



# Inside Microsoft AI innovation

**Mark Russinovich**

CTO and Technical Fellow, Microsoft Azure

@markrussinovich

March 19, 2024

# Microsoft AI innovations

Infrastructure

AI applications

My AI research

# Infrastructure



**AI infrastructure**

AI accelerators

Offloads

## TOP10 System - November 2023

$R_{\max}$  and  $R_{\text{peak}}$  values are in PFlop/s. For more details about other fields, check the TOP500 description.

$R_{\text{peak}}$  values are calculated using the advertised clock rate of the CPU. For the efficiency of the systems you should take into account the Turbo CPU clock rate where it applies.

Rank	System	Cores	$R_{\max}$ (PFlop/s)	$R_{\text{peak}}$ (PFlop/s)	Power (kW)
1	<b>Frontier</b> - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	<b>Aurora</b> - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	4,742,808	585.34	1,059.33	24,687
3	<b>Eagle</b> - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Microsoft Azure United States	1,123,200	561.20	846.84	
4	<b>Supercomputer Fugaku</b> - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899

# Infrastructure

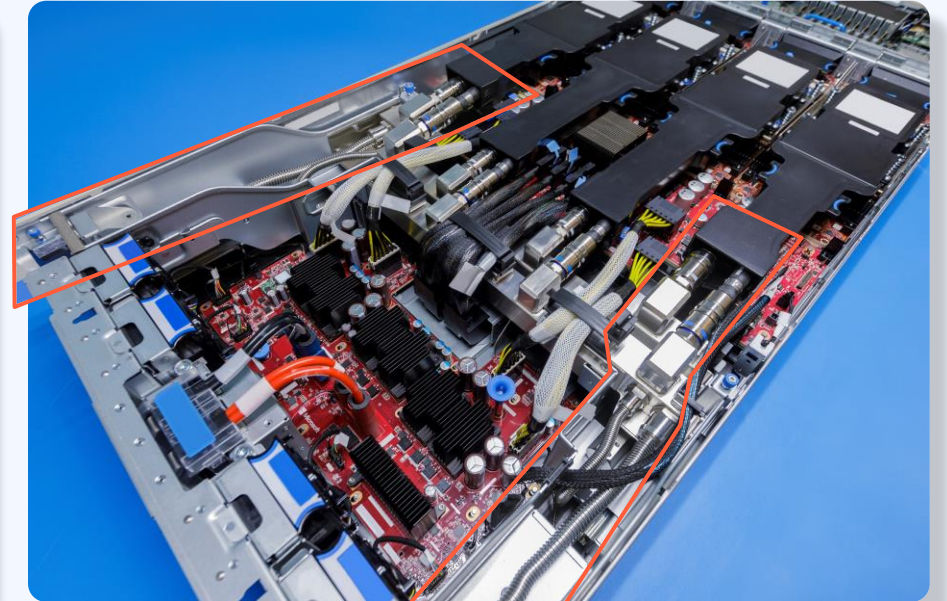


**AI infrastructure**

AI accelerators

Offloads

# Azure Maia



# Infrastructure



AI infrastructure

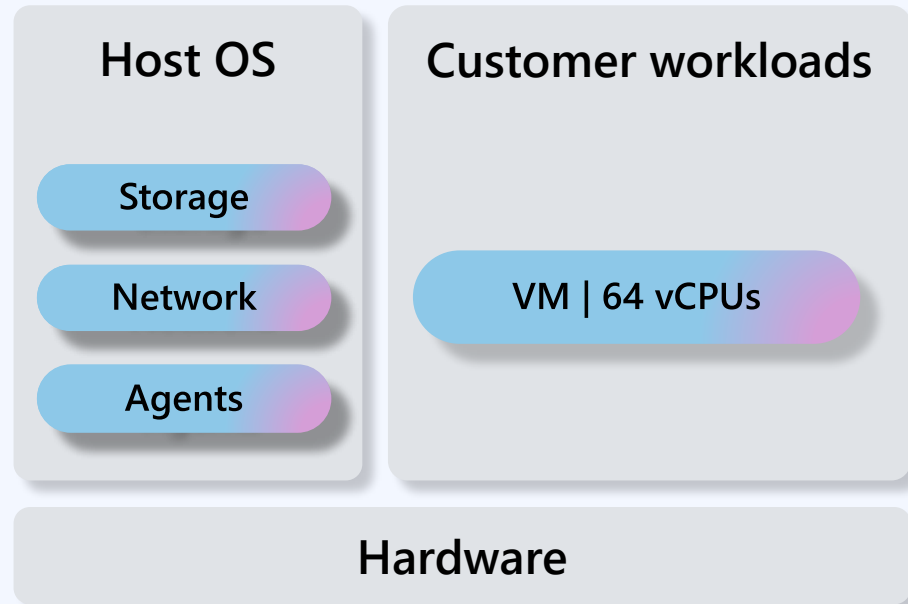
AI accelerators

Offloads

# Infrastructure acceleration through offload

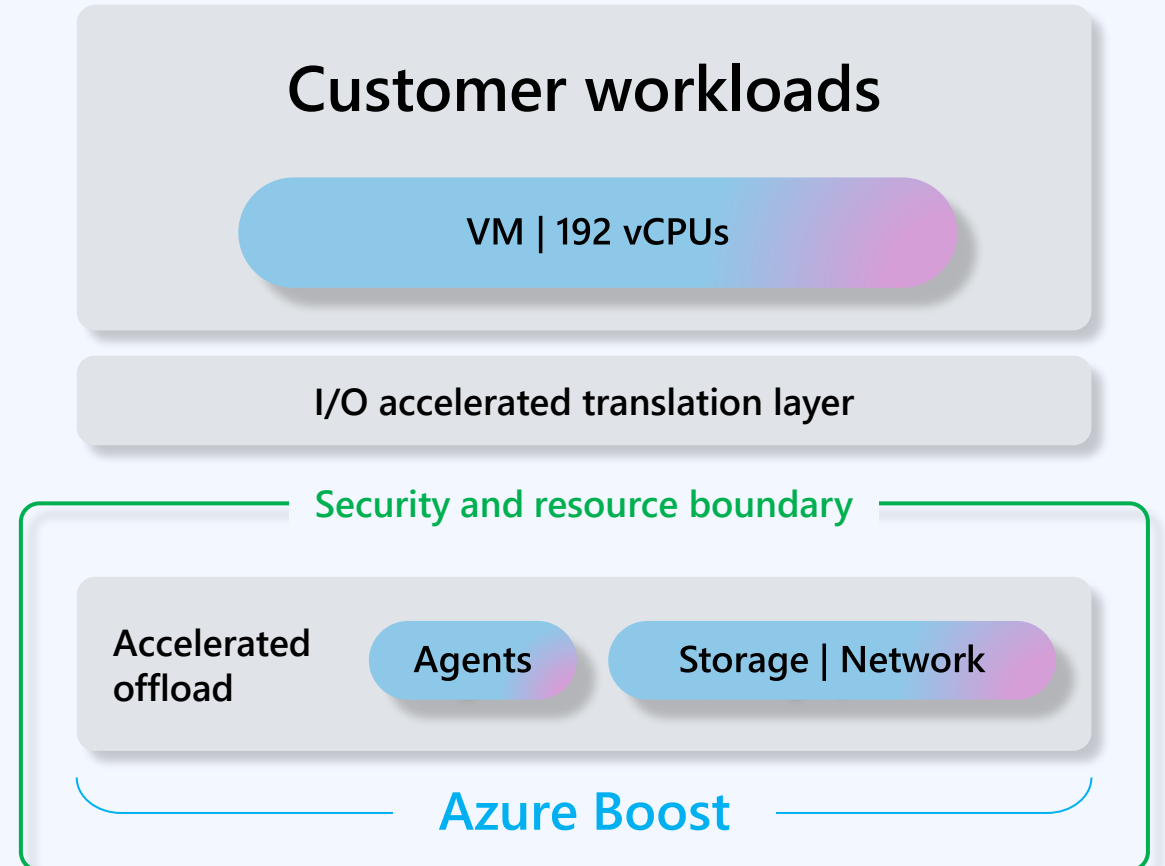
## Traditional infrastructure

Network and storage I/O processed in Host OS

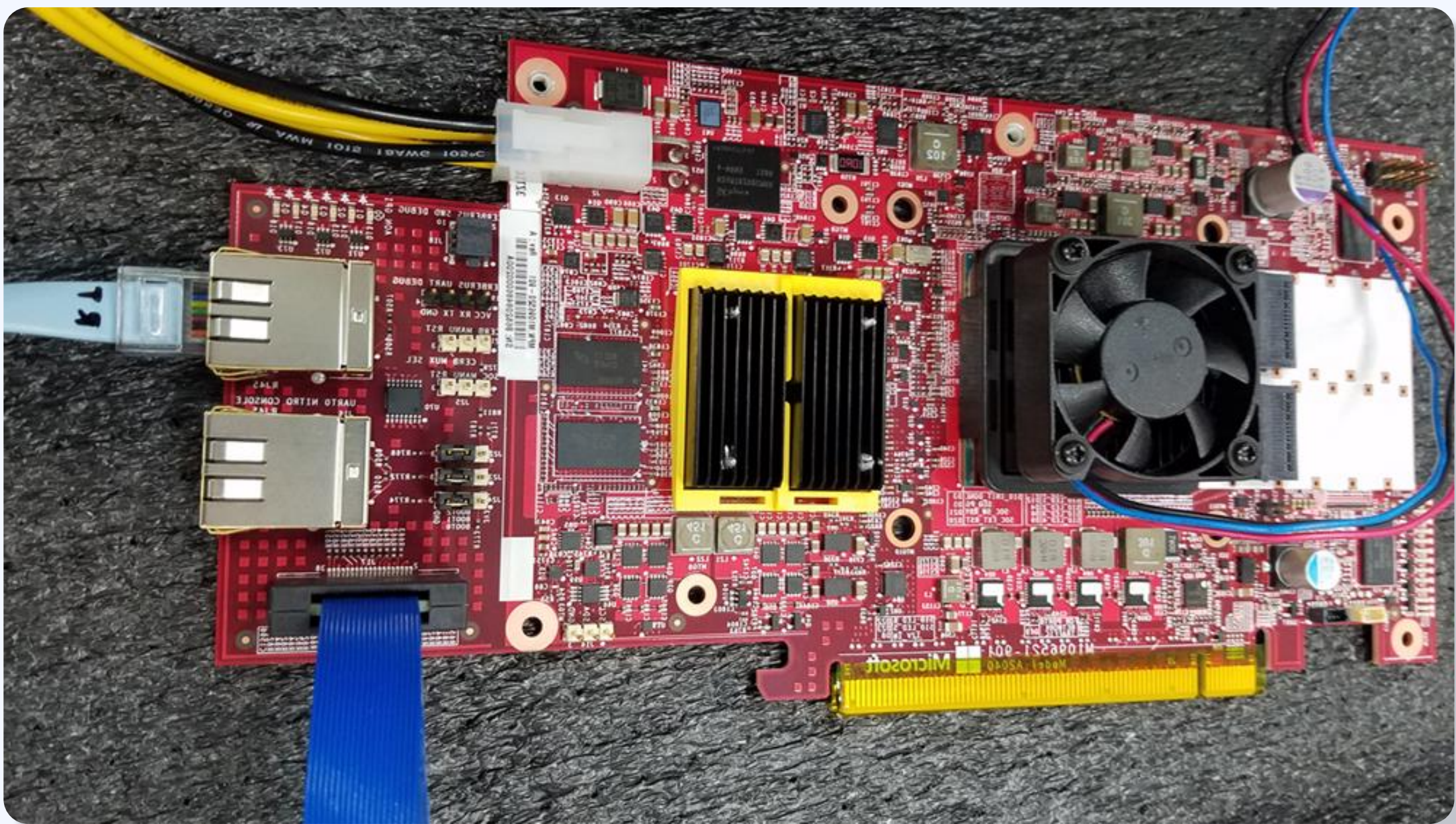


## Offloaded infrastructure

Server is managed through dedicated offload card







# AI applications

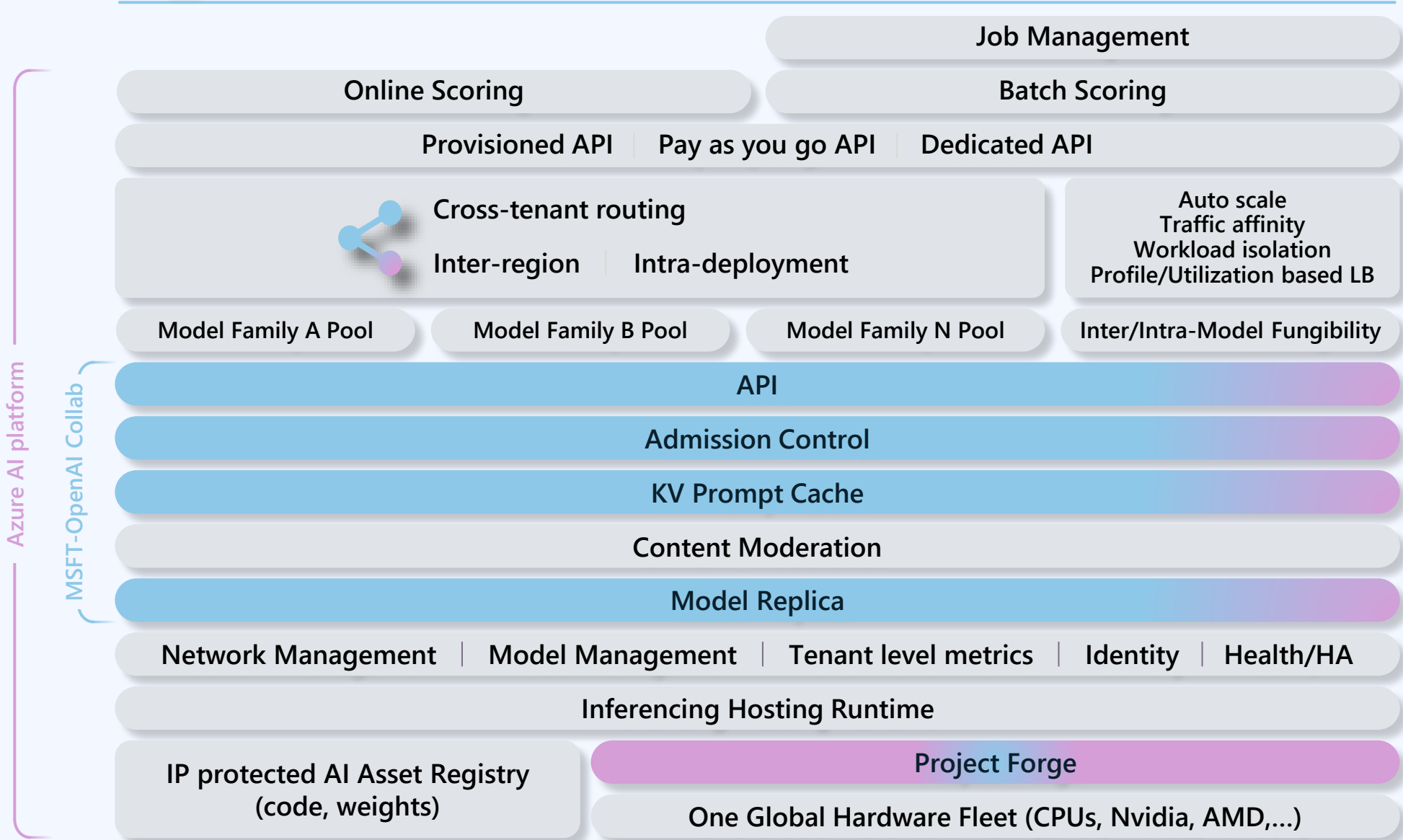
**AI workload management**

Frontier models and SLMs

AI security and safety

Confidential computing

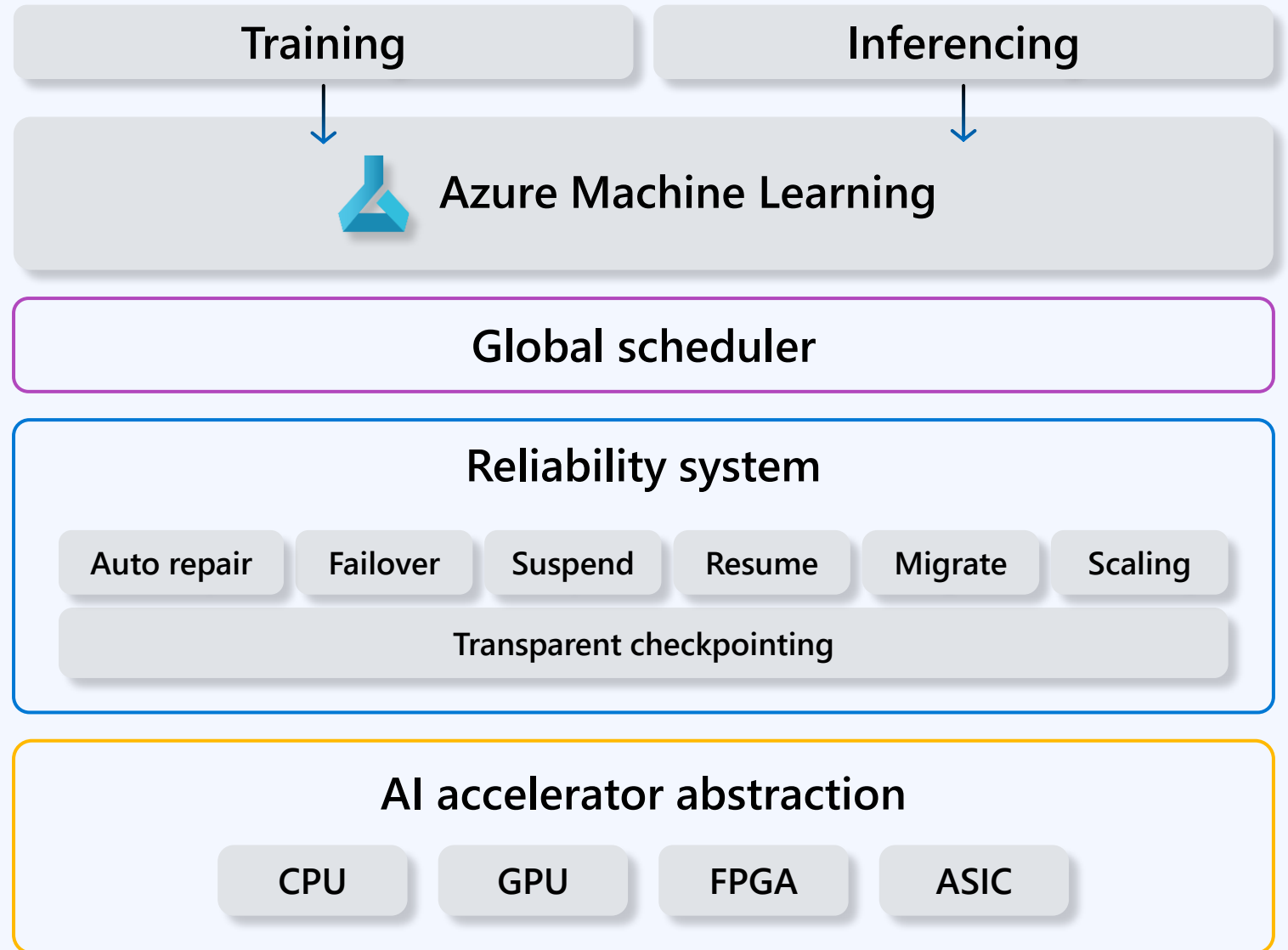
LLM serving  
platform  
architecture



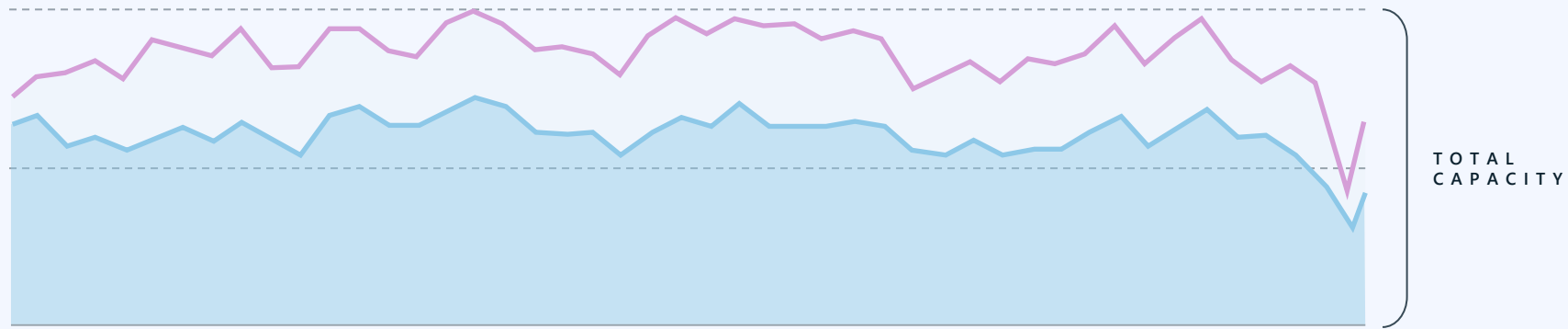
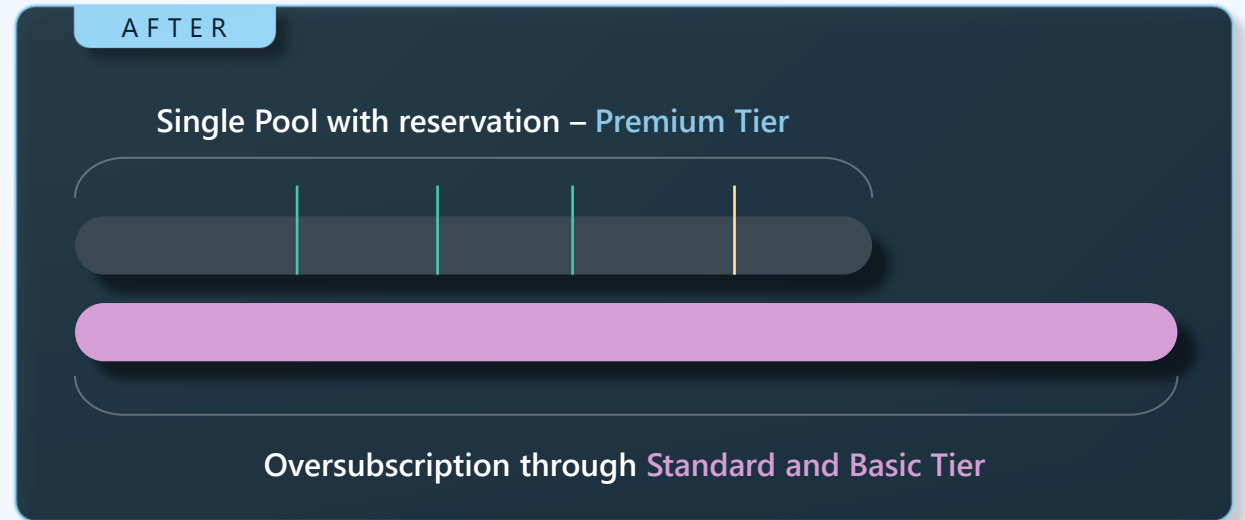
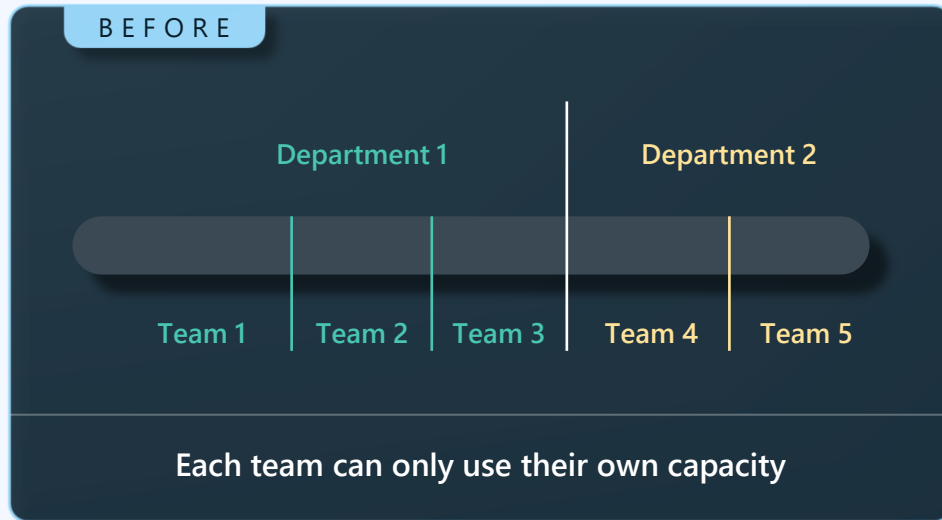


# Project Forge

- ✓ Azure-wide, serverless workload-aware global scheduling
- ✓ Highly reliable and efficient AI infrastructure
- ✓ Infrastructure and environment abstraction for workloads



# A Single Pool for all



- Total Usage  
(Premium + Standard + Basic)
- Reserved Capacity Usage  
(Premium)

Training utilization with/without Single Pool

# AI applications



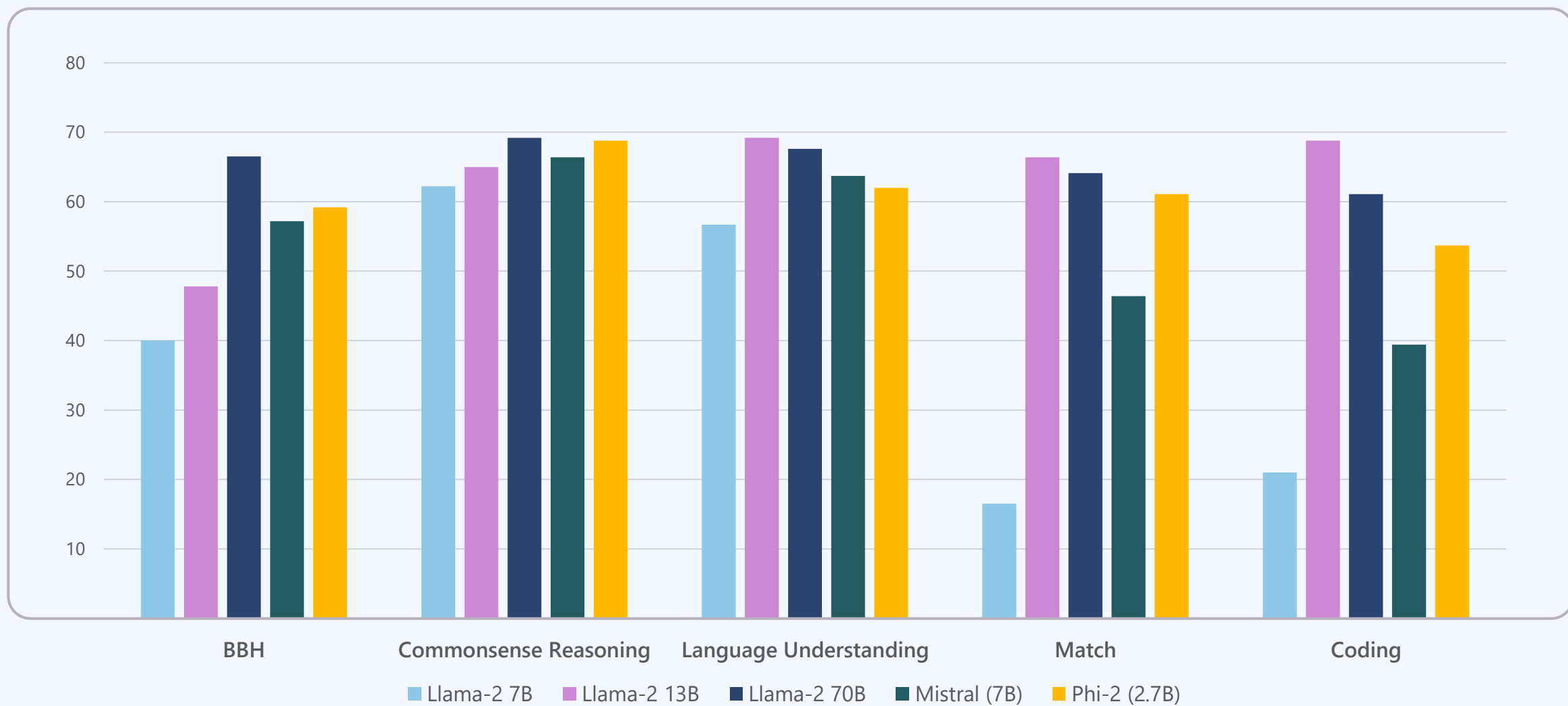
**AI workload management**

Frontier models and SLMs

AI security and safety

Confidential computing

# Phi-2 rivals models 5-10x its size



# AI applications

**AI workload management**

**Frontier models and SLMs**

AI security and safety

Confidential computing



# Microsoft's AI Principles



Fairness



Reliability  
& Safety



Privacy &  
Security



Inclusiveness



Transparency



Accountability

# AI applications

AI workload management

Frontier models and SLMs

AI security and safety

Confidential computing

# Confidential computing

## Existing encryption



### Data at rest

Encrypt inactive data when stored in blob storage, database, etc.



### Data in transit

Encrypt data that is flowing between untrusted public or private networks

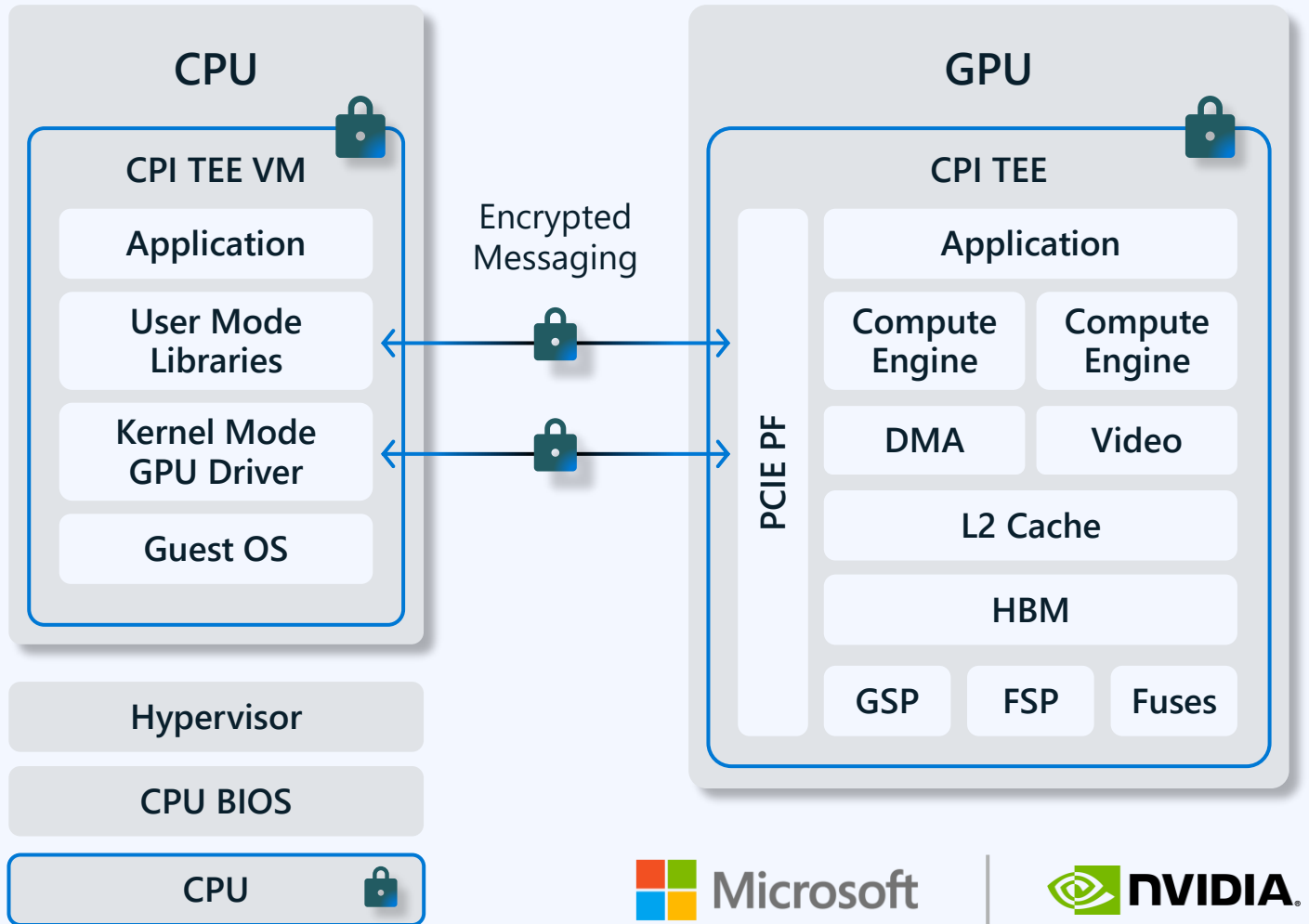
## Confidential computing



### Data in use

Protect/encrypt data that is in use, while in RAM, and during computation

# Azure Confidential GPU VMs powered by NVIDIA



# **My AI research**

File

Edit

Selection

View

Go

Run

...

←

→

unlearn\_new [SSH: 10.2.0.8]

demo.ipynb U

notebooks\_exploration > demo.ipynb > import torch

+ Code

+ Markdown

Run All

Restart

Clear All Outputs

Variables

Outline

...

Python 3.10.12

3

1

1

import torch

import torchvision.models as models

import torchvision.transforms as transforms

# load the model

model = models.resnet18(pretrained=True)

model.eval()

[ ]

Python

# write a function to get all internal representations for BatchNorm layers outputs of the model

# return a list of tensors

[ ]

Python

# visualize the outputs distribution of each layer as 5\*4 subplots

[ ]

Python

SSH: 10.2.0.8

main\*

0

0

0

Spaces: 4

Cell 1 of 3

# Who's Harry Potter? Approximate Unlearning in LLMs

Ronen Eldan\* and Mark Russinovich†

October 2, 2023



Model



Dataset



Jobs



Settings



## Welcome!



Style

Same Prompt ▾



Layout

1 x 2 ▾



Model ☒



Temperature ☐

0



Maximum Length

3000



Stop Seq (JSON)

["<|endoftext|>",



Top P

1



Show Probs ☐



+ Add images



Retry ☒

▶ Start

▶ Start

llama-2-7b-chat ▾

📁 Save

▶ Start

Llama2-minus-Ha ▾

📁 Save

Who is Harry Potter?

Who is Harry Potter?