Azure OpenAI

TheTechPlatform
Where Tech Meets.

Azure OpenAI Series (Virtual)

Saturday, 23 Mar, 2024

Unleashing the Power of Artificial Intelligence in the Cloud– Part 3

Aroh Shukla

Regional Microsoft Cloud Architect

Microsoft MVP Alumni, MCT

TheTechPlatform
Where Tech Meets.

Aroh Shukla

MVP Alumni, MCT

Global Speaker

- Passionate to **learn**.
- Passionate to **share knowledge**.
- Passionate to work on **Microsoft Technologies**

Aroh.Shukla@gmail.com

/arohshukla

@aaroh_bits

TheTechPlatform
Where Tech Meets.

AGENDA SLIDE

Part 1





01

Recap

What we have covered in last 2 sessions

02

Fine tuning

Adjusting pre-trained model for task.

03

Securing Azure OpenAI Service

Different ways to secure OpenAI workloads

04

Azure AI Content Safety

Different methods for content safety

05

Learning Resources

Different methods for content safety

TheTechPlatform
Where Tech Meets.

Azure Open AI

Sessions Roadmap

Part 01

Exploring Azure OpenAI: a rewarding journey into integrating Azure services with OpenAI tech. Learn about ChatGPT, DALL-E2, and more to establish a strong AI foundation and unleash creativity.

Part 2

Gain basic Azure OpenAI insight: access, models, use cases, responsible AI, pricing, deployment, and text model utilization.

Part 3

Part 3 enhances Part 2's learning: refine models, ensure security, employ content safety, demo applications. Relevant for those eager to leverage Azure OpenAI for personal AI solutions..

TheTechPlatform
Where Tech Meets.

Session 1 (03 Feb 2024)

1. What is ChatGPT
2. Azure OpenAI vs OpenAI
3. AI Terminologies
4. History of ChatGPT
5. Compare ChatGPT plans
6. DEMOS

- Slide Deck -
- Video Link: YouTube – Part 1

TheTechPlatform
Where Tech Meets.

Session 2 (16 Mar 2024)

1. How to access the Azure OpenAI
2. Different Models with Pros & Cons
3. Use cases
4. Pricings & Quotas
5. DEMOS

- Slide Deck -
- Video Link: YouTube – Part 2

TheTechPlatform
Where Tech Meets.

Session 3 (23 Mar 2024)

1. Recap
2. Fine tuning
3. Securing Azure OpenAI Service
4. Azure AI Content Safety
5. DEMOS

- Slide Deck – UPDATE LATER
- Video Link: UPDATED LATER

TheTechPlatform
Where Tech Meets.

Fine Tuning (Why?)

- Built in models were trained on **publicly available data**
- With Azure OpenAI service you can train the model **on your own data**
- Allows creating chat bots that **work on the organizational data**
- A huge **added value**

To fine tune OR NOT to fine tune

- **When you need a LLM tailor to a specific task or domain**

Consider fine tuning **IF**

- You want to teach the model a **new skill**
- You want to teach the model how to do **something with examples**
- You want to **reduce latency**

TheTechPlatform
Where Tech Meets.

Fine Tunning - now available with Azure OpenAI Service

Fine tuning: You'll now be able to use **Azure OpenAI Service, or Azure Machine Learning**, to fine tune Babbage/Davinci-002 and GPT-3.5-Turbo.

Babbage-002 & Davinci-002:

- GPT3 based: smaller, lower latency
- Understand & generate natural language or code
- Completion support

GPT-35-Turbo:

- Most capable & cost effective GPT-3.5 model
- More sophisticated capabilities
- Chat support

Fine Tunning - now available with Azure OpenAI Service

Model	Hourly Hosting	Input tokens	Output Tokens
Babbage-002	$1.70	$0.0004 / 1k	$0.0004 / 1k
Davinci-002	$3.00	$0.0020 / 1k	$0.0020 / 1k
GPT-35-Turbo	$7.00	$0.0015 / 1k	$0.0020 / 1k

Model ID	Fine-Tuning Regions	Max Request (tokens)	Training Data (up to)
`babbage-002`	North Central US, Sweden Central	16,384	Sep 2021
`davinci-002`	North Central US, Sweden Central	16,384	Sep 2021
`gpt-35-turbo` (0613)	North Central US, Sweden Central	4096	Sep 2021

Fine Tunning – Cost Analysis

- Fine tuning is done as **managed service**
- Paying for **active training time** for **successful fine-tuning runs**

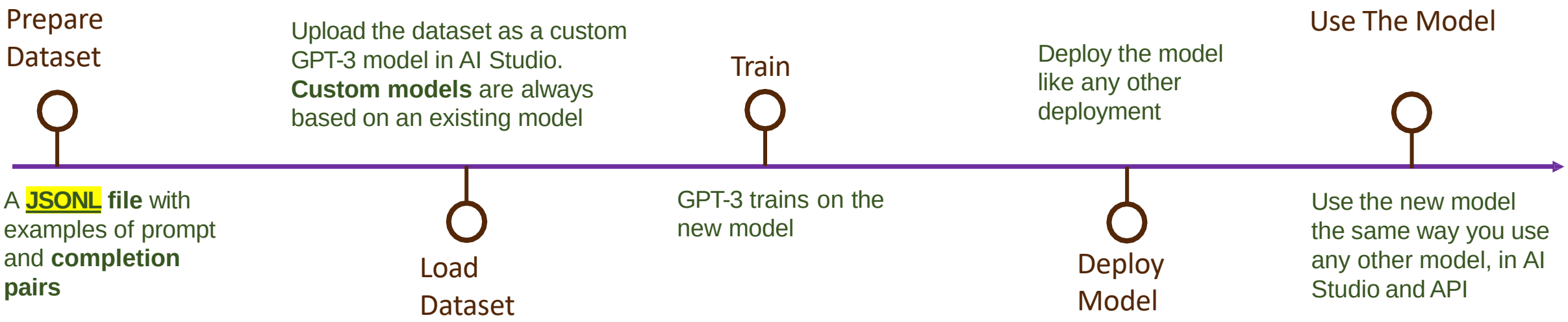
TRAINING

Model	Cost
Babbage-002	$34 / hour
Davinci-002	$65 / hour
GPT-35-Turbo	$102/ hour

- **HOSTING**

Model	Hourly Hosting	Inputs tokens	Output tokens
Babbage-002	$1.7	$0.0004 / 1k	$0.0004 / 1k
Davinci-002	$3.00	$0.0020 / 1k	$0.0020 / 1k
GPT-35-Turbo	$7.00	$0.0015 / 1k	$0.0020 / 1k

The Fine Tunning Process



TheTechPlatform
Where Tech Meets.

The Fine Tunning – Resources

1. Customize a model with fine-tuning (preview)
2. Advancing AI - Fine Tuning LLMs with Azure OpenAI
3. Azure OpenAI 101: An introduction to Building Custom AI Models #python #chatgpt #azure

NOTE: Fine Tuning is **costly exercise**

TheTechPlatform
Where Tech Meets.

DEMO 1: How to perform fine tuning the Azure OpenAI

TheTechPlatform
Where Tech Meets.

Securing Azure OpenAI Service

- Azure has **powerful security controls that helps securing** cloud

  resources

- These controls can be applied also to **OpenAI Service**

Securing Azure OpenAI Service

- OpenAI Service can be protected using:

Network Security

Microsoft Entra ID (formerly Azure AD)

Network Security

- Access to OpenAI service can be limited **to specific networks**
- Can be also denied from **public network** and allowed only from **VNets using private IP**
- This is done using **Private Endpoint**
- The **most secure network access**

Microsoft Entra ID (Azure AD Identity)

- So far we **used API key** to access OpenAI
- Not **very secure**
- Anyone who has the key can access the resource
- It’s more secure to **use Azure AD identity**
- **Only authorized users** and services can access the resource
- No need **to use api-key**

TheTechPlatform
Where Tech Meets.

DEMO 2: How to secure the Azure OpenAI Service

TheTechPlatform
Where Tech Meets.

Azure AI Content Safety

- Helps create **secure online experience**
- Detects **hateful, violent, sexual, self-harm content**
- Works using **REST API**
- Can be integrated into any user-generated web app
- Works in real time

What is Azure AI Content Safety

- A lot of user-generated content on the web

Social Networks	Education
User Rating	Chats
Online Marketplace	Forums
Gaming	More…

What is Azure AI Content Safety

- User-generated content requires moderation
- Looks for harmful content

Hate

Sexual

Violence

Self harm

What is Azure AI Content Safety

- Mostly manual
- Slow
- Costly
- Error prone

What is Azure AI Content Safety

- Azure AI Content Safety automates **text and image moderation**
- Analyzes the content and **returns score in the four categories**
- Real time analysis using **REST API**
- **Azure AI Content Safety Studio** for testing content and monitoring activity
- Can be used with **any web app**

Azure AI Content Safety Cost

Instance	Features	Price
Free – Web	Text	5,000 text records per month[1]
	Image	5,000 images per month
Standard – Web	Text Jailbreak risk detection Protected material detection	**$0.75** per 1,000 text records[1]
	Image	**$1.50** per 1,000 images

1A text record in the **S tier contains up to 1,000 characters** as measured by Unicode code points. If a text input into the Content Safety API is more than 1,000 characters, it counts as one text record for each unit of 1,000 characters. For instance, if a text input sent to the API contains 7,500 characters, it will count as 8 text records. If a text input sent to the API contains 500 characters, it will count as 1 text record**.**

Azure AI Content Safety pricing

TheTechPlatform
Where Tech Meets.

DEMO 3: How to use Azure AI Content Safety

TheTechPlatform
Where Tech Meets.