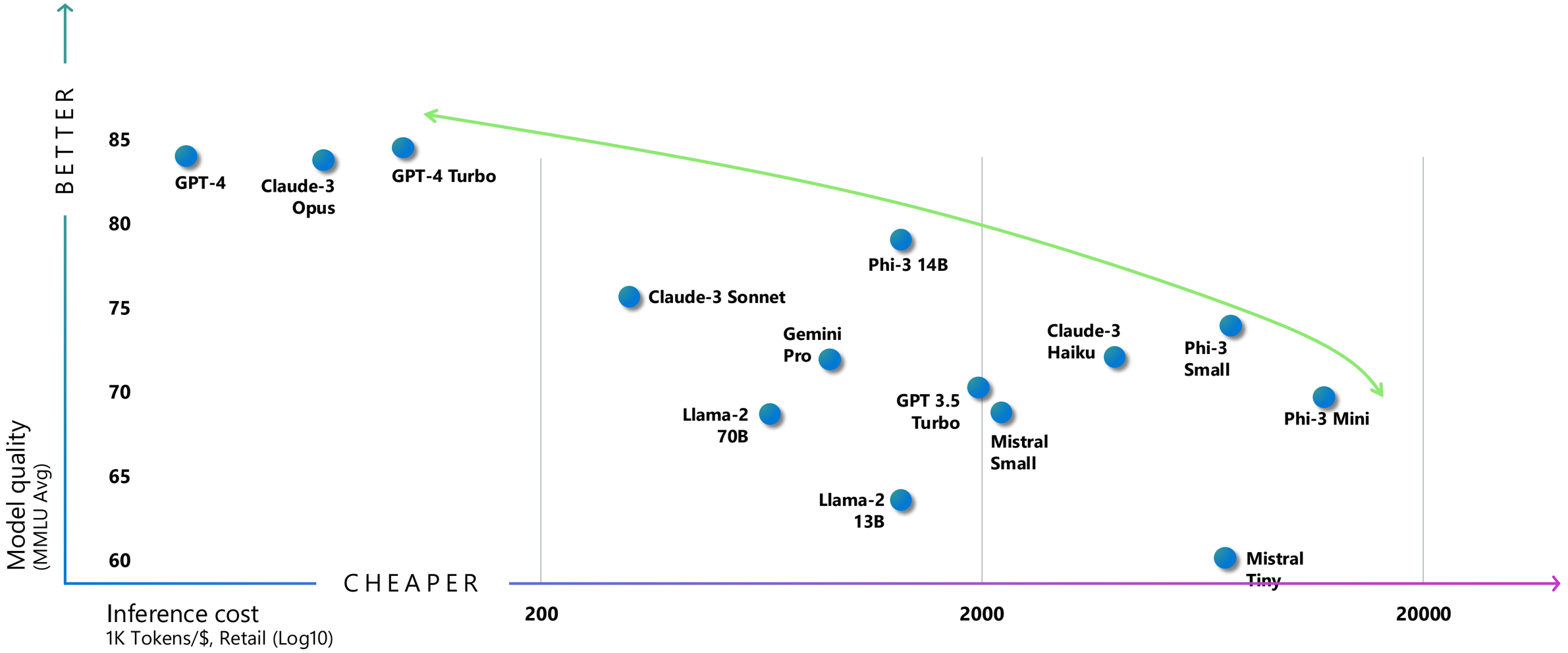
Phi-3 with NVIDIA NIM

Kinfey Lo

Microsoft Senior Cloud Advocate

小模型开始入局


BETTER
85
80
75
70
65
60
Model quality
(MMLU Avg)
GPT-4
Claude-3
Opus
GPT-4 Turbo
Phi-3 14B
Claude-3 Sonnet
Gemini
Pro
Claude-3
Haiku
Phi-3
Small
GPT 3.5
Turbo
Llama-2
70B
Mistral
Small
Phi-3 Mini
Llama-2
13B
Mistral
Tiny
CHEAPER
Inference cost
1K Tokens/$, Retail (Log10)
200
2000
20000

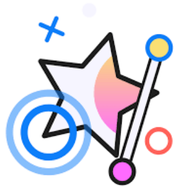
我们选择小模型的理由

性能

1.更少的算力需求
2.部署在更小的设备甚至边缘计算场景上

无障碍

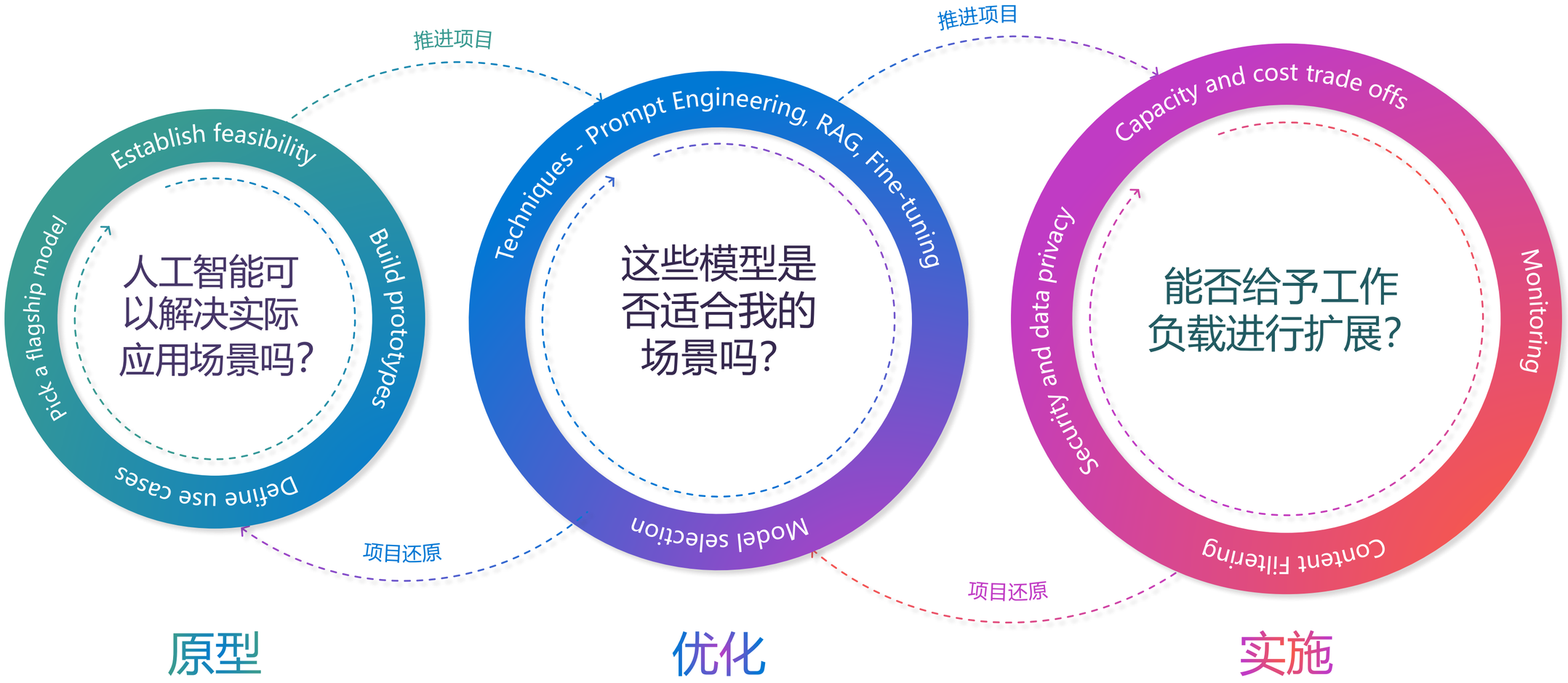
1.更多开发者和组织可以使用
2.具备一定的业务能力，便于企业开发人员使用



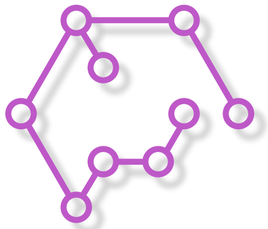
定制化

1.针对特定领域和任务进行微调
2.所有权

对开发团队的挑战

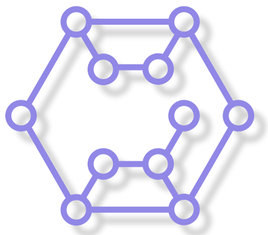

推进项目
推进项目
Establish feasibility
Build prototypes
Define use cases
pick a flagship model
人工智能可以解决实际应用场景吗?
Techniques - Prompt Engineering, RAG, Fine-tuning
Model selection
这些模型是否适合我的场景吗?
Capacity and cost trade offs
Monitoring
Content Filtering
Security and data privacy
能否给予工作负载进行扩展?
项目还原
项目还原
原型
优化
实施

Microsoft Phi 3 Family



Phi-3-mini

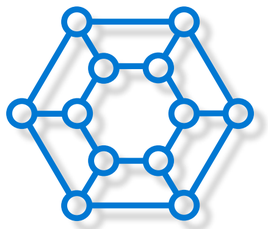
Phi-3-mini 是一种 3.8B 语言模型，可在 Microsoft Azure AI Studio、Hugging Face 和 Ollama 上使用。 Phi-3 模型在关键基准测试中显着优于相同和更大尺寸的语言模型（请参阅下面的基准数据，越高越好）。 Phi-3-mini 的性能优于两倍大小的型号，Phi-3-small 和 Phi-3-medium 的性能优于更大的型号，包括 GPT-3.5



Phi-3-small & medium

仅 7B 参数的 Phi-3-small 在各种语言、推理、编码和数学基准测试中击败了 GPT-3.5T。

具有 14B 参数的 Phi-3-medium 延续了这一趋势，并且性能优于 Gemini 1.0 Pro。

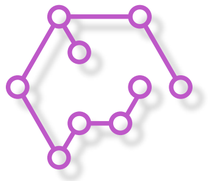


Phi-3-Vision

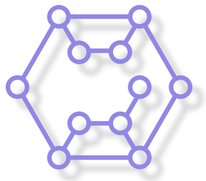
仅 4.2B 参数的 Phi-3-vision 延续了这一趋势，并且在一般视觉推理任务、OCR、表格和图表理解任务中优于 Claude-3 Haiku 和 Gemini 1.0 Pro V 等较大模型。

Microsoft Phi 进化史

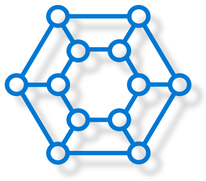
microsoft/phi-1
microsoft/phi-1.5
microsoft/phi-2



Phi-3-mini



Phi-3-small & medium



Phi-3-Vision

从单兵作战，变成成团作团作战，并且不断演变

Benchmarks	Original	June 2024 Update
Instruction Extra Hard	5.7	5.9
Instruction Hard	5.0	5.2
JSON Structure Output	1.9	60.1
XML Structure Output	47.8	52.9
GPQA	25.9	29.7
MMLU	68.1	69.7
Average	**25.7**	**37.3**

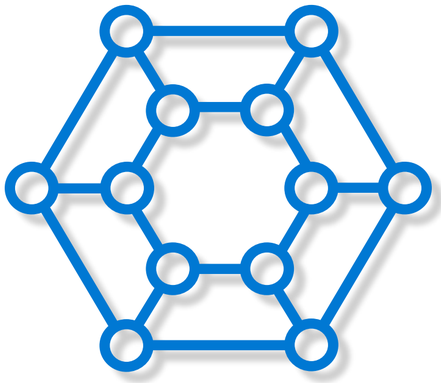
RULER: a retrieval-based benchmark for long context understanding

Model	4K	8K	16K	32K	64K	128K	Average
Original	86.7	78.1	75.6	70.3	58.9	43.3	**68.8**
June 2024 Update	92.4	91.1	90.8	87.9	79.8	65.6	**84.6**

RepoQA: a benchmark for long context code understanding

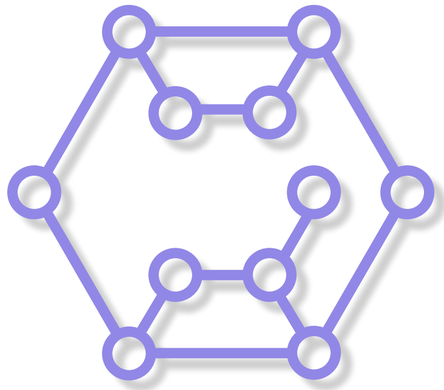
Model	Python	C++	Rust	Java	TypeScript	Average
Original	27	29	40	33	33	**32.4**
June 2024 Update	85	63	72	93	72	**77**

量化模型(INT4,FP16,FP32)



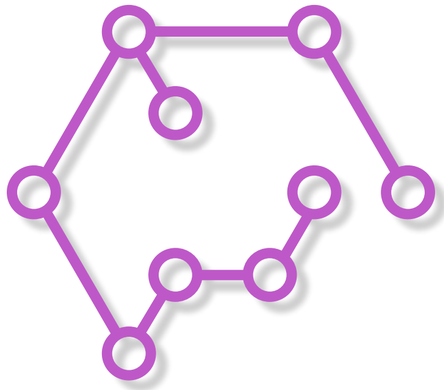
Phi-3-mini-V

(3.8B + 0.3B)



Phi-3-small

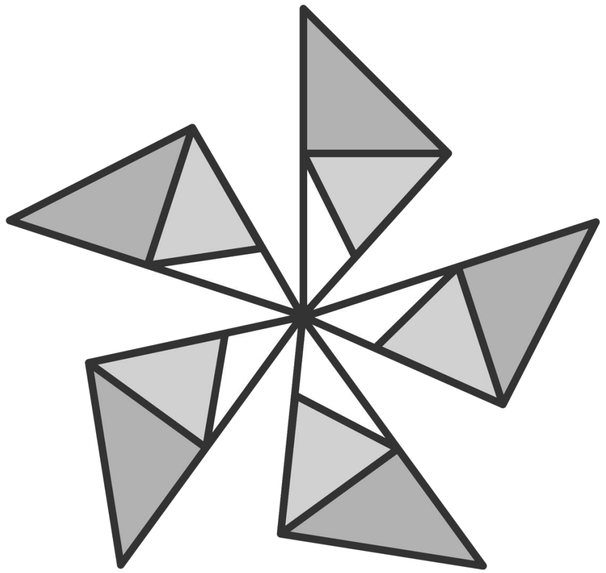
(7B)



Phi-3-mini

(3.8B)

GGUF



ONNX Runtime

Play Phi-3

Blog Discord GitHub

Search models

Models Sign in Download

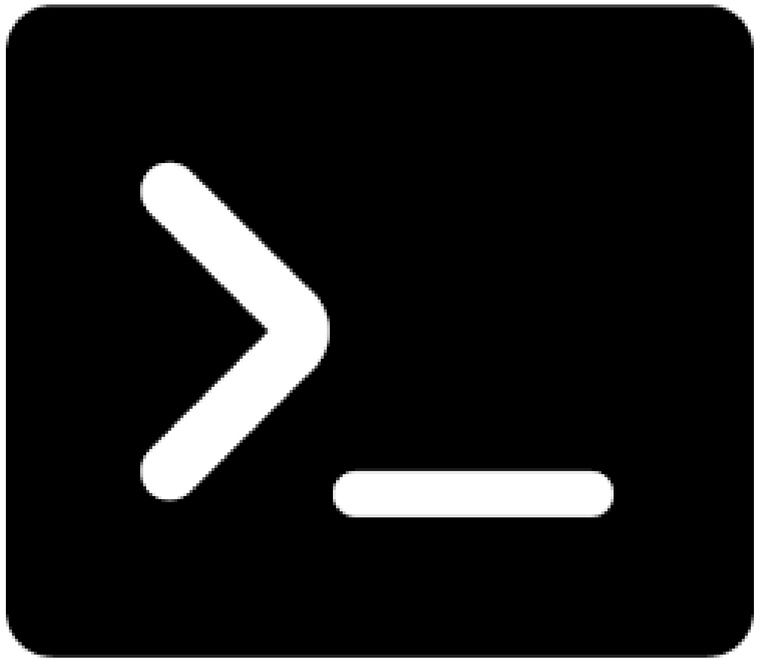
Get up and running with large language models.

Run Llama 3, Phi 3, Mistral, Gemma, and other models. Customize and create your own.

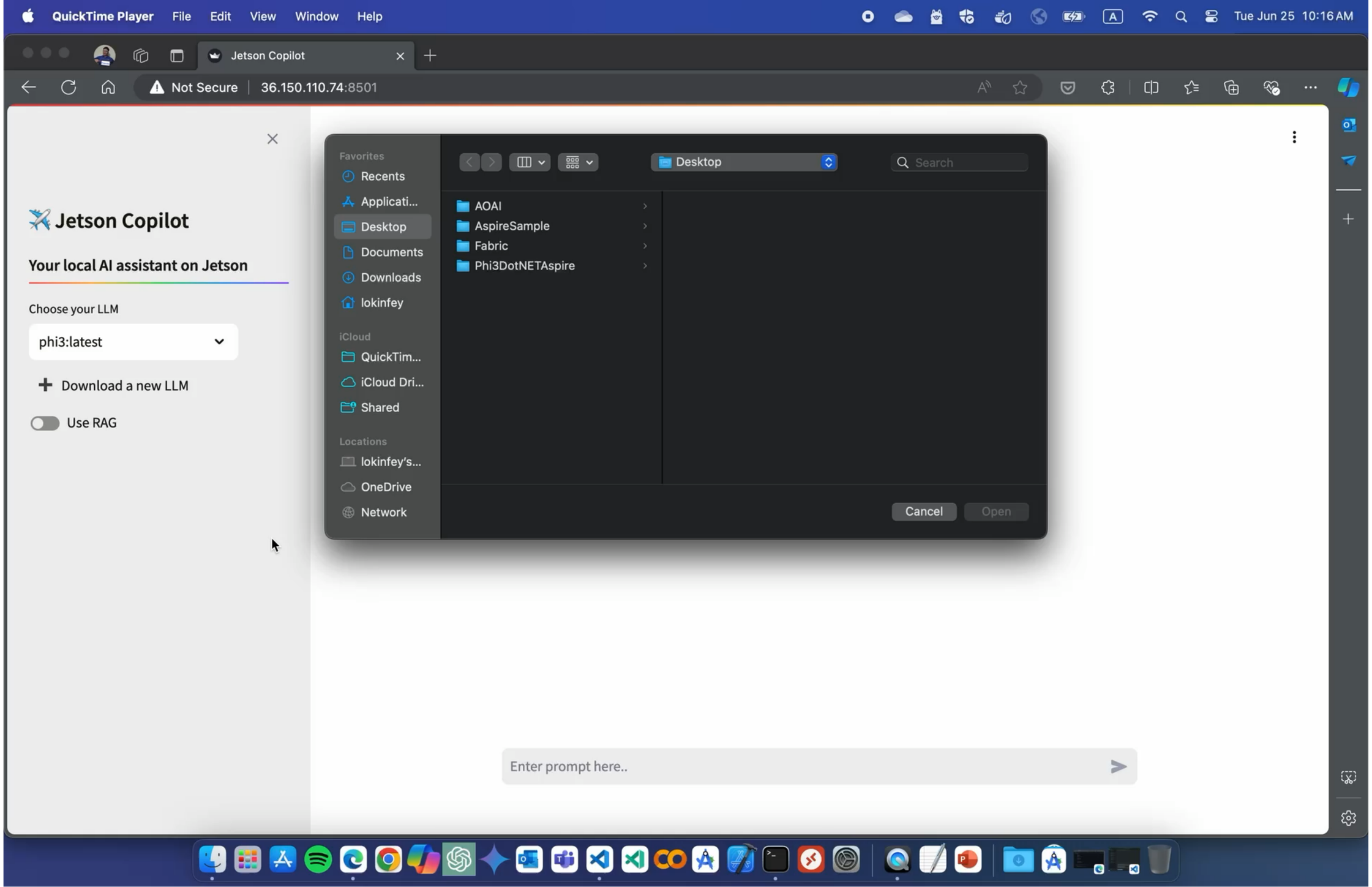
Download ↓

Available for macOS, Linux, and Windows (preview)

https://ollama.com/



ollama run phi3

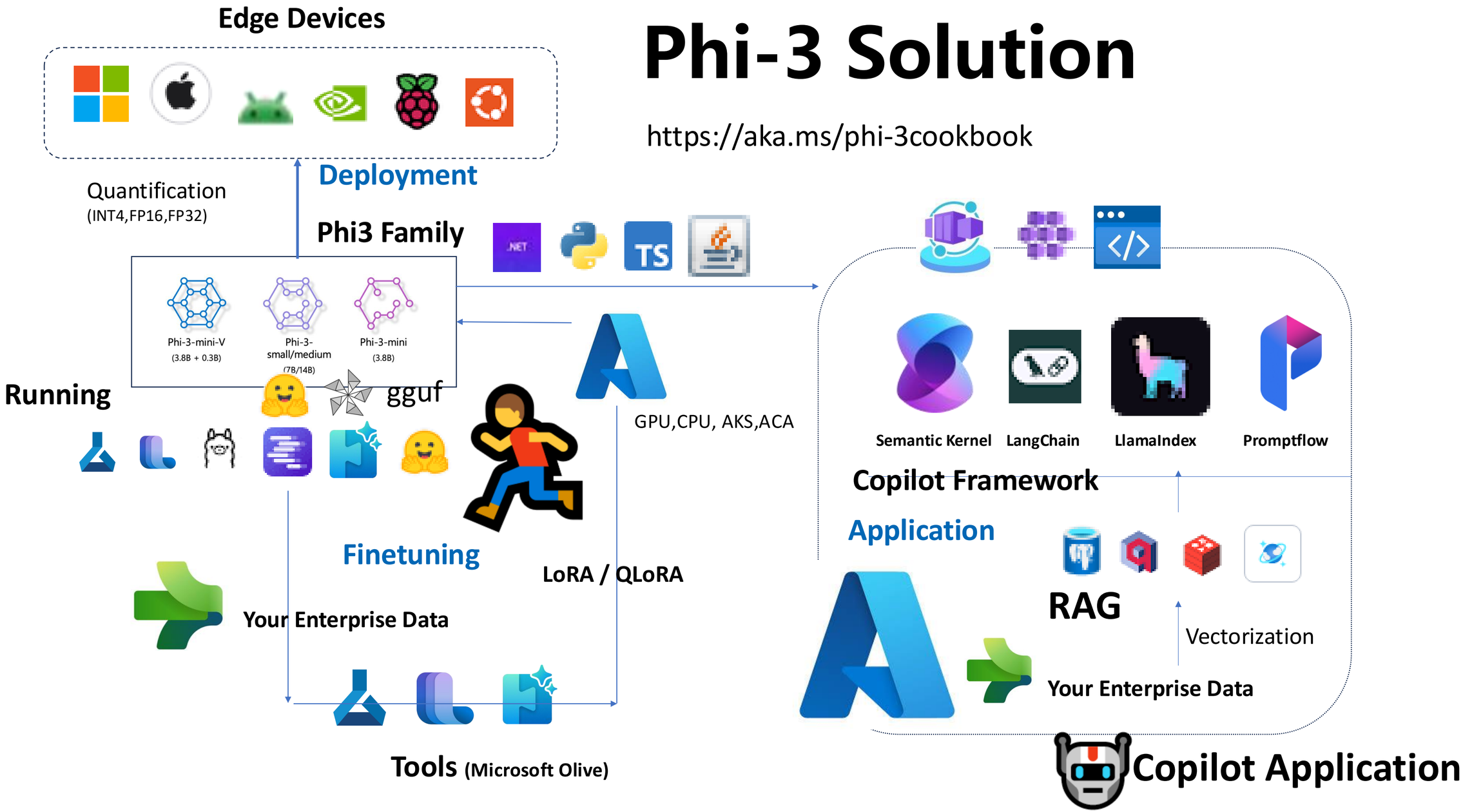

QuickTime Player File Edit View Window Help
Tue Jun 25 10:16 AM
Jetson Copilot
Not Secure | 36.150.110.74:8501
Jetson Copilot
Your local AI assistant on Jetson
Choose your LLM
phi3:latest
Download a new LLM
Use RAG
Favorites
Recents
Applicati...
Desktop
Documents
Downloads
lokinfey
iCloud
QuickTim...
iCloud Dri...
Shared
Locations
lokinfey's...
OneDrive
Network
Desktop
Search
AOAI
AspireSample
Fabric
Phi3DotNETAspire
Cancel
Open
Enter prompt here..

Running Phi3 in Jetson

```
Model: NVIDIA Jetson Xavier NX Developer Kit - Jetpack 5.1.2 [L4T 35.4.1]
 1  [|||||                                                    8.1%] 1.4GHz 4  [||||||
 2  [||||||||||                                              15.0%] 1.4GHz 5  [||||||
 3  [||||||||                                                12.0%] 1.4GHz 6  [||||||
Mem [|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||4.0G/6.7G] FAN [||||||||||||||||||||||||||
Swp [                                                        32.0M/4.0G] Jetson Clocks: [running]
Emc [204MHz:::::::::::::::::::::::::::::::::::::::::::::::::::::1.9GHz] 1.9GHz   0% NV Power[8]: MODE_20W_6CORE
                                                                        Uptime: 0 days 0:23:5
CPU [
```

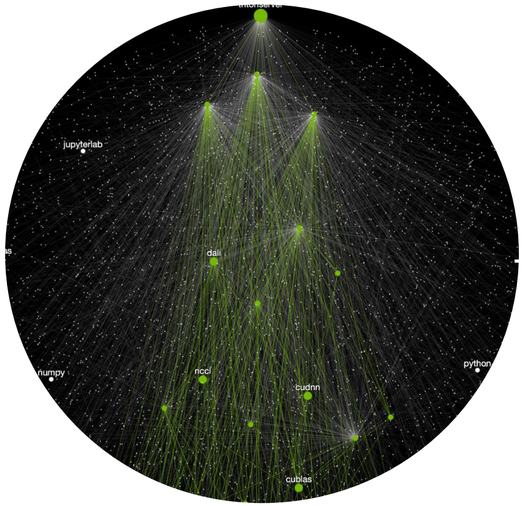
Phi-3 Solution

https://aka.ms/phi-3cookbook



开发者需要更简洁的方式

NIM 基于云原生的部署



Triton has 417 packages/libraries
across OSS, 3rd party and NVIDIA

Prompt

Event

Industry Standard APIs

Text | Speech | Image | Video | 3D | Biology

Cloud Native Container

K8s Support | Metrics & Monitoring | Identity | Secret Management | Liveness Probe

Triton Inference Server

cuDF | CV-CUDA | DALI | NCCL

Pre Processing

Post Processing

Inflight Batching

TensorRT Engine

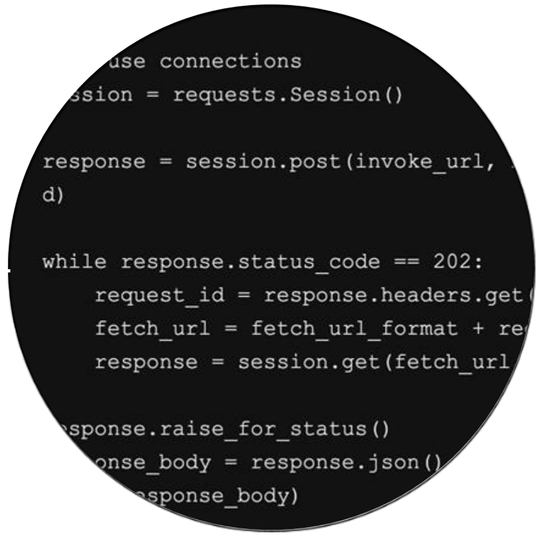
cuBLAS | cuDNN | In-Flight Batching | Memory Optimization | FP8 Quantization

Customization Cache

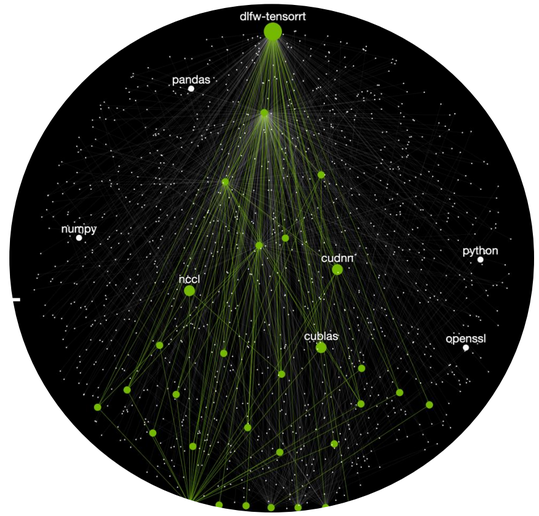
LORA | P-tuning

AI Model

Text-to-Text | Text-to-Image | Text-to-3D | Multimodal | ASR | Text-to-Speech



```
use connections
ssion = requests.Session()

response = session.post(invoke_url,
d)

while response.status_code == 202:
    request_id = response.headers.get(
    fetch_url = fetch_url_format + re
    response = session.get(fetch_url

sponse.raise_for_status()
onse_body = response.json()
sponse_body)
```



TensorRT has 333 packages/libraries
across OSS, 3rd party and NVIDIA

NIM 在 Azure 上的部署

NIM on Azure Kubernetes Service (AKS)

To deploy NIM on AKS successfully, ensure you have the right GPU and driver version. The default GPU driver in Azure Kubernetes Services (AKS) is usually outdated for the latest NVIDIA software, and Microsoft does not yet have an official solution for this issue.

To resolve this, use the preview version of the CLI to create the AKS cluster. The Prerequisites section explains how to set up your local environment to enable AKS creation with the preview CLI.

After you are ready to create AKS, the next thing is to choose the right GPU instance. Only L40S, A100, H100 GPU work for NIM but not all system configurations. Create AKS section has more details about this.

Prerequisites

Please follow Pre-rquirement instruction to get ready for AKS creation.

Create AKS

Please follow Create AKS instruction to create AKS.

Deploy NIM

Please follow Deploy NIM instruction to create AKS.

nim-deploy/cloud-service-providers/azure/aks at main · NVIDIA/nim-deploy (github.com)

NVIDIA NIM support Microsoft Phi 3 Family

Discover

MODELS

Reasoning

Vision

Visual Design

Retrieval

Speech

Biology

Simulation

INDUSTRIES

Gaming

Healthcare

Industrial

Showing results for "Phi-3"

microsoft / **phi-3-mini-128k-instruct**

Lightweight, state-of-the-art open LLM with strong math and logical reasoning skills.

chat text-to-text

microsoft / **phi-3-mini-4k-instruct**

Lightweight, state-of-the-art open LLM with strong math and logical reasoning skills.

chat text-to-text

microsoft / **phi-3-medium-128k-instruct**

Cutting-edge lightweight open language model exceling in high-quality reasoning.

chat text-to-text

microsoft / **phi-3-medium-4k-instruct**

Cutting-edge lightweight open language model exceling in high-quality reasoning.

chat text-to-text

microsoft / **phi-3-small-128k-instruct**

Long context cutting-edge lightweight open language model exceling in high-quality reasoning.

chat text-to-text

microsoft / **phi-3-small-8k-instruct**

Cutting-edge lightweight open language model exceling in high-quality reasoning.

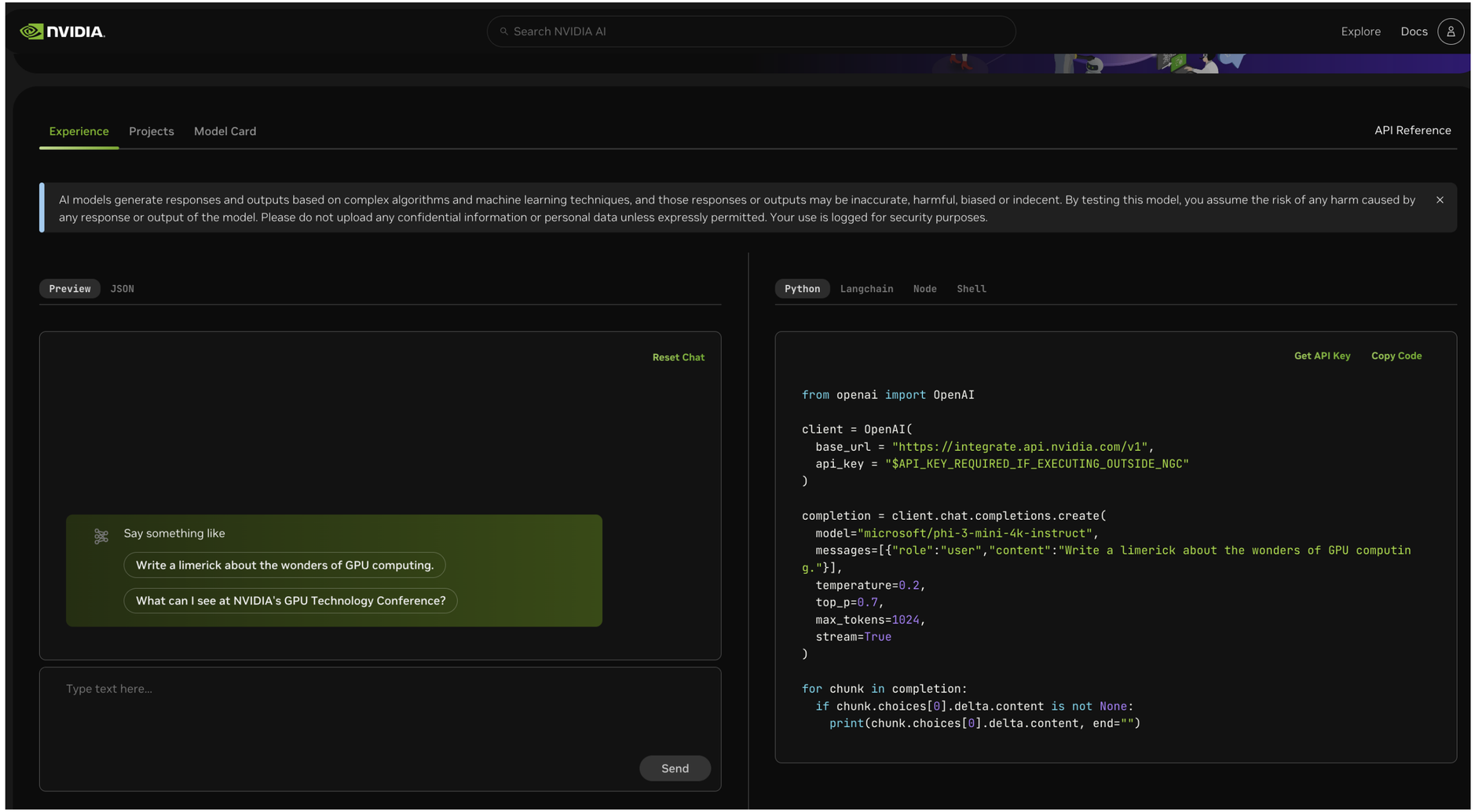
chat text-to-text

microsoft / **phi-3-vision-128k-instruct**

Cutting-edge open multimodal model exceling in high-quality reasoning from images.

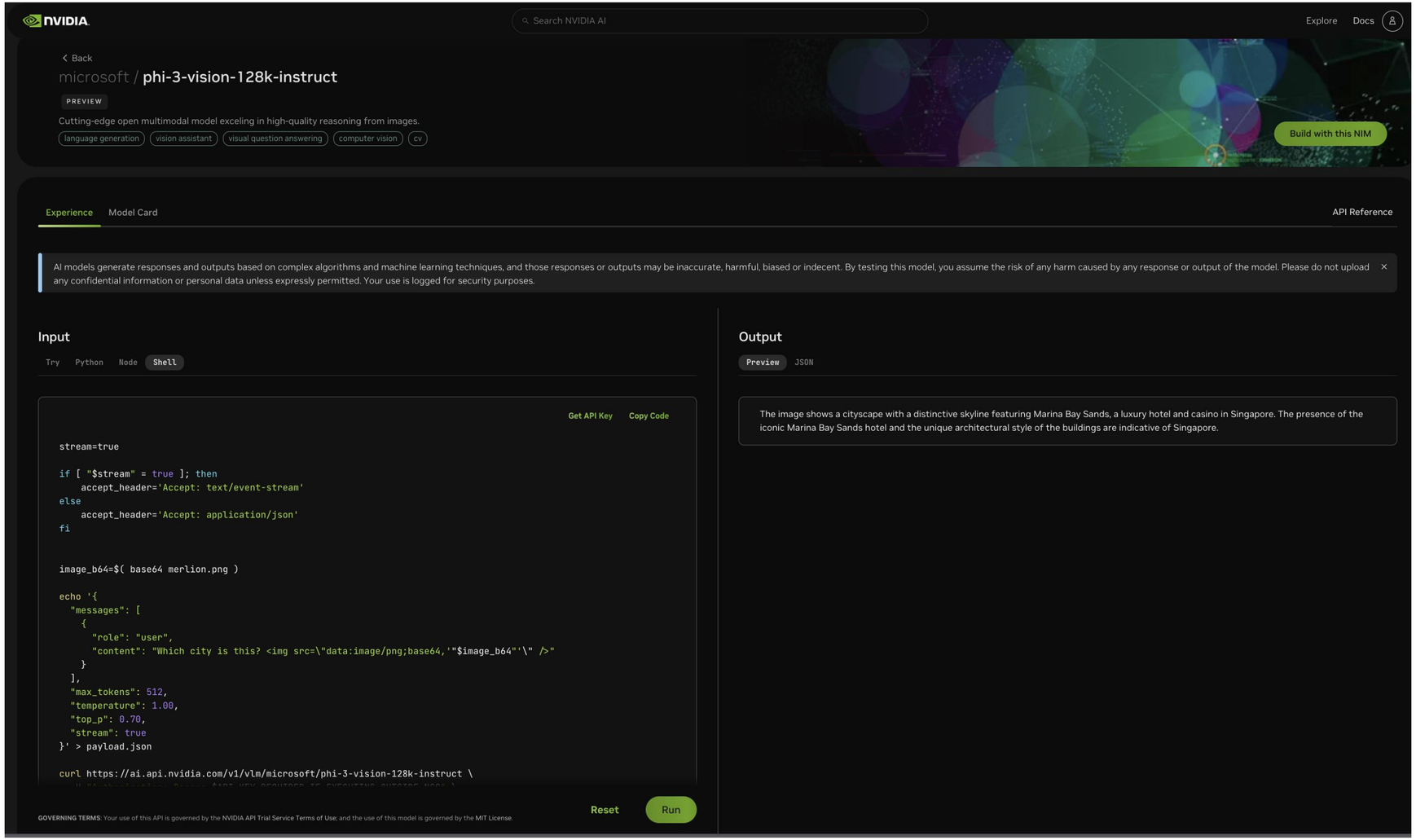
image cv

NVIDIA NIM API – phi-3-mini-4k



https://build.nvidia.com/microsoft/phi-3-mini-4k

NVIDIA NIM API – phi-3-vision-128k



NVIDIA NIM API - Samples

```
client = OpenAI(
  base_url = "https://integrate.api.nvidia.com/v1",
  api_key = "Your NVIDIA NIM API Key"
)
```

[3] ✓ 0.0s Python

```
completion = client.chat.completions.create(
  model="microsoft/phi-3-mini-128k-instruct",
  messages=[{"role":"user","content":"Write about Microsoft and Nvidia partnership"}],
  temperature=0.4,
  top_p=0.7,
  max_tokens=2048,
  stream=True
)
```

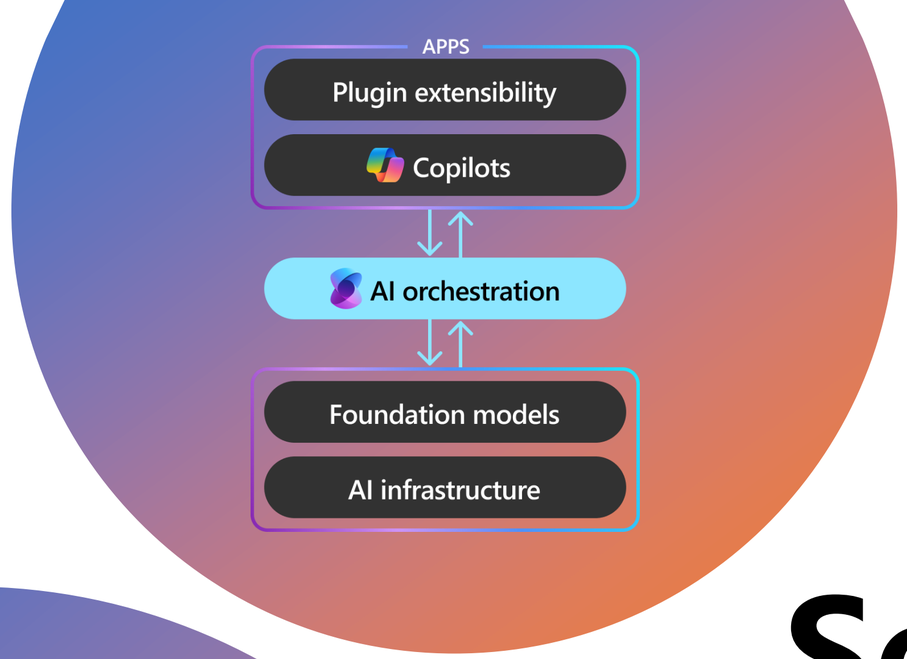
[4] ✓ 0.7s Python

+ Code + Markdown

```
for chunk in completion:
  if chunk.choices[0].delta.content is not None:
    print(chunk.choices[0].delta.content, end="")
```

[5] ✓ 12.7s Python

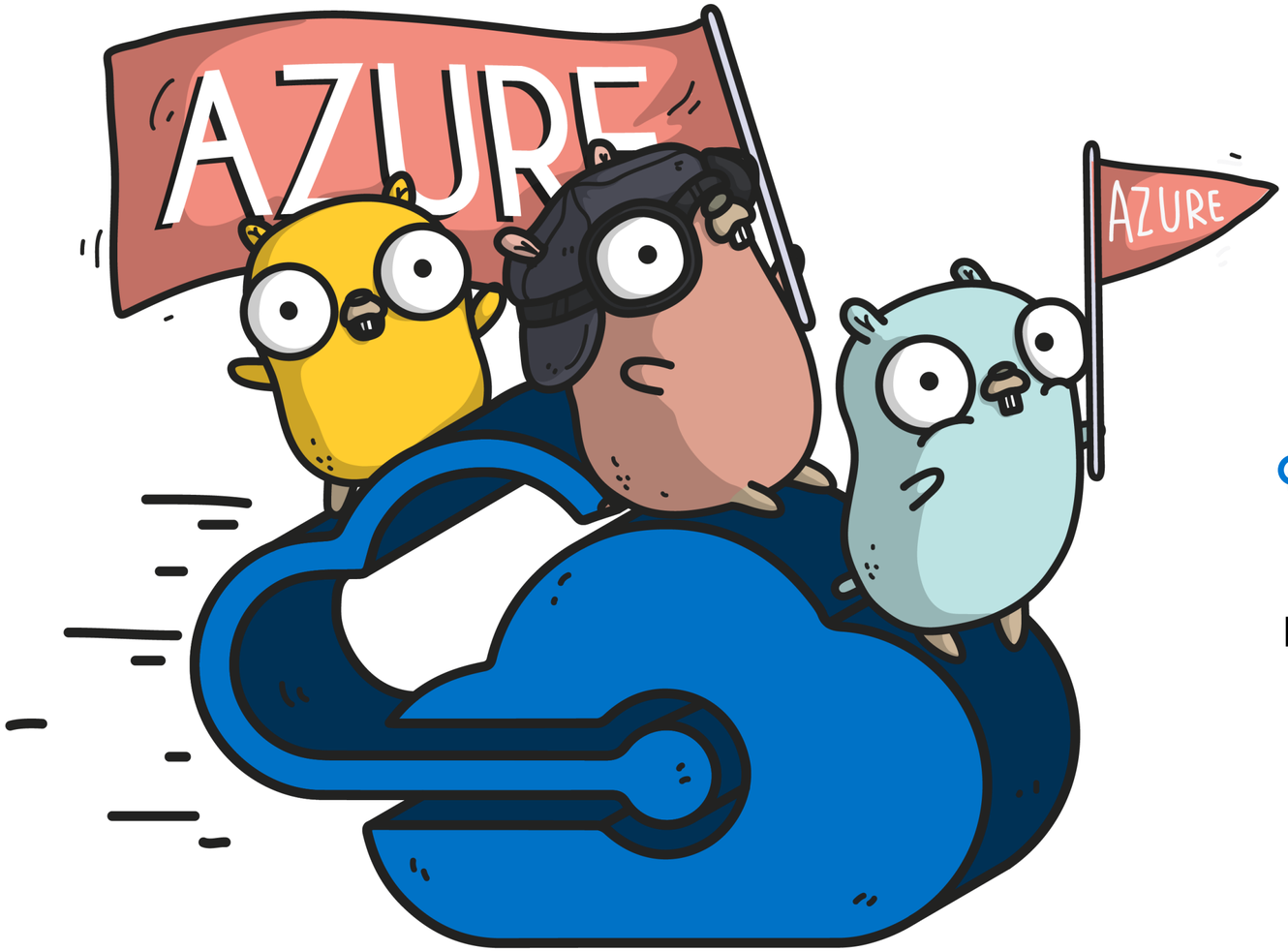
```
Microsoft and Nvidia have formed a strategic partnership to advance the development of artificial intelligence (AI) and deep learning technologies. The collaboration aims to leverage Microsoft's expertise i

The partnership will focus on several key areas, including the development of AI-powered applications, cloud-based AI services, and the creation of new AI-driven hardware solutions. Microsoft and Nvidia wil

One of the primary goals of the partnership is to accelerate the development of AI-powered applications that can help solve some of the world's most pressing challenges. For example, the collaboration will

In addition to developing AI-powered applications, Microsoft and Nvidia will also work together to create new AI-driven hardware solutions. This includes the development of next-generation GPUs that can del

To support the development of AI-powered applications and hardware solutions, Microsoft and Nvidia will leverage their respective strengths in software and cloud computing. Microsoft's Azure cloud platform

Overall, the Microsoft and Nvidia partnership represents a significant step forward in the development of AI and deep learning technologies. By combining Microsoft's expertise in software and cloud computin
```



Semantic Kernel

这个高度可扩展的开源框架使您能够在现有的 C#、Python 和 Java 代码之上利用最新的 AI 模型。您将能够构建可自动化业务流程的自定义 AI 代理。

- https://aka.ms/SemanticKernelCookBook



Phi-3-mini-V
(3.8B + 0.3B)

Phi-3-small
(7B)

Phi-3-mini
(3.8B)

Begin to talk about **Copilot Solution**

NVIDIA NIM API – Semantic Kernel

```
var nvidiaNIMChat = new CustomChatCompletionService();
nvidiaNIMChat.ModelUrl = "https://integrate.api.nvidia.com/v1/chat/completions";
nvidiaNIMChat.ModelName = "microsoft/phi-3-mini-128k-instruct";
nvidiaNIMChat.ApiKey = "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";
```

[8] ✓ 0.0s csharp - C# Script Code

```
using Microsoft.Extensions.DependencyInjection;
```

[9] ✓ 0.0s csharp - C# Script Code

```
var builder = Kernel.CreateBuilder();
builder.Services.AddKeyedSingleton<IChatCompletionService>("nvidiaNIMChat", nvidiaNIMChat);
var kernel = builder.Build();
```

[10] ✓ 0.0s csharp - C# Script Code

```
var chat = kernel.GetRequiredService<IChatCompletionService>();
var history = new ChatHistory();
history.AddUserMessage("hi, who are you?");
```

[11] ✓ 0.0s csharp - C# Script Code

```
var result = await chat.GetChatMessageContentsAsync(history);
```

[12] ✓ 2.4s csharp - C# Script Code

```
result[^1].Content
```

[13] ✓ 0.0s csharp - C# Script Code

```
Hello! I'm Phi, an AI developed by Microsoft. I'm here to help you with questions, provide information, and assist with a wide range of tasks. How can I help you today?
```


NVIDIA
THANK YOU