# Build Your Own AI Lab

A hands-on guide to home and cloud-based AI labs and infrastructure

Omar Santos

https://hackertraining.org

/santosomar

@santosomar

hackerrepo.org

# About Omar

Omar Santos is a Distinguished Engineer at Cisco; leading the AI Security Engineering team within Cisco's Security & Trust Organization. He has an extensive background in vulnerability research, incident response, and ethical hacking. He is the co-chair of the Coalition for Secure AI (CoSAI) and served in the board of the OASIS Open standards organization. Omar is also the chair of the OpenEoX and the Common Security Advisory Framework (CSAF) technical committee. His work led the creation of the CSAF ISO standard.

Omar's industry contributions extend to numerous organizations, including the Forum of Incident Response and Security Teams (FIRST), OASIS, the Industry Consortium for Advancement of Security on the Internet (ICASI). Omar is the co-chair of the FIRST PSIRT Special Interest Group (SIG) and was the lead of the DEF CON Red Team Village for several years.

Omar is the author of over 25 books, 21 video courses, and over 50 academic research papers. Omar is a renowned expert in ethical hacking, vulnerability research, incident response, and AI security. Omar's work in cybersecurity is also recognized through multiple granted patents. Prior to Cisco, Omar served in the United States Marines focusing on the deployment, testing, and maintenance of Command, Control, Communications, Computer, and Intelligence (C4I) systems.

# ABOUT THIS TRAINING

- Instruction on set up and optimization of AI labs to research and experiment in a secure environment
- Emphasis on real-world applications, hands-on projects, and case studies that allow you to apply learned concepts directly
- Focus on open-source large language models (LLMs) and generative AI

# Agenda

# AGENDA

Segment 1: Introduction and Foundations

- Overview of AI labs: home-based vs. cloud-based
- Setting up home-based AI labs
- Choosing the right hardware
- CPUs, GPUs, TPUs, and NPUs
- Building or buying pre-built systems
- Operating systems (Linux, Windows, macOS)
- Essential software (Python, Anaconda, Jupyter)
- Installing Ollama
- Do you need to install AI frameworks (TensorFlow, PyTorch, Hugging Face)?
- Securing the home AI lab, network setup, and optimization

# AGENDA

Segment 2: Cloud-Based AI Labs

- Advantages and disadvantages of cloud AI labs
- Utilizing cloud AI services and tools
  - Amazon Bedrock
  - Amazon SageMaker
  - Google Vertex AI
  - Microsoft Azure AI Foundry and AI Foundry Agent Service
  - OpenAI Agent Builder and OpenAI Agent Kit

# AGENDA

Segment 3: Integrating and Leveraging AI Environments

- Hybrid AI Labs: Combining home and cloud resources
- Synchronizing data and projects
- Leveraging the strengths of both environments
- Running open-source models available on Hugging Face (Llama, Phi, Qwen, DeepSeek, FoundationSec, Gemma, etc.)
- Demonstrations on building a home system to run these models
- Development environments (Jupyter Notebooks, IDEs, coding agents)
- Data management and storage
- Experiment tracking

# AGENDA

Segment 4: Advanced Topics and Practical Applications

- High-Performance Computing and Edge AI
- Introduction to HPC for AI
- Running AI models on edge devices
- Integrating AI with IoT systems
- Real-World Case Studies

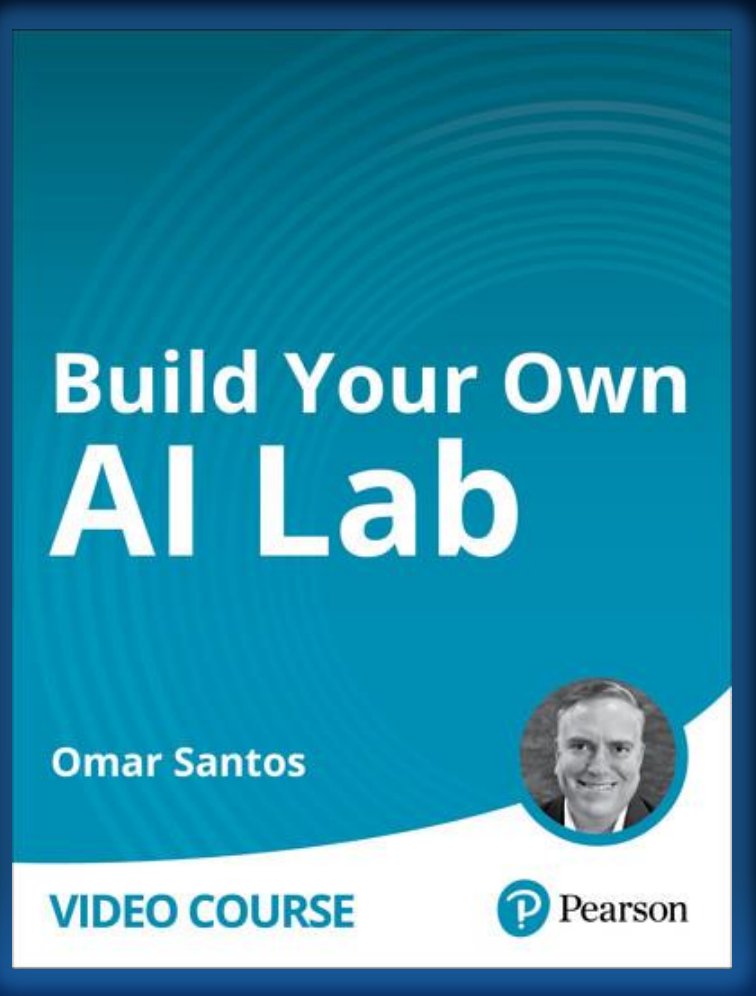
Q&A and Course Wrap-Up

# GitHub Repositories

- ✨ Cybersecurity and AI Comprehensive GitHub Repository: hackerrepo.org
- ✨ The GitHub Repository Used for this course: https://github.com/santosomar/build-your-ai-lab

# Video On-Demand Follow Up

https://www.oreilly.com/videos/build-your-own/9780135439616



# Hacker Training

https://hackertraining.org

AI and Cybersecurity Video-on-Demand Training, Books, and Other Resources

# Overview of AI Labs:
## Home-based vs. Cloud-based

Clearly outline what you want to achieve with your *AI lab*, whether it's learning basic ML algorithms, working on data science projects, or exploring “Small Language Models (SLMs)”, AI agents, and Retrieval Augmented Generation (RAG)

https://github.com/santosomar/build-your-ai-lab

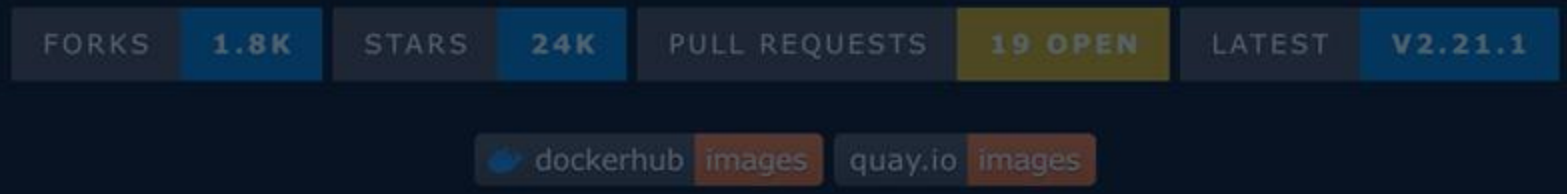
| | Home-based AI Labs | Cloud-based AI Labs |
|---|---|---|
| Cost | - Initial Investment: High (hardware, infrastructure)<br>- Ongoing Costs: Maintenance, upgrades | - Pricing Model: Pay-as-you-go<br>- Lower Initial Costs: No need for physical hardware |
| Scalability | - Limited Scalability: Constrained by physical resources<br>- Expansion Costs: Significant for scaling | - Highly Scalable: Easily adjust resources based on demand<br>- Flexible Pricing: Scale up or down as needed |
| Security | - Full control over data and security measures<br>- Vulnerabilities: Dependent on in-house security expertise | - Managed Security: Providers offer security features<br>- Data Privacy: Reliant on provider's compliance and policies |
| Maintenance | - Responsibility: Requires in-house maintenance and updates<br>- Downtime: Potential longer downtimes during maintenance | - Provider Managed: Maintenance and updates handled by the cloud provider<br>- Minimal Downtime: Typically, high uptime guarantees |
| Customization | - High Customization: Tailor hardware and software to specific needs<br>- Flexibility: Greater control over configurations | - Limited Customization: Dependent on provider's offerings<br>- Standardized Environments: May not meet all specific requirements |

https://github.com/santosomar/build-your-ai-lab

# Setting up Home-based AI labs

# Local AI

https://localai.io



README | MIT license | Security

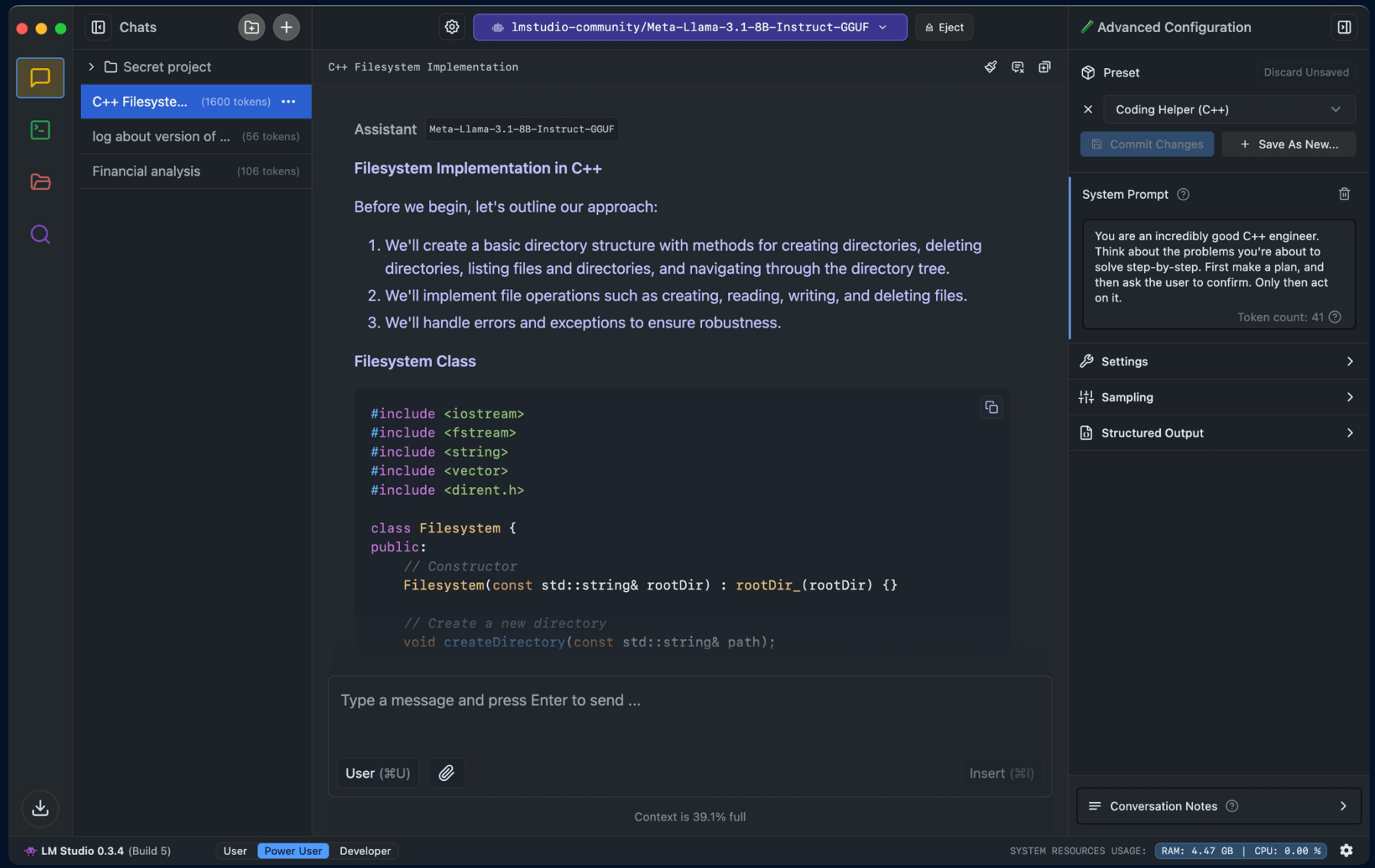
# Ollama

Ollama

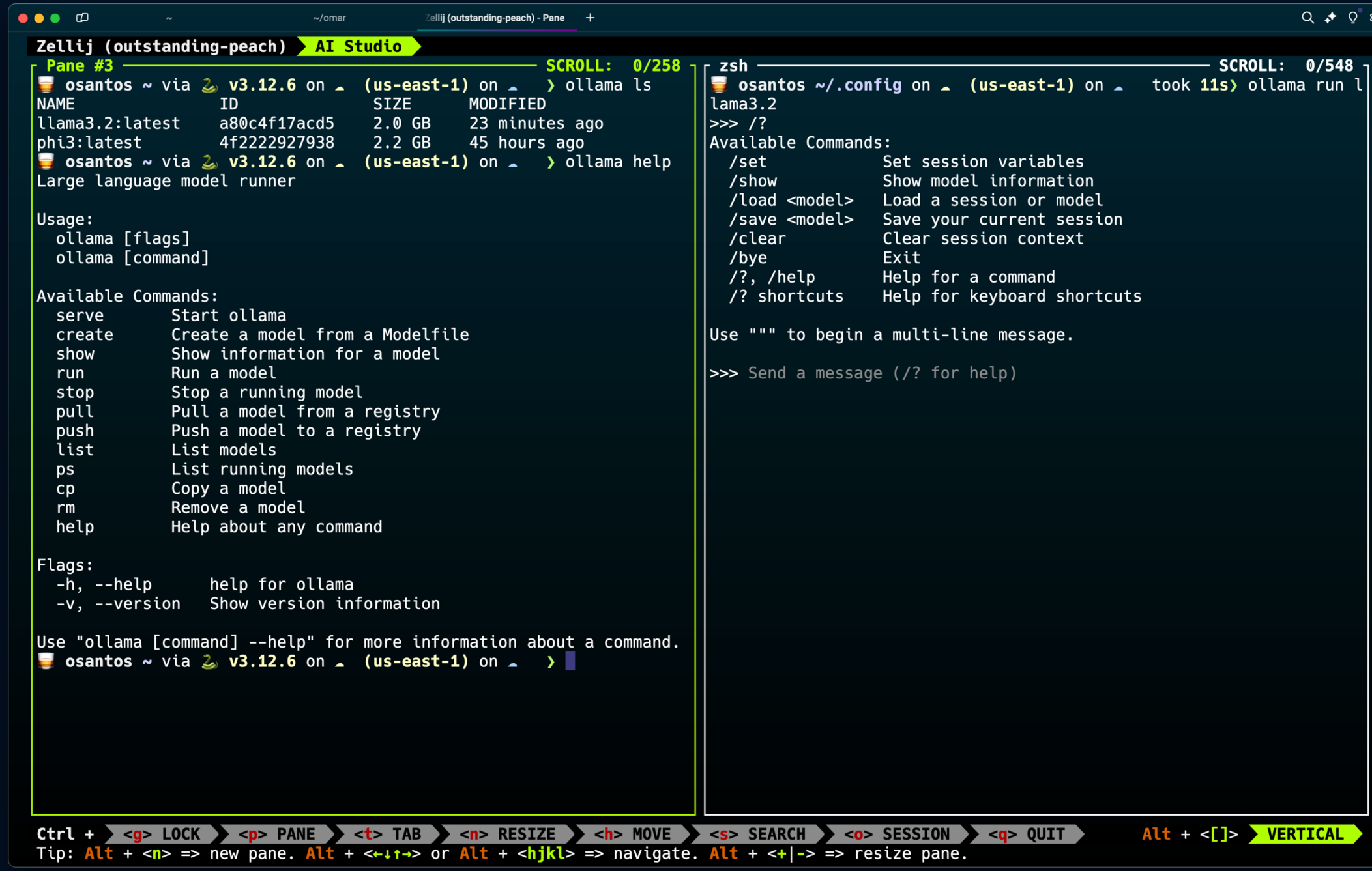
Get up and running with large language models.

https://ollama.com

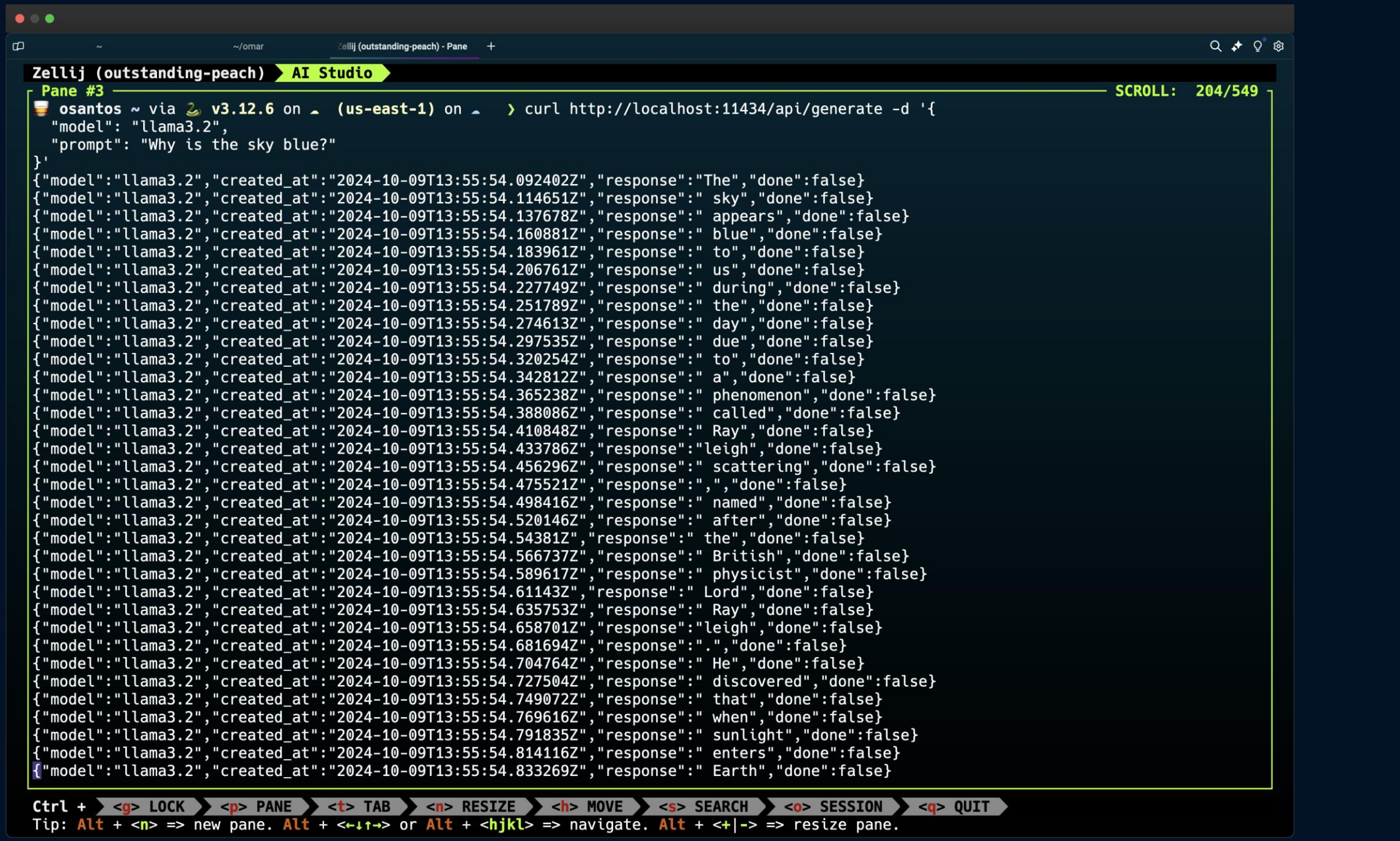
LM Studio




https://lmstudio.ai

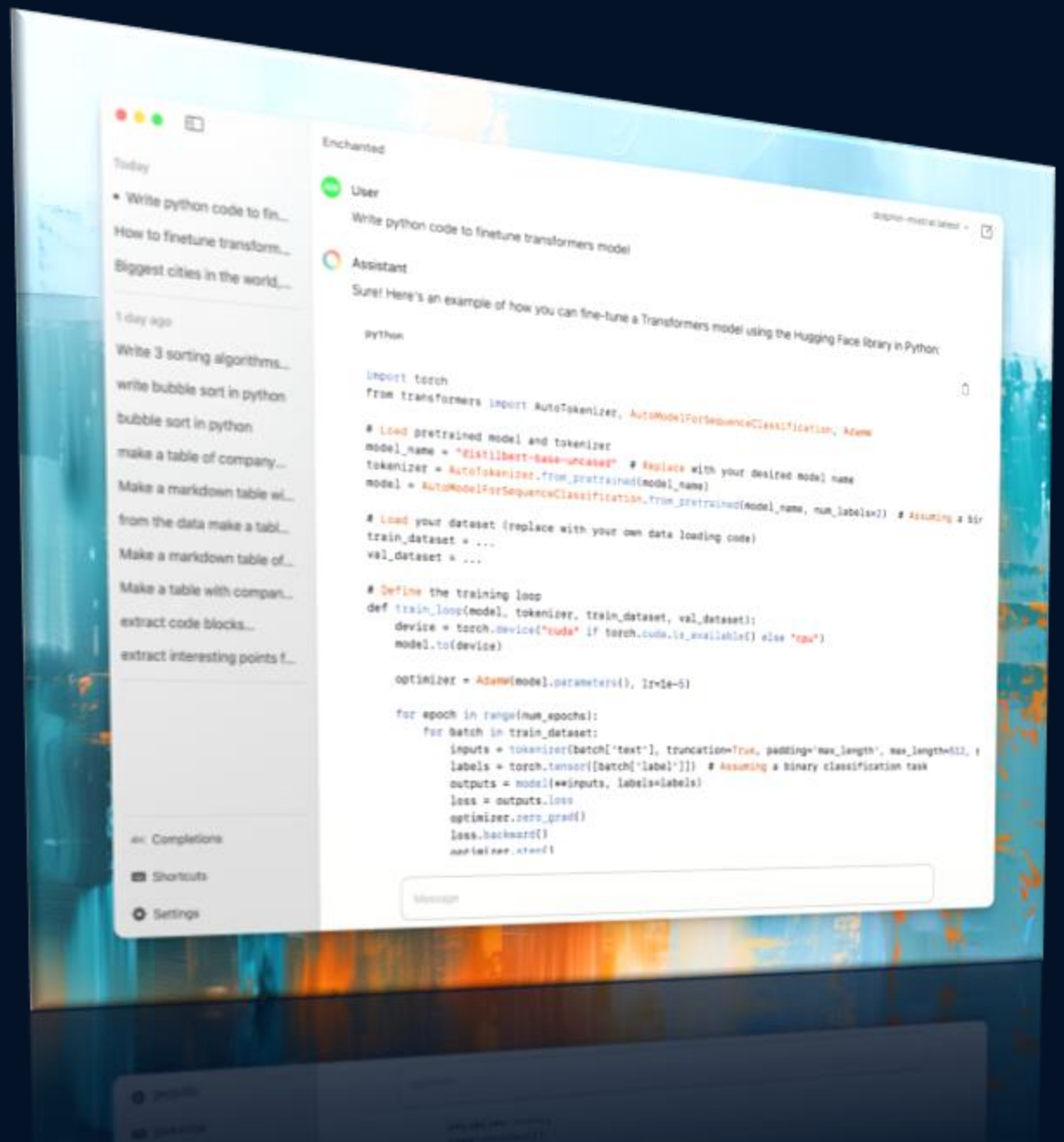

~
~/omar
Zellij (outstanding-peach) - Pane
Zellij (outstanding-peach) AI Studio
Pane #3 SCROLL: 0/258
osantos ~ via v3.12.6 on (us-east-1) on ❯ ollama ls
NAME ID SIZE MODIFIED
llama3.2:latest a80c4f17acd5 2.0 GB 23 minutes ago
phi3:latest 4f2222927938 2.2 GB 45 hours ago
osantos ~ via v3.12.6 on (us-east-1) on ❯ ollama help
Large language model runner
Usage:
ollama [flags]
ollama [command]
Available Commands:
serve Start ollama
create Create a model from a Modelfile
show Show information for a model
run Run a model
stop Stop a running model
pull Pull a model from a registry
push Push a model to a registry
list List models
ps List running models
cp Copy a model
rm Remove a model
help Help about any command
Flags:
-h, --help help for ollama
-v, --version Show version information
Use "ollama [command] --help" for more information about a command.
osantos ~ via v3.12.6 on (us-east-1) on ❯
zsh SCROLL: 0/548
osantos ~/.config on (us-east-1) on took 11s❯ ollama run l
lama3.2
>>> /?
Available Commands:
/set Set session variables
/show Show model information
/load <model> Load a session or model
/save <model> Save your current session
/clear Clear session context
/bye Exit
/?, /help Help for a command
/? shortcuts Help for keyboard shortcuts
Use """ to begin a multi-line message.
>>> Send a message (/? for help)
Ctrl + <g> LOCK <p> PANE <t> TAB <n> RESIZE <h> MOVE <s> SEARCH <o> SESSION <q> QUIT
Alt + <[]> VERTICAL
Tip: Alt + <n> => new pane. Alt + <←↓↑→> or Alt + <hjkl> => navigate. Alt + <+|-> => resize pane.


~
~/omar
Zellij (outstanding-peach) - Pane
Zellij (outstanding-peach) AI Studio
Pane #3
SCROLL: 0/307
osantos ~ via v3.12.6 on (us-east-1) on ❯ ollama show llama3.2
Model
architecture llama
parameters 3.2B
context length 131072
embedding length 3072
quantization Q4_K_M
Parameters
stop "<|start_header_id|>"
stop "<|end_header_id|>"
stop "<|eot_id|>"
License
LLAMA 3.2 COMMUNITY LICENSE AGREEMENT
Llama 3.2 Version Release Date: September 25, 2024
osantos ~ via v3.12.6 on (us-east-1) on ❯
Ctrl + <g> LOCK <p> PANE <t> TAB <n> RESIZE <h> MOVE <s> SEARCH <o> SESSION <q> QUIT
Tip: Alt + <n> => new pane. Alt + <←↓↑→> or Alt + <hjkl> => navigate. Alt + <+|-> => resize pane.


~
~/omar
Zellij (outstanding-peach) - Pane
Zellij (outstanding-peach) AI Studio
Pane #3
SCROLL: 204/549
osantos ~ via v3.12.6 on (us-east-1) on ❯ curl http://localhost:11434/api/generate -d '{
"model": "llama3.2",
"prompt": "Why is the sky blue?"
}'
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.092402Z","response":"The","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.114651Z","response":" sky","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.137678Z","response":" appears","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.160881Z","response":" blue","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.183961Z","response":" to","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.206761Z","response":" us","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.227749Z","response":" during","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.251789Z","response":" the","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.274613Z","response":" day","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.297535Z","response":" due","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.320254Z","response":" to","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.342812Z","response":" a","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.365238Z","response":" phenomenon","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.388086Z","response":" called","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.410848Z","response":" Ray","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.433786Z","response":"leigh","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.456296Z","response":" scattering","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.475521Z","response":",","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.498416Z","response":" named","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.520146Z","response":" after","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.54381Z","response":" the","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.566737Z","response":" British","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.589617Z","response":" physicist","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.61143Z","response":" Lord","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.635753Z","response":" Ray","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.658701Z","response":"leigh","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.681694Z","response":".","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.704764Z","response":" He","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.727504Z","response":" discovered","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.749072Z","response":" that","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.769616Z","response":" when","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.791835Z","response":" sunlight","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.814116Z","response":" enters","done":false}
{"model":"llama3.2","created_at":"2024-10-09T13:55:54.833269Z","response":" Earth","done":false}
Ctrl + <g> LOCK <p> PANE <t> TAB <n> RESIZE <h> MOVE <s> SEARCH <o> SESSION <q> QUIT
Tip: Alt + <n> => new pane. Alt + <←↓↑→> or Alt + <hjkl> => navigate. Alt + <+|-> => resize pane.

# Ollama Community Integrations

https://github.com/ollama/ollama?tab=readme-ov-file#community-integrations




# Enchanted

https://github.com/AugustDev/enchanted

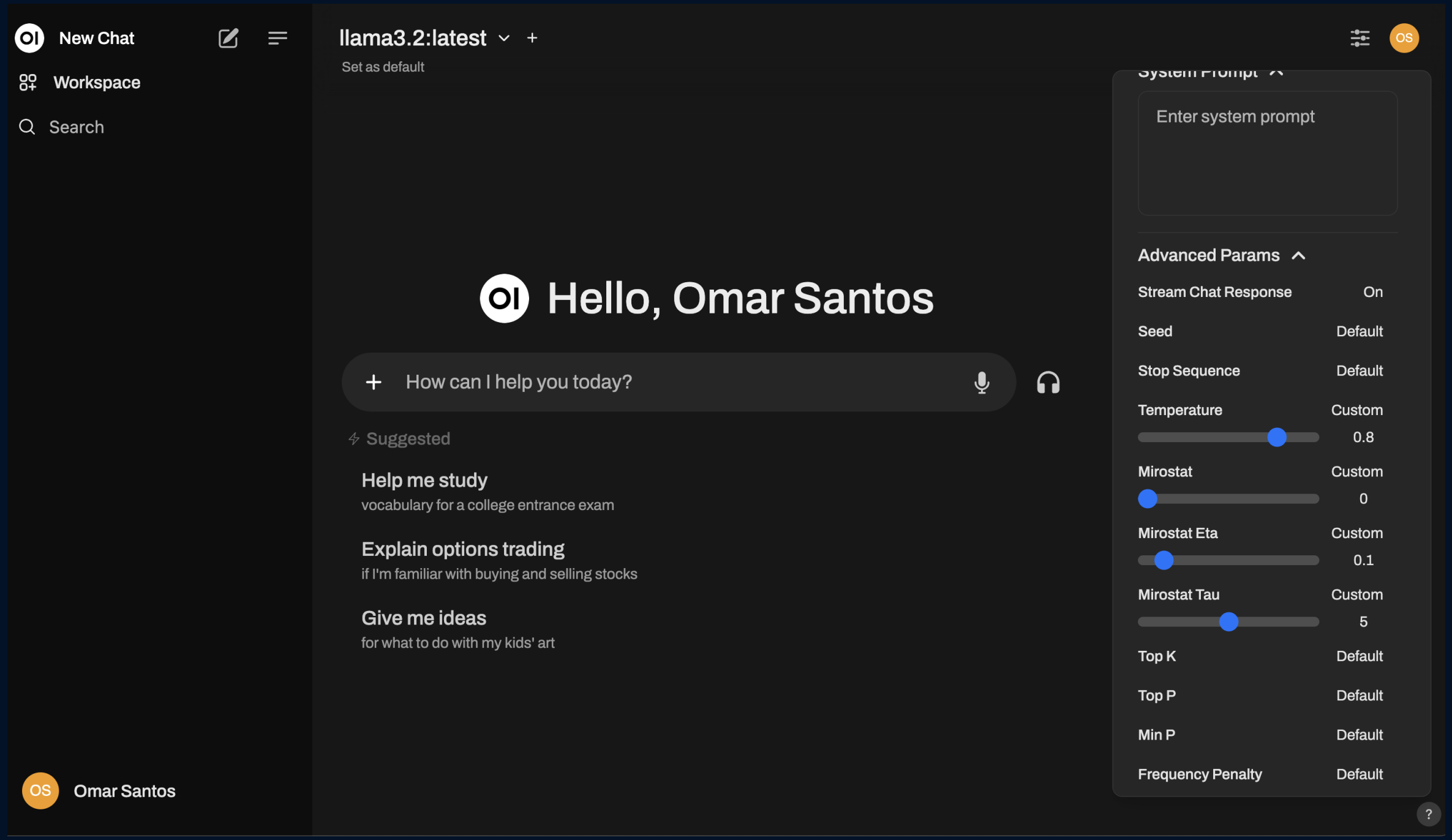
# Open WebUI 👋

Stars 42k | Forks 5k | Watchers 211 | repo size 115 MB | languages 11 | svelte 55.2%

last commit today | hits 21284 / 5028282 | Discord Open WebUI | Sponsor ❤

Open WebUI is an extensible, feature-rich, and user-friendly self-hosted WebUI designed to operate entirely offline. It supports various LLM runners, including Ollama and OpenAI-compatible APIs. For more information, be sure to check out our Open WebUI Documentation.

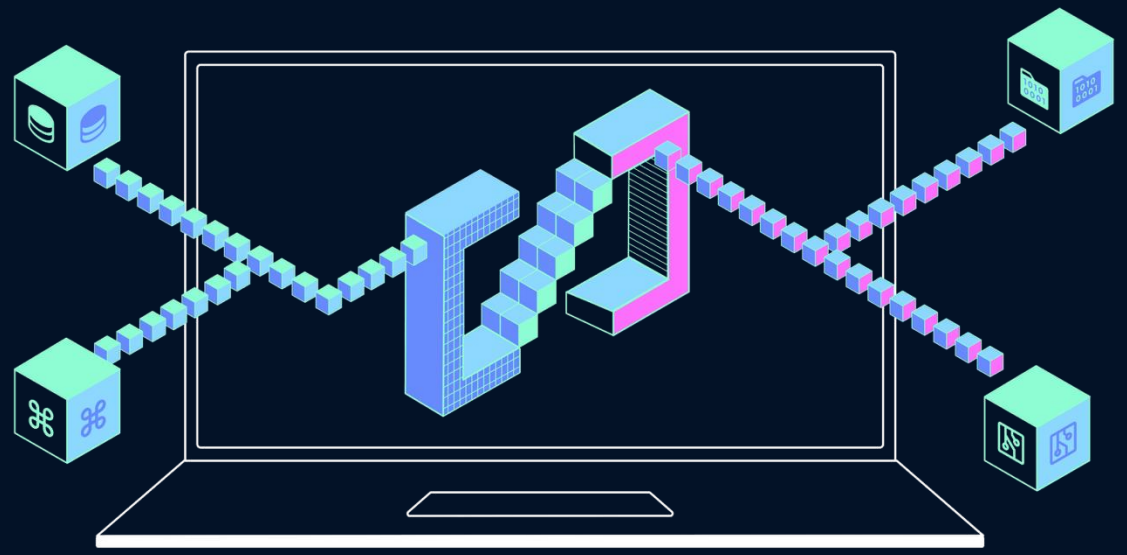

https://github.com/open-webui/open-webui


New Chat
Workspace
Search
llama3.2:latest
Set as default
OS
System Prompt
Enter system prompt
Advanced Params
Stream Chat Response
On
Seed
Default
Stop Sequence
Default
Temperature
Custom
0.8
Mirostat
Custom
0
Mirostat Eta
Custom
0.1
Mirostat Tau
Custom
5
Top K
Default
Top P
Default
Min P
Default
Frequency Penalty
Default
Hello, Omar Santos
How can I help you today?
Suggested
Help me study
vocabulary for a college entrance exam
Explain options trading
if I'm familiar with buying and selling stocks
Give me ideas
for what to do with my kids' art
OS
Omar Santos
?

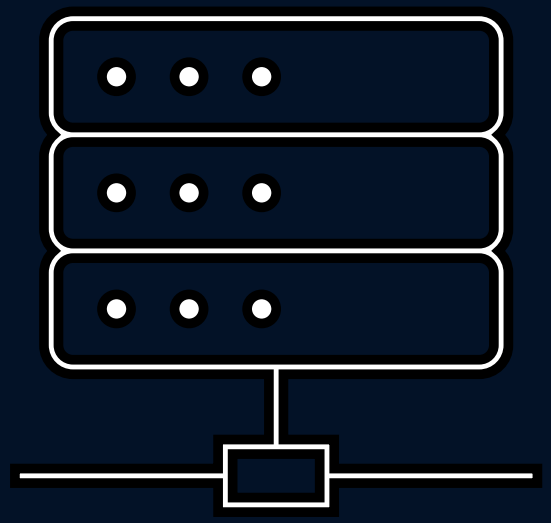
# RAG + Ollama + LlamaIndex

```python
from typing import List, Union, Generator, Iterator
from schemas import OpenAIChatMessage
import os

from pydantic import BaseModel


class Pipeline:

    class Valves(BaseModel):
        LLAMAINDEX_OLLAMA_BASE_URL: str
        LLAMAINDEX_MODEL_NAME: str
        LLAMAINDEX_EMBEDDING_MODEL_NAME: str

    def __init__(self):
        self.documents = None
        self.index = None

        self.valves = self.Valves(
            **{
                "LLAMAINDEX_OLLAMA_BASE_URL": os.getenv("LLAMAINDEX_OLLAMA_BASE_URL", "http://localhost:11434"),
                "LLAMAINDEX_MODEL_NAME": os.getenv("LLAMAINDEX_MODEL_NAME", "llama3"),
                "LLAMAINDEX_EMBEDDING_MODEL_NAME": os.getenv("LLAMAINDEX_EMBEDDING_MODEL_NAME", "nomic-embed-text"),
            }
        )

    async def on_startup(self):
        from llama_index.embeddings.ollama import OllamaEmbedding
```

# Anything LLM



https://anythingllm.com



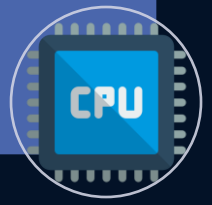
# Choosing the right hardware

https://github.com/santosomar/build-your-ai-lab/blob/main/segment-1-introduction-and-foundations/03-choosing-hardware.md

# ESSENTIAL HARDWARE

A multi-core processor with high clock speeds is crucial for data processing and running complex algorithms

A good CPU




NVIDIA's CUDA-compatible GPUs, like the RTX or Quadro series, are ideal for accelerating AI computations, especially deep learning tasks. (more on NVIDIA Project DIGITS later)

High-Performance GPU


At least 16GB is recommended, but 32GB or more is ideal for handling large datasets

Ample RAM

Combine SSDs (for speed) and HDDs (for capacity) to ensure sufficient storage for datasets and software

Storage

- Reliable routers and switches for a stable network connection

Networking Equipment

EVGA
500 G+
GEFORCE RTX 4090
GAMING OC 24G
BYTE
AMD RYZEN
AMD 3D V-CACHE TECHNOLOGY

# GPUs

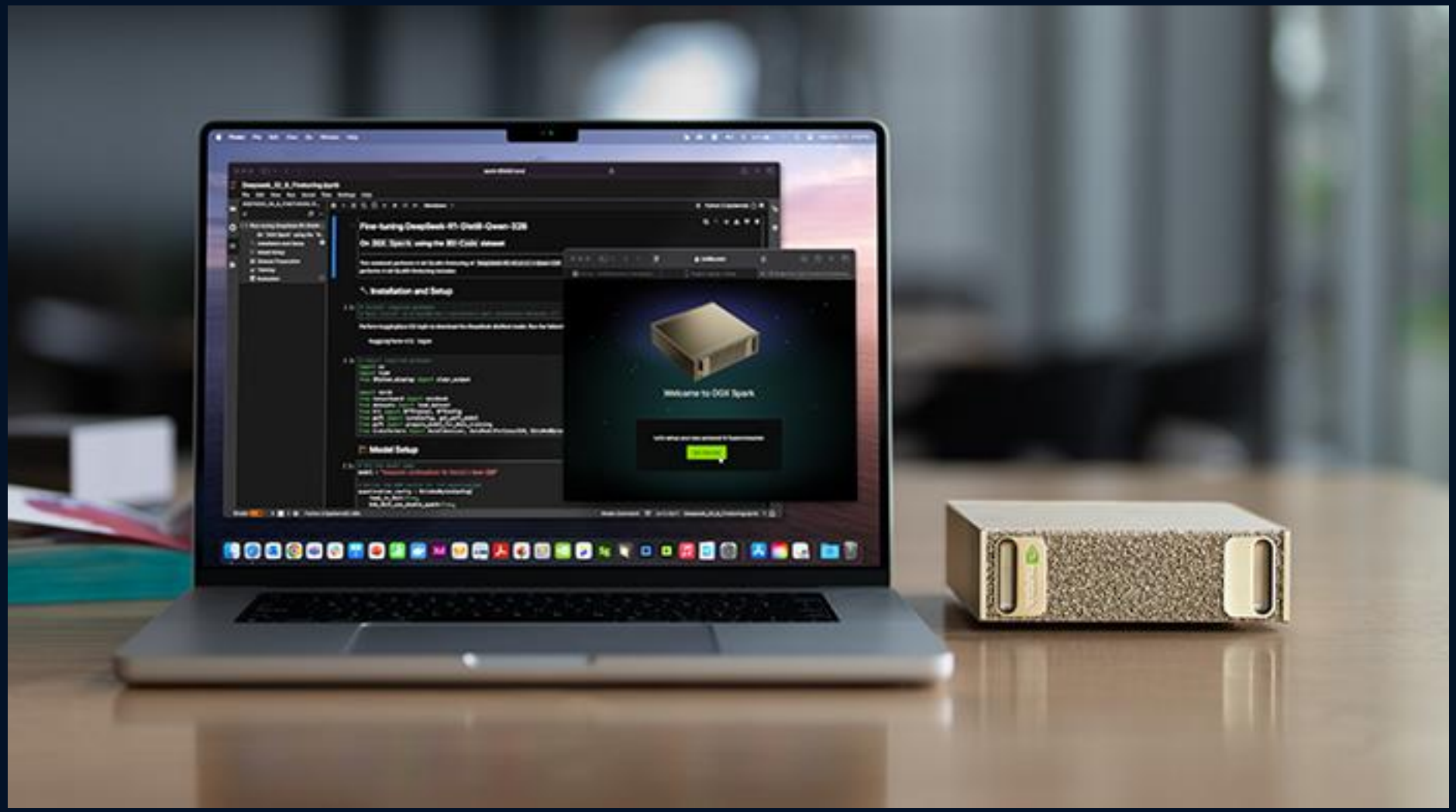
# CPUs


ASUS ProArt ProArt X670E-CREATOR WIFI
G.SKILL
RGB DDR5
TRIDENT Z5 RGB
reddot winner 2022
BC Best Choice AWARD
Intel XMP 3.0 Ready
TRIDENT Z5 RGB

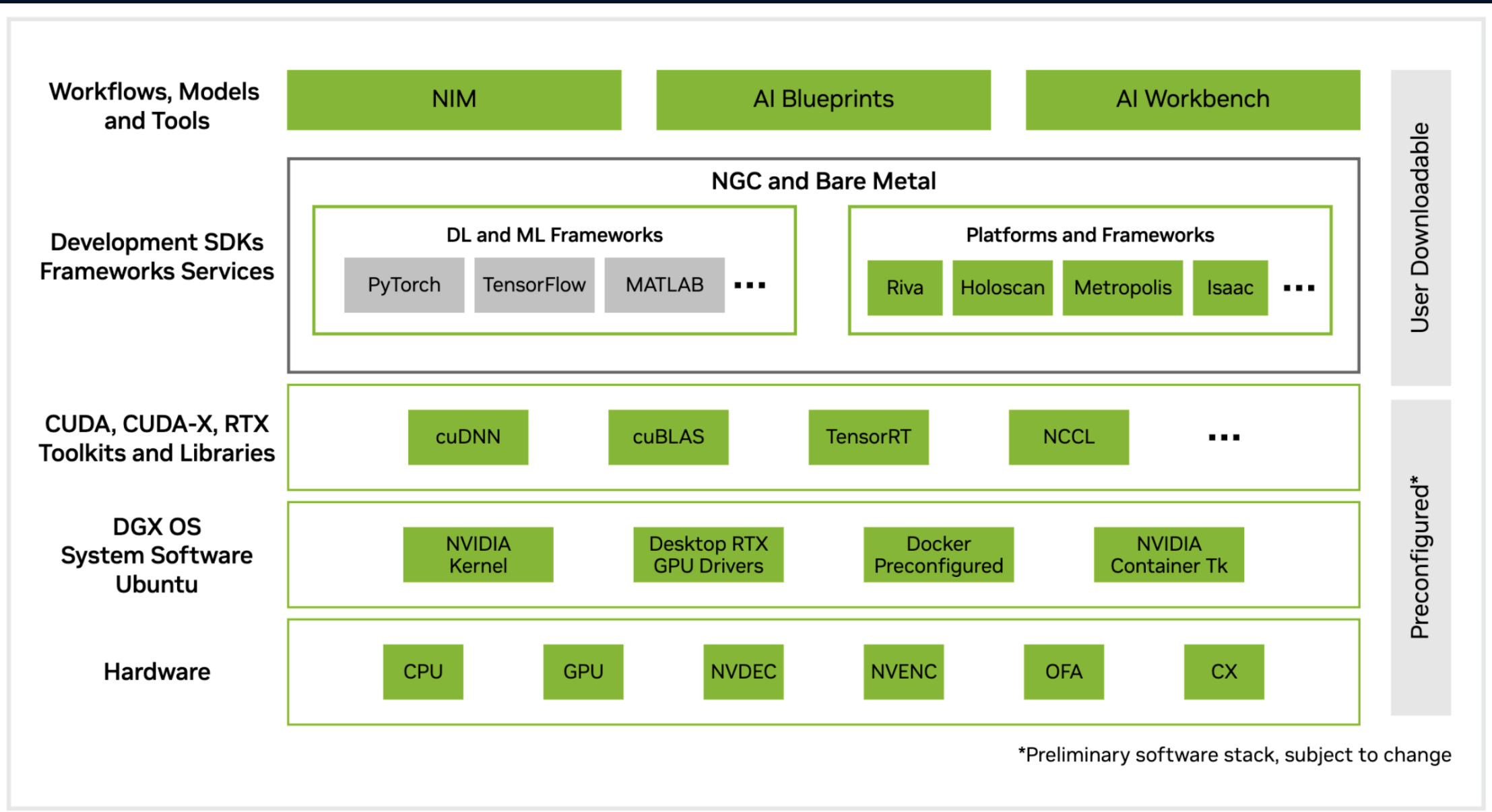
# RAM


ASUS ProArt ProArt X670E-CREATOR WIFI
ADVANCE WITH AMD
AMD
2TB
SAMSUNG
V-NAND SSD
990 PRO
PCIe 4.0 NVMe M.2 SSD
Blistering speed, endless victory
990 PRO
PCIe®4.0 NVMe™ M.2 SSD
Read Speed
7450MB/s
SSD 990 PRO
CORSAIR
STORAGE

# **NVIDIA** DGX Spark

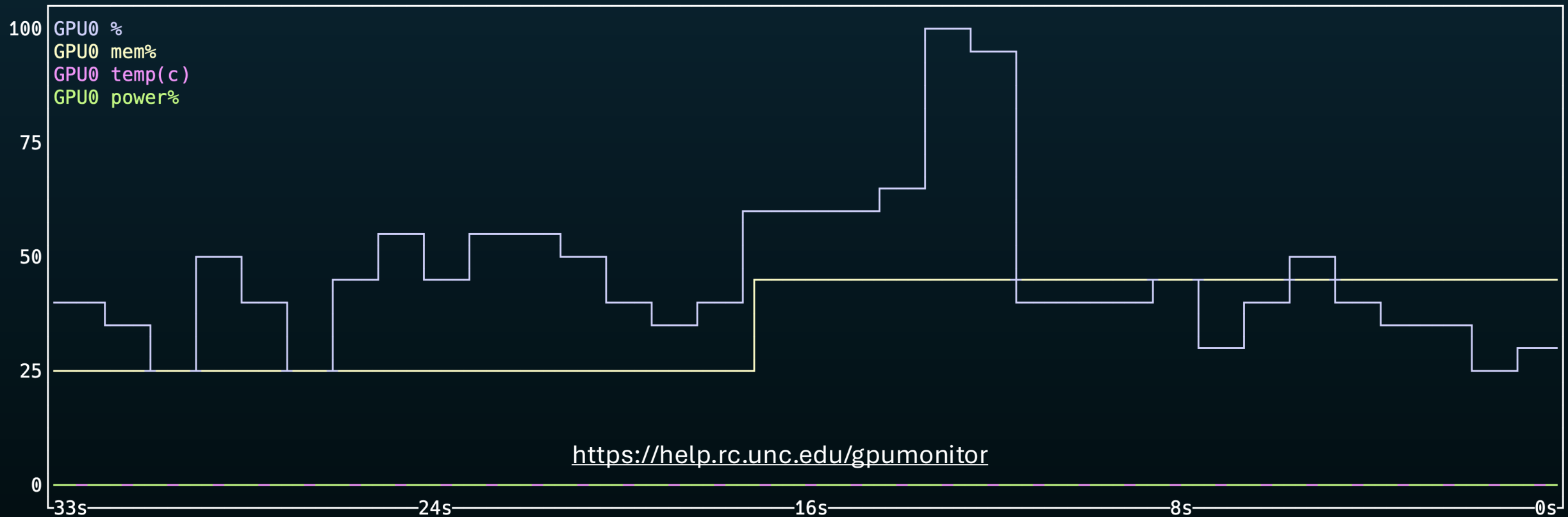



https://marketplace.nvidia.com/en-us/developer/dgx-spark




NVIDIA DGX Spark software stack

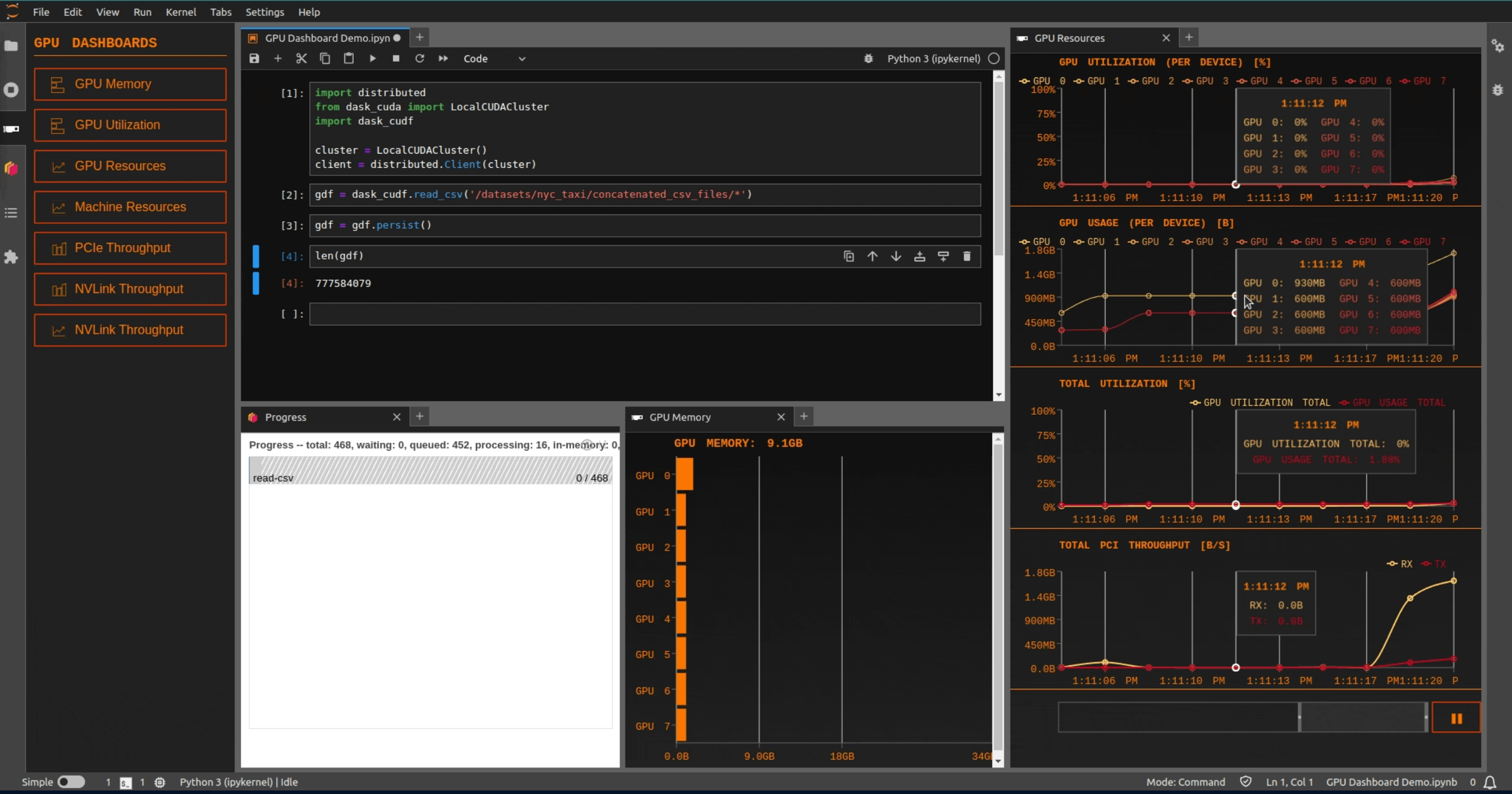
https://nvdam.widen.net/s/tlzm8smqjx/workstation-datasheet-dgx-spark-gtc25-spring-nvidia-us-3716899-web



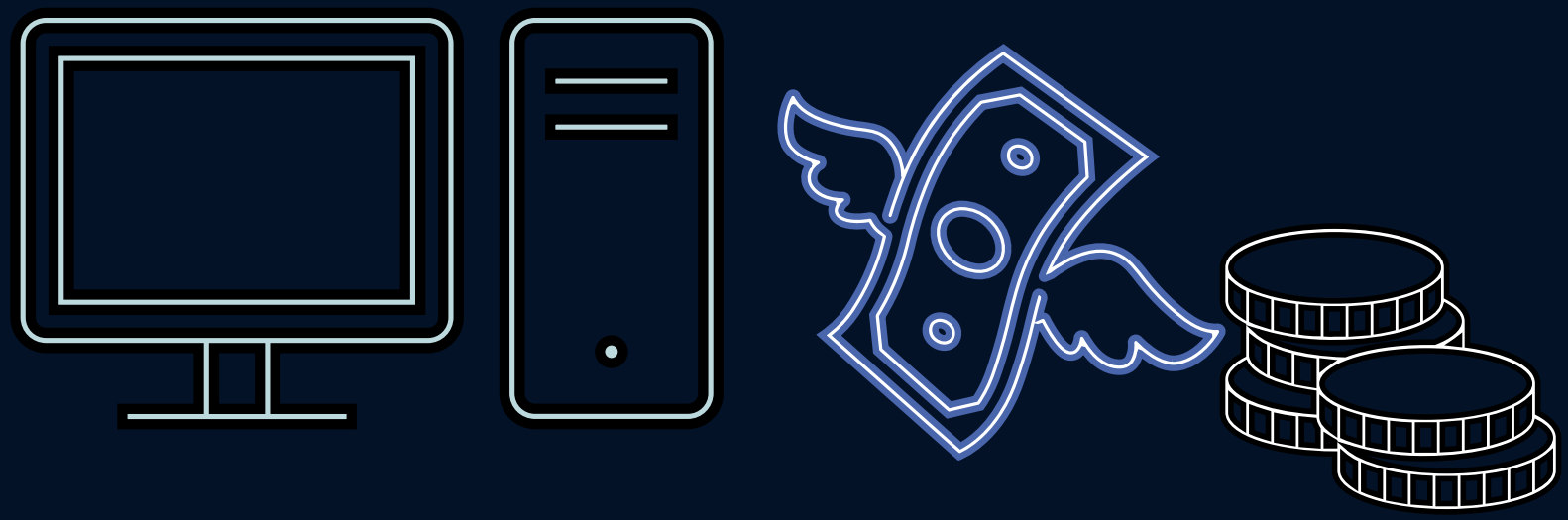

| PID | USER | DEV | TYPE | GPU | GPU MEM | CPU | HOST MEM | Command |
|---|---|---|---|---|---|---|---|---|
| 83278 | osantos | 0 | Both | G+C | 0% | 0% | 26MiB | /System/Library/CoreServices/screencaptureui.app/Contents/MacOS/scre |
| 371 | _windowserver | 0 | Both | G+C | 0% | N/A | N/A | |
| 572 | osantos | 0 | Both | G+C | 0% | 1% | 16MiB | /System/Library/PrivateFrameworks/UniversalAccess.framework/Versions |
| 383 | root | 0 | Both | G+C | 0% | N/A | N/A | |
| 613 | osantos | 0 | Both | G+C | 0% | 0% | 36MiB | /System/Library/CoreServices/NotificationCenter.app/Contents/MacOS/N |
| 623 | osantos | 0 | Both | G+C | 0% | 0% | 37MiB | /System/Library/CoreServices/ControlCenter.app/Contents/MacOS/Contro |
| 376 | osantos | 0 | Both | G+C | 0% | 0% | 29MiB | /System/Library/CoreServices/loginwindow.app/Contents/MacOS/loginwin |
| 654 | osantos | 0 | Both | G+C | 0% | 0% | 10MiB | /usr/libexec/replayd |
| 628 | osantos | 0 | Both | G+C | 0% | 1% | 63MiB | /System/Library/CoreServices/Finder.app/Contents/MacOS/Finder |
| 797 | | | | | | | | |
| 647 | | | | | | | | |
| 822 | | | | | | | | |
| 871 | | | | | | | | |
| 810 | | | | | | | | |
| 884 | | | | | | | | |

# Monitoring GPU and CPU Resources

F2Setup



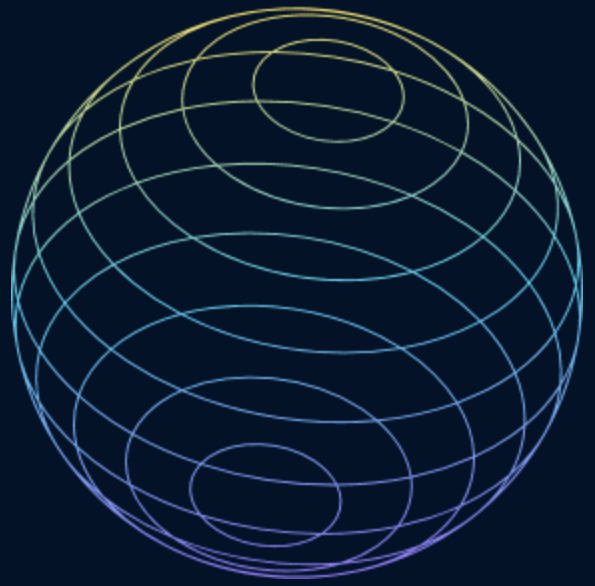

https://github.com/rapidsai/jupyterlab-nvdashboard



# Building or buying pre-built systems

https://github.com/santosomar/build-your-ai-lab/blob/main/segment-1-introduction-and-foundations/05-build-vs-buy.md

# Operating systems (Linux, Windows, macOS)

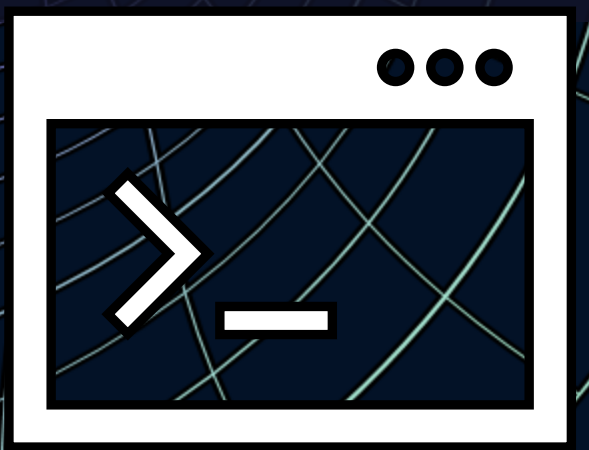


https://github.com/santosomar/build-your-ai-lab/blob/main/segment-1-introduction-and-foundations/06-operating-systems.md

# Essential Software

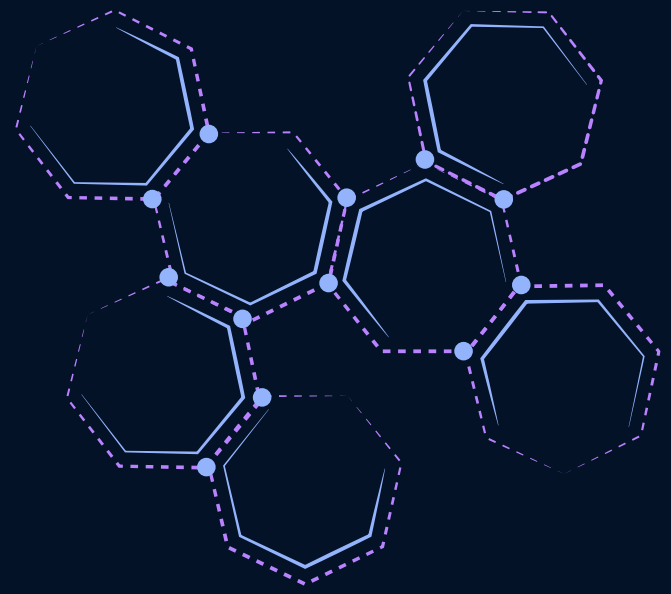
## (Python, Anaconda, Jupyter?
## Or Ollama, AnythingLLM, etc?)

https://github.com/santosomar/build-your-ai-lab/tree/main/segment-1-introduction-and-foundations



Do you need to install AI frameworks (TensorFlow, PyTorch, Hugging Face)?

# Installing Ollama and Ollama Labs



# Cloud-based Labs

https://github.com/santosomar/build-your-ai-lab/tree/main/segment-2-cloud-based-ai-labs

# Azure AI Foundry

# Find the right model to build your custom AI solution

Show filters

## Announcements

### Meta Llama 3.2 models are here!

Llama 3.2 11B Vision Instruct and 90B Vision Instruct are here for your image reasoning...

View models | Read blog

### News from Cohere!

Cohere's collection now includes Command R 08-2024 and Command R+ 08-2024.

View models

### Experience the o1 models

The o1 series feature an enhanced reasoning abilities to solve science and...

Try limited access | Read blog

### ALLaM-2-7B: latest Arabic LLM

ALLaM-2-7B is here! A robust 7B LLM model crafted to boost Arabic language...

View models | Read blog

All filters | Collections | Deployment options | Inference tasks | Fine-tuning tasks | Licenses

Search

Models 1791

- gpt-4o-realtime-preview — Audio generation
- openai-whisper-large-v3 — Speech recognition
- openai-whisper-large — Speech recognition
- gpt-4 — Chat completion
- gpt-35-turbo — Chat completion
- o1-preview — Chat completion
- o1-mini — Chat completion
- gpt-4o-mini — Chat completion
- gpt-4o — Chat completion
- gpt-4-32k — Chat completion
- gpt-35-turbo-instruct — Chat completion
- gpt-35-turbo-16k — Chat completion
- dall-e-3 — Text to image
- dall-e-2 — Text to image
- whisper — Speech recognition
- tts-hd — Text to speech

# Prompt catalog

## Browse prompt samples for common use cases

Choose a sample prompt to see how it works or as a starting point for your project. Then customize it for your scenario and evaluate how it performs before integrating into your app.

Prompt Samples

Search

### Prompts

**Applied filters**

**Generate User Questions On A Pr...**
To submit your application to evaluation, you can ...
Chat completions

**Travel Assistant**
Provide the travel information.
Summarization

**Social Media Post Analysis**
Analyze short videos to generate tags, content su...
Summarization

**Property Listing Video**
Provide a summary and highlights of the property
Summarization

**Insurance Report**
Car Insurance Damage report writing
Summarization

**Advertising Summary**
Summarize an advertisement and identify issues (i....
Summarization

**Real Estate Agents Assistant**
Provide overview descriptions for houses.
Summarization

**Listing Assistant**
Generate enticing vacation rental listings from ima...
Summarization

**Image Tagging Assistant**
Identify and list prevalent tags associated with the...
Summarization

**Image Description Assistant**
Image Content Description Prompt

**Defect Detector**
Find defects on a test image based a reference im...

**Apple Cycle Analyst**
Examine the test image to determine the specific s...

### Filters

**Modalities**

Chat | Image | Video | Completion

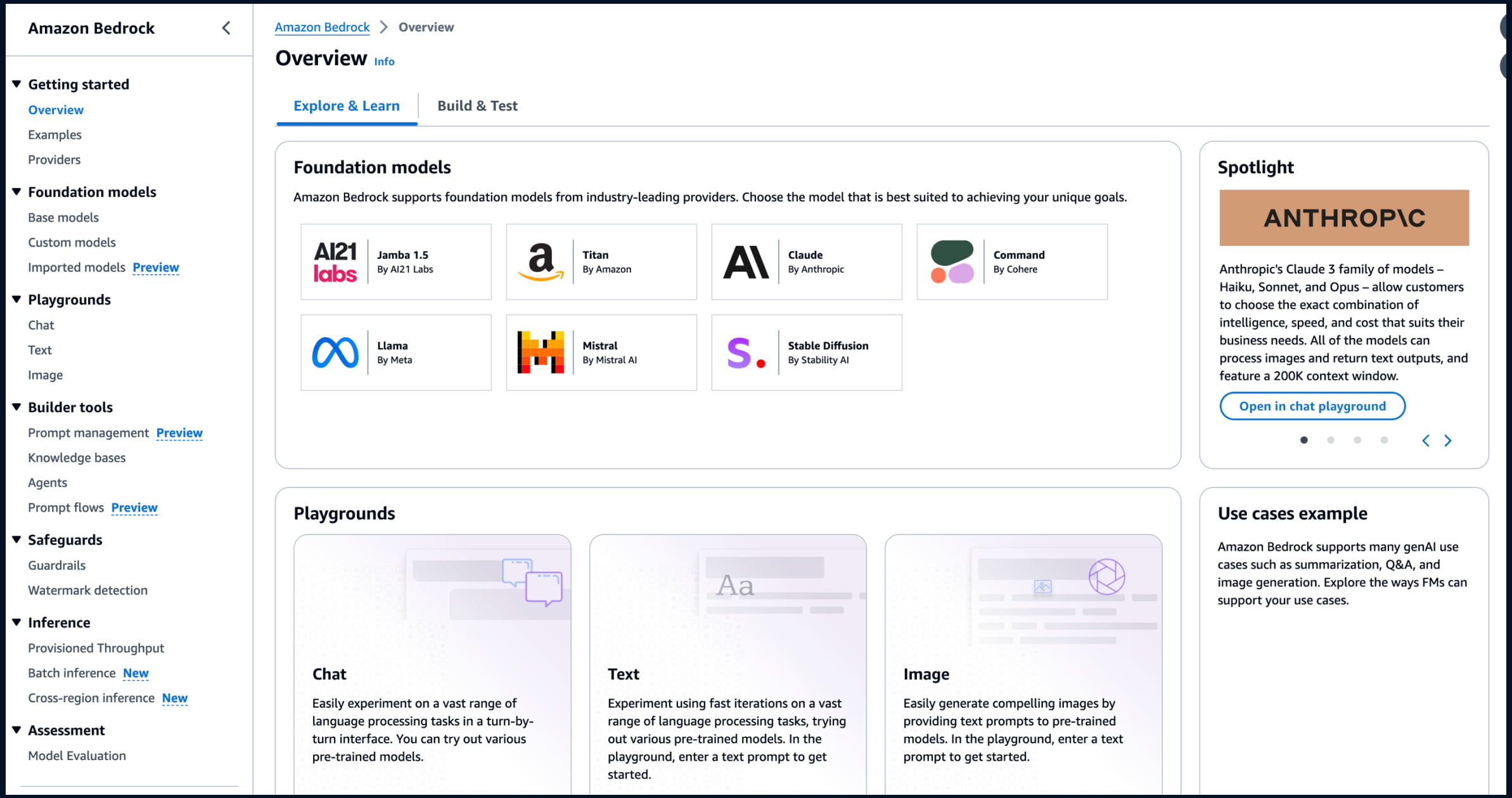
**Industries**

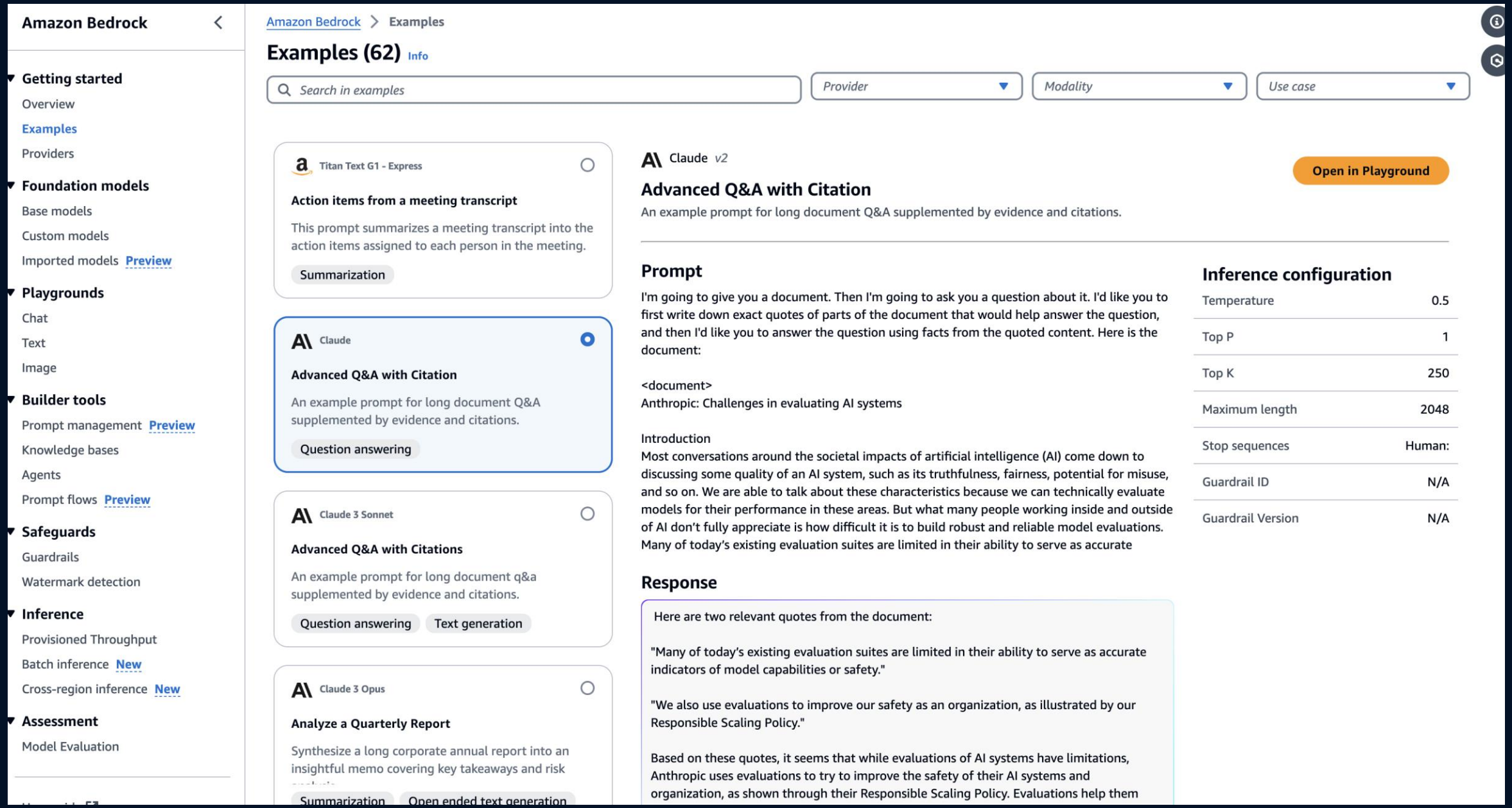
Retail | Education

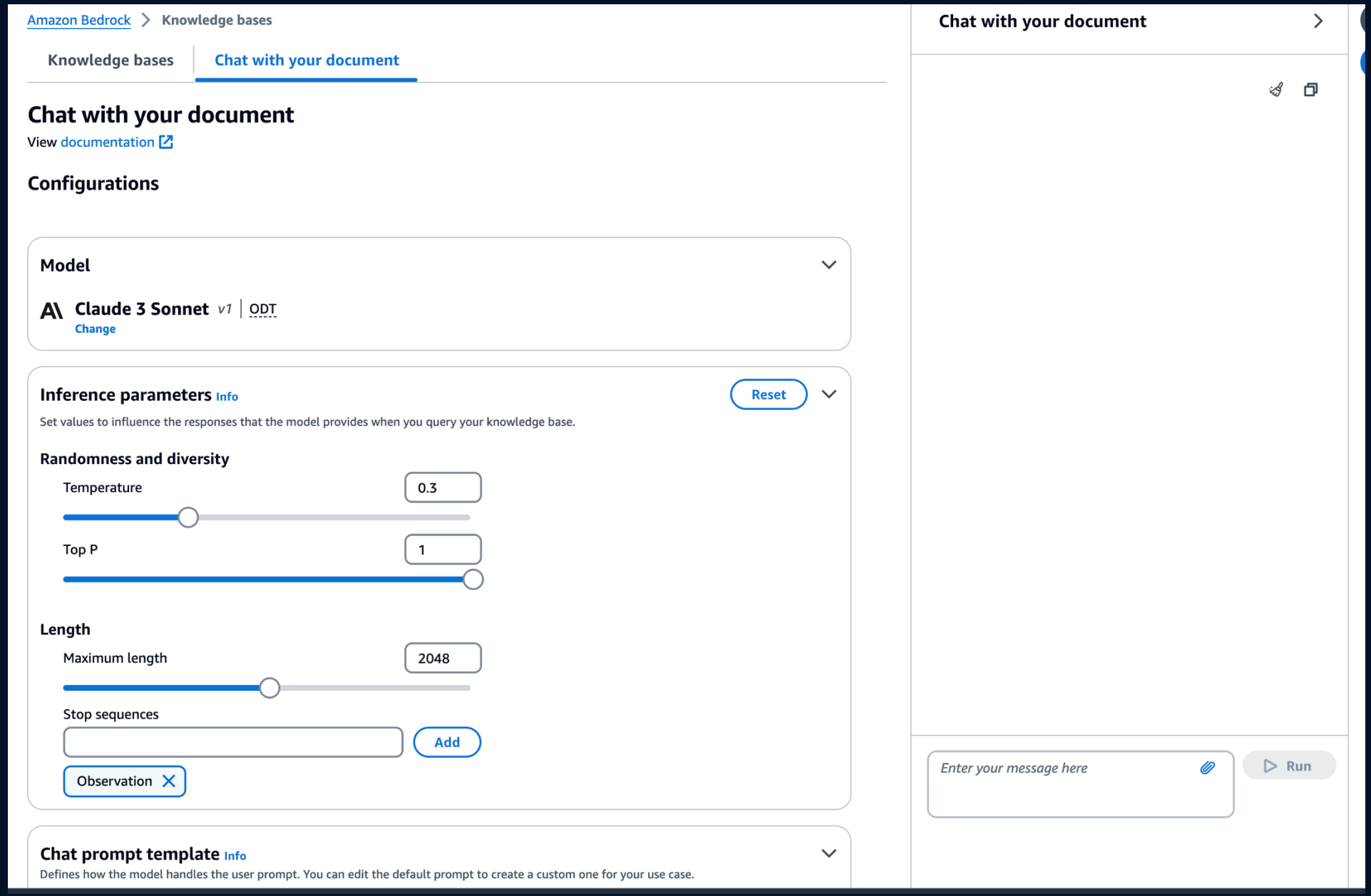
**Tasks**

Chat completions
Summarization
Content creation
Reasoning & insights
Natural language to code
Recommendation | Explain code

# Amazon Bedrock & SageMaker


Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Builder tools
Prompt management Preview
Knowledge bases
Agents
Prompt flows Preview
Safeguards
Guardrails
Watermark detection
Inference
Provisioned Throughput
Batch inference New
Cross-region inference New
Assessment
Model Evaluation
Amazon Bedrock > Overview
Overview Info
Explore & Learn
Build & Test
Foundation models
Amazon Bedrock supports foundation models from industry-leading providers. Choose the model that is best suited to achieving your unique goals.
Jamba 1.5
By AI21 Labs
Titan
By Amazon
Claude
By Anthropic
Command
By Cohere
Llama
By Meta
Mistral
By Mistral AI
Stable Diffusion
By Stability AI
Spotlight
ANTHROP\C
Anthropic's Claude 3 family of models – Haiku, Sonnet, and Opus – allow customers to choose the exact combination of intelligence, speed, and cost that suits their business needs. All of the models can process images and return text outputs, and feature a 200K context window.
Open in chat playground
Playgrounds
Chat
Easily experiment on a vast range of language processing tasks in a turn-by-turn interface. You can try out various pre-trained models.
Text
Experiment using fast iterations on a vast range of language processing tasks, trying out various pre-trained models. In the playground, enter a text prompt to get started.
Image
Easily generate compelling images by providing text prompts to pre-trained models. In the playground, enter a text prompt to get started.
Use cases example
Amazon Bedrock supports many genAI use cases such as summarization, Q&A, and image generation. Explore the ways FMs can support your use cases.


Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Builder tools
Prompt management Preview
Knowledge bases
Agents
Prompt flows Preview
Safeguards
Guardrails
Watermark detection
Inference
Provisioned Throughput
Batch inference New
Cross-region inference New
Assessment
Model Evaluation
Amazon Bedrock > Examples
Examples (62) Info
Search in examples
Provider
Modality
Use case
Titan Text G1 - Express
Action items from a meeting transcript
This prompt summarizes a meeting transcript into the action items assigned to each person in the meeting.
Summarization
Claude
Advanced Q&A with Citation
An example prompt for long document Q&A supplemented by evidence and citations.
Question answering
Claude 3 Sonnet
Advanced Q&A with Citations
An example prompt for long document q&a supplemented by evidence and citations.
Question answering
Text generation
Claude 3 Opus
Analyze a Quarterly Report
Synthesize a long corporate annual report into an insightful memo covering key takeaways and risk
Summarization
Open ended text generation
Claude v2
Advanced Q&A with Citation
An example prompt for long document Q&A supplemented by evidence and citations.
Open in Playground
Prompt
I'm going to give you a document. Then I'm going to ask you a question about it. I'd like you to first write down exact quotes of parts of the document that would help answer the question, and then I'd like you to answer the question using facts from the quoted content. Here is the document:
<document>
Anthropic: Challenges in evaluating AI systems
Introduction
Most conversations around the societal impacts of artificial intelligence (AI) come down to discussing some quality of an AI system, such as its truthfulness, fairness, potential for misuse, and so on. We are able to talk about these characteristics because we can technically evaluate models for their performance in these areas. But what many people working inside and outside of AI don't fully appreciate is how difficult it is to build robust and reliable model evaluations. Many of today's existing evaluation suites are limited in their ability to serve as accurate
Response
Here are two relevant quotes from the document:
"Many of today's existing evaluation suites are limited in their ability to serve as accurate indicators of model capabilities or safety."
"We also use evaluations to improve our safety as an organization, as illustrated by our Responsible Scaling Policy."
Based on these quotes, it seems that while evaluations of AI systems have limitations, Anthropic uses evaluations to try to improve the safety of their AI systems and organization, as shown through their Responsible Scaling Policy. Evaluations help them
Inference configuration
Temperature 0.5
Top P 1
Top K 250
Maximum length 2048
Stop sequences Human:
Guardrail ID N/A
Guardrail Version N/A


Amazon Bedrock
Knowledge bases
Knowledge bases
Chat with your document
Chat with your document
View documentation
Configurations
Model
Claude 3 Sonnet v1 | ODT
Change
Inference parameters Info
Reset
Set values to influence the responses that the model provides when you query your knowledge base.
Randomness and diversity
Temperature
0.3
Top P
1
Length
Maximum length
2048
Stop sequences
Add
Observation
Chat prompt template Info
Defines how the model handles the user prompt. You can edit the default prompt to create a custom one for your use case.
Chat with your document
Enter your message here
Run

## About model

Claude 3.5 Sonnet raises the industry bar for intelligence, outperforming competitor models and Claude 3 Opus on a wide range of evaluations, with the speed and cost of our mid-tier model, Claude 3 Sonnet.

Supported use cases

Claude 3.5 excels at complex tasks like customer support, coding, data analysis, and visual processing. It streamlines workflows, generates insights, and produces high-quality, natural-sounding content.

Model attributes

Code generation, text generation, complex reasoning and analysis

Model version

v1

Max tokens

200k

Languages

English, Spanish, Japanese, and multiple other languages

Pricing

View pricing

Model ID

anthropic.claude-3-5-sonnet-20240620-v1:0

Model ARN

arn:aws:bedrock:us-east-1::foundation-model/anthropic.claude-3-5-sonnet-20240620-v1:0

### API request

Copy code

```
{
  "modelId": "anthropic.claude-3-5-sonnet-20240620-v1:0",
  "contentType": "application/json",
  "accept": "application/json",
  "body": {
    "anthropic_version": "bedrock-2023-05-31",
```

# Agent builder Info

Manual | Assistant | Test | Prepare | Save | Save and exit

## Agent details

**Agent name**

omar-agent-1

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

**Agent description** - *optional*

*Enter description*

The description can have up to 200 characters.

**Agent resource role**

- (•) Create and use a new service role
- ( ) Use an existing service role

AmazonBedrockExecutionRoleForAgents_WI3YLYL9IS

**Select model**

Anthropic | Claude 3.5 Sonnet

Next-gen AI assistant trained on helpful, honest, and harmless AI systems, Claude can help with summarization, search, creative writing, Q&A, coding, as well as take direction.

**Instructions for the Agent**

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

*Enter instructions*

This instruction must have a minimum of 40 characters.

### ▼ Additional settings

**Code Interpreter** Preview

Code Interpreter enables agents to handle tasks that involve writing, running, testing, and troubleshooting code in a secure environment.

- ( ) Enabled
- (•) Disabled

## Test Agent

Using ODT Change

*Enter your message here*

Run

Step 2 - *optional*
Configure content filters

Step 3 - *optional*
Add denied topics

Step 4 - *optional*
Add word filters

Step 5 - *optional*
Add sensitive information filters

Step 6 - *optional*
Add contextual grounding check

Step 7
Review and create

## Harmful categories

Enable to detect and block harmful user inputs and model responses. Use a higher filter strength to increase the likelihood of filtering harmful content in a given category.

Enable harmful categories filters

### Filters for prompts

Reset all

Hate — None | Low | Medium | High

Insults — None | Low | Medium | High

Sexual — None | Low | Medium | High

Violence — None | Low | Medium | High

Misconduct — None | Low | Medium | High

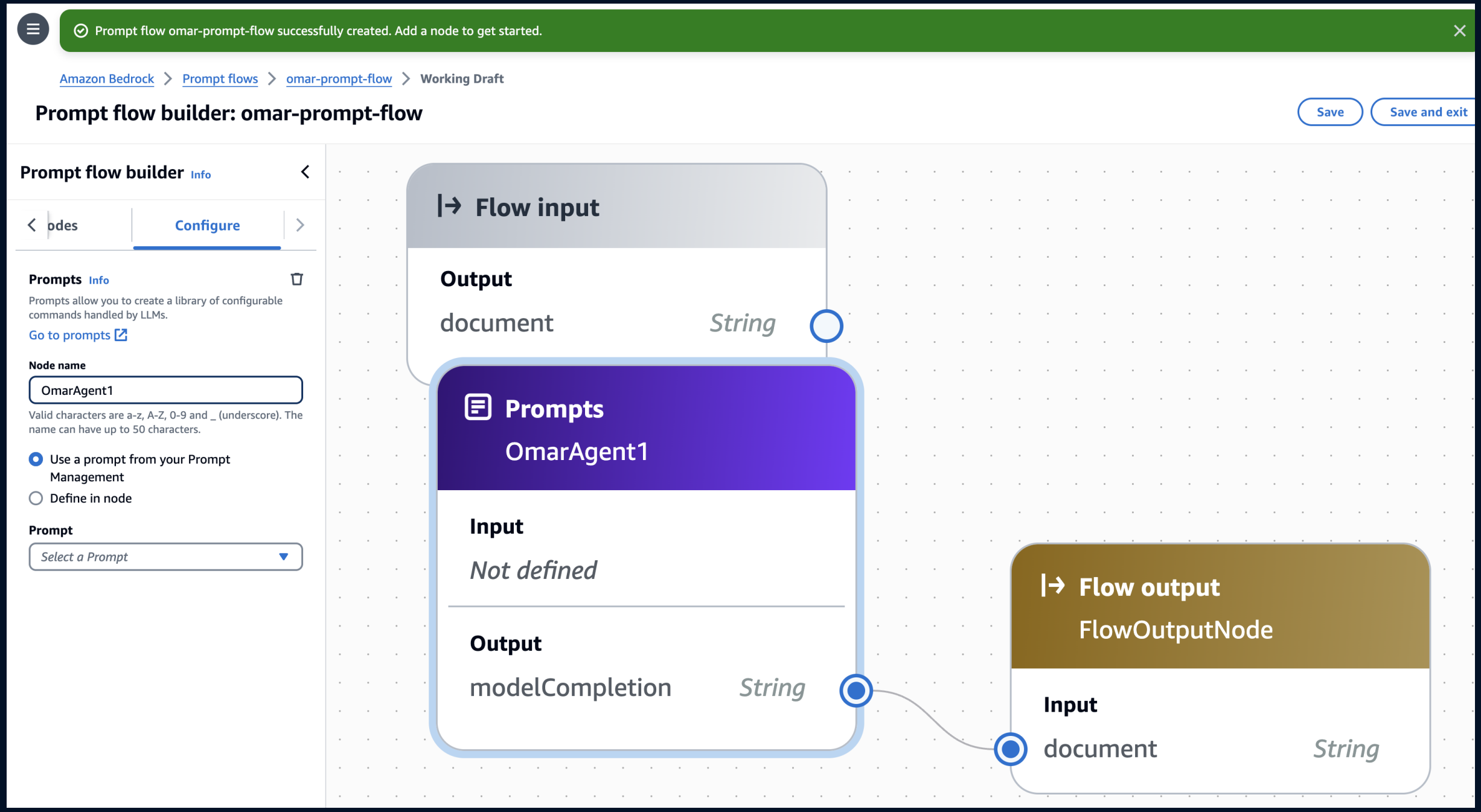
Use the same harmful categories filters for responses

## Prompt attacks

Enable to detect and block user inputs attempting to override system instructions. To avoid misclassifying system prompts as a prompt attack and ensure that the filters are selectively applied to user inputs, use input tagging.

Enable prompt attacks filter

Prompt Attack — None | Low | Medium | High


Prompt flow omar-prompt-flow successfully created. Add a node to get started.
Amazon Bedrock
Prompt flows
omar-prompt-flow
Working Draft
Prompt flow builder: omar-prompt-flow
Save
Save and exit
Prompt flow builder Info
odes
Configure
Prompts Info
Prompts allow you to create a library of configurable commands handled by LLMs.
Go to prompts
Node name
OmarAgent1
Valid characters are a-z, A-Z, 0-9 and _ (underscore). The name can have up to 50 characters.
Use a prompt from your Prompt Management
Define in node
Prompt
Select a Prompt
Flow input
Output
document
String
Prompts
OmarAgent1
Input
Not defined
Output
modelCompletion
String
Flow output
FlowOutputNode
Input
document
String

# Google Vertex AI

# Model Garden

EXPLORE GENERATIVE AI | VIEW MY ENDPOINTS & MODELS | DEPLOY FROM HUGGING FACE

## Modalities

| | |
|---|---|
| Language | 64 |
| Vision | 87 |
| Tabular | 7 |
| Document | 6 |
| Speech | 1 |
| Video | 6 |
| Multimodal | 19 |

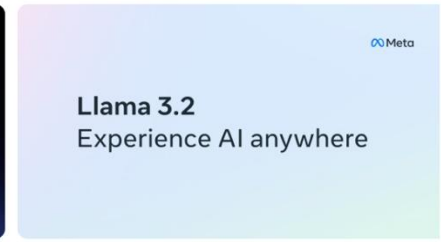
## Tasks

| | |
|---|---|
| Generation | 72 |
| Classification | 64 |
| Detection | 43 |
| Extraction | 27 |
| Recognition | 24 |
| Translation | 21 |
| Embedding | 7 |
| Segmentation | 10 |
| Retrieval | 2 |
| Open vocabulary detection | 2 |

Search models

Browse, customize, and deploy machine learning models with **Model Garden**. Choose from models created by Google and other providers.




Sort by: **Trending** Newest Last Update

## Foundation models

SHOW ALL (89)

Pre-trained multi-task models that can be further tuned or customized for specific tasks.

### Gemini 1.5 Pro

Created from the ground up to be multimodal (text, images, videos) and to scale across a wide range of tasks

### Gemini 1.5 Flash

The best performing Gemini model with features for a wide range of tasks

### Gemini 1.0 Pro

Designed to balance quality, performance, and cost for tasks such as content generation, editing, summarization, and classification

### Gemini 1.0 Pro Vision

Created to be multimodal (text, images, code) and to scale across a wide range of tasks

### Imagen for Ger Edit

Use text prompts images, edit exist parts of an image more.

## Featured partners

Pipelines

UPLOAD CREATE RUN LEARN REFRESH

TEMPLATE GALLERY YOUR TEMPLATES SCHEDULES RUNS

**Type**

| | |
|---|---|
| Component | 58 |
| Pipeline | 12 |

**Integration**

| | |
|---|---|
| Vertex AI | 39 |
| BigQuery | 26 |
| Dataflow | 1 |
| Dataproc | 4 |

Search

## AutoML for Tabular Classification / Regression

Complete AutoML Tables pipeline. Includes feature engineering, architecture search, and hyper-parameter tuning.

Pipeline

CREATE RUN VIEW DETAILS

Public preview

## Learn to Learn Forecasting Pipeline on Tabular Workflows

The AutoML Forecasting pipeline.

Pipeline

CREATE RUN VIEW DETAILS

## LLM Text Generation Evaluation

LLM Text Generation Evaluation pipeline. This pipeline supports evaluating large language models, publisher or managed models, performing the following generative tasks: `summarization`, `question-answering`, and `text-generation`.

Pipeline

CREATE RUN VIEW DETAILS

## Model-Based LLM Side-By-Side Evaluation

Determines the SxS winrate between two models.

Pipeline

CREATE RUN VIEW DETAILS

Public preview

## Reinforcement Learning from AI Feedback

Performs reinforcement learning from AI feedback.

Pipeline

CREATE RUN VIEW DETAILS

Public preview

## Reinforcement Learning from Human Feedback

Performs reinforcement learning from human feedback.

Pipeline

CREATE RUN VIEW DETAILS

Public preview

## Sequence to Sequence Forecasting Pipeline on Tabular Workflows

Public preview

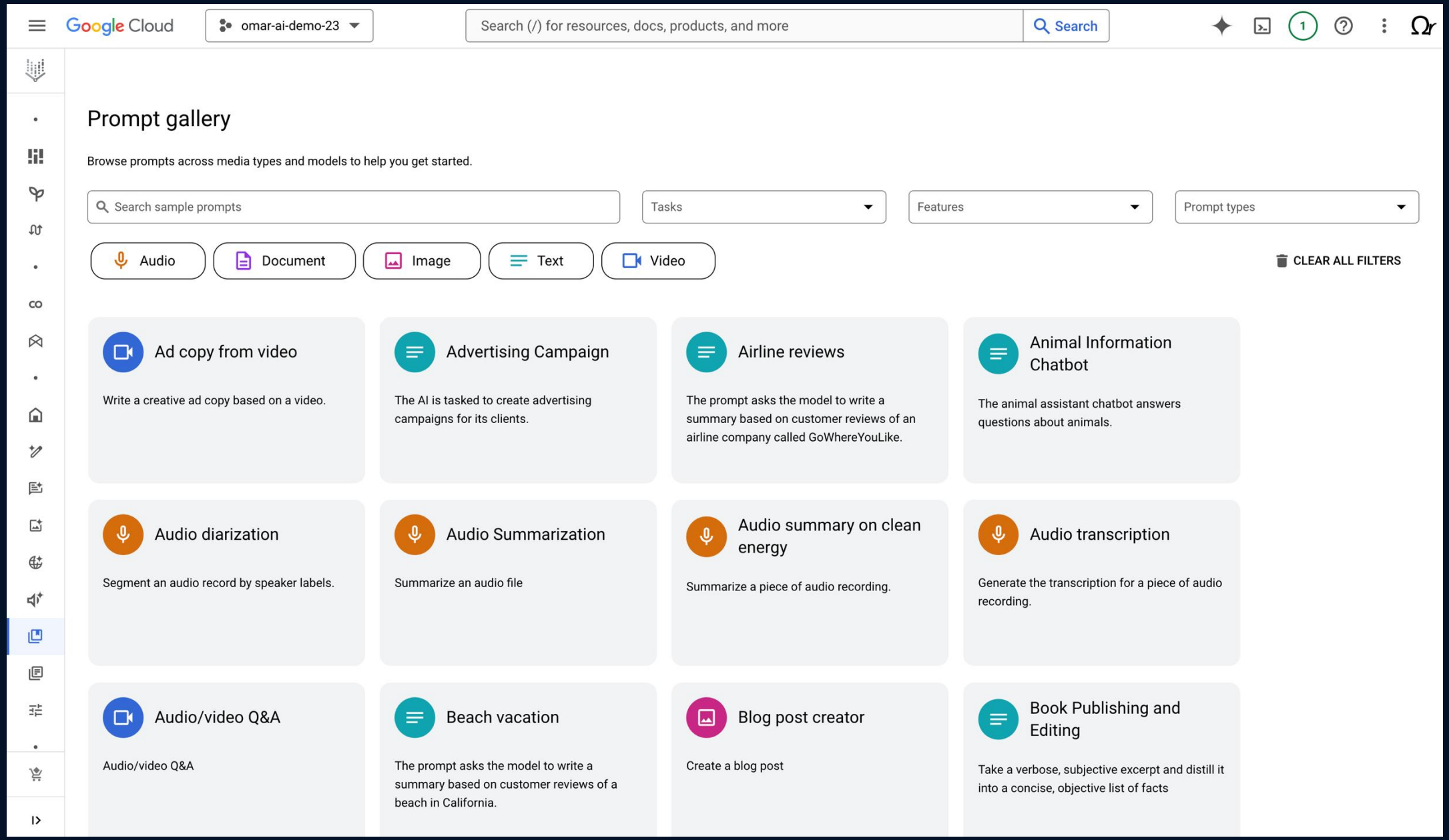
## Starry Net Forecasting Pipeline

Starry Net is a state-of-the-art forecaster used internally

Public preview

## TabNet Training Pipeline on Tabular Workflows

https://console.cloud.google.com/vertex-ai/pipelines/vertex-ai-templates?hl=en&inv=1&invt=AbeU1Q&project=omar-ai-demo-23


Google Cloud
omar-ai-demo-23
Search (/) for resources, docs, products, and more
Search
Prompt gallery
Browse prompts across media types and models to help you get started.
Search sample prompts
Tasks
Features
Prompt types
Audio
Document
Image
Text
Video
CLEAR ALL FILTERS
Ad copy from video
Write a creative ad copy based on a video.
Advertising Campaign
The AI is tasked to create advertising campaigns for its clients.
Airline reviews
The prompt asks the model to write a summary based on customer reviews of an airline company called GoWhereYouLike.
Animal Information Chatbot
The animal assistant chatbot answers questions about animals.
Audio diarization
Segment an audio record by speaker labels.
Audio Summarization
Summarize an audio file
Audio summary on clean energy
Summarize a piece of audio recording.
Audio transcription
Generate the transcription for a piece of audio recording.
Audio/video Q&A
Audio/video Q&A
Beach vacation
The prompt asks the model to write a summary based on customer reviews of a beach in California.
Blog post creator
Create a blog post
Book Publishing and Editing
Take a verbose, subjective excerpt and distill it into a concise, objective list of facts

Google Cloud | omar-ai-demo-23 | Search (/) for resources, docs, products, and more | Search

Data sources and monitoring | History | Compare | Notes | API reference | Save | Get code

**System instructions** *You are focusing on enhancing machine learning systems by providing the requested code enhancements. You always briefly ment...* Edit

**Prompt** Insert Media | Add examples | Add variable | Clear Prompt

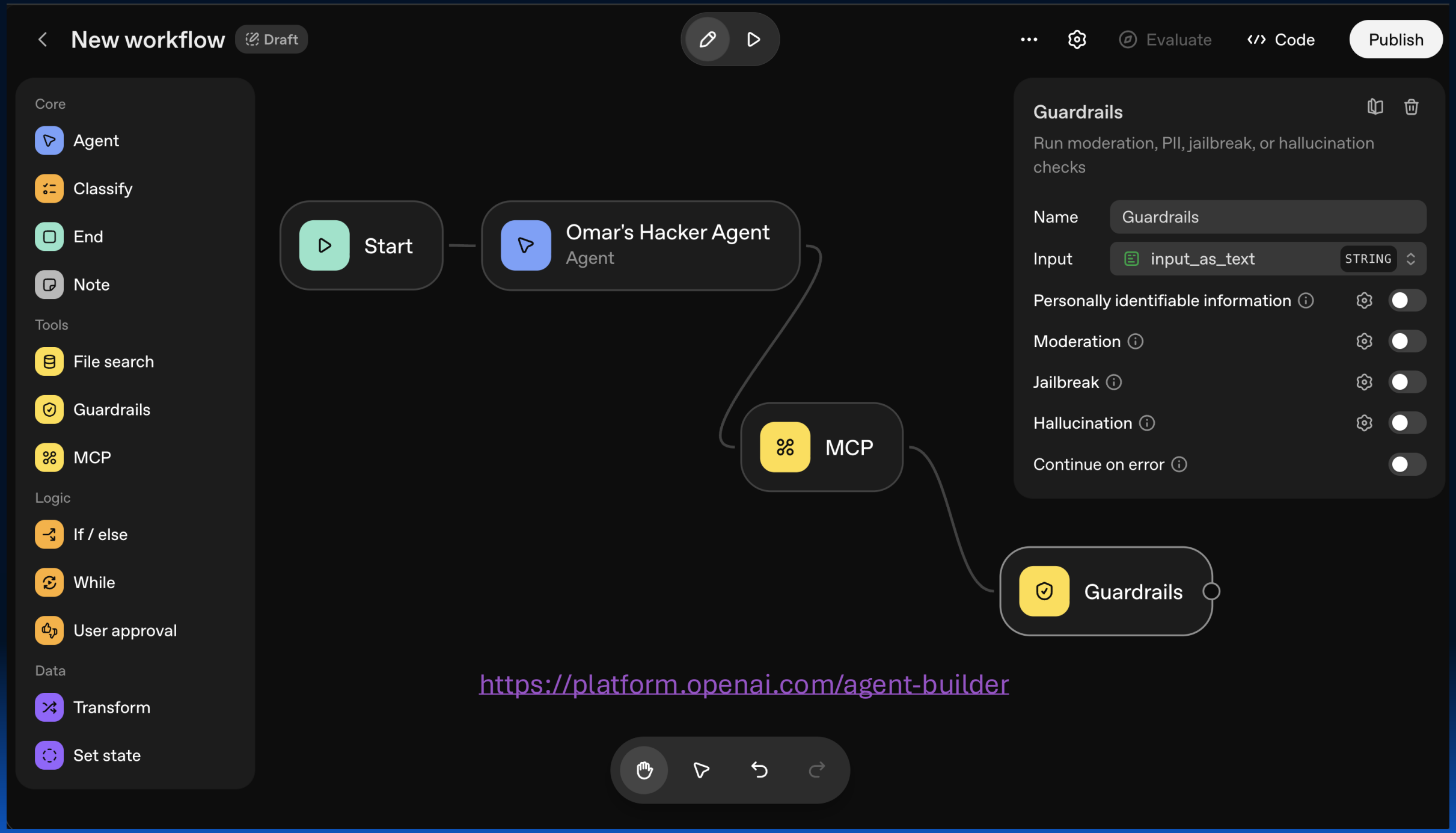
I am working on a sentiment analysis project that processes customer feedback using TensorFlow and Keras. Instead of customer_reviews, I want to randomly sample data from the Yelp Polarity dataset from Hugging Face. Sample only the training data, not the test or validation data. Do the sampling before the tokenization. I also want to integrate resource usage monitoring. Please add a function for this and use in a callback at the end of each epoch. It should monitor and log CPU usage and memory usage.

Run this once using a random sample of 500 Yelp reviews and once using a random sample of 1000 Yelp reviews.

Here is my code:

```

800 tokens

**Response** Refine prompt | Markdown

The model will generate a response after you click Submit

Model: gemini-1.5-flash-002

Region *: us-central1 (Iowa)

Temperature: 0 — 2 — 0.2

Output token limit: 1 — 8192 — 2048

Grounding: Source: Google Search

Customize

Add stop sequence

Press Enter after each sequence

Output format: Plain text | Edit

Safety Filter Settings

Advanced

Reset parameters

Share feedback
```

# OpenAI Agent Builder


New workflow
Draft
Evaluate
Code
Publish
Core
Agent
Classify
End
Note
Tools
File search
Guardrails
MCP
Logic
If / else
While
User approval
Data
Transform
Set state
Start
Omar's Hacker Agent
Agent
MCP
Guardrails
Guardrails
Run moderation, PII, jailbreak, or hallucination checks
Name
Guardrails
Input
input_as_text
STRING
Personally identifiable information
Moderation
Jailbreak
Hallucination
Continue on error
https://platform.openai.com/agent-builder

# OpenAI Agent SDK

If you're familiar with uv, using the tool would be even similar:

```
uv init
uv add openai-agents
```

For voice support, install with the optional `voice` group: `uv add 'openai-agents[voice]'`.

For Redis session support, install with the optional `redis` group: `uv add 'openai-agents[redis]'`.

## Hello world example

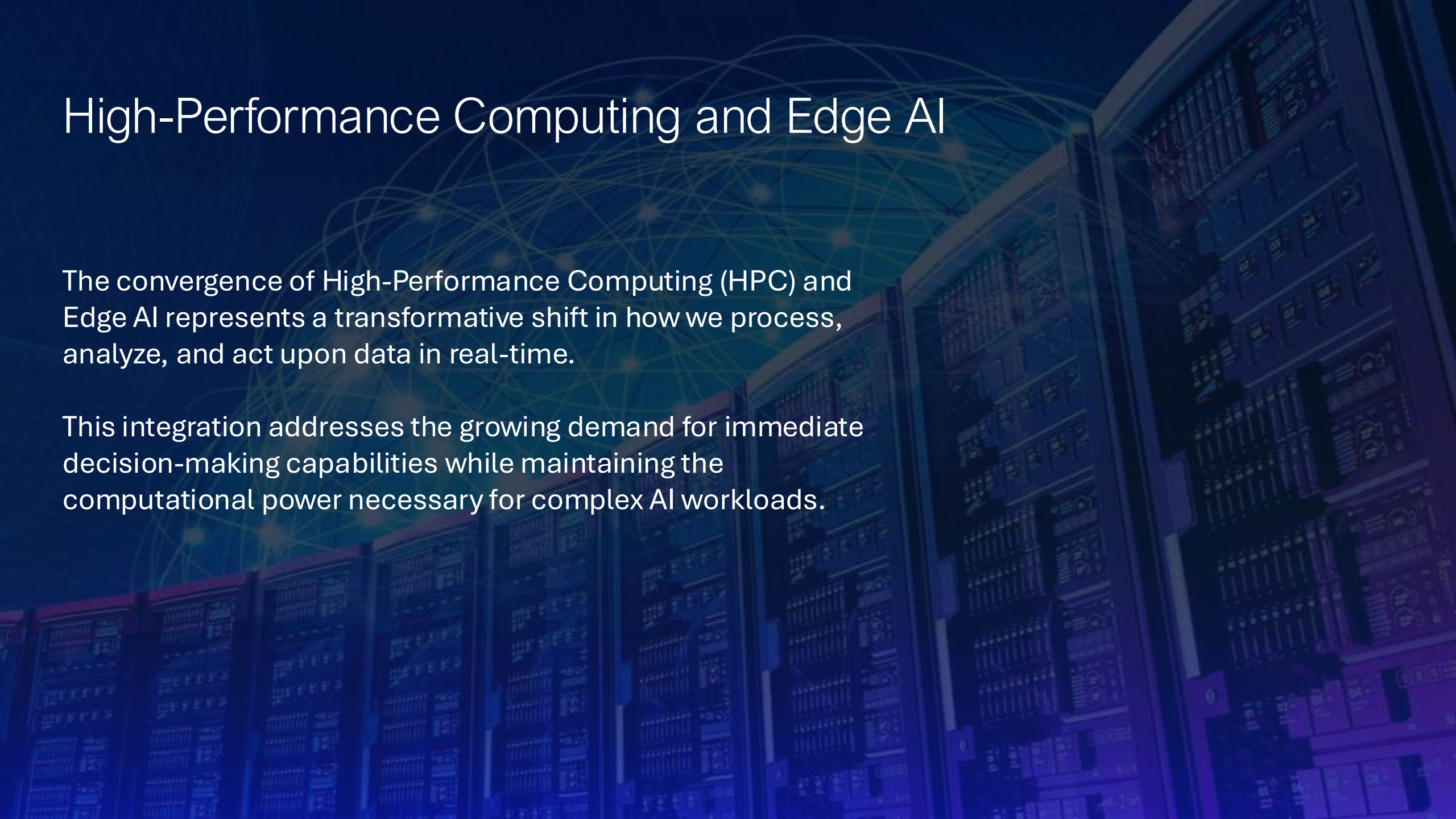
```python
from agents import Agent, Runner

agent = Agent(name="Assistant", instructions="You are a helpful assistant")

result = Runner.run_sync(agent, "Write a haiku about recursion in programming.")
print(result.final_output)

# Code within the code,
# Functions calling themselves,
# Infinite loop's dance.
```

(*If running this, ensure you set the* `OPENAI_API_KEY` *environment variable*)

https://github.com/openai/openai-agents-python

# High-Performance Computing and Edge AI

# High-Performance Computing and Edge AI



The convergence of High-Performance Computing (HPC) and Edge AI represents a transformative shift in how we process, analyze, and act upon data in real-time.

This integration addresses the growing demand for immediate decision-making capabilities while maintaining the computational power necessary for complex AI workloads.

# What is High-Performance Computing?

High-Performance Computing refers to the use of powerful computing systems, including supercomputers, GPU clusters, and specialized accelerators, to perform complex computations at speeds far exceeding traditional computing systems. HPC systems leverage:

- Parallel Processing: Distributing computational tasks across multiple processors simultaneously
- High-Speed Interconnects: Ultra-fast networking technologies like InfiniBand for rapid data transfer
- Specialized Hardware: GPUs, TPUs, and custom accelerators optimized for specific workloads
- Distributed Memory Systems: Large-scale memory architectures that support massive datasets

https://github.com/santosomar/build-your-ai-lab

# Resources

# Recommended Resources

- ✨ AI and Cybersecurity Resources in O'Reilly: https://hackertraining.org
- Hacking Scenarios (Labs) in O'Reilly: https://hackingscenarios.com
- My Personal Blog: https://becomingahacker.org
- My Cisco Blog: https://blogs.cisco.com/author/omarsantos
- Upcoming Live Cybersecurity and AI Training in O'Reilly: Register before it is too late

# Q&A

Please remember to complete the survey

# Thank you!

 /santosomar

 @santosomar

 hackerrepo.org