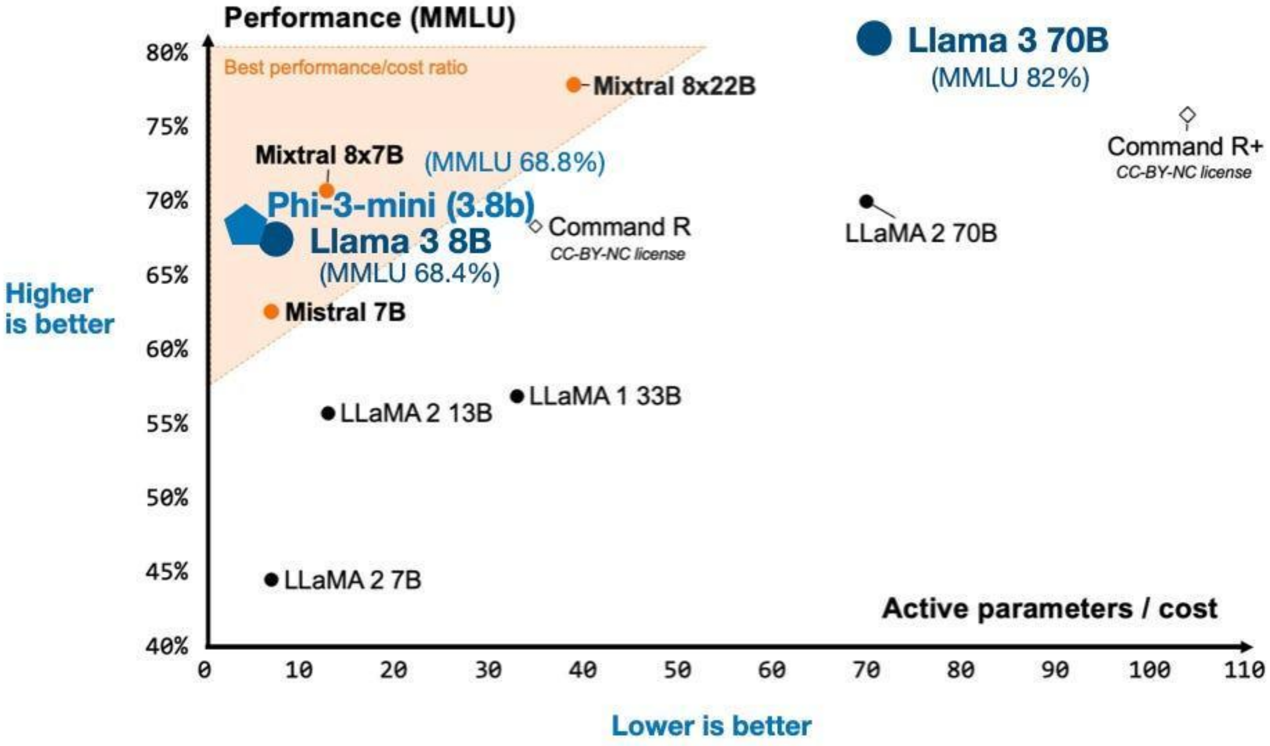
SIMFORM

https://medium.com/@sumudithalanz/unlocking-seamless-access-how-to-harness-your-self-hosted-llm-anywhere-with-ollama-web-ui-0ef687aae604

5 Leading Small Language Models of 2024

Llama 3	Meta	8 billion parameters
Phi-3	Microsoft	3.8 billion - 7 billion parameters
Gemma	Google	2 billion - 7 billion parameters
Mixtral 8x7B	Mistral AI	7 billion parameters
OpenELM	Apple	0.27 billion - 3 billion parameters

datasciencedojo
data science for everyone



Llama 3.1 can fit into 6gb of RAM

Resources

Laptop requirements: a system with at least 8GB of RAM and a modern multi-core processor is recommended

- Download Ollama – a llm platform: https://ollama.com/
- Download docker container to run ollama: https://www.docker.com/products/docker-desktop/
- Pull lightweight LLM: https://ollama.com/library/phi3
- Run a docker image of an Web UI- documentation: https://github.com/open-webui/open-webui

Ollama Web UI Overview

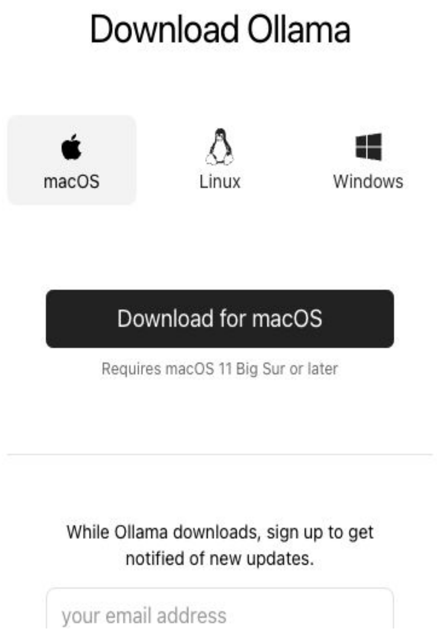
- **Ollama =** Open-source Large Language Model Management platform.
- **Local RAG Integration**: Enhances chat with Retrieval Augmented Generation.
- **Web Search for RAG**: Uses multiple providers for integrated search results, e.g. serper, Serply, DuckDuckGo
- **Web Browsing**: Integrates websites into chats with # command followed by a URL.
- **Image Generation**: Adds dynamic visuals using various APIs, e.g. ComfyUI, OpenAI.
- **Multiple Models**: Engages multiple models for better responses.
- **RBAC**: Restricts access to authorized users and admins, i.e. local log-in
- **Multilingual Support**: Offers Open WebUI in various languages, seeks contributors.
- **Pipeline building** and much more…

Demonstration

Summarize email threads and read a financial statement

Exercise Time

Download Ollama Directly



Download Ollama

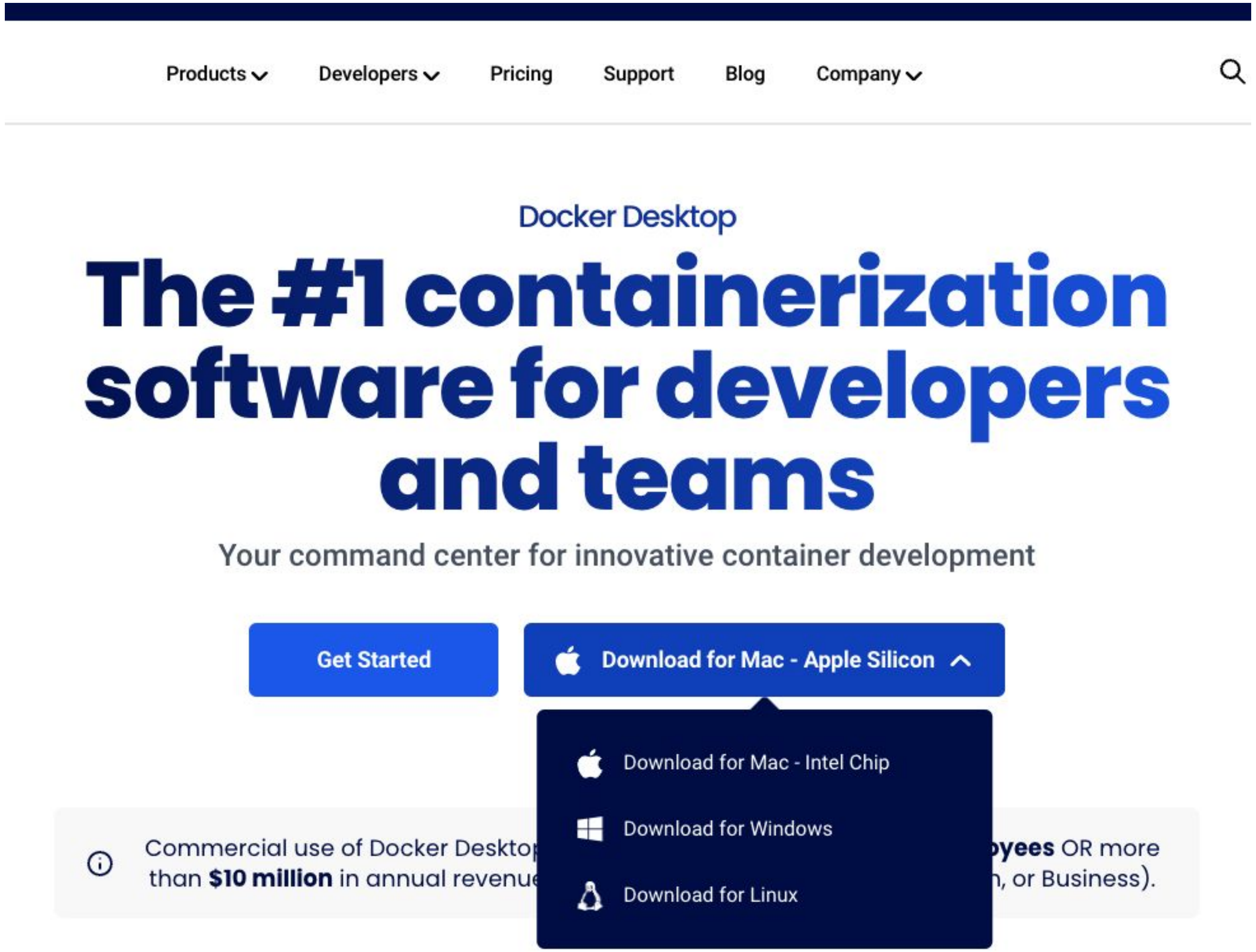
Requires macOS 11 Big Sur or later

While Ollama downloads, sign up to get notified of new updates.

your email address

Docker

- Docker is your resource manager
- Contain ollama + web ui within your system
- Download: https://www.docker.com/products/docker-desktop/
- **Open it on your mac**



Docker Desktop

The #1 containerization software for developers and teams

Your command center for innovative container development

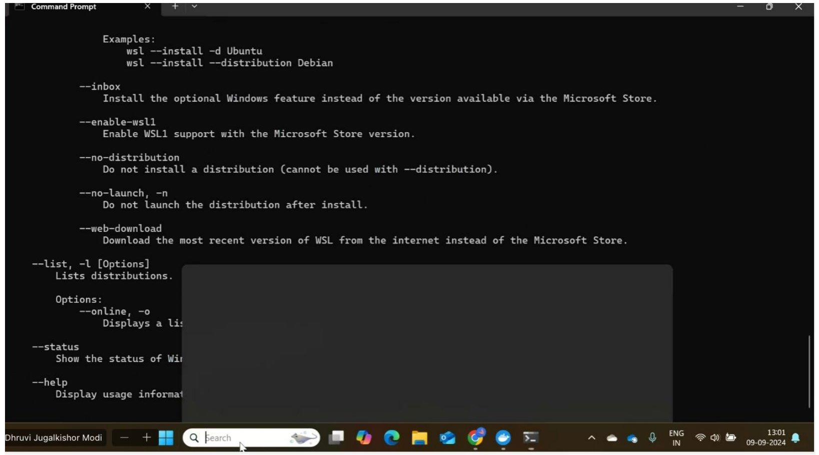
Get Started
Download for Mac - Apple Silicon
Download for Mac - Intel Chip
Download for Windows
Download for Linux

Commercial use of Docker Deskto than **$10 million** in annual revenu

oyees OR more n, or Business).

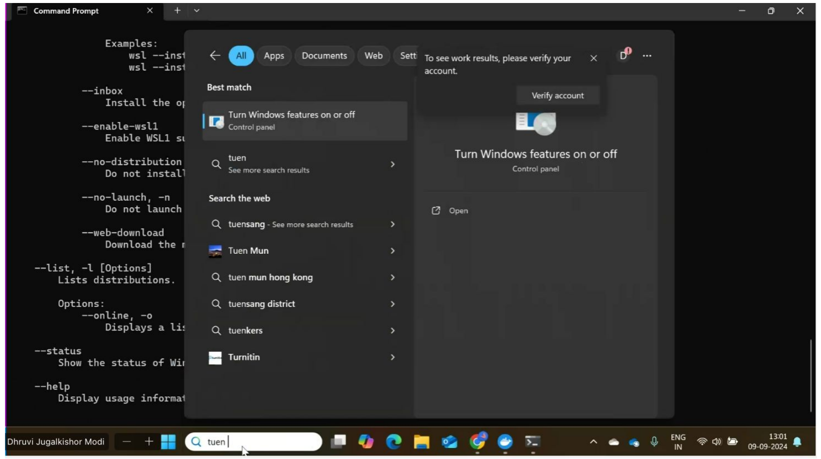
Extra Settings in Windows

1. Go to your search bar



```
Command Prompt

        Examples:
            wsl --install -d Ubuntu
            wsl --install --distribution Debian

    --inbox
        Install the optional Windows feature instead of the version available via the Microsoft Store.

    --enable-wsl1
        Enable WSL1 support with the Microsoft Store version.

    --no-distribution
        Do not install a distribution (cannot be used with --distribution).

    --no-launch, -n
        Do not launch the distribution after install.

    --web-download
        Download the most recent version of WSL from the internet instead of the Microsoft Store.

--list, -l [Options]
    Lists distributions.

    Options:
        --online, -o
            Displays a lis

--status
    Show the status of Wir

--help
    Display usage informat
```

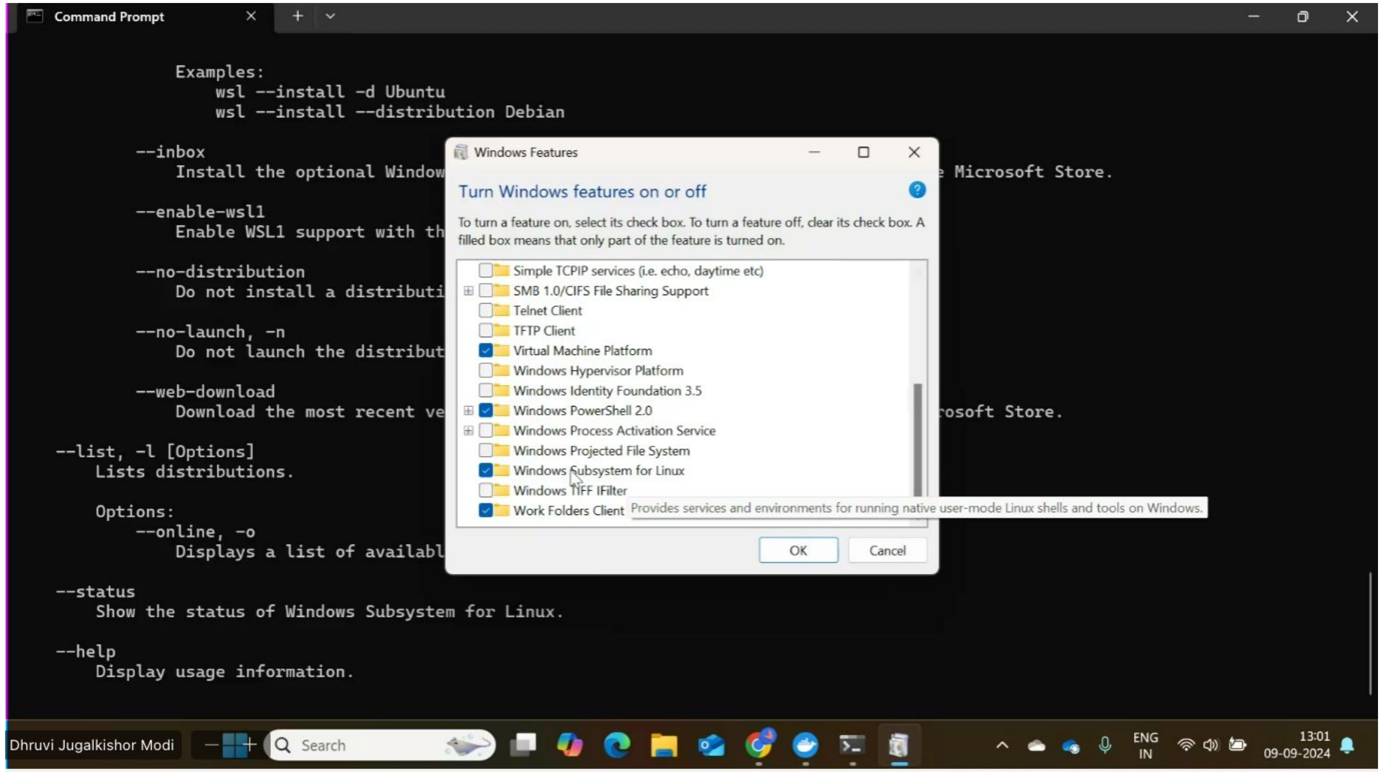
2. Type “Turn Windows features on or off”



```
Command Prompt

        Examples:
            wsl --inst
            wsl --inst

    --inbox
        Install the op

    --enable-wsl1
        Enable WSL1 su

    --no-distribution
        Do not instal

    --no-launch, -n
        Do not launch

    --web-download
        Download the r

--list, -l [Options]
    Lists distributions.

    Options:
        --online, -o
            Displays a lis

--status
    Show the status of Wir

--help
    Display usage informat
```

To see work results, please verify your account.

3. Check Windows Subsystem for Linux



```
Command Prompt

        Examples:
            wsl --install -d Ubuntu
            wsl --install --distribution Debian

    --inbox
        Install the optional Window                                   e Microsoft Store.

    --enable-wsl1
        Enable WSL1 support with th

    --no-distribution
        Do not install a distributi

    --no-launch, -n
        Do not launch the distribut

    --web-download
        Download the most recent ve                                   osoft Store.

--list, -l [Options]
    Lists distributions.

    Options:
        --online, -o
            Displays a list of availabl

--status
    Show the status of Windows Subsystem for Linux.

--help
    Display usage information.
```

Windows Features

Turn Windows features on or off

To turn a feature on, select its check box. To turn a feature off, clear its check box. A filled box means that only part of the feature is turned on.

- ☐ Simple TCPIP services (i.e. echo, daytime etc)
- ☐ SMB 1.0/CIFS File Sharing Support
- ☐ Telnet Client
- ☐ TFTP Client
- ☑ Virtual Machine Platform
- ☐ Windows Hypervisor Platform
- ☐ Windows Identity Foundation 3.5
- ☑ Windows PowerShell 2.0
- ☐ Windows Process Activation Service
- ☐ Windows Projected File System
- ☑ Windows Subsystem for Linux
- ☐ Windows TIFF IFilter
- ☑ Work Folders Client

Provides services and environments for running native user-mode Linux shells and tools on Windows.

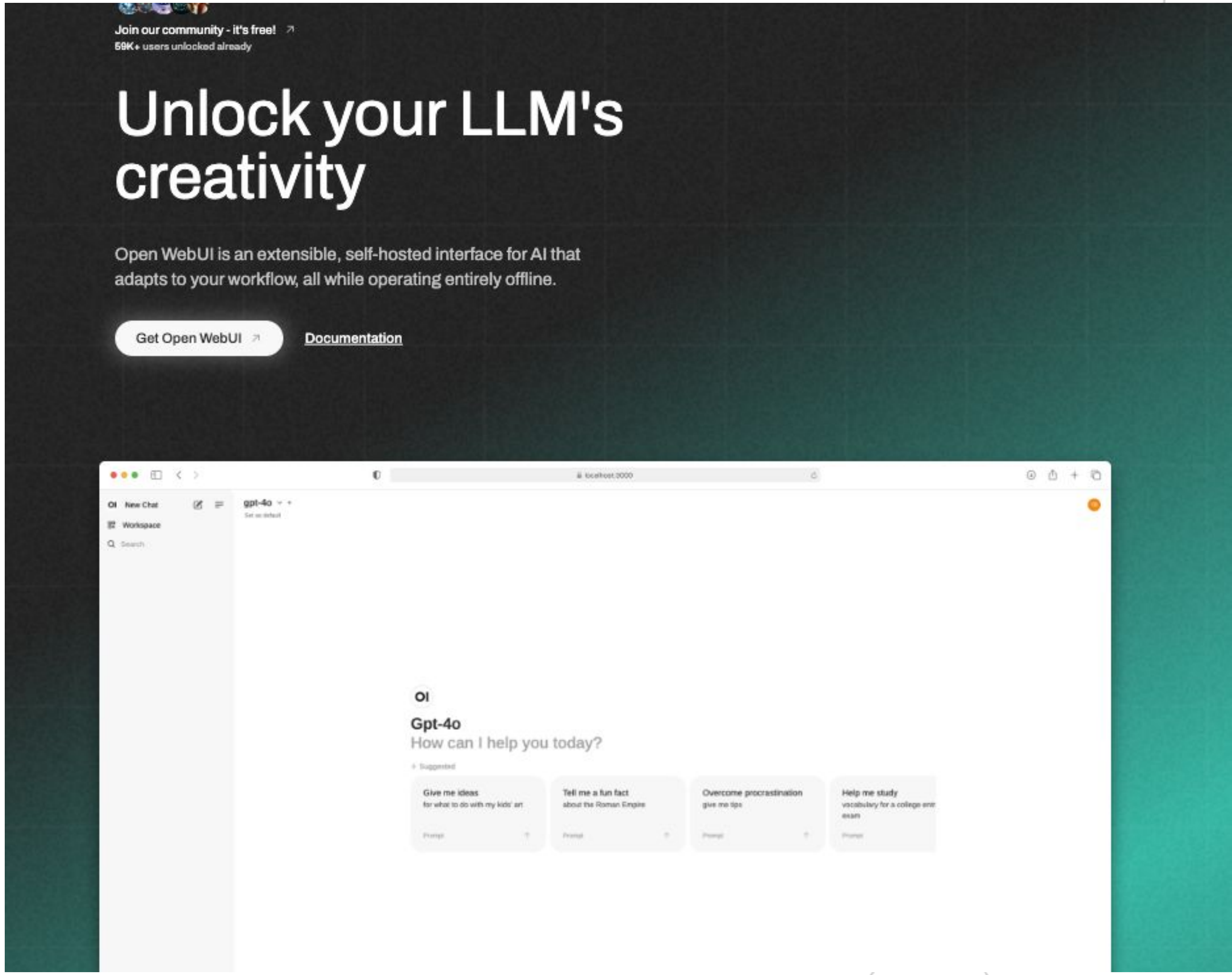
AI IN ACTION
Building Your Essential AI Toolkit

https://youtu.be/XgRGI0Pw2mM?si=GPOQbHl6ib884k6N

Pinokio alternative: https://www.youtube.com/watch?v=VbfHAHCAYT4

Open WebUI

- Docker will run Open WebUI
- Documentation: https://github.com/open-webui/open-webui
- Copy and paste this line to your terminal (mac): docker run -d -p 3000:8080 --add host=host.docker.internal:host-gateway -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:main

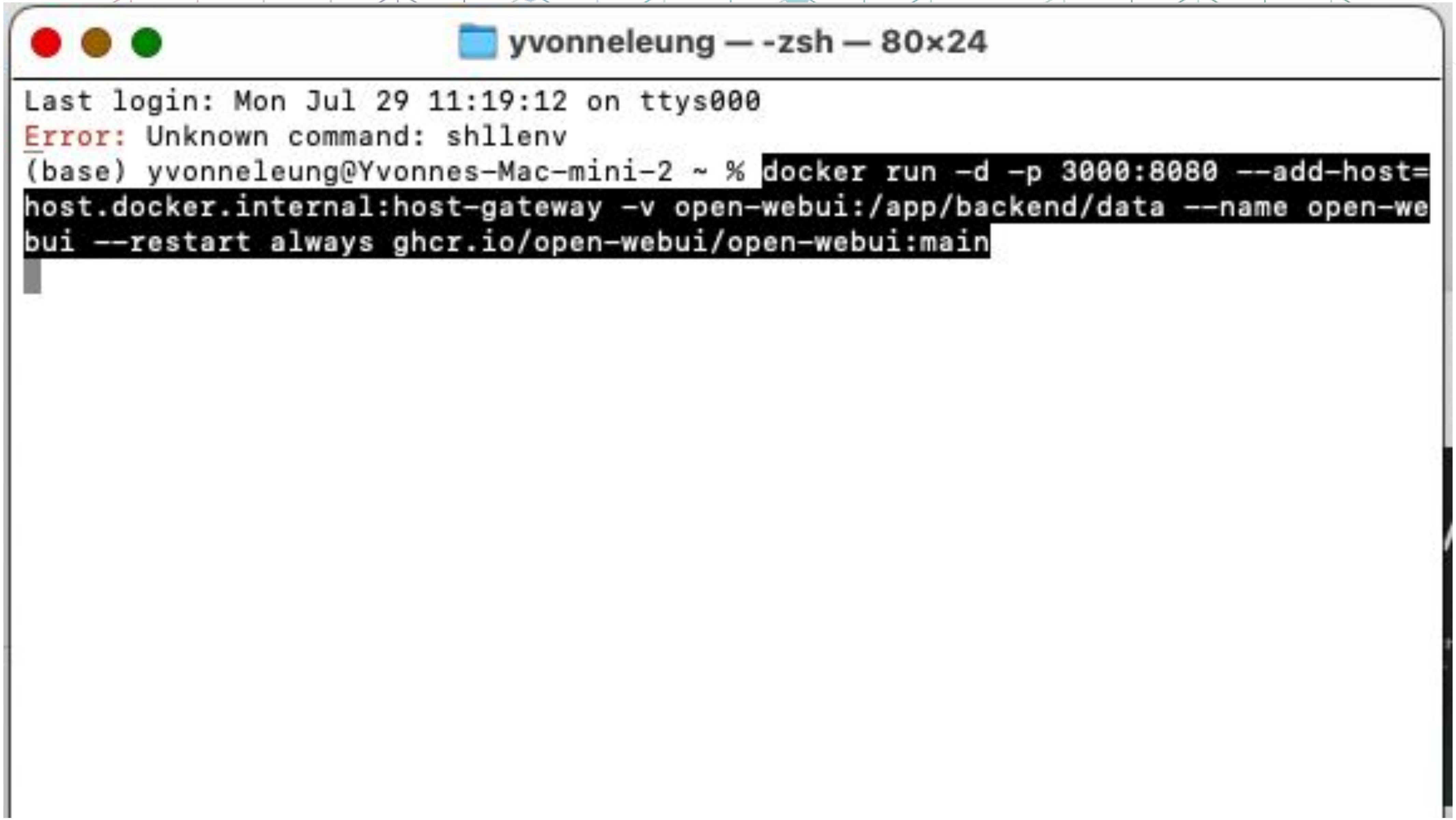


Join our community - it's free!

59K+ users unlocked already

Unlock your LLM's creativity

Open WebUI is an extensible, self-hosted interface for AI that adapts to your workflow, all while operating entirely offline.

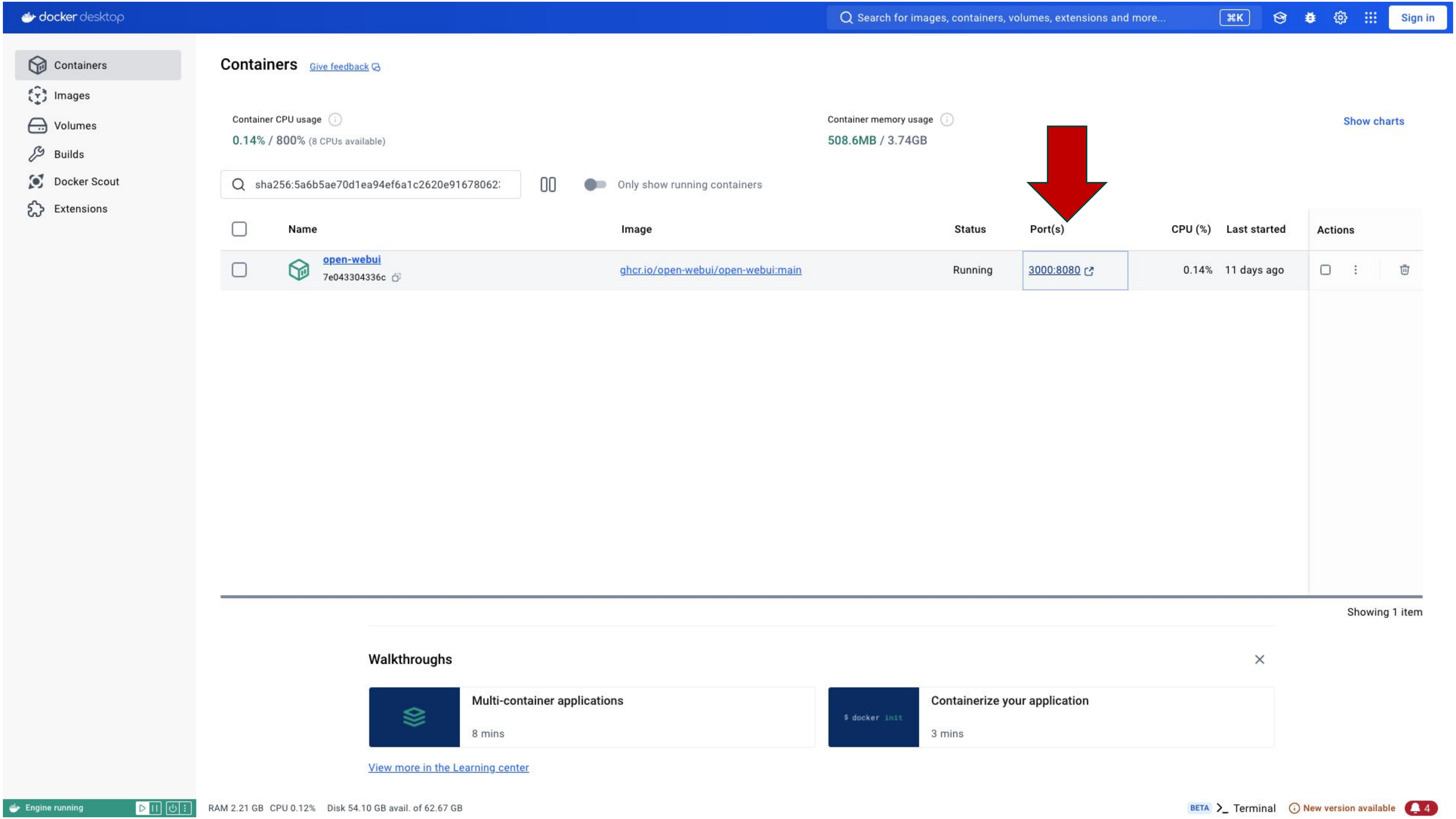


```
Last login: Mon Jul 29 11:19:12 on ttys000
Error: Unknown command: shllenv
(base) yvonneleung@Yvonnes-Mac-mini-2 ~ % docker run -d -p 3000:8080 --add-host=
host.docker.internal:host-gateway -v open-webui:/app/backend/data --name open-we
bui --restart always ghcr.io/open-webui/open-webui:main
```

GPU support (Windows)

In your command prompt, please type:

```
docker run -d -p 3000:8080 --gpus=all -v ollama:/root/.ollama -v
open-webui:/app/backend/data --name open-webui --restart always
ghcr.io/open-webui/open-webui:ollama
```



Containers

Container CPU usage
0.14% / 800% (8 CPUs available)

Container memory usage
508.6MB / 3.74GB

sha256:5a6b5ae70d1ea94ef6a1c2620e91678062:

□	Name	Image	Status	Port(s)	CPU (%)	Last started	Actions
□	open-webui 7e043304336c	ghcr.io/open-webui/open-webui:main	Running	3000:8080	0.14%	11 days ago	

Showing 1 item

Multi-container applications
8 mins

$ docker init

Containerize your application
3 mins

RAM 2.21 GB CPU 0.12% Disk 54.10 GB avail. of 62.67 GB

Pull a light-weight model

In your terminal, type:
ollama pull phi3

Alternatively,
Open Web UI > Settings > Admin Settings >Models

https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/phi-3-vision-catalyzing-multimodal-innovation/ba-p/4170251

AI IN ACTION
Building Your Essential AI Toolkit

Enable Google Search as a tool

Personal Search Engine Setting:

- https://programmablesearchengine.google.com/controlpanel/overview?cx=114db67d48a6742bc

Get an API Key:

- https://developers.google.com/custom-search/v1/introduction
- https://console.cloud.google.com/apis/credentials/key/f2d1963b-0bc7-43ab-9edf-470290352fe1?authuser=0&project=omega-iterator-384117

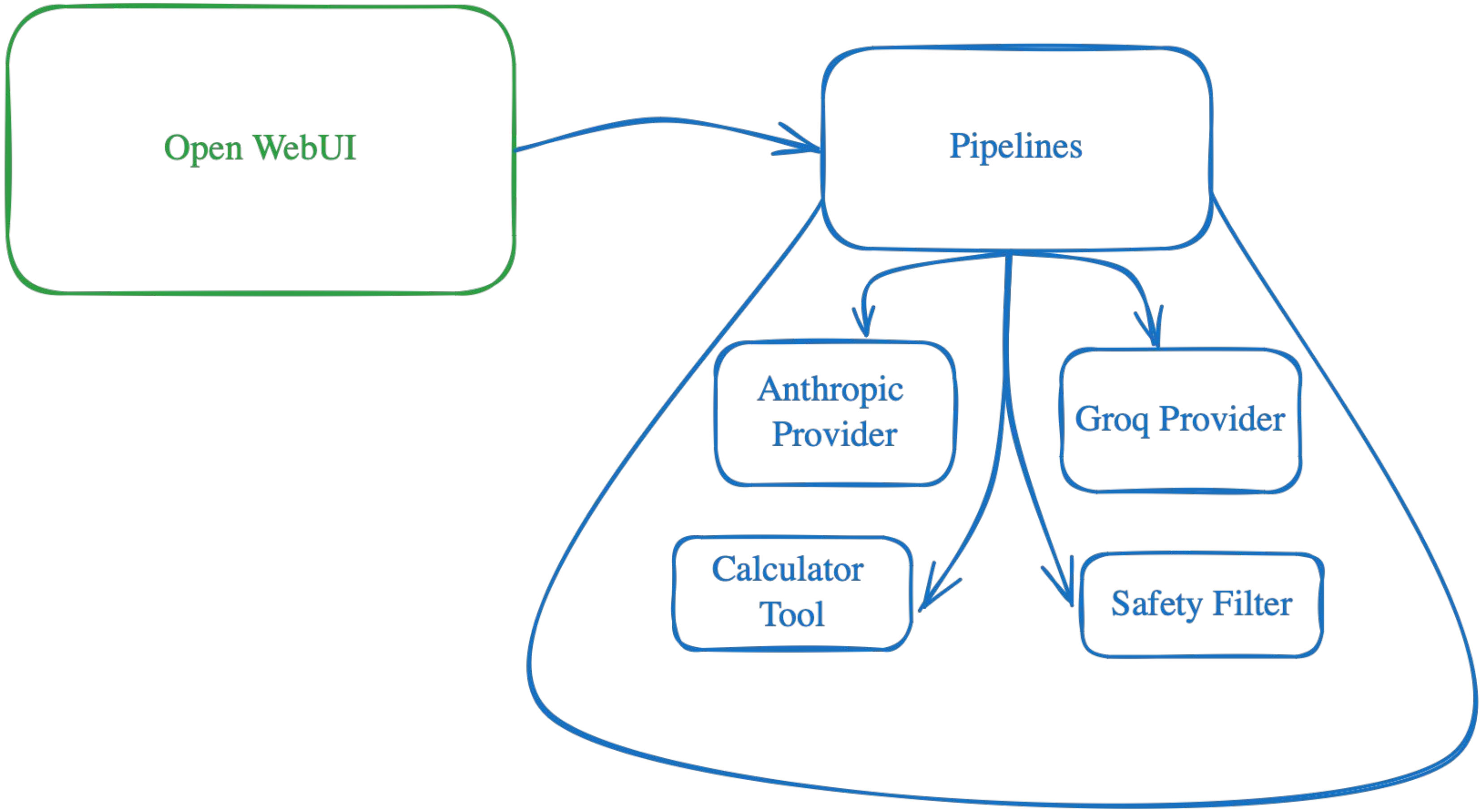
Agent and Chain of Thought Reasoning

Discover a model: the multi-agent
https://openwebui.com/m/stewart/multi-agent:latest

Discover a function: the Reflection function
https://openwebui.com/m/stewart/multi-agent:latest

Build a pipeline using Open WebUI

- Tools
- Functions
- LLMs
- Prompts



AI IN ACTION
Building Your Essential AI Toolkit

Demonstration

Write an app in python code

Conclusions

- We understand what are Ollama + Web UI + Docker and how they work
- We demonstrated how to summarize many emails
- We asked Phi3 to read a corporate financial statements
- We built an agent with reflective reasoning and taking action capabilities
- In a no-code environment, there are many options to customize your personal llm based chatbots depending on your needs!

Q&A

Why Ollama is safe?

1. **Local Data Storage**: Ollama's models run locally, meaning all user-generated data is stored on your device. This approach enhances data privacy and security by eliminating the need to transmit data over the internet[1].
2. **Offline Operation**: Open WebUI is designed to operate entirely offline, which further reduces the risk of data breaches or unauthorized access[2].
3. **Customization and Control**: Running these applications locally allows for greater customization and control over your data and interactions, making them a compelling choice for developers and enterprises[3].
4. **Compatibility and Efficiency**: Both Ollama and Open WebUI are compatible with various large language models and can leverage local hardware, such as GPUs, to improve processing efficiency[4].

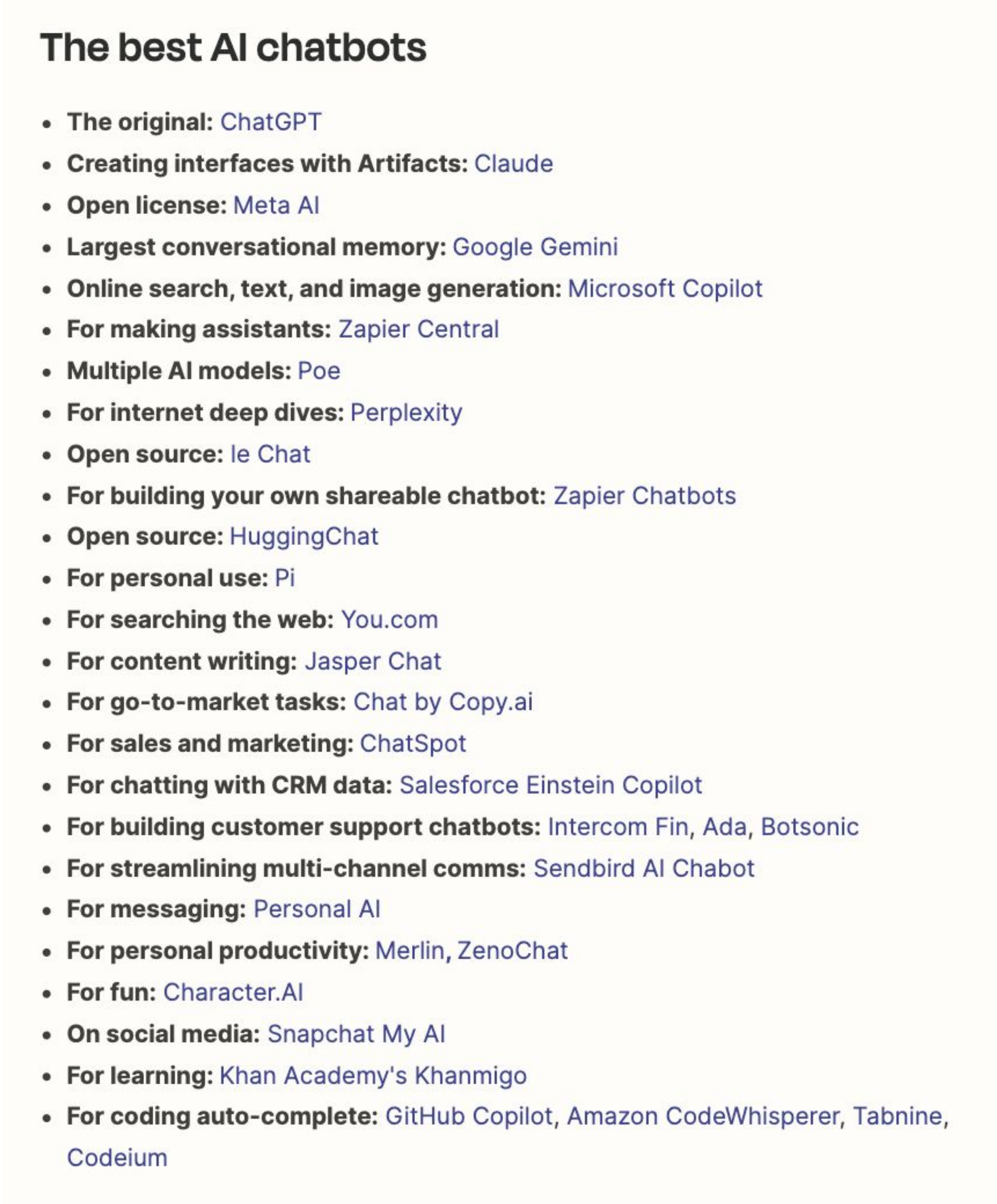
References:
https://github.com/open-webui/open-webui/blob/main/README.md

https://ai-box.eu/large-language-models/open-webui-ollama/1237/

https://dev.to/tylerjrbuell/supercharge-your-productivity-with-ollama-open-web-ui-and-large-language-models-51eo

https://namrata23.medium.com/run-llms-locally-or-in-docker-with-ollama-ollama-webui-379029060324

https://www.youtube.com/watch?v=-qMvSi_BGOc

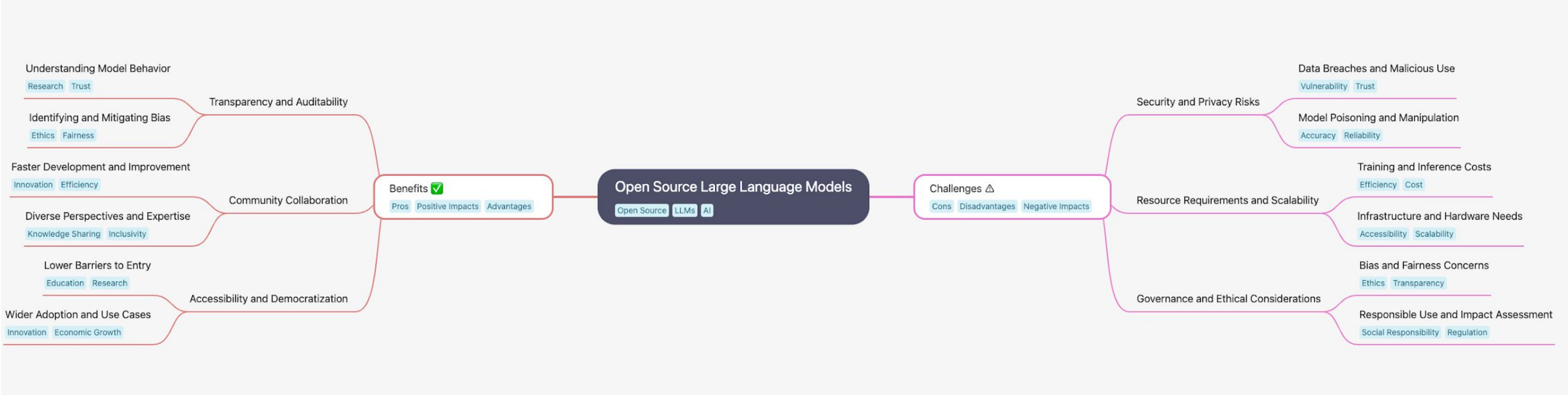


The best AI chatbots

- **The original:** ChatGPT
- **Creating interfaces with Artifacts:** Claude
- **Open license:** Meta AI
- **Largest conversational memory:** Google Gemini
- **Online search, text, and image generation:** Microsoft Copilot
- **For making assistants:** Zapier Central
- **Multiple AI models:** Poe
- **For internet deep dives:** Perplexity
- **Open source:** le Chat
- **For building your own shareable chatbot:** Zapier Chatbots
- **Open source:** HuggingChat
- **For personal use:** Pi
- **For searching the web:** You.com
- **For content writing:** Jasper Chat
- **For go-to-market tasks:** Chat by Copy.ai
- **For sales and marketing:** ChatSpot
- **For chatting with CRM data:** Salesforce Einstein Copilot
- **For building customer support chatbots:** Intercom Fin, Ada, Botsonic
- **For streamlining multi-channel comms:** Sendbird AI Chabot
- **For messaging:** Personal AI
- **For personal productivity:** Merlin, ZenoChat
- **For fun:** Character.AI
- **On social media:** Snapchat My AI
- **For learning:** Khan Academy's Khanmigo
- **For coding auto-complete:** GitHub Copilot, Amazon CodeWhisperer, Tabnine, Codeium

AI IN ACTION
Building Your Essential AI Toolkit

Benefits and Challenges of Open Source



- Open Source Large Language Models (Open Source, LLMs, AI)
  - Benefits ✅ (Pros, Positive Impacts, Advantages)
    - Transparency and Auditability
      - Understanding Model Behavior (Research, Trust)
      - Identifying and Mitigating Bias (Ethics, Fairness)
    - Community Collaboration
      - Faster Development and Improvement (Innovation, Efficiency)
      - Diverse Perspectives and Expertise (Knowledge Sharing, Inclusivity)
    - Accessibility and Democratization
      - Lower Barriers to Entry (Education, Research)
      - Wider Adoption and Use Cases (Innovation, Economic Growth)
  - Challenges ⚠ (Cons, Disadvantages, Negative Impacts)
    - Security and Privacy Risks
      - Data Breaches and Malicious Use (Vulnerability, Trust)
      - Model Poisoning and Manipulation (Accuracy, Reliability)
    - Resource Requirements and Scalability
      - Training and Inference Costs (Efficiency, Cost)
      - Infrastructure and Hardware Needs (Accessibility, Scalability)
    - Governance and Ethical Considerations
      - Bias and Fairness Concerns (Ethics, Transparency)
      - Responsible Use and Impact Assessment (Social Responsibility, Regulation)

Trouble Shooting

If something goes wrong, you can remove the existing Docker container and volume, and then start the installation process again.

1. **Stop and Remove the Container**:
   ```
     docker stop open-webui
             docker rm open-webui
   ```
2. **Remove the Docker Volume**:
   ```
   docker volume rm open-webui
   ```
3. **Remove the WebUI image from the Docker Destop**
4. **Reinstall Open WebUI**: Run the original Docker command to reinstall Open WebUI:
   ```
   docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway
   -v open-webui:/app/backend/data --name open-webui --restart always
   ghcr.io/open-webui/open-webui:main
   ```

This will give you a fresh installation of Open WebUI. After reinstalling, you can set up your admin password and other configurations as needed.

Let me know if you need any further assistance!