

Building Al apps: Technical use cases and patterns

### About me

### Yanyong Prommajarn

Full Stack Developer





A **global tech innovation consultancy** dedicated to achieving your business goals through next-level product-centric software delivery.

We provide expertise across the full product journey from ideation to launch and scale, using Sustainable IT practices as a commitment to crafting **tech as a force for good**.





### GitHub Verified Partner









### DOWNLOAD PRESENTATION





Technical Patterns and challenges for AI-powered app development



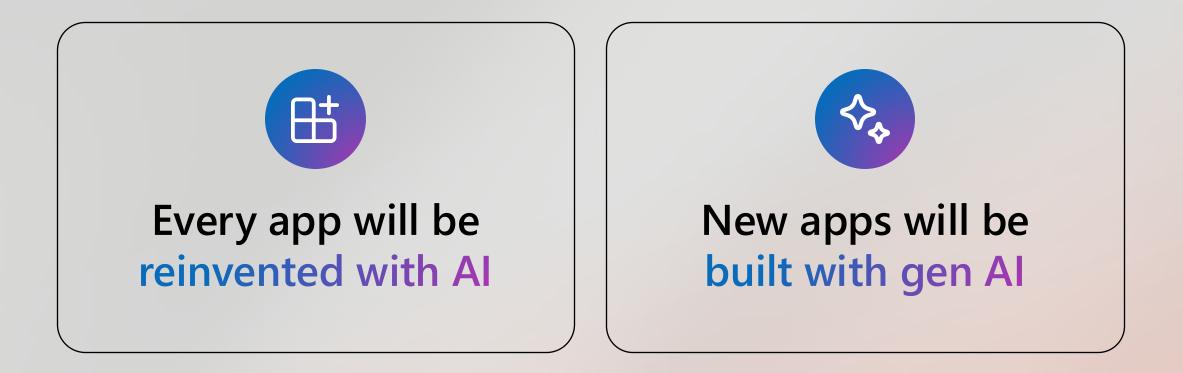


Tools to help you get started easily



Operationalizing AI apps in production

# Intelligence is the new baseline for modern apps



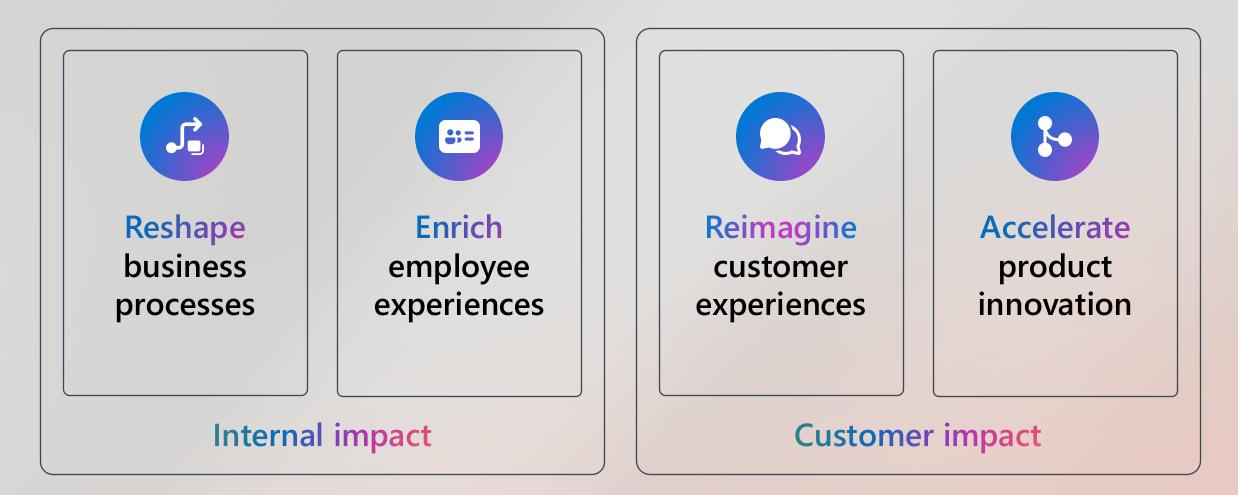
### Delivering on intelligent apps transformation requires

Identify intelligent app use cases aligned to business outcomes

Modernization of existing application and data estate for AI Platform operations for intelligent app deployment

### How you invest in AI matters

### Prioritize projects that deliver repeatable impact



### Reinvent customer experiences with intelligent apps

# Personalization & product discovery

Deliver personalized shopping experiences for customers by analyzing data on their behaviors and preferences and leveraging semantic search capabilities

Engage customers and present fashion recommendations based on trend data

asos

### Content generation & marketing

Create personalized marketing content across platforms, from social media posts to product descriptions

> CarMax uses AI to produce content for car research web pages

CARMAX

### Service & support

Provide personalized and interactive responses to answer customer questions and facilitate routine tasks

> 30% cost reduction for customer support interaction

### ally

#### Build your own copilot

Go beyond bots to chat with your data using natural language, generate and summarize content, surface information over vast amounts of data, and provide engaging experiences

TomTom's Digital Cockpit understands 95% of complex requests, improving response time from 12 seconds to 2.5 seconds



### **Reshape business processes with Intelligent apps**

#### Build your own copilot

Streamline employee access to information about internal policies, and internal documents

70 percent of KPMG's employees have adopted KymChat, making over 120,000 requests since its release



#### Transaction processing and anomaly detection

Handle high volumes of transactions swiftly, accurately, and reliably

Manulife data scientists now take only days to set up an environment, making it much easier to detect fraud

**Manulife** 

#### Information discovery and knowledge mining

Transform unstructured data, such as orders, contracts, applications, and forms, into structured digital information

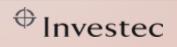
H&R Block tax professionals to find detailed client history in seconds and helps them sort through stacks of forms in seconds

#### H&R BLOCK

Document intelligence and summarization

Automate extraction, aggregation, and summarization of data from various sources, such as webpages, contact center logs and internal documents

Using conversation intelligence to surface keywords and other relevant conversation data to help spot trends and quickly get a pulse on their customers.

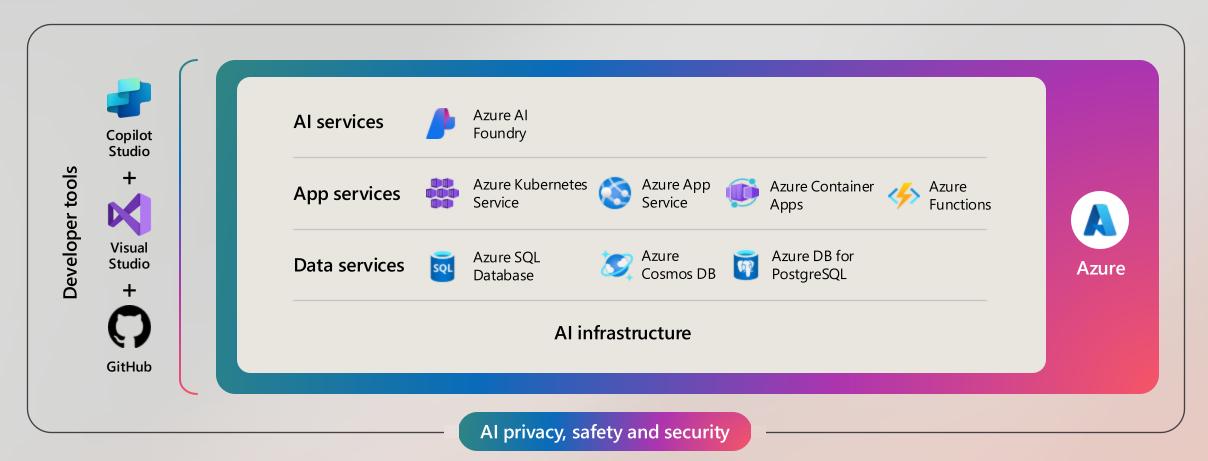


### **AI** application platform

**Extend Microsoft Copilot** 

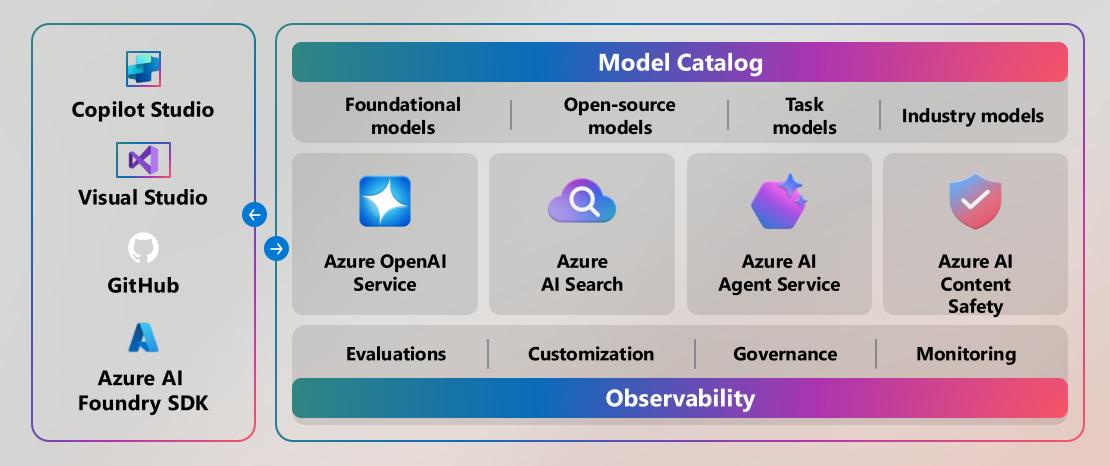
Build your own copilot

Innovate and automate with AI

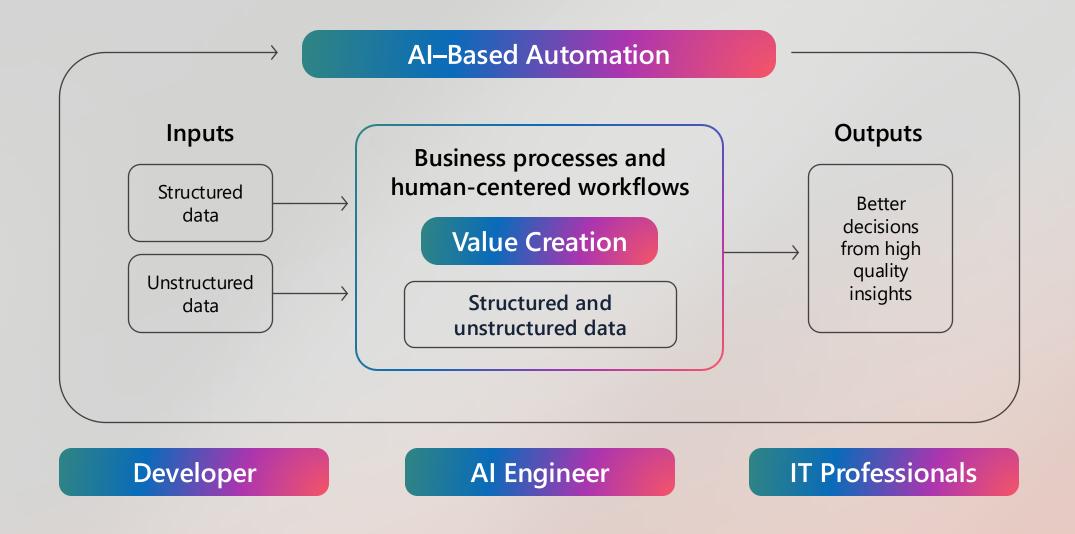




# Azure Al Foundry



### **Challenges organizations face today**



### End to end developer experience





### End to end developer experience





**Available** 



# **Azure OpenAl SDK**

### Azure OpenAI clients for Python, .NET, JavaScript, Java and Go

Azure OpenAl Client in OpenAl library for Python Azure OpenAl companion packages to OpenAl libraries for .NET and JS

Standalone Azure OpenAl libraries for Java and Go

https://aka.ms/oai/sdk

**General** availability



# **OpenAl library for .NET**

### An official stable production grade client library for OpenAl

Supports latest OpenAl features soon after release

Use *Realtime* (preview) and *Structured Outputs* (stable) in your .NET apps Stable companion package for Azure OpenAI with Azure value adds

https://aka.ms/oai/net

**Public preview** 

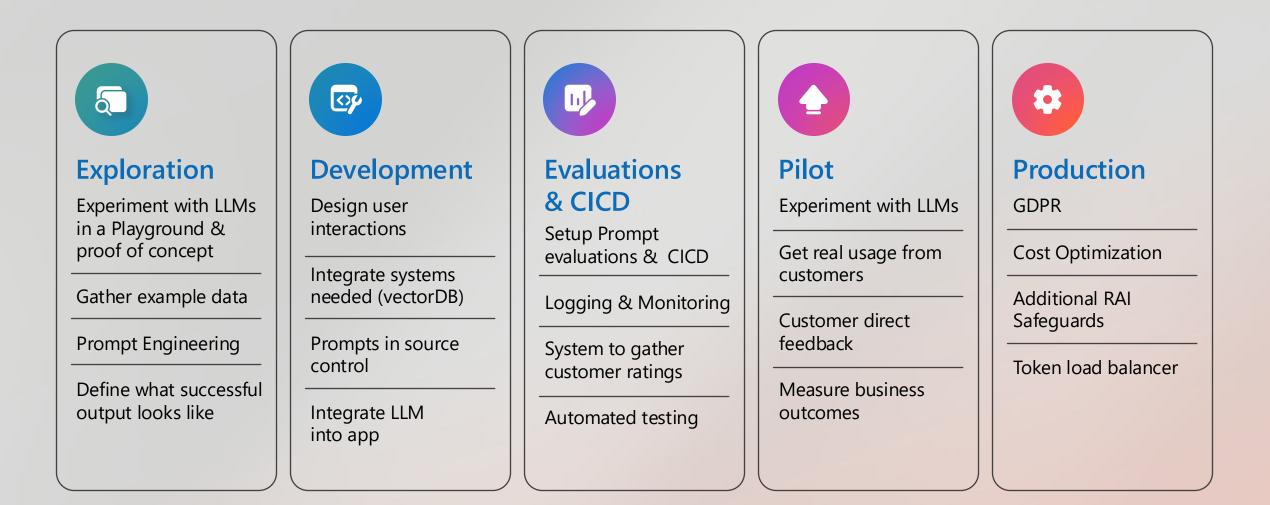
## **Azure AI Foundry SDK**

### **Unified toolchain for AI application development**

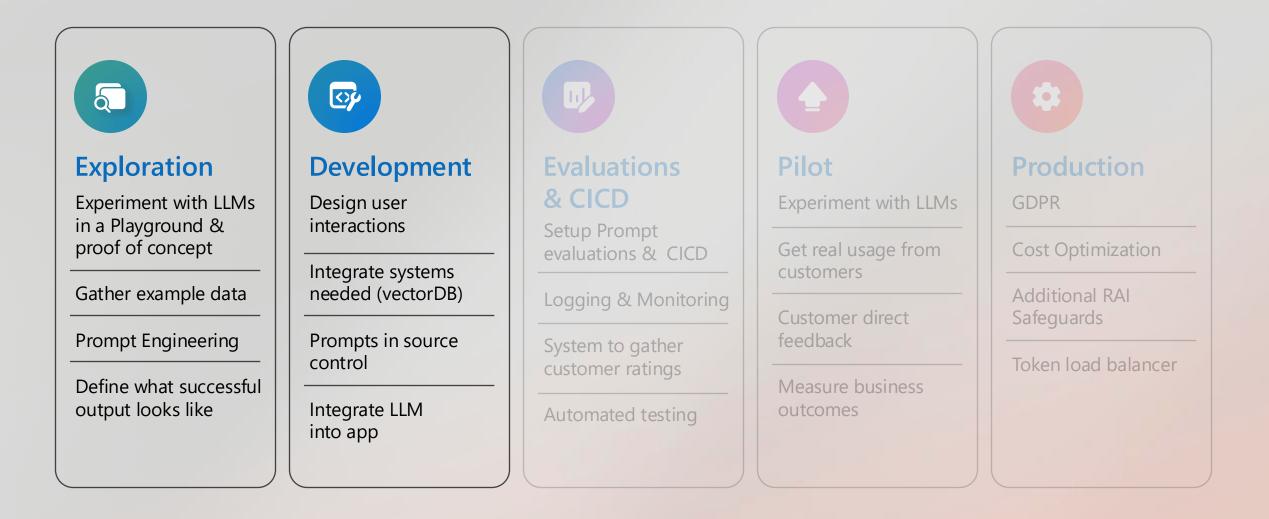
Access our leading models through a single interface Easily integrate Azure AI capabilities into your applications Develop faster with a simplified coding experience

ai.azure.com

### **Building an AI App**



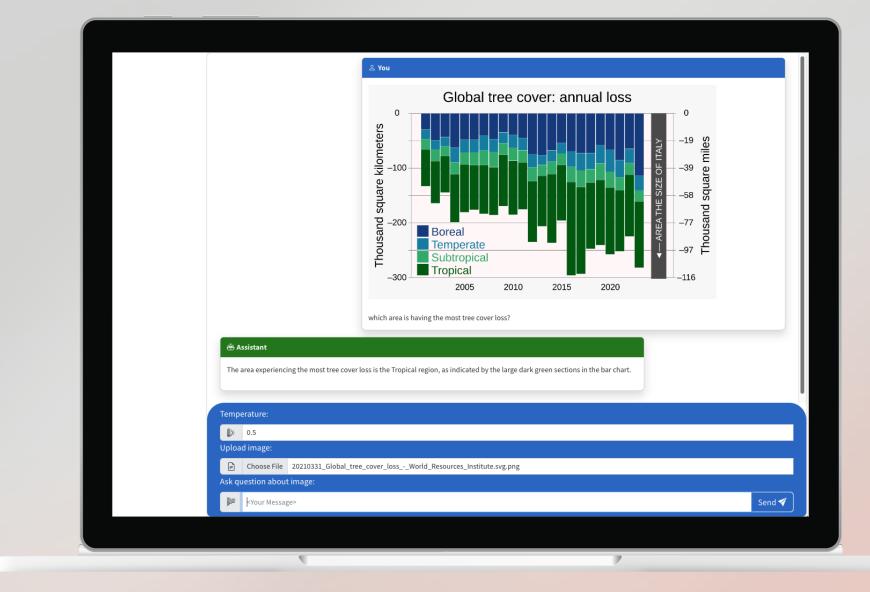
### **Building an AI App**



## **Explore & Build**

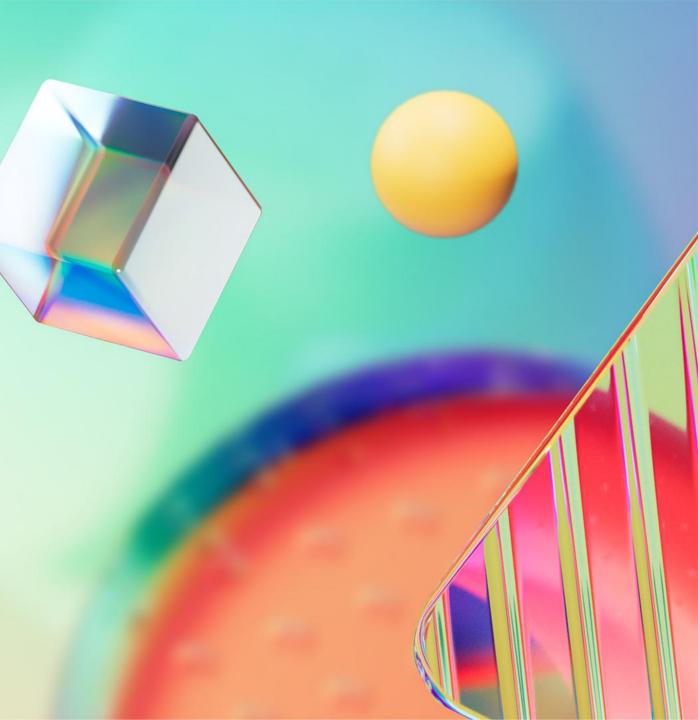
Learn how to build an app using vision capabilities of gpt-40

Build proof of concept and explore prompts



### Demo

Toei Yanyong



#### Announcement

Azure AI application templates gallery

# Quick start your Al development

Al application templates to help you go quickly from idea to application.

Each application repo includes both application code and infrastructure as code files.

Deploy each application to Azure using VS Code & GitHub Actions.



Al app templates					10	Submit fee		
Build w	ith Al							
Duna w								
	application patterns ode and GitHub Action							
Featured AI te	emplates							
Build a Python chat		Deploy a JavaS	cript RAG application that	Deploy a C#	RAG application with	h		
images using OpenA	N State	uses Pinecone		LlamaIndex	Deploy a C# RAG application with LlamaIndex			
A demonstration of chatting of OpenAl vision models like gpt	with uplicaded images using L		t amat, consectatur adipticing elit. Proin uspendisse ubvios ribh justis, sed varia		pp for DialGPT-style experien using Retrinvel Augmented (Kr			
(Descenter (Descenter)	Contraction Date	Description (down	and an other					
(3)			10.00			18.01		
New to Microsoft Adure? By fai								
New to Microsoft Agues? by for								
					Most popular			
Filter by					Most popular			
Filter by			Deploy a Python RAG applicati		Most popular			
Filter by	C. Search templates Showing 15 templates Deploy a Python RA		serverless functions	on with Deg	ioy a Python RAG applicat			
Filter by Language	C. Search templates Showing 15 templates <b>Deploy a Python R4</b> Ide or mean successful	IG application	serverless functions Use remain augmented presenter (RAG) many a chart application.	on with Deg ents	Roy a Python RAG applicat rprise minut augmented generation (Rd + + that application	tion for C) to mote pr		
Filter by Language O ACT C 4 D Second	C Search templates Showing 15 templates Deploy a Python RA template a new magnetic	IG application	serverless functions	on with Deg ents	ioy a Python RAG applicat	tion for C) to mote pr		
Filter by Linguage ONT CF E Assocrat	C Search templates Showing 15 templates Deploy a Python RA template a new magnetic	IG application	serverless functions for original approximation (MA) mapping a chart application.	on with Deg ents	loy a Python RAG applicat rprise mana commend generation RAG a that applications	tion for C) to mote set		
Filter by Language O ACT C A D Second	Sharch templates     Showing 15 templates     Deploy a Python R4     Beploy a net approach     Beploy a net approach     Cot mayor a period	AG application annuals (MAC to relate and annual) area (F1 No. 0.10)	Servedess Functions Desenved augusted presenter (MA) import of the destination The market of the destination The destinat	on with Deg minimum and D the following the second D the following the second D	loy a Python RAG applicat rprise mana commend generation RAG a that applications	tion for Citie motor of Ma		
Filter by Language AUE C C C C S Senarcept Filter Fighton Task Cut completion	C Search templates Showing 15 templates Deploy a Python RA Internet and applaces Comment of applaces Comment of applaces	MG application annuation (SAG) to make and annual (P 1 No 10 SAG) bot for an e-	Serverless Functions Searching and advances Second Se Second Second Sec	on with Deg minimum and D the following the second D the following the second D	loy a Python RAG applicat rprise mana commend generation RAG a that applications	tion for Citie motor of Ma		
Filter by Language Def NET CF Def CF Def Net Python Task Char completion Cuestion answering.	C. Search templates Showing 15 templates Deploy a Python RA Issue and approach of the mean ap	AG application memory RAC to react and and a set of the Context to the for an e-	Servedess Functions Desenved augusted presenter (MA) import of the destination The market of the destination The destinat	on with Deg minimum and Deg in with Deg on with	loy a Python RAG applicat rprise mana commend generation RAG a that applications	tion for Citie metere ris		
Filter by Language AUE C C C C S Senarcept Filter Fighton Task Cut completion	C. Search templates Showing 15 templates Deploy a Python RA Issue and approach of the mean ap	NG application measure (NA) to reast and any (TA) (O) 100 (P) TA: (O) 100 (D) To ran e- mension (NA) to reast and	Serverless Functions Serverless Functions Construction Serverless Construction Cons	an with to math an with an with to math to math	kay a Python RAG applicat rafte mana da agendado. International Internat	tion for Channes of The Phile (Chan Start)		
Filter by Linguage DATP Co Co DataCorpt Data Python Task Data completion Data	C Search templates Showing 15 templates Deploy a Python RA Is the mean agriculture In the mean agriculture Unit mean agriculture Unit mean agriculture	NG application measure (NA) to reast and any (TA) (O) 100 (P) TA: (O) 100 (D) To ran e- mension (NA) to reast and	Anverless functions Internet and approximate generators (MA) approximate approximate generators (MA) Deploy a Python RAG application Poly States (MA) Deploy a Python RAG application Poly (MA) Deploy a Python RAG application Deploy a Python RAG application	an with to math an with an with to math to math	koy a Python RAG applicat reprize an observation provided provided in the particle operations.	tion for Classest and Plass (2.3 S eard and they		

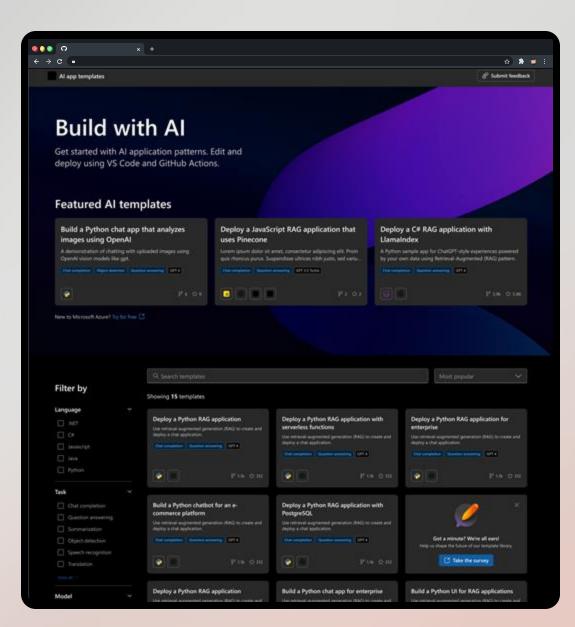
#### http://aka.ms/aiapps

### **Workflow Deployment**

azd init -t azure-search-openai-javascript

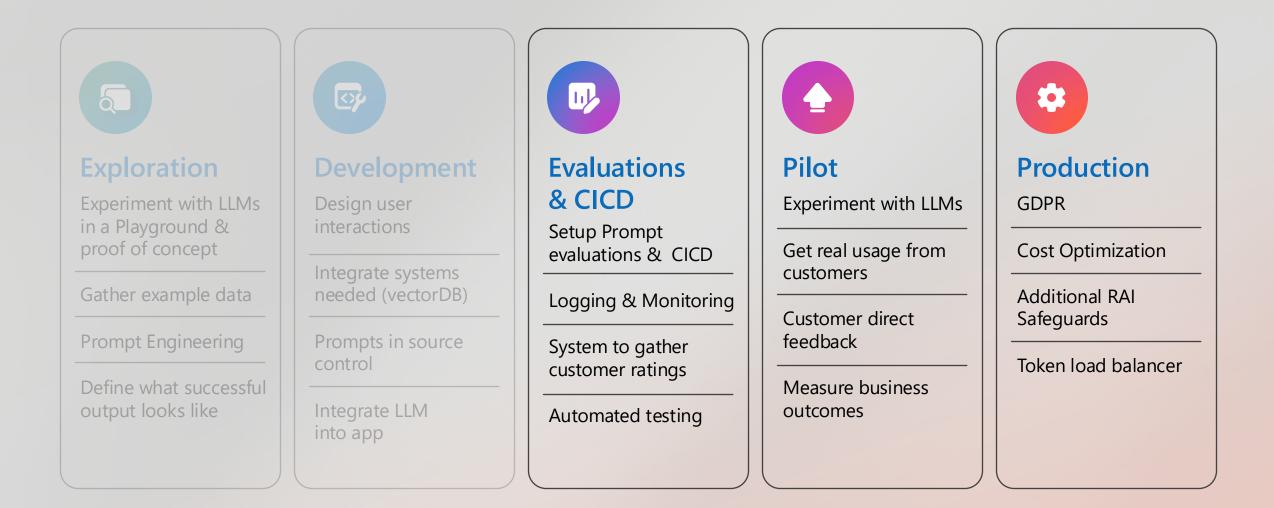
azd login

azd up // main.bicep



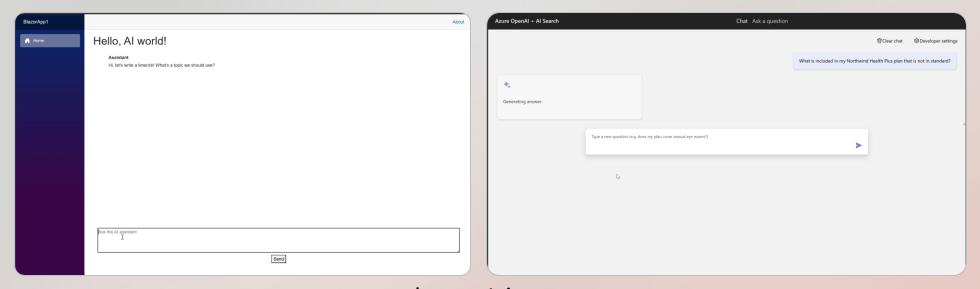
http://aka.ms/aiapps

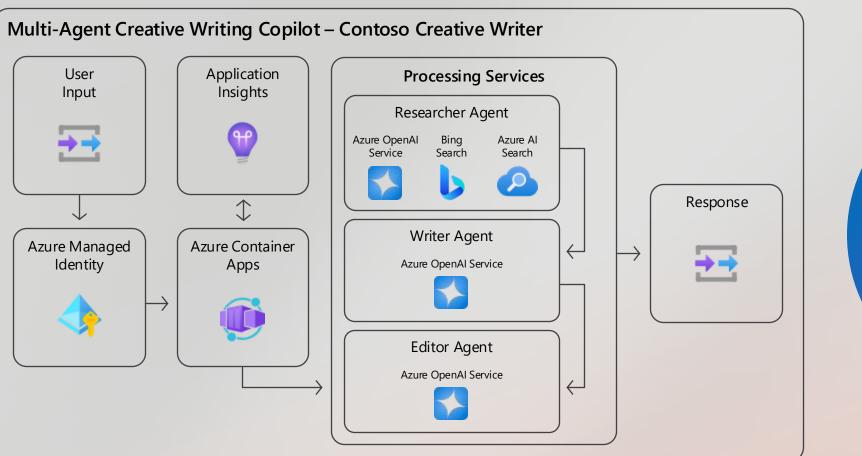
### **Building an AI App**



### **Operationalizing AI Apps**

	Contoso Creative Writer This tool helps you write articles for the Contoso blog.
	Create Your Article         Research         Carry ou find things should 1 find?         Carry ou find the extension of the find and sheet folds are doing in the writer?         Products         Mater products should 1 look ar?
Temperature:       D       0.5       Uplased image:       O: Choose Name (See 100 m)       Act question Social Image:       P: onur Inseage:       D: onur Inseage:	Assignment What kind of writing should 1 do? Write a fun and engaging anticle hall includes the research and product information. The article should be between 800 and 1000 words. Make sure to cles sources in the article as your methods the product information. The article as your methods the product information in the article as your methods the article as your methods the product information in the article as your methods the product information in the product in the product information in the article as your methods the product in t

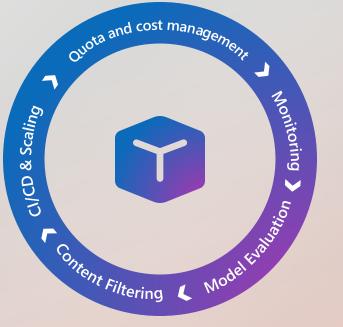




**Operationalizing AI Apps: Demo Scenario** 

Al-powered Content Generation – Contoso Creative Writer





#### **Operationalization Loop**

Azure AI Foundry

#### azd init -t contoso-creative-writer

https://aka.ms/ai-apps-docs/multiagents-gs Information Classification: CONFIDENTIAL (sensitive business information, the level of protection is dictated by legal agreements)

### Demo Scenario Operationalizing AI Apps in Azure



# Demo – Operationalizing Al Apps

Toei Yanyong

### **Demo: Governance with AI Hubs**

All hubs + projects 🗘 👘 😳

Project 🗸 😡 ai-project-ay6ykgl3d772i

#### Management center

8 All hubs + projects

💯 Quota

Hub (ai-hub-ay6ykgl3d772i) 🔿

🛛 🗐 Overview SS Users

 $\bigcirc$  Models + endpoints

S Connected resources

-∿- Compute

Project (ai-project-ay6ykgl...) 🗘 B Overview 음의 Users

 $\bigcirc$  Models + endpoints

of Connected resources

Go to project

#### ai-hub-ay6ykgl3d772i 🧷 + New project () Refresh **O** Reset view = Filter 0 Name Hub Description 1 ai-project-experiments ai-hub-ay6ykgl3d772i ai-project-enterprise-chat

ai-project-ay6ykgl3d772i

ai-hub-ay6ykgl3d772i

ai-hub-ay6ykgl3d772i

### Description

Name

Subscription

**Hub properties** 

Columns

Project for the Al Content

Ignite Al Hub - Provisioned from an Al App Template

Location swedencentral Resource Group MCAPS-Hybrid-REQ-66283-2023-dgartne 🖸 🛽 rg-ignite-cc-demo2 🖸 🛛

Manage in Azure Portal [3]

Groups and application permissions

Get API endpoints and keys

View subscription quota

ai-hub-ay6ykgl3d772i

Users 3

View all  $\rightarrow$ ai-hub-ay6ykgl3d772i Azure Al Administrator Dan Gartner (You) Owner Dan.Gartner@microsoft.com

Manage in Azure Portal

+ New user

25/Page ∨

## **Demo: Developer Playgrounds**

Azure Al Foundry | Management center Hub (ai-hub-ay6ykgl3d772i) **Hub** overview

ai-hub-ay6ykgl3d772i 🧷

4 6

Ŀ

Management center

8 All hubs + projects

Hub (ai-hub-ay6ykgl3d772i) 🗘

& Connected resources

Project (ai-project-ay6ykgl...) 🗘

 $\bigcirc$  Models + endpoints S Connected resources

💯 Quota

0verview

-\/- Compute

B Overview

R Go to project

용 Users

SS Users

#### + New project () Refresh **⑦** Reset view Columns **Filter** Name Hub Description 0 ai-project-experiments ai-hub-ay6ykgl3d772i ai-project-enterprise-chat ai-hub-ay6ykgl3d772i ai-project-ay6ykgl3d772i Project for the Al Content ai-hub-ay6ykgl3d772i

25/Page  $\vee$ 

#### Description Ignite Al Hub - Provisioned from an Al App Template Hub properties Location Name ai-hub-ay6ykgl3d772i swedencentral Subscription **Resource Group** MCAPS-Hybrid-REQ-66283-2023-dgartne 🖸 🚨 rg-ignite-cc-demo2 🖸 🛛 Manage in Azure Portal [2] Get API endpoints and keys View subscription quota Users 3 View all $\rightarrow$ ai-hub-ay6ykgl3d772i Azure Al Administrator Dan Gartner (You) Owner Dan.Gartner@microsoft.com Groups and application permissions + 1 + New user

All hubs + projects 🗘 🚮 🕄 😳

 $\sim$ ai-project-ay6ykgl3d772i

DG

Project

### **Demo: Tracing & LLM Observability**

#### **Contoso Creative Writer**

This tool helps you write articles for the Contoso blog.

#### **Create Your Article**

Research

What kinds of things should I find?

Can you find the latest camping trends and what folks are doing in the winter?

Products

What products should I look at?

Can you use a selection of tents and sleeping bags as context?

2

#### Assignment

What kind of writing should I do?

Write a fun and engaging article that includes the research and product information. The article should be between 800 and 1000 words. Make sure to cite sources in the article as you mention the research not at the end.



#### **Your Article**

**贤** 1.0.0

### **Model Evaluation**

Azure Al Foundry / ai-project-ay6ykgl3d772i 0 / Evaluation										
E	Assess and cor	npare AI applicatior	n performan	ce						⑦ Help
Overview	ew Automated evaluations Manual evaluations Evaluator library									
තී Model catalog										
$\mathcal{Q}_1$ Playgrounds										
Al Services	+ New evaluation	C Refresh X Cancel		mpare 😴 View options 🗸 Default			0	Show all runs	ා 🕞 Switch to c	lashboard view
Build and customize 🛛 🔨	> Evaluation pr	200000								
Code preview		ocess								
日 Fine-tuning	Q Search								<del>∓</del> Filter	🕼 Columns
Prompt flow	Evaluations		Status	Created on ↓	Groundedness	Relevance	Retrieval score	Coherence	Similarity	Fluency
Assess and improve			Status		Gloundeuness	Relevance	Retrieval score	contenence	Similarity	Fluency
So Tracing PREVIEW	evaluate-api	-multi-modal-eval-dataset-cdd383-	Completed	Nov 14, 2024 7:09 PM						
😳 Evaluati	Remote Eval	uation	📀 Completed	Nov 14, 2024 7:09 PM	5	5		4.67		4.33
Safety + security	jolly_cheese_	fy3zt824	Completed	Nov 14, 2024 7:08 PM	5	5		4.67		4.33
My assets ^	evaluation_s	ad_pear_qvjljynwhp	Completed	Nov 14, 2024 6:57 PM						-
🖯 Data + indexes	evaluate-api	-multi-modal-eval-dataset-6744e7	Completed	Nov 13, 2024 3:12 PM						
🖹 Web apps	Remote Eval	uation	Completed	Nov 13, 2024 3:12 PM	4.33	5		5		4.33
	patient_man	go_wp8wrhx4	🤣 Completed	Nov 13, 2024 3:11 PM	4.33	4.67		5		4
	wheat_ant_Ir	h46h0s	🥝 Completed	Nov 13, 2024 2:59 PM	4.33	4.67		4.33		4.33
	evaluation_to	pugh_head_w6qs3pv3f7	Completed	Nov 13, 2024 2:03 PM	5	5		5		5
										•
→ Management center	sid - /subscriptions/5d770805-5r10-4907-b			C Page 1 of 1	> >> 25/Pa	ge 🗸				

https://ai.azure.com/build/evaluation?wsid=/subscriptions/5d779805-fc10-4f97-ba05-f9db44a06981/resourceGroup...

#### Announcement

Azure AI evaluation and experimentation

# Scale AI applications using CI/CD workflows

Streamline workflows with GitHub Actions for evaluation (in public preview) and A/B experimentation (in private preview)

Evaluate full impact with out of the box AI model metrics and custom metrics

GitHub Copilot for Azure plugin that assists with experimentation, creates metrics, powers decisions and more



A/B experimentation private preview sign-up: https://aka.ms/genAI-CI-CD-private-preview

#### × + <br/> 🚖 🛊 🐞 E **Experiment** analysis ♀ Tip For interactive navigation of the latest analysis results for this experiment, view the Experiment Analysis workbook. Heature flag: system\_prompt Allocation ID: gss\_ciSW6lGL8kZUxG4x Analysis period: 0.1 days (10/18/2024 01:23 - 10/18/2024 03:00 UTC) Allocation Assignment Data quality Treatment effect Variant 🥏 Type Control 50% 1.019 n/a n/a professional

#### Metric results

adventurous

Treatment

#### 🖓 Тір

Hover your cursor over a treatment effect badge to display the metric value and the p-value of the statistical test.

1.022

SRM check Pass

Change Detected

#### Important (3 of 5 conclusive) Metric professional 🥜 adventurous 🥜 Degraded +37.0% Average chat call duration [ms] 1.604 Average chat usage tokens Degraded +48.9% 178.6 Inconclusive -1.1% Number of chat calls Total chat usage tokens Degraded +47.2% 125,350 Users with LLM error conclusive +49.6% ► Metric details Usage (0 of 2 conclusive) Cost (2 of 2 conclusive)

50%

#### Azure AI evaluation and experimentation

# Scale AI applications using CI/CD workflows

Ξ 🧔 gartdan / contoso-writer-ignite δ						Q	Type 🕖 to search	8	+ • 💿 n 🖻 🛞
<> Code 💮 Issues 👬 Pull requests 2	♦ Actions	2 🗠 Insights 🕸 Settings							
← Evaluate <b>⊘ Evaluate</b> #4									Re-run all jobs
🙃 Summary	Manually triggered 2 days ago	Status Total duration	Billable time	Artifacts					
Jobs	🥘 gartdan -0- 4ə42f5b main	Success 3m 32s	4m	1					
🥪 evaluate									
Run details Ö Usage	evaluate.yml on: workflow_dispatch								
🕑 Workflow file									
	📀 evaluate 3m 2								
									[3] — +
	evaluate summary								
	Promptflow Evaluation Results								
	research_context			gpt_relevance	gpt_fluency	gpt_coherence	gpt_groundedness		
	0 Can you find the latest can	nping trends and what folks are doi	ing in the winter?	1	5	5	5		
	1 Can you find the latest tren	nds in hiking shoes?		5			5		
	2 Find information about the	e best snow camping spots in the w	vorld	5	5	5	5		

# **Get started today!**

### Al App Template Gallery



### **GitHub Copilot for Azure**



aka.ms/aiapps

#### aka.ms/GetGitHubCopilotForAzure

# How did we do?

Tell us your thoughts about our sessions.





© Copyright Microsoft Corporation. All rights reserved.