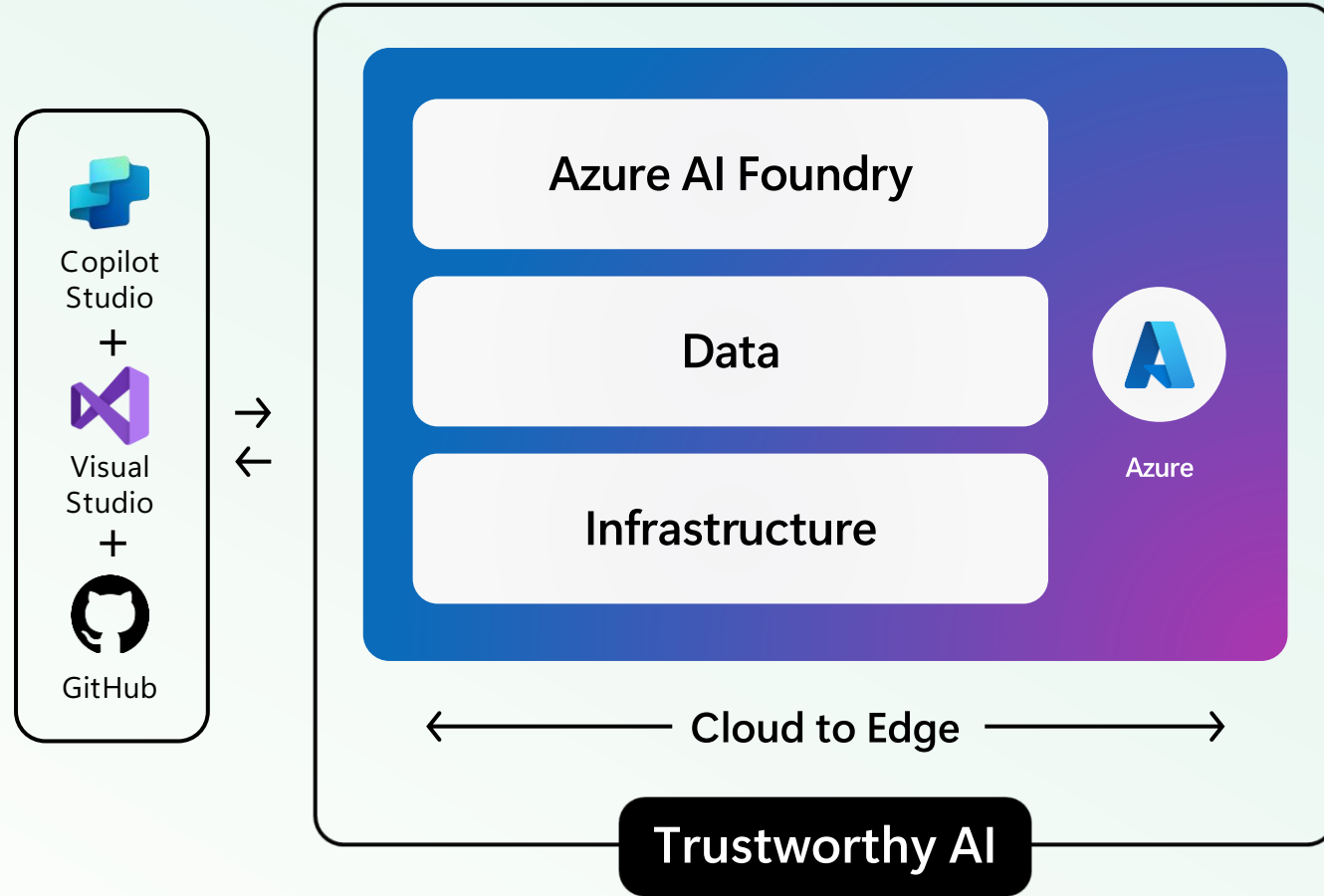
The background is a solid teal color. It features several thin, white, curved lines that sweep across the frame, primarily concentrated in the top-left and bottom-right corners, creating a sense of motion or design.

Design, customize, and manage AI
apps & agents

Copilot & AI stack





Visual Studio



GitHub



Copilot Studio

World's **most loved** developer tools



Azure AI Foundry



Copilot Studio



Visual Studio



GitHub



Azure AI
Foundry SDK



Model Catalog

Foundational models

Open-source models

Task models

Industry models



Azure
OpenAI Service



Azure
AI Search



Azure AI
Agent Service



Azure AI
Content Safety



Azure
Machine Learning

Evaluations

Customization

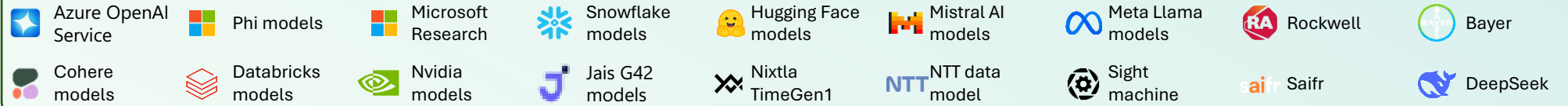
Governance

Monitoring

Observability

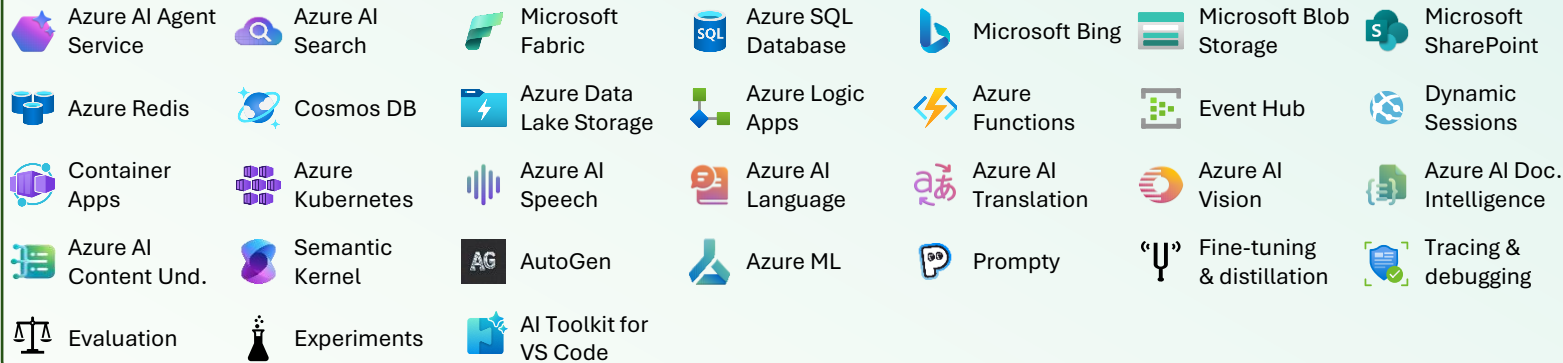
Azure AI Foundry ecosystem

Design with the best models

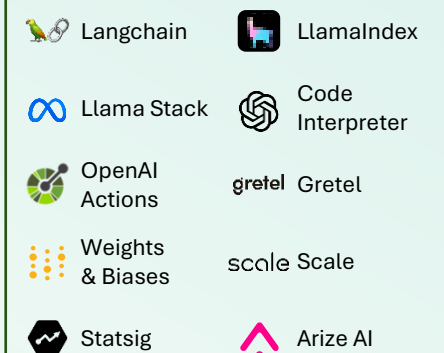


Customize with a comprehensive agent toolchain

1P offerings



3P offerings



Manage performance in production



Safeguard with Trustworthy AI



Copilot Studio

Visual Studio

GitHub

Azure AI Foundry SDK

What does an agent do?



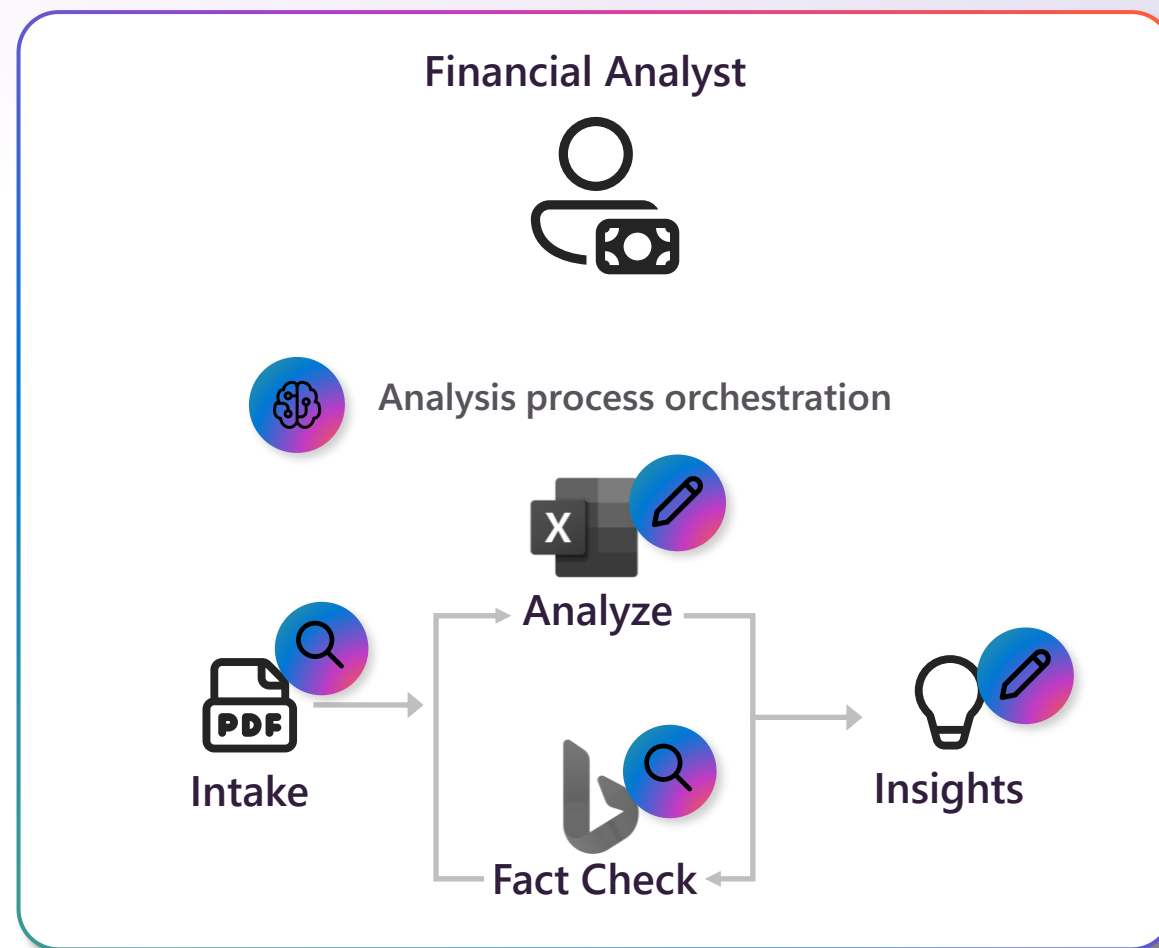
Reason over a provided business process



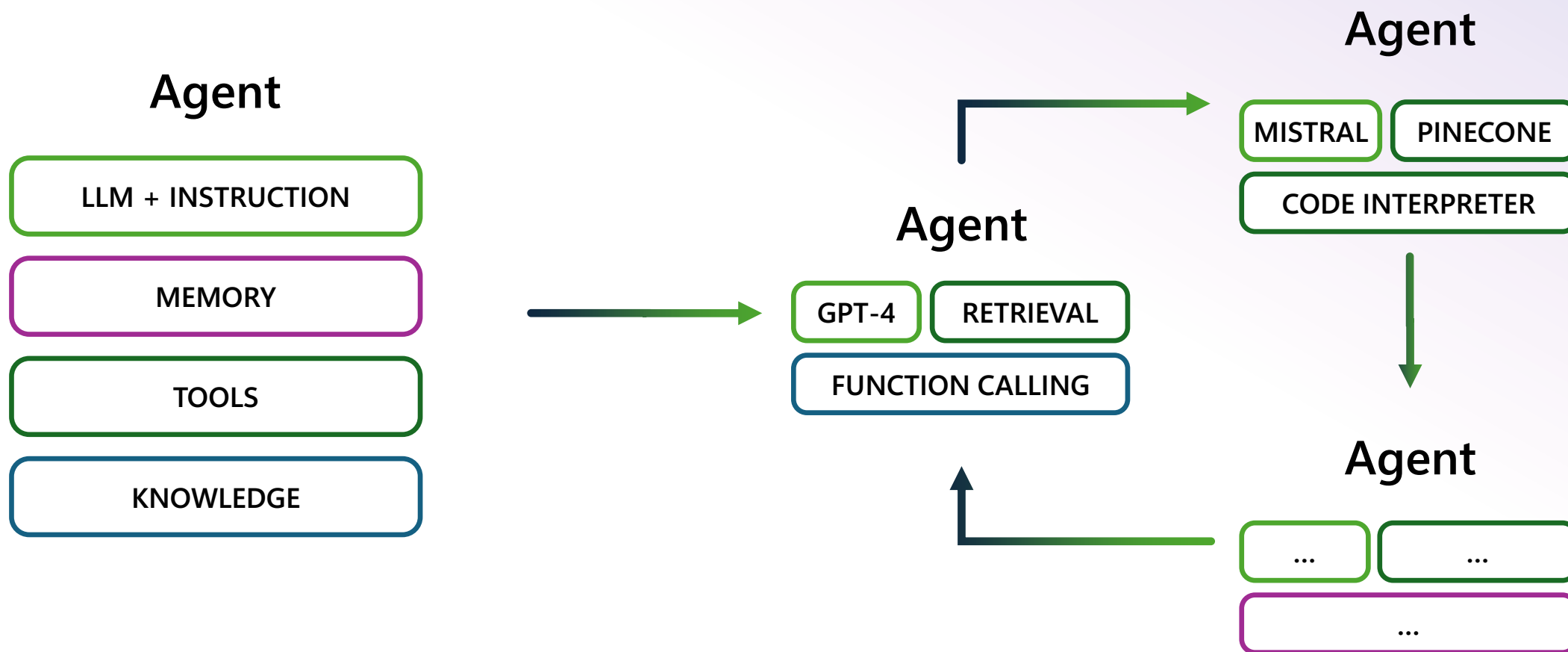
Retrieve context to complete the process



Act on behalf of the user



A single and multi-agent world



Foundry Demo

- What is foundry?
 - AI Studio... now full stack
 - Model Catalog (benchmarks, costs, playgrounds...)
 - Fine tuning
 - Content filters
 - Agents, etc.
- Today – focusing on 2 aspects:
 - Catalog
 - Fine Tuning
- Why?
 - Foundry is a session of its own, so today we're focused
 - Our theme is Token per watt per dollar
 - Using cost-effective models as agents
 - Fine tuning costs up front, but reduces token count in the long run
- Here's the catalog
 - It's a catalog of catalogs
 - Costs, benchmarks, comparisons
- Here's Fine tuning
 - Deploy to Platform or Self-managed
 - Once tuned, here's URL and key... Nico is going to use these

Foundry Screenshots

Jump into a project in Azure AI Foundry

[View all projects](#)[+ Create project](#)[? Help](#)

Project	Created on	Location	Hub	Description
Fine-Tune-Demo-Pr...	Mar 12, 2025 4:43 PM	eastus2	scottyg-ai-hub2	
scottyg-ai-demo-pr...	Aug 12, 2024 3:29 PM	eastus	Scottyg-AI-Hub	
Fine-Tune-Demo-Pr...	Mar 12, 2025 4:43 PM	eastus2	scottyg-ai-hub2	
scottyg-ai-demo-pr...	Aug 12, 2024 3:29 PM	eastus	Scottyg-AI-Hub	

Work outside of a project

Focused on Azure OpenAI Service?

Build specifically with Azure OpenAI Service models and features.

[Let's go](#)

Chat playground

When did Mona say that planning project and what is the timeline r Planning Document?

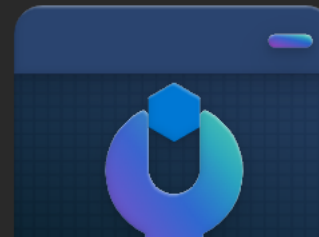
Chatbot

Mona said that Summit Center project is set to and site preparation for the new arena in Atlanti go into the following year.

1 | 2024 Project Plannin | +4

Explore Azure AI Services

Discover the latest in Speech, Language, Vision, and more.

[Try now](#)

Find it fast



Quota management

Actively manage the allocation of rate limits across deployments and increase quotas for resources.



Model catalog and benchmarks

Explore the latest models and see how select models compare to each other.



Safety and security

Learn the end-to-end process of incorporating safety and security into your AI solution.



Content Understanding

Explore how you can transform content of any modality into task specific structured data.

Help

Find the right model to build your custom AI solution

What's new?

DeepSeek-V3 is here!



We present DeepSeek-V3, a strong Mixture-of-Experts (MoE) language mod...

[Check out model](#)

Introducing GPT-4.5 Preview



The latest GPT model that excels at diverse text and image tasks

[Check out model](#) [Read blog](#)

Phi-4-Mini is here!



Enhanced quality, reasoning, efficiency, and speed, all packed into a compact size.

[Check out model](#) [Read blog](#)

Phi-4-Multimodal is here!

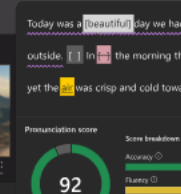


First multimodal SLM with 3 inputs (text, audio, image) in a unified architecture.

[Check out model](#) [Read blog](#)

Explore all the Azure AI Services now available in AI Foundry

Try out powerful AI Service capabilities for free and build customized solutions using Speech, Language, Vision, Content Safety, Translator, and Document Intelligence APIs.

[Check out Speech service models](#) [Learn more about AI Services](#)

Overview

Model catalog

Playgrounds

AI Services

Build and customize

Agents PREVIEWTemplates PREVIEW

Fine-tuning

Prompt flow

Assess and improve

Tracing PREVIEW

Evaluation

Safety + security

My assets

Models + endpoints

Data + indexes

Web apps

Collections

Industry

Capabilities

Deployment options

Inference tasks

Fine-tuning tasks

Licenses

[Compare models](#)

Collections

Search

- ☐ Curated by Azure AI 247
- ☐ Benchmark results
- ☐ Azure OpenAI Service 31
- ☐ Microsoft 40
- ☐ Meta 45
- ☐ Mistral 15

Models 1867


















 o3-mini
Chat completion DeepSeek-V3
Chat completion DeepSeek-R1
Chat completion Phi-4-mini-instruct
Chat completion Phi-4-multimodal-instruct
Chat completion Phi-4
Chat completion gpt-4o-mini-realtime-preview
Audio generation o1
Chat completion o1-mini
Chat completion gpt-4o
Chat completion gpt-4o-mini
Chat completion gpt-4o-audio-preview
Audio generation gpt-4o-realtime-preview
Audio generation

Muse


Cohere-rerank-v3.5

Stable-Diffusion-3.5-Large

Stable-Image-Ultra

-  Overview
-  Model catalog
-  Playgrounds
-  AI Services
- Build and customize 
 -  Agents PREVIEW
 -  Templates PREVIEW
-  Fine-tuning
-  Prompt flow
- Assess and improve 
 -  Tracing PREVIEW
 -  Evaluation
-  Safety + security
- My assets 
 -  Models + endpoints
 -  Data + indexes
 -  Web apps

Fine-Tune-Demo-Project

Add a project description (optional) 

 Help

Endpoints and keys

[View all endpoints](#)



Looking for API keys?

To access keys, your role must be at least Azure AI Developer or higher. Update your permissions and try again, or ask your admin for help. [Learn more about roles](#)

[Learn more about roles](#)

Included capabilities


Azure AI inference


Azure OpenAI Service

Azure AI Services

Use the following endpoint to call your Azure OpenAI Service models:


Azure OpenAI Service endpoint

<https://scott-m72he1ia-eastus2.openai.azure.com/> 


 API documentation

Project details


Project connection string

eastus2.api.azureml.ms;926c329b-9565-433... 


Subscription

Visual Studio Enterprise Subscription 

Subscription ID

926c329b-9565-4331-9e4f-3038a4d278c2  

Location

eastus2 

Manage project settings

- Add users
- View quota
- Connect resources
- Track costs

[Open in management center](#)

Nail the basics with these steps



Stage 1

Define and explore

Choose the right model →

Experiment in the playground →



Stage 2

Build and customize

Work directly in code →

Fine-tune for your use case →

Orchestrate in prompt flow →



Stage 3

Assess and improve

Create a trace →

Evaluate app quality →

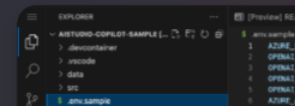
Implement safety and security →

Recent resources


Name	Resource type	Quick actions	Created by
fine_tune_jsonl...	Dataset		Scotty Gresham

Work from a template

Start building with code samples for core scenarios



[Management center](#)

- 
- Overview
 - Model catalog
 - Playgrounds
 - AI Services
 - Build and customize
 - Agents PREVIEW
 - Templates PREVIEW
 - Fine-tuning**
 - Prompt flow
 - Assess and improve
 - Tracing PREVIEW
 - Evaluation
 - Safety + security
 - My assets
 - Models + endpoints
 - Data + indexes
 - Web apps

Fine-tune a model by training it on your own data

[? Help](#)

Optimize pre-trained models for specific tasks by training it on a smaller, task specific dataset to improve its performance and accuracy. Because this method tends to require fewer examples in the prompts, generally less text is sent—and tokens processed—per call.


Generative AI fine-tuning AI Service fine-tuning


Select a model to fine-tune


Some models can only be fine-tuned in specific regions. [Learn more about regional constraints for fine-tuning](#)


Note: Fine-tuning of models other than Azure OpenAI Service models is in preview.


Models 44


 Collections


 Fine-tuning tasks


 Show description


 Search


 **Phi-4-mini-instruct**
Chat completion


 **Phi-4**
Chat completion

 **gpt-4o**
Chat completion

 **gpt-4o-mini**
Chat completion

 **tsuzumi-7b**
Chat completion

 **Ministral-3B**
Chat completion

 **gpt-4**

[Prev](#)

[Next](#)

Phi-4-mini-instruct

 Task: Chat completion

Phi-4-mini-instruct is a lightweight open model built upon synthetic data and filtered publicly available websites - with a focus on high-quality, reasoning dense data. The model belongs to the Phi-4 model family and supports 128K token context length. The model underwent an enhancement process, incorporating both supervised fine-tuning and direct preference optimization to support precise instruction adherence and robust safety measures.

Phi-4-mini-instruct is a dense decoder-only Transformer model with 3.8B parameters, offering key improvements over Phi-3.5-Mini, including a 200K vocabulary, grouped-query attention, and shared embedding. It is designed for chat-completion prompts, generating text based on user input, with a context length of 128K tokens. This static model was trained on an offline dataset with a June 2024 data cutoff. It supports many languages, including Arabic, Chinese, Czech, Danish, Dutch, English, Finnish, French, German, Hebrew, Hungarian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Turkish, Ukrainian.


The model is intended for broad multilingual commercial and research use. The model provides uses for general purpose AI systems and applications which require 1)

Next

Cancel





- Overview
- Model catalog
- Playgrounds
- AI Services
- Build and customize
 - Agents PREVIEW
 - Templates PREVIEW
 - Fine-tuning**
 - Prompt flow
- Assess and improve
 - Tracing PREVIEW
 - Evaluation
 - Safety + security
- My assets
 - Models + endpoints
 - Data + indexes
 - Web apps

Fine-tune a model by training it on your own data

 Help

Optimize pre-trained models for specific tasks by training it on a smaller, task specific dataset to improve its performance and accuracy. Because this method tends to require fewer examples in the prompts, generally less text is sent—and tokens processed—per call.

Generative AI fine-tuning AI Service fine-tuning

 Fine-tune model  Refresh  Deploy  Reset view

Model name	Base model	Status	Created on
My project (1)			

Fine-tuning options PREVIEW

Choose a serving method for fine-tuning this model:

Serverless API

This option lets you fine-tune the model on a fully-managed service that does not require you to host or manage infrastructure.







Managed compute


This option lets you fine-tune the model on Azure infrastructure that you host and manage.





Cancel

 Overview Model catalog Playgrounds AI Services




Build and customize ^

 Agents PREVIEW




</> Templates PREVIEW

 Fine-tuning Prompt flow


Assess and improve ^

 Tracing PREVIEW Evaluation Safety + security

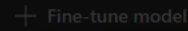
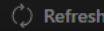
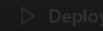
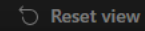
My assets ^

 Models + endpoints Data + indexes Web apps

Fine-tune a model by training it on your own data

 Help

Optimize pre-trained models for specific tasks by training it on a smaller, task specific dataset to improve its performance and accuracy. Because this method tends to require fewer examples in the prompts, generally less text is sent—and tokens processed—per call.

Generative AI fine-tuning AI Service fine-tuning   

Fine-tune Phi-4-mini-instruct PREVIEW

1 Basic settings

2 Training data

3 Validation data optional

4 Task parameters optional

5 Review

Customize this models using your own training data.

Fine-tuned model name *

①

Phi-4-mini-instruct-Finetune

Description**Tags**

Name















Value

Add


Next

Submit






Cancel


-  Overview
-  Model catalog
-  Playgrounds
-  AI Services
- Build and customize ^
-  Agents PREVIEW
-  Templates PREVIEW
-  Fine-tuning
-  Prompt flow
- Assess and improve ^
-  Tracing PREVIEW
-  Evaluation
-  Safety + security
- My assets ^
-  **Models + endpoints**
-  Data + indexes
-  Web apps








Manage deployments of your models and services

 Help

Model deployments Service endpoints

 Deploy model  Refresh  Edit  Open in playground  Reset view

 Columns

Name	Model name	Model version	State	Model retirement date	Content filter	Deployment type
 scott-m72he1ia-eastus2_aoi Azure AI Services  Get endpoint						
gpt-4o	gpt-4o	2024-11-20	Succeeded		DefaultV2 	Global Standard
Phi-4	Phi-4	3	Succeeded		Default 	Global Standard
 Serverless-Phi-4-mini-instruct-Finetune Serverless  Get endpoint						
 Phi-4-mini-instruct-Finet...	Phi-4-mini-instruct-Finetune		Succeeded		Default	

Manage deployments of your models and services

Help

Model deployments Service endpoints

+ Deploy model Refresh Edit Open in playground Reset view

Columns

Name	Model name	Model version	State	Model retirement date	Content filter	Deployment type
scott-m72he1ia-eastus2_aoi						
	gpt-4o				DefaultV2 ⓘ	Global Standard
	Phi-4				Default ⓘ	Global Standard
Serverless-Phi-4-mini-instruct						
	Phi-4-mini-instruct				Default	

Serverless API deployment for Phi-4-mini-instruct-Finetune

Overview Pricing and terms



Phi-4-mini-instruct-Finetune is provided by Microsoft as a First Party Consumption Service.
Learn more about Models as a Service.

Project Name

fine-tune-demo-project

Serverless API is not available now for this model on any of the azure regions, please check again later.

Deployment name *

Phi-4-mini-instruct-Finetune-izg

Content filter (preview)

Enabled

Content filtering uses default configuration and is billed through Azure AI Content Safety. Learn more


Deploy

Cancel

- Overview
- Model catalog
- Playgrounds
- AI Services
- Build and customize
- Agents PREVIEW
- Templates PREVIEW
- Fine-tuning
- Prompt flow
- Assess and improve
- Tracing PREVIEW
- Evaluation
- Safety + security
- My assets
- Models + endpoints
- Data + indexes
- Web apps

Details Consume

Consume

[Open in playground](#) Refresh Edit

Deployment info

Name

Phi-4-mini-instruct-Finetune-seg

Provisioning state

Succeeded

Last updated on

Mar 13, 2025 10:15 AM

Created by

scottygresham@hotmail.com

Created on

Mar 13, 2025 10:15 AM

Model

Phi-4-mini-instruct-Finetune

Endpoint

Target URI

<https://Phi-4-mini-instruct-Finetune-seg.eastus2.model...>

Key

Compute type

Consumption

Swagger URI

<https://Phi-4-mini-instruct-Finetune-seg.eastus2.model...>

10