Microsoft

AI at Scale

Introduction to Foundation models in Azure

Keita Onabuta

Senior Customer Engineer for AI & Machine Learning
Azure CXP – FastTrack for Azure
Microsoft Corporation

Logistics

- Duration about 45min
- Question – please post questions in the chat
- This slide deck will be shared after the session
- No recording
  - If you need our support to build and deploy Azure solution, please let us know.

Agenda

1. Introduction to Foundation models
2. Foundation models in Azure
   - Demo #1 : Prompt Flow
   - Demo #2 : Training large scale model
3. Enterprise Search
   - Demo #3 : Enterprise Search with Azure AI
4. Recap & Call to action

1. Introduction to Foundation models

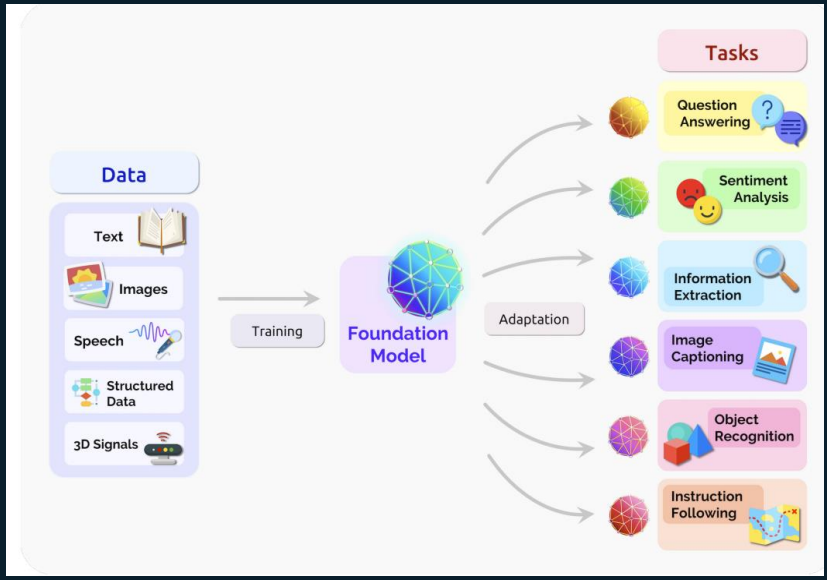
Foundation models

What are Foundation models?

- In recent years, a new successful paradigm for building AI systems has emerged: Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model.

Why do we care?

- Foundation models have demonstrated impressive behavior, but can fail unexpectedly, harbor biases, and are poorly understood. Nonetheless, they are being deployed at scale.



arxiv.org/pdf/2108.07258.pdf

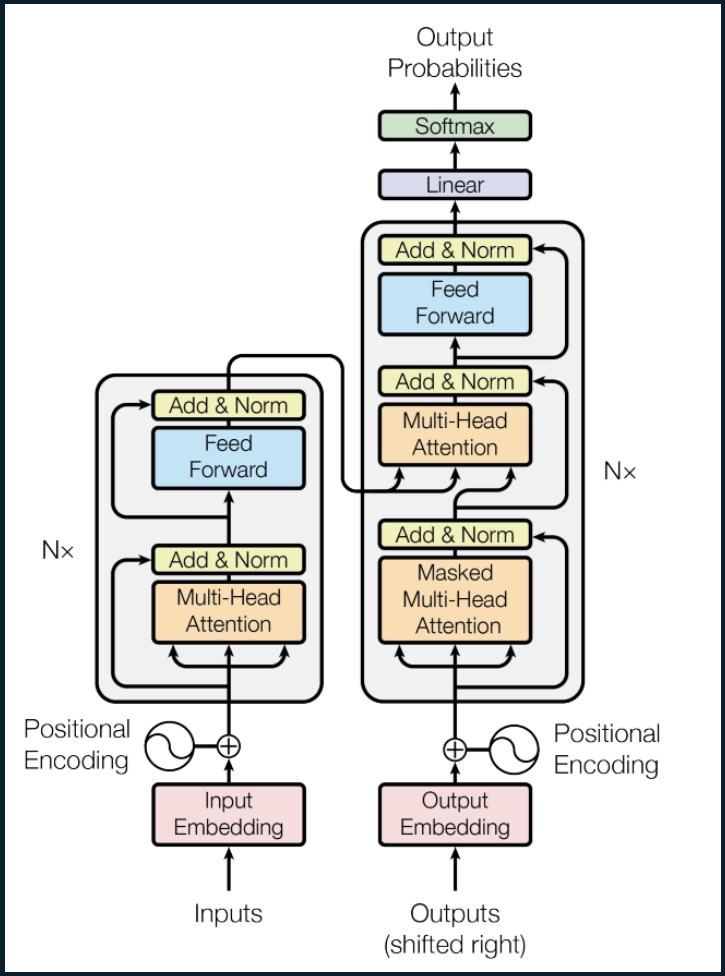
Stanford CRFM

Critical components for Foundation models

- Transformers
- Scale
- In-context learning

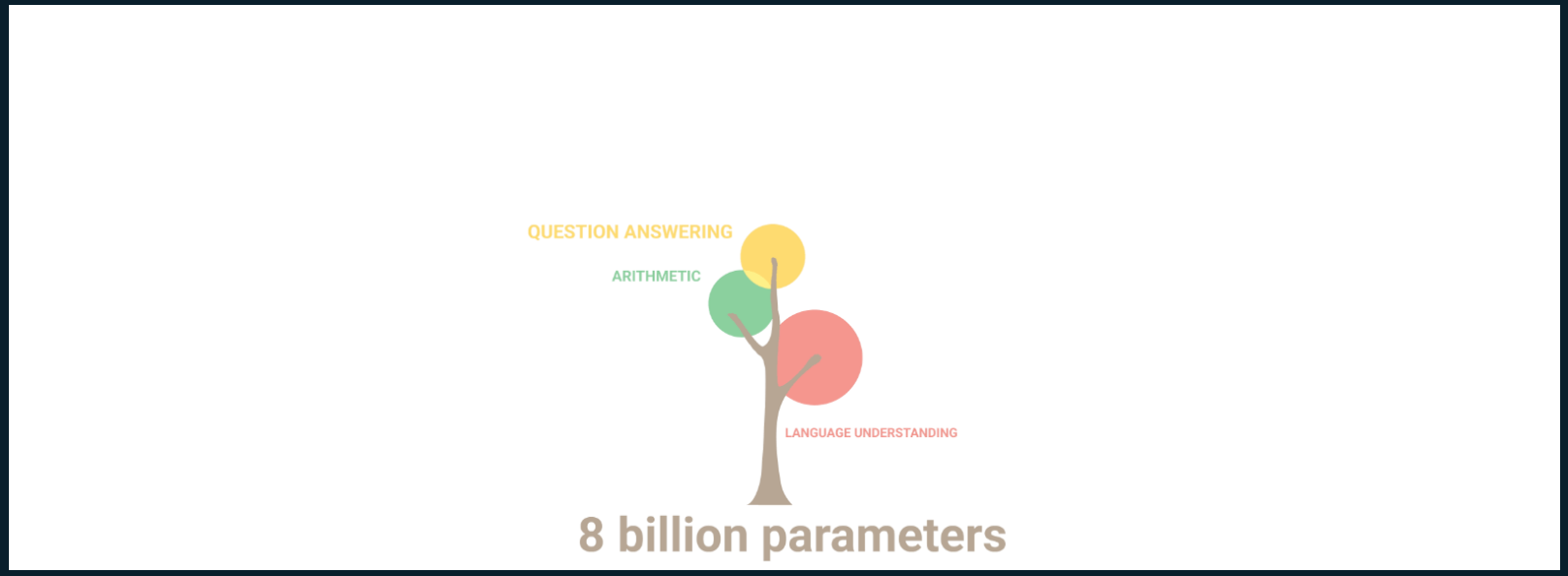
Transformers

- Paper : Attention is All You Need
- Easy to scale and parallelize
  - fast training
  - more data for training
- Dominating the field of NLP and moving over beyond NLP as well



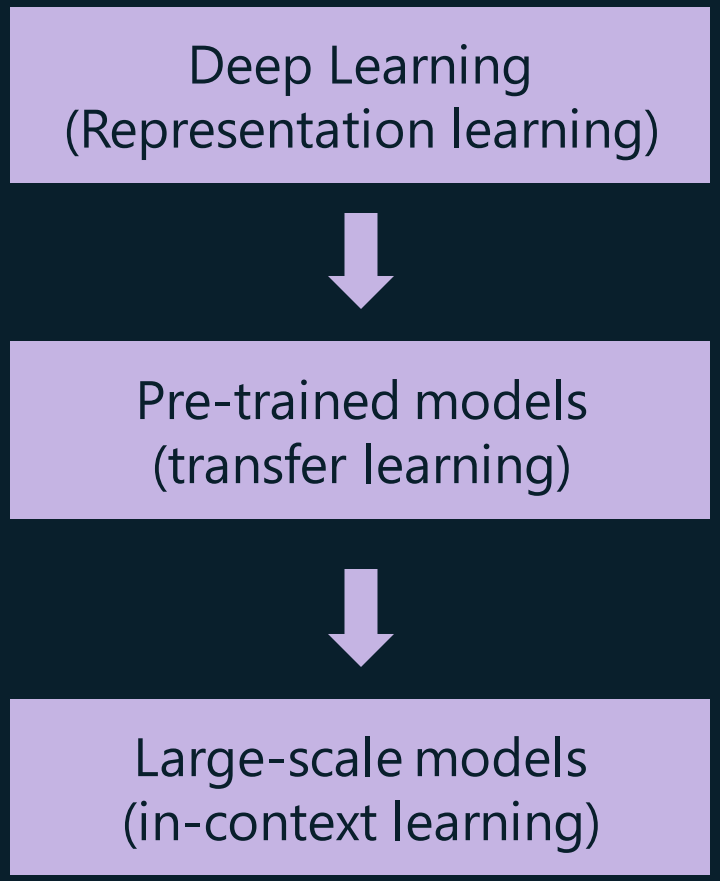
Scale

- Scale leads to emerging capabilities
- Many capabilities emerge unpredictably only whe models reach a critical size.



Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance – Google AI Blog (googleblog.com)

In-context learning



- Fully supervised
- architecture design

- Pre-train and fine-tune
- no architecture design

- Pre-train and prompt
- Zero/Few-shot in-context learning

Why in-context learning?

- Models are applied to new tasks out of the box.
- Amazing performance with no or few examples.
- Tasks are adapted to models instead of models adapting to tasks.
- Humans can interact with the models in natural language.
- Blurring the line between ML users and developers.

2. Foundation models in Azure

Microsoft AI portfolio

Business Users
& Citizen Developers

Developers &
Data Scientists

Applications

Microsoft 365 | Microsoft Dynamics 365 | Microsoft Edge | Microsoft Bing | Windows | XBOX

Power Platform

Power BI

Power Apps

Power Automate

Power Virtual Agents

Azure AI

Applied AI Services

Bot Service

Cognitive Search

Form Recognizer

Video Indexer

Metrics Advisor

Immersive Reader

Cognitive Services

Vision

Speech

Language

Decision

Azure OpenAI Service

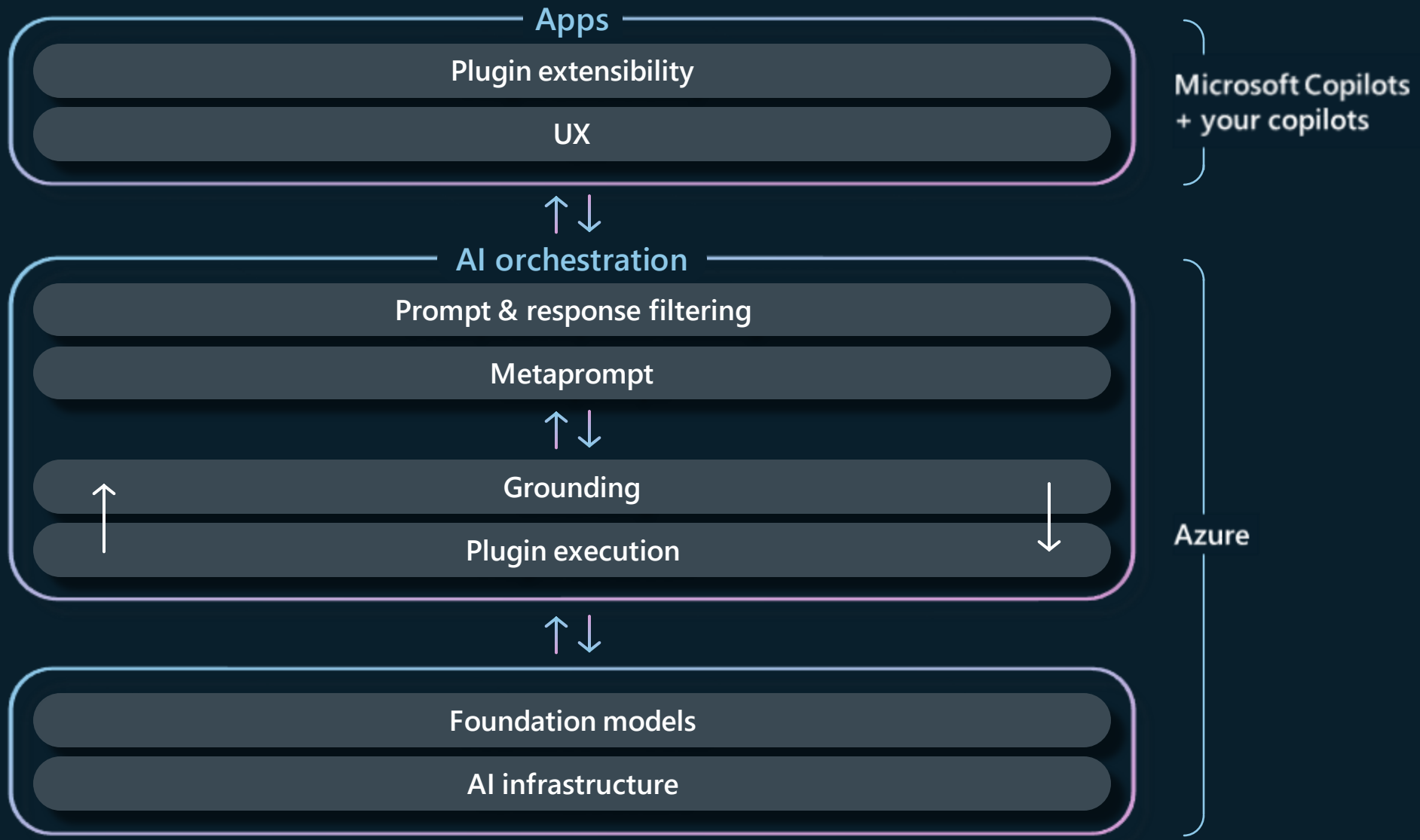
ML Platform

Azure Machine Learning

Microsoft runs on Azure AI

- Microsoft 365 Copilot
- Microsoft Security Copilot
- Dynamics 365 Copilot
- Power Platform Copilot
- Bing chat
- GitHub Copilot
- Nuance
- LinkedIn
- And more!

Copilot stack



Microsoft outlines framework for building AI apps and copilots; expands AI plugin ecosystem - Source

Foundation models in Azure AI

Hosted Foundation models

- Azure OpenAI Service
  - Fine tuning supported for some modelss.
- Azure Machine Learning – Model Catalog
  - Open source models and Huggin Face models

BYO Foundation models

- Azure Machine Learning
  - Azure Container for PyTorch

Azure OpenAI Service

Large pretrained foundation AI models
custom-tunable with your parameters and your data

Summarization
Reasoning over data

Writing tools
Code generation

ChatGPT
The Era of Copilots

GPT-3 (GA)	DALL•E 2 (preview)	ChatGPT (GA)	GPT-4 (GA)

Foundation of enterprise security, privacy and compliance

Models

GPT-3

Davinci

- summarizing for specific audience
- Ganerating creative content

Curie

- Answering questions
- Complex, nuanced classification

Babbage

- Semantic search ranking
- Moderately complex classification

Ada

- Simple classification
- Parsing and formatting text

GPT-3.5

GPT 3.5 Turbo

- Primary Chat
- Summarizing
- Code generation

GPT-4

GPT 4 - 32k

- Evaluation of GPT 3.5

GPT 4 - 8k

- Evolution of GPT 3.5

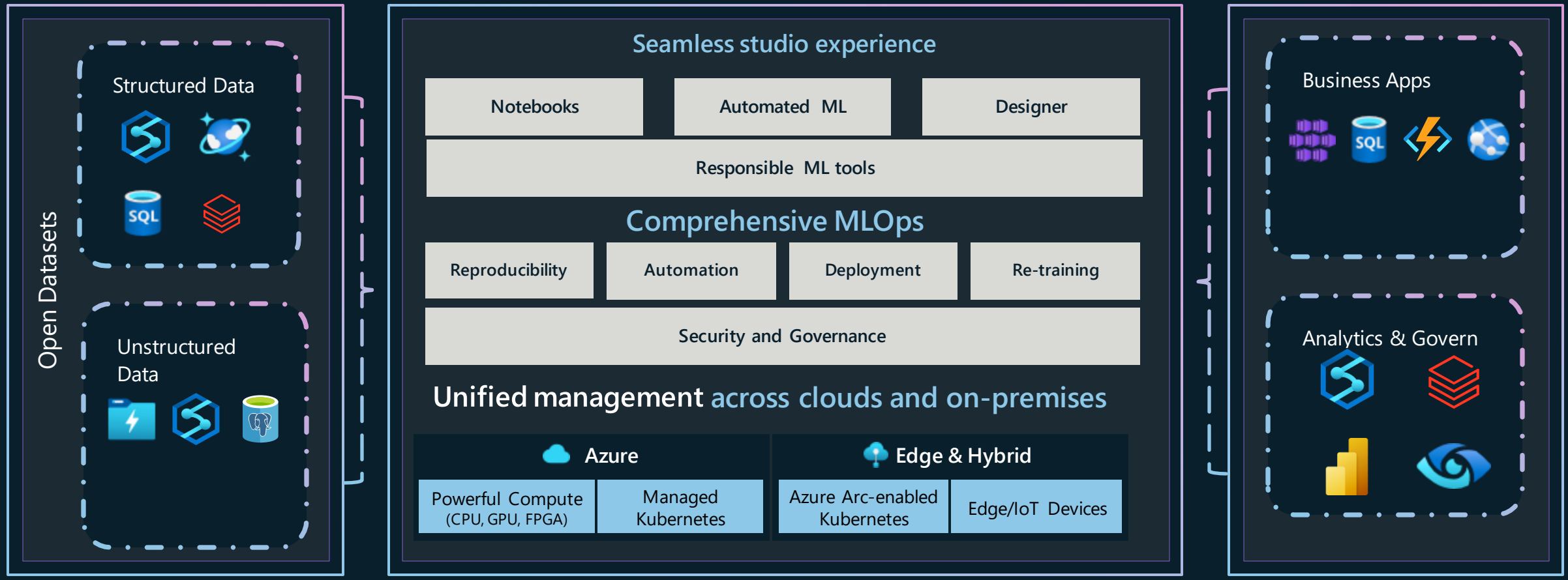
Azure OpenAI Service models - Azure OpenAI | Microsoft Learn

Recent updates (Microsoft Build 2023)

- Azure OpenAI Service on your data (Public Preview)
- Plugins for Azure OpenAI Service (Coming soon)
- Configurable Content Filters
- Provisioned Throughtput (Limited Availablity in June)

Azure Machine Learning

AI Platform for data scientists, machine learning engineers and prompt engineers!



Features for Foundation models

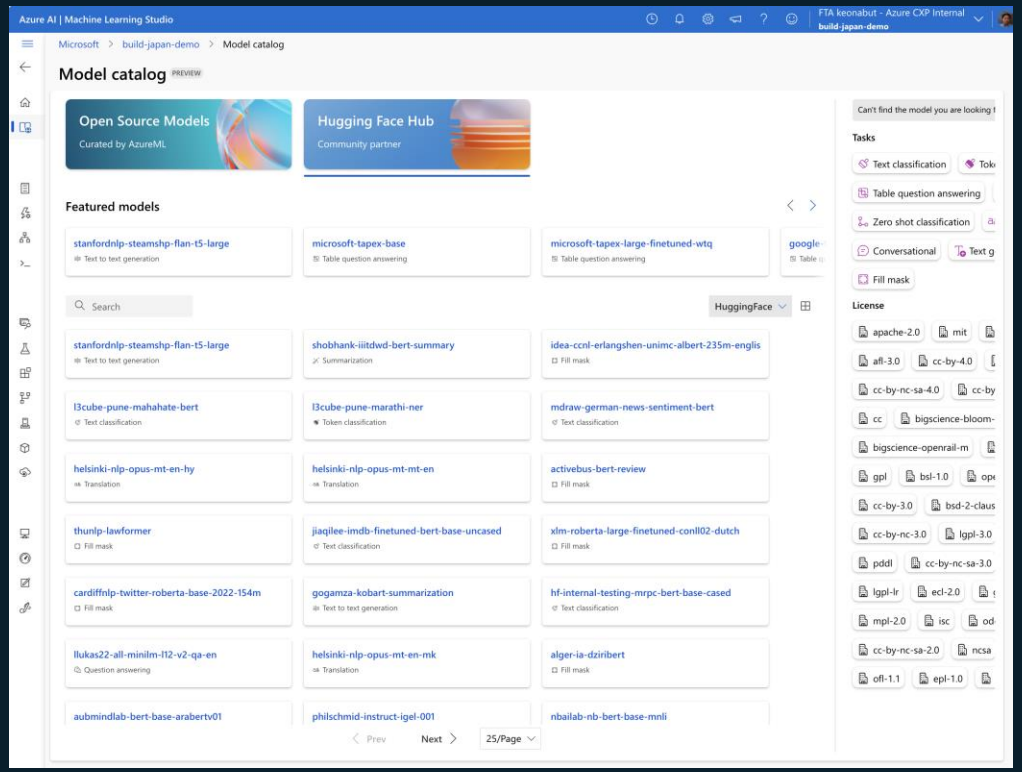
- Compute Cluster
- Model Catalog
  - Open source models and Hugging face models
- Prompt Flow
- Scalable deploymenrt – Managed Online Endpoint & Batch Endpoint
- Azure Container for PyTorch
- Model monitoring

Model Catalog (Public Preview)

Hub for Foundation models in AzureML

Open Source Models

- The most popular open source third-party models curated by Azure Machine Learning. Evaluate, fine tune and deploy models for out of the box usage and are optimized for use in Azure Machine Learning.



Hugging Face hub

- Thousands of models from HuggingFace hub for real time inference with online endpoints.

Azure OpenAI Service

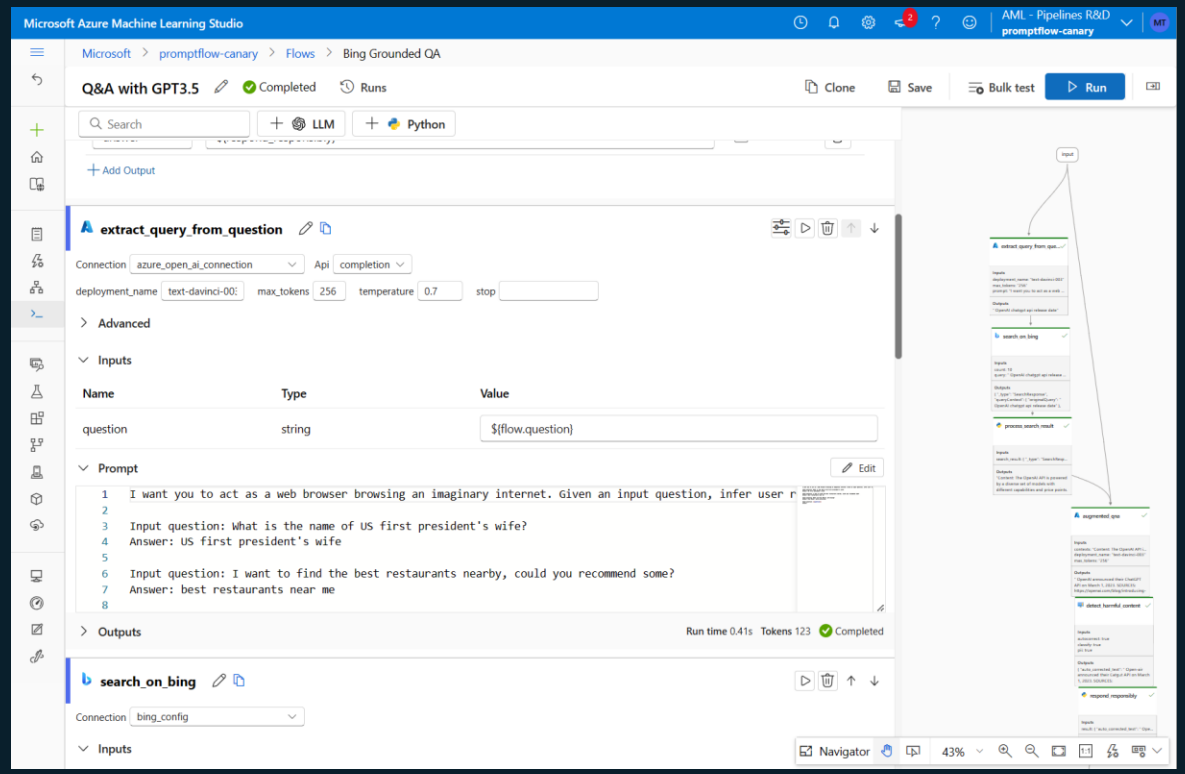
- coming soon

[Foundation Models in Azure Machine Learning - Azure Machine Learning | Microsoft Learn]

Prompt flow (Private Preview)

devops for prompt engineering

- Create AI workflows that consume various language models and data sources using the frameworks and APIs of your choice
- One platform to quickly iterate through build, tune, & evaluate for your GenAI workflow
- Evaluate the quality of AI workflows with pre-built and custom metrics
- Easy historical tracking and team collaboration
- Easy deployment and monitoring

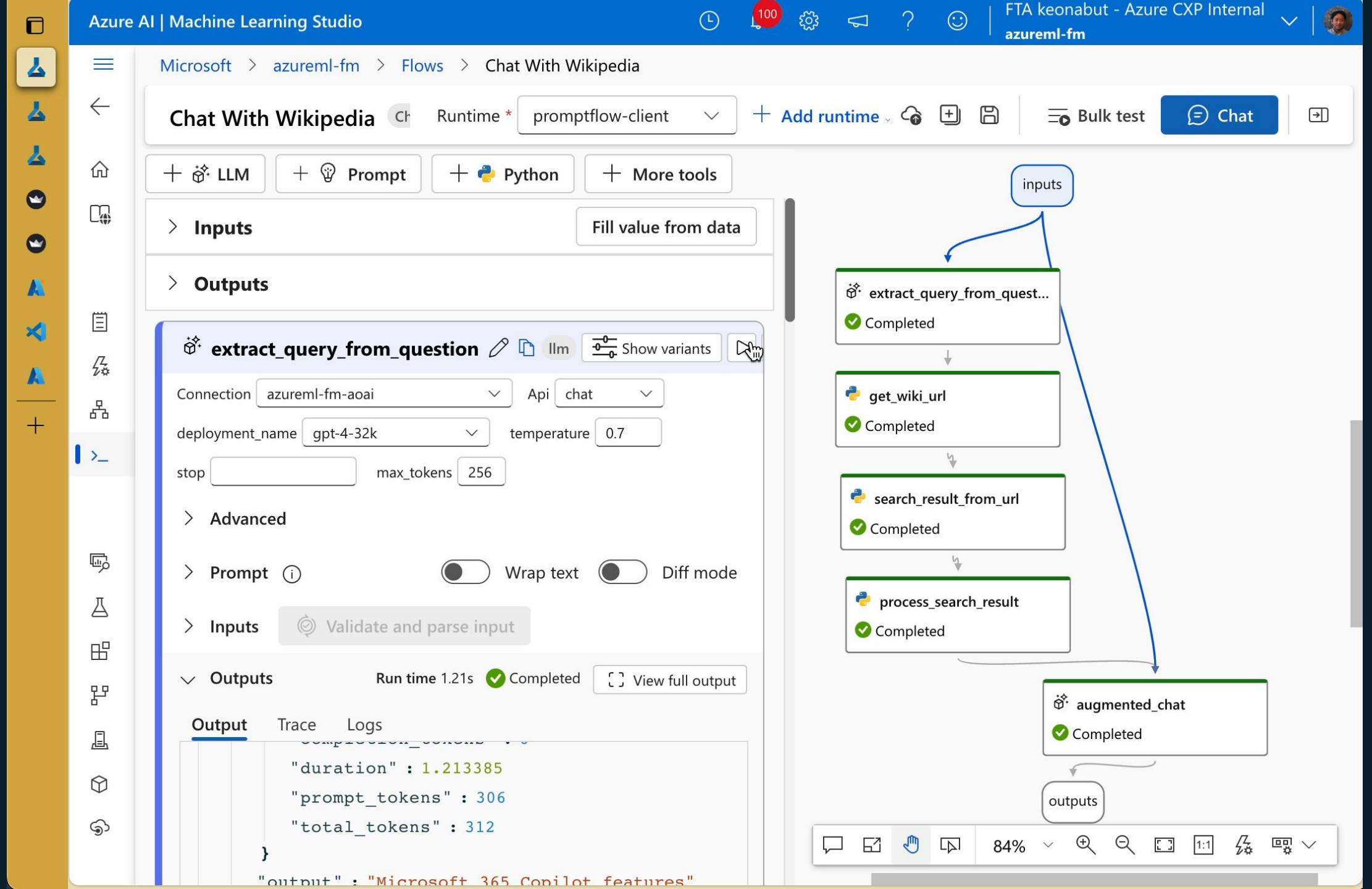


[Harness the power of Large Language Models with Azure Machine Learning prompt flow - Microsoft Community Hub]

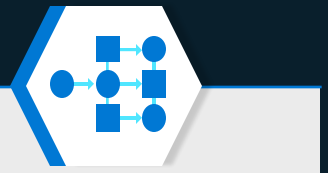
Demo #1 : Prompt flow

- flow - “Chat with Wikipedia”
- Chat with this flow
- Deploy to Managed Online Endpoint
- Call deployed API from Streamlit application

https://github.com/kyoro1/BuildJapan_2023/


Azure AI | Machine Learning Studio
FTA keonabut - Azure CXP Internal
azureml-fm
Microsoft > azureml-fm > Flows > Chat With Wikipedia
Chat With Wikipedia
Runtime * promptflow-client
Add runtime
Bulk test
Chat
LLM
Prompt
Python
More tools
Inputs
Fill value from data
Outputs
extract_query_from_question
llm
Show variants
Connection azureml-fm-aoai
Api chat
deployment_name gpt-4-32k
temperature 0.7
stop
max_tokens 256
Advanced
Prompt
Wrap text
Diff mode
Inputs
Validate and parse input
Outputs
Run time 1.21s
Completed
View full output
Output
Trace
Logs
"duration" : 1.213385
"prompt_tokens" : 306
"total_tokens" : 312
}
"output" : "Microsoft 365 Copilot features"
inputs
extract_query_from_quest...
Completed
get_wiki_url
Completed
search_result_from_url
Completed
process_search_result
Completed
augmented_chat
Completed
outputs
84%

Azure Container for PyToch



Optimized training framework

Set up, develop, and accelerate PyTorch models on large workloads

Up-to-date stack

Latest compatible versions of Ubuntu, Python, PyTorch, Cuda\ROCm, etc.

Ease of use

Installed and validated against dozens of Microsoft workloads to reduce setup costs and accelerate time to value.

Latest training optimization technologies

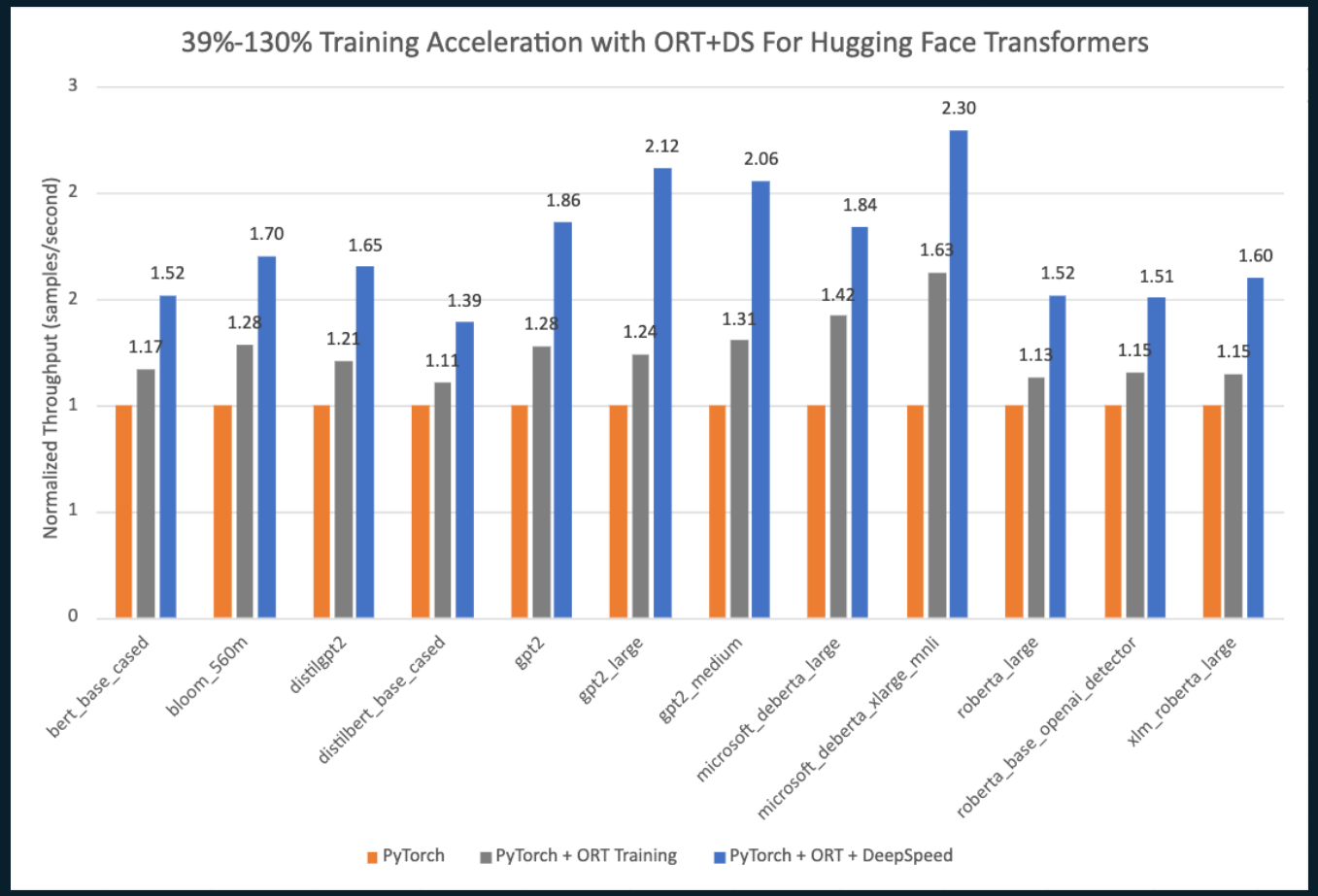
ONNX Runtime, ORT MoE, DeepSpeed, Nebula, MSCCL, and others.

Native integration with Azure

Customer support

[Azure Container for PyTorch - Azure Machine Learning | Microsoft Learn]

Benchmark



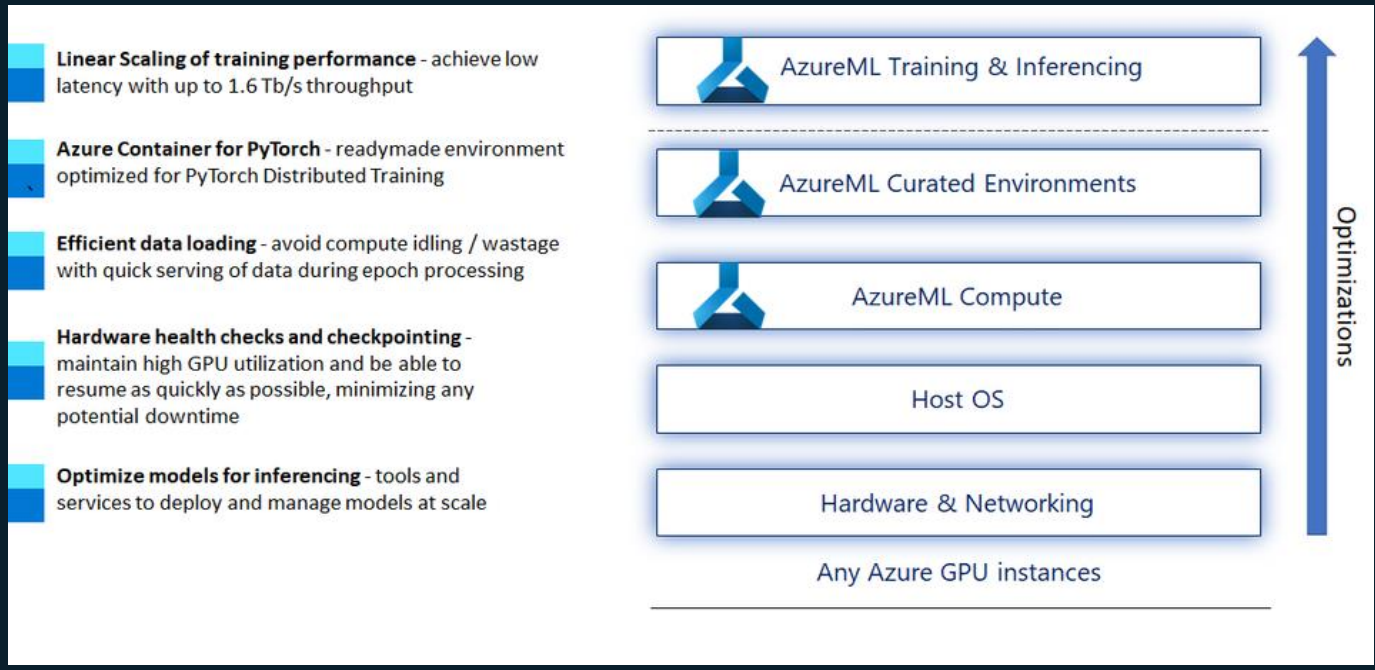
39% - 150% Training Accelereration with ORT+DS For Hugging Face Transformers

[azureml-examples/best-practices/largescale-deep-learning/Training]

Large-scale model in Azure Machine Learning

Best place for high performance deep learning

Microsoft provide best practices for large scale training workloads to get highly efficient optimized performance using state of art technologies.



Training large models in Azure Machine Learning (microsoft.com)

azureml-examples/best-practices/largescale-deep-learning at main · Azure/azureml-examples · GitHub

Key consideration

Azure ML Datastore

- Written in Rust (high speed and high memory efficiency, Avoid issues with Python GIL)
- Multi-process (parallel) data loading etc

Linear scaling with Infiniband Enabled SKUs

- Most of the HPC VM sizes & N-series size designated with ‘r’ are RDMA-capable.

DeepSpeed & ONNXRuntime Training for training optimization

Nebula for fast checkpoint

DeepSpeed-MII for optimized inference

Demo #2 : Train large-scale model

- DeepSpeed configuration file
- Python code for finetuning model
- Azure Container for PyTorch (ACPT) as a curated Environment
- Creating custom Environment based on ACPT
- Training the large-scale model with custom Environment

microsoft / onnxruntime-training-examples

Type / to search

Code | Issues 4 | Pull requests | Actions | Projects | Security | Insights

onnxruntime-training-examples

Public

Edit Pins | Couldn't load subscription status. Retry | Fork 52 | Starred 209

master | 82 branches | 0 tags | Go to file | Add file | Code

prathikr and Prathik Rao add stable diffusio... | 97cc0ff yesterday | 175 commits

ImageClassification-fi...	add readme for aml-vision/image-classification...	last week
QnA-finetune	nebula working (#117)	2 months ago
StableDiffusion-finetu...	add stable diffusion inference (#149)	yesterday
T5	Add files via upload	last month
huggingface	removed GPT2 special treatment (#89)	2 months ago
on_device_training	Fix on-device training example (#141)	3 weeks ago
optimum @ 4056d24	Add Optimum submodule support + add workin...	last year
transformers @ de46c...	Add Optimum submodule support + add workin...	last year
whisper-finetune	Update README.md	last month
.gitignore	add code for youtube demo (#94)	7 months ago

About

Examples for using ONNX Runtime for model training.

Readme
MIT license
Code of conduct
Security policy
Activity
209 stars
40 watching
52 forks
Report repository

Releases

No releases published

3. Enterprise Search

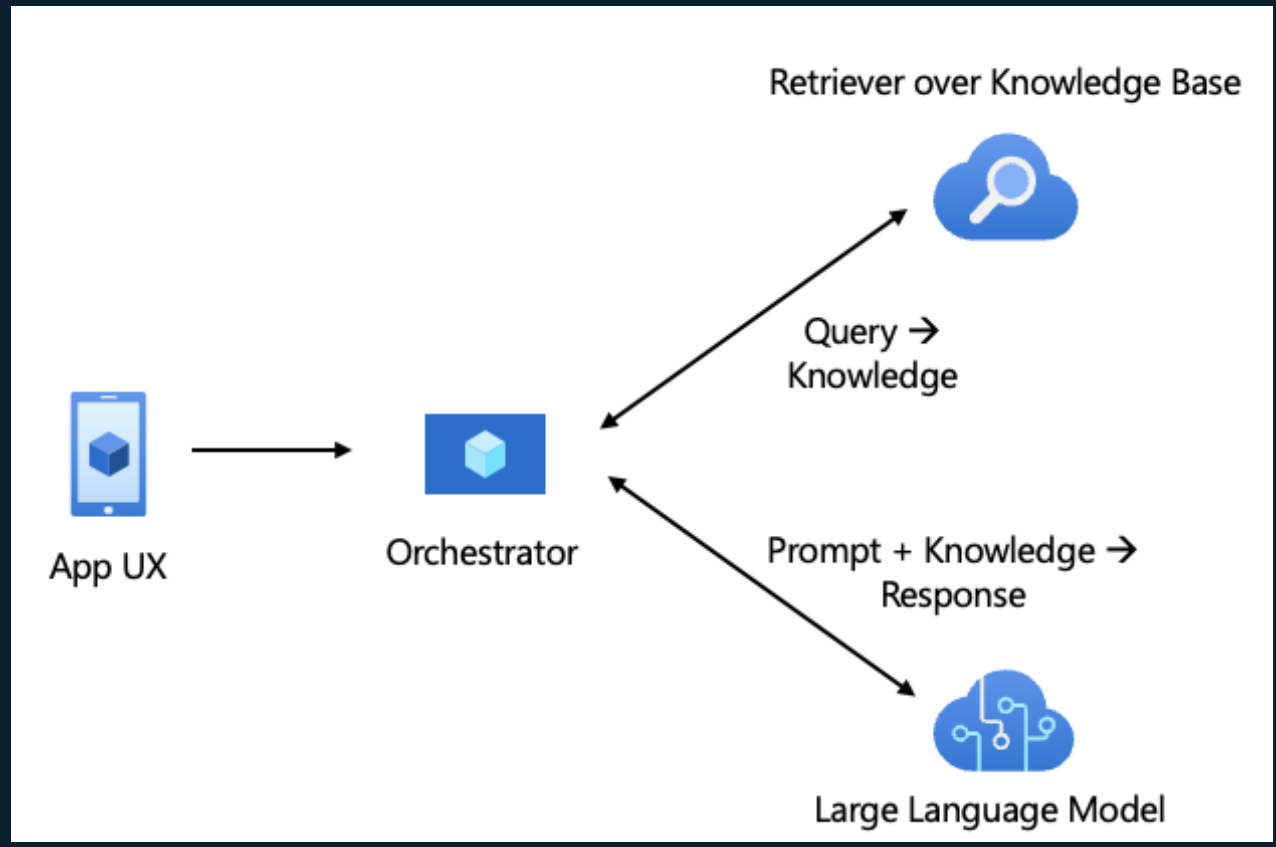
Example of application

- Enterprise search
- Code generation or transformation
- Robotics
- Writing ad
- And more!

Retrieval Augmented Generation

Complement LLMs knowledge by retrieving information relevant to the question

1. Find the most relavant infromation from a large data.
2. Inline these information in a prompt along with instructions and the question itself.

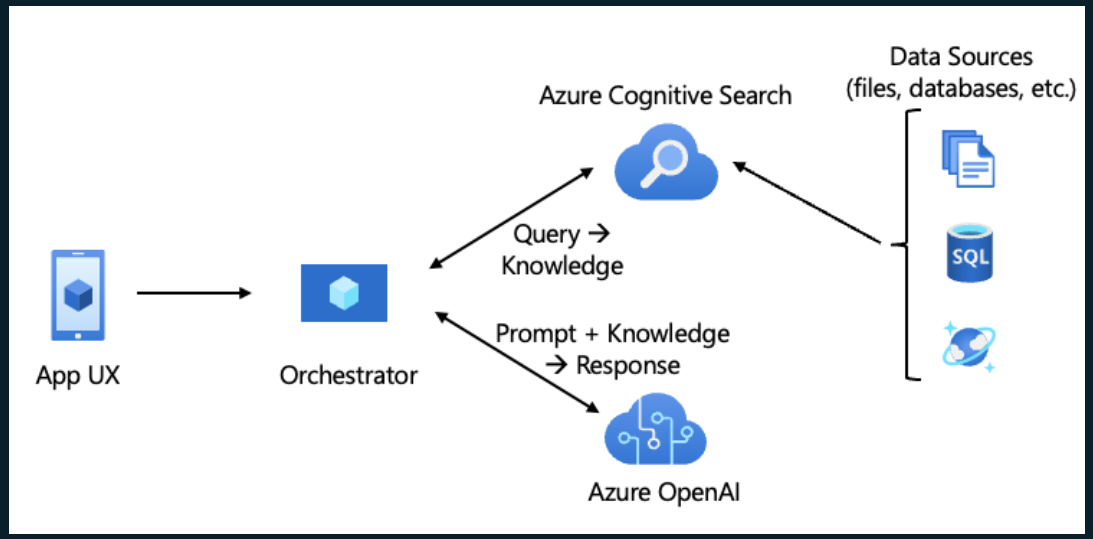


Enterprise search with Azure Cognitive Search

Azure Cognitive Search

- Azure's complete retrieval solution
- Data ingestion, enterprise-grade security, partitioning and replication for scaling, support for 50+ written languages, and more

Basic architecture



Azure-Samples/azure-search-openai-demo: A sample app for the Retrieval-Augmented Generation pattern running in Azure, using Azure Cognitive Search for retrieval and Azure OpenAI large language models to power ChatGPT-style and Q&A experiences. (github.com)

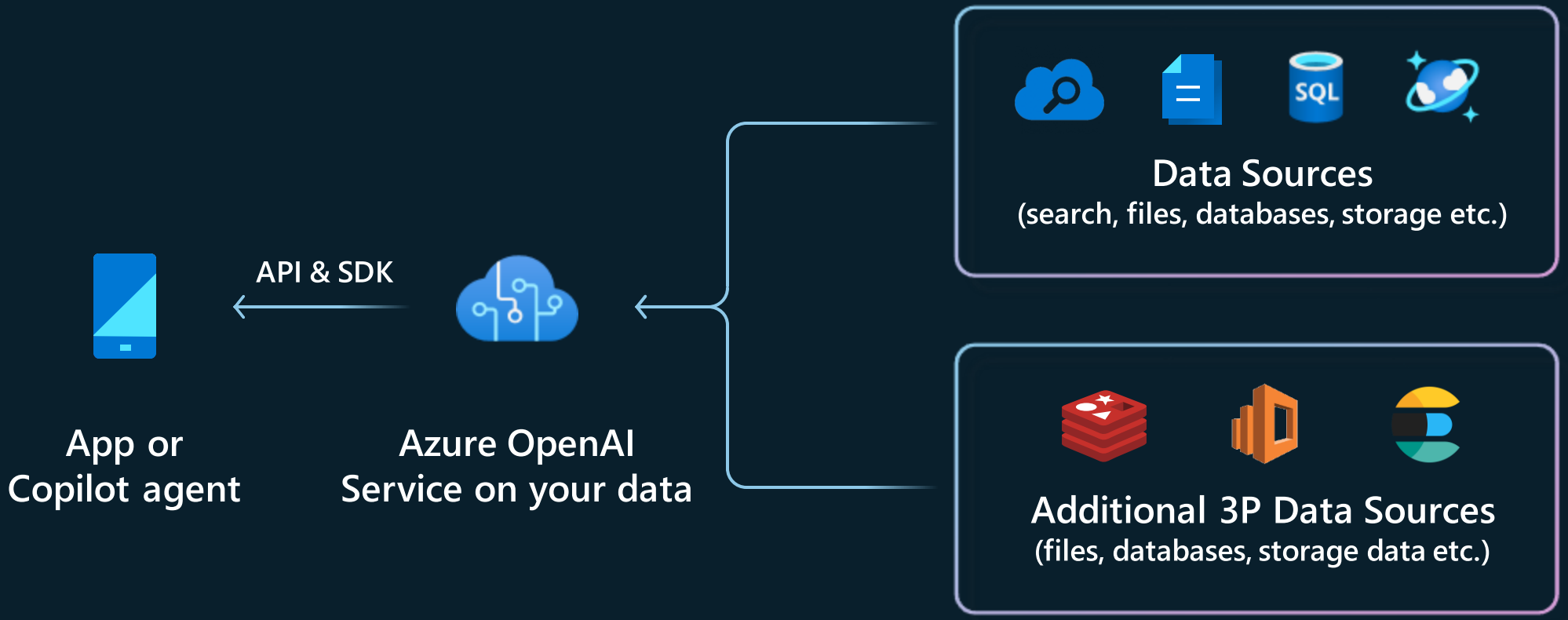
Vector search in Azure Cognitive Search (Private Preview)

Power your retrieval-augmented generation applications

- Use vector or hybrid search
- Use Azure OpenAI embeddings or bring your own
- Deeply integrate with Azure
- Scale with replication and partitioning
- Build generative AI apps and retrieval plugins

Azure/cognitive-search-vector-pr: The official documentation and code samples for the Vector search feature (preview) in Azure Cognitive Search. (github.com)

Azure OpenAI Service on your data (Public Preview)



Introducing Azure OpenAI Service On Your Data in Public Preview - Microsoft Community Hub

Demo #3 : Enterprise search

https://youtu.be/6SNfeVop4zM

4. Recap & Call to action

Recap

1. Introduction to Foundation models
2. Foundation models in Azure
   - Demo #1 : Prompt Flow
   - Demo #2 : Training large scale model
3. Enterprise Search
   - Demo #3 : Enterprise Search with Azure AI
4. Recap & Call to action

Call to action!

Learn Build skills on Microsoft Learn

Develop AI solutions with Azure OpenAI - Training | Microsoft Learn

Connect Join the AI Tech Community to connect, learn, and engage with thousands of members around the world

Artificial Intelligence and Machine Learning - Microsoft Community Hub

Explore Stay up to date with the latest news, announcements and release notes

Azure updates | Microsoft Azure

Release note for CLI v2 and Python SDK v2.

Join Use FastTrack for Azure program to accelerate your project!

FastTrack for Azure – Technical Enablement FAQ | Microsoft Azure

Q&A

Thank you!