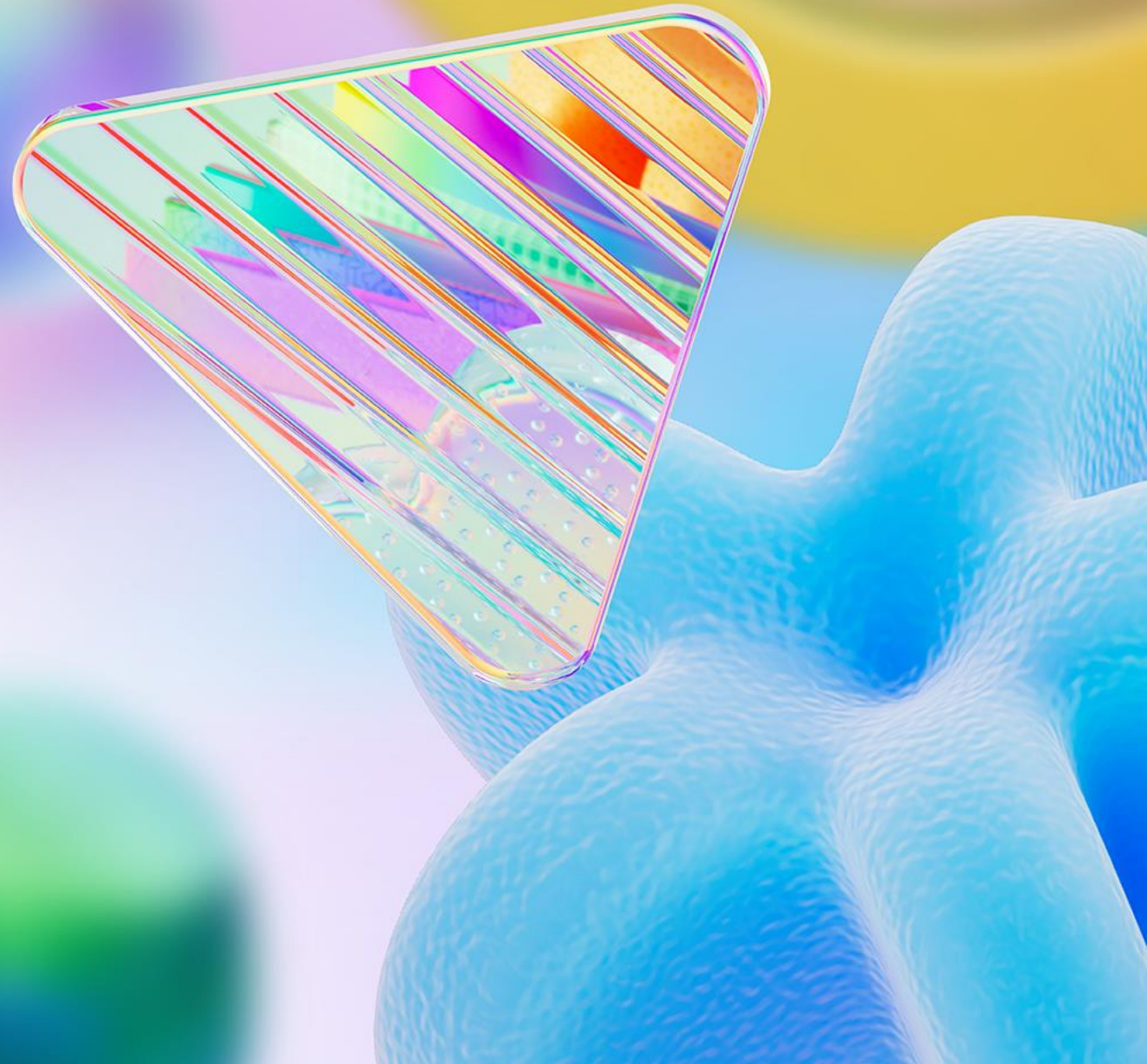




Building AI apps: Technical use cases and patterns



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**.

50
Nationalities

18
Offices

10
Countries



25%
Organic Growth

\$70M
Turnover

700+
Technology experts from across the globe

100%
Independently-owned

0
Debt

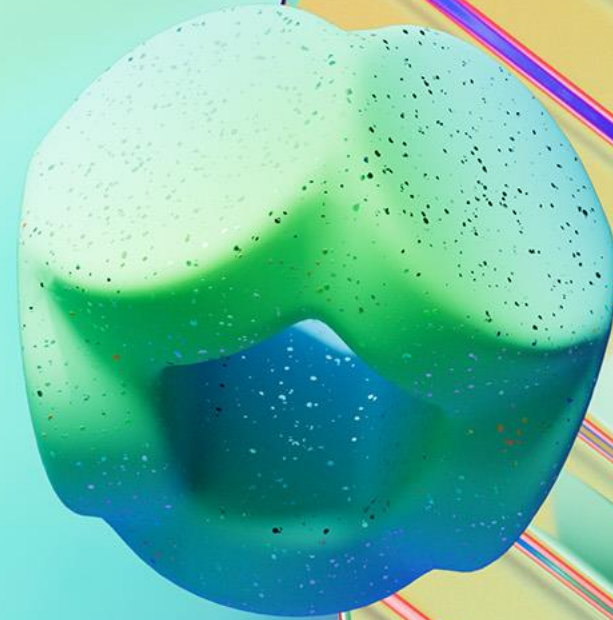
Information Classification: CONFIDENTIAL (sensitive business information, the level of protection is dictated



GitHub Verified Partner



DOWNLOAD
PRESENTATION



Agenda



Technical Patterns and challenges for AI-powered app development



Tools to help you get started easily

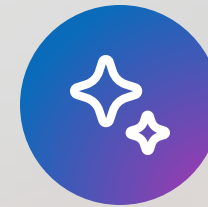


Operationalizing AI apps in production

Intelligence is the new baseline for modern apps



Every app will be
reinvented with AI



New apps will be
built with gen AI

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



Reshape
business
processes



Enrich
employee
experiences

Internal impact



Reimagine
customer
experiences



Accelerate
product
innovation

Customer 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

tomtom

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



Information discovery and knowledge mining

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

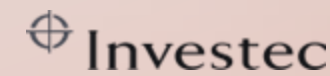
H&R Block tax professionals to find detailed client history in seconds and helps them sort through stacks of forms in seconds



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.

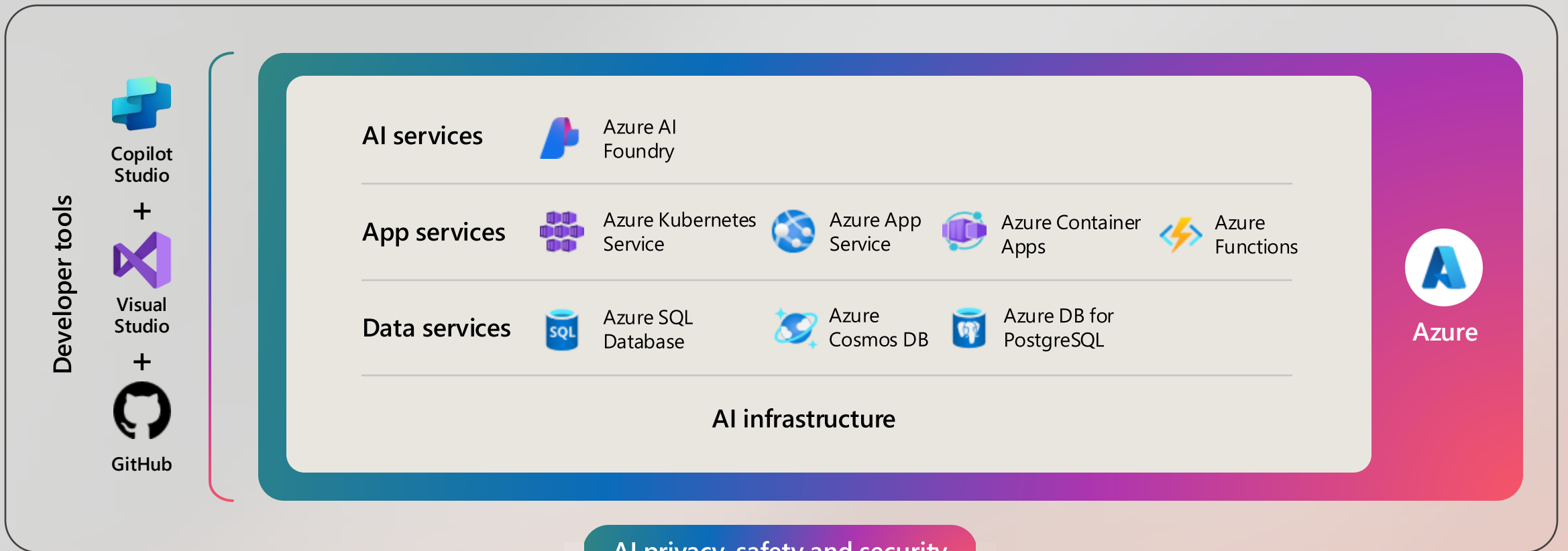


AI application platform

Extend Microsoft Copilot

Build your own copilot

Innovate and automate with AI



Announcing

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**

Evaluations

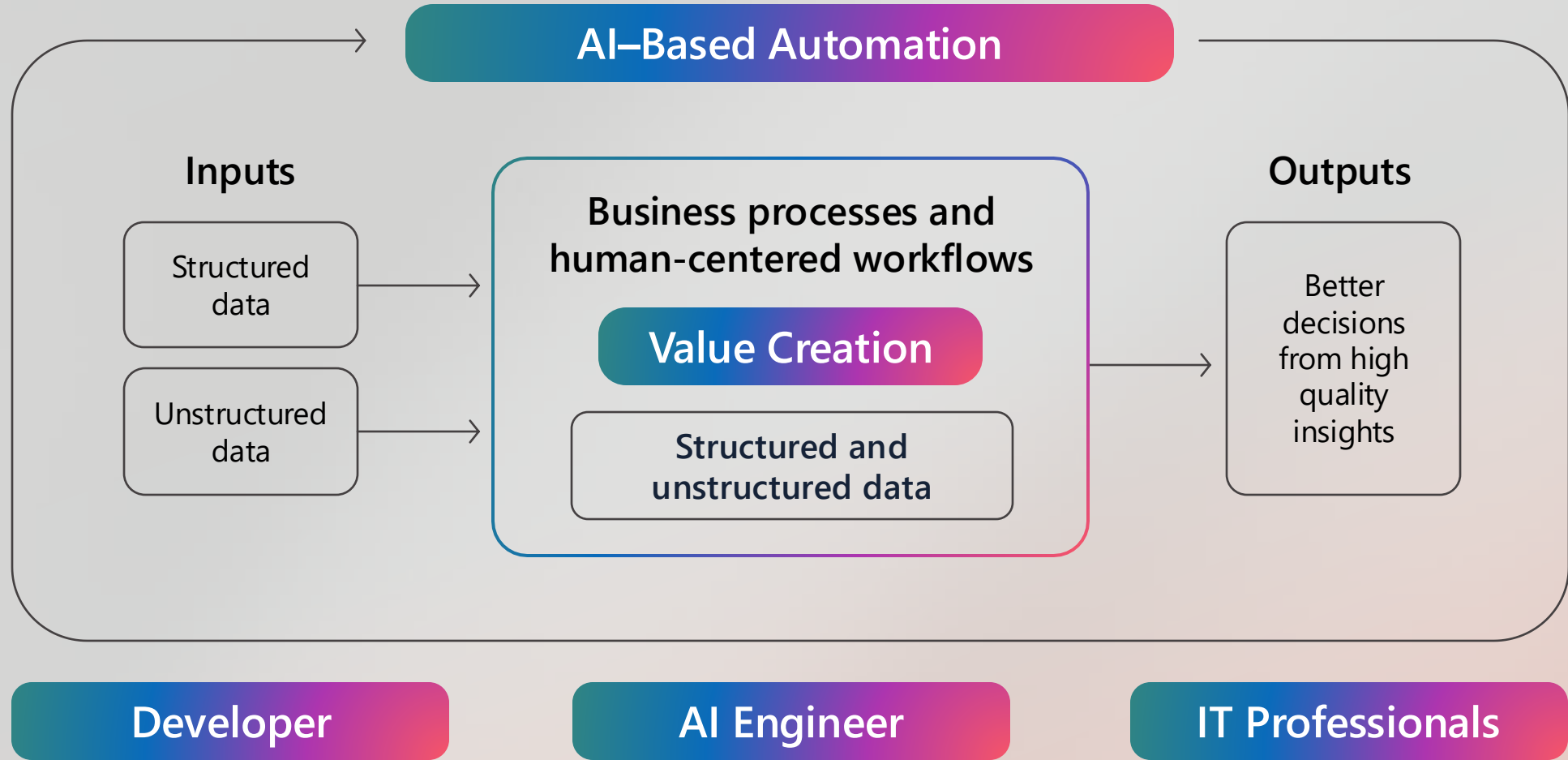
Customization

Governance

Monitoring

Observability

Challenges organizations face today



End to end developer experience



End to end developer experience



Available



Azure OpenAI SDK

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

Azure OpenAI Client
in OpenAI library for
Python

Azure OpenAI companion
packages to OpenAI
libraries for .NET and JS

Standalone Azure OpenAI
libraries for Java and Go

<https://aka.ms/oai/sdk>

General availability



OpenAI library for .NET

An official stable production grade client library for OpenAI

Supports latest
OpenAI 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



Exploration

Experiment with LLMs in a Playground & proof of concept

Gather example data

Prompt Engineering

Define what successful output looks like



Development

Design user interactions

Integrate systems needed (vectorDB)

Prompts in source control

Integrate LLM into app



Evaluations & CICD

Setup Prompt evaluations & CICD

Logging & Monitoring

System to gather customer ratings

Automated testing



Pilot

Experiment with LLMs

Get real usage from customers

Customer direct feedback

Measure business outcomes



Production

GDPR

Cost Optimization

Additional RAI Safeguards

Token load balancer

Building an AI App



Exploration

Experiment with LLMs in a Playground & proof of concept

Gather example data

Prompt Engineering

Define what successful output looks like



Development

Design user interactions

Integrate systems needed (vectorDB)

Prompts in source control

Integrate LLM into app



Evaluations & CICD

Setup Prompt evaluations & CICD

Logging & Monitoring

System to gather customer ratings

Automated testing



Pilot

Experiment with LLMs

Get real usage from customers

Customer direct feedback

Measure business outcomes



Production

GDPR

Cost Optimization

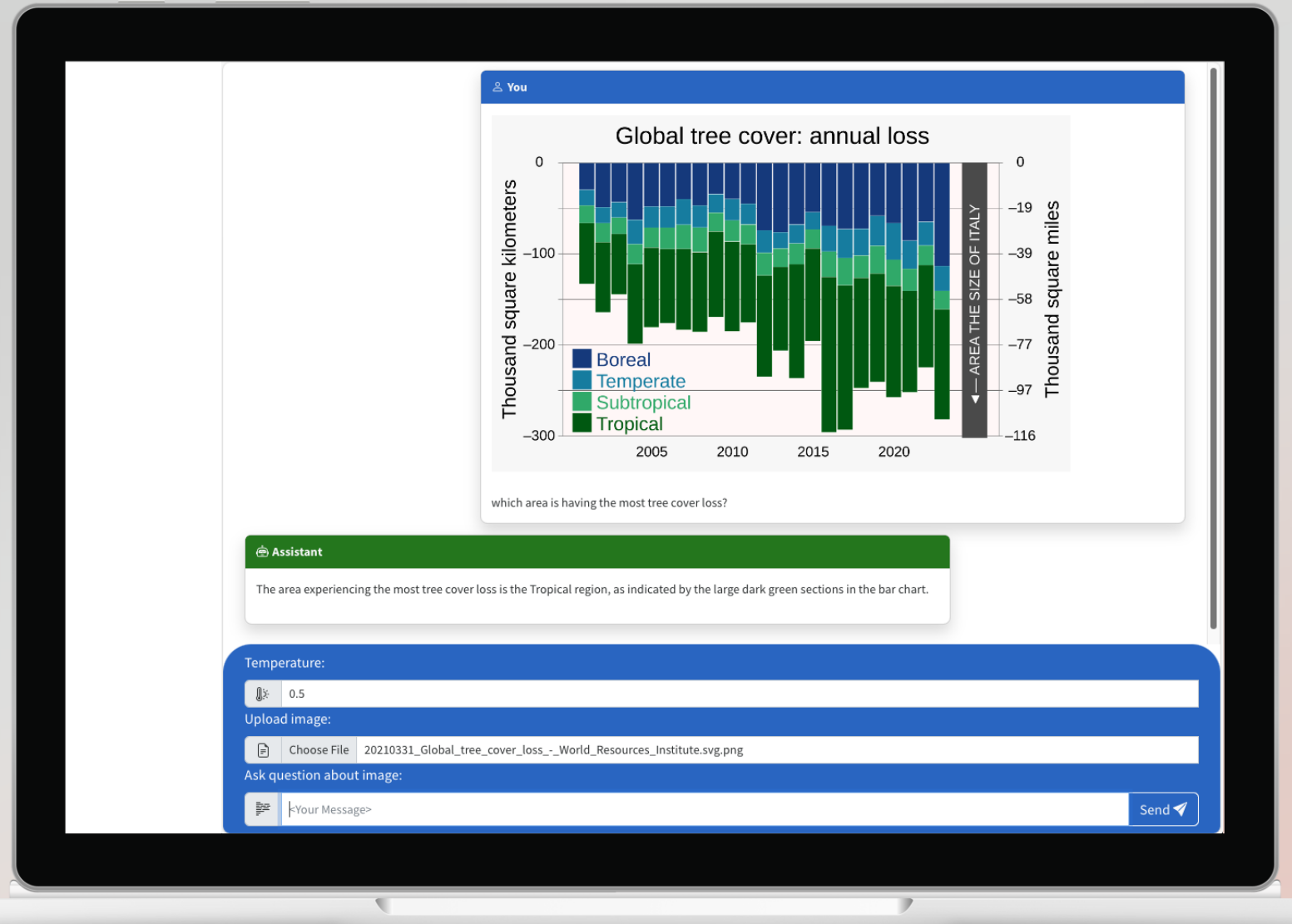
Additional RAI Safeguards

Token load balancer

Explore & Build

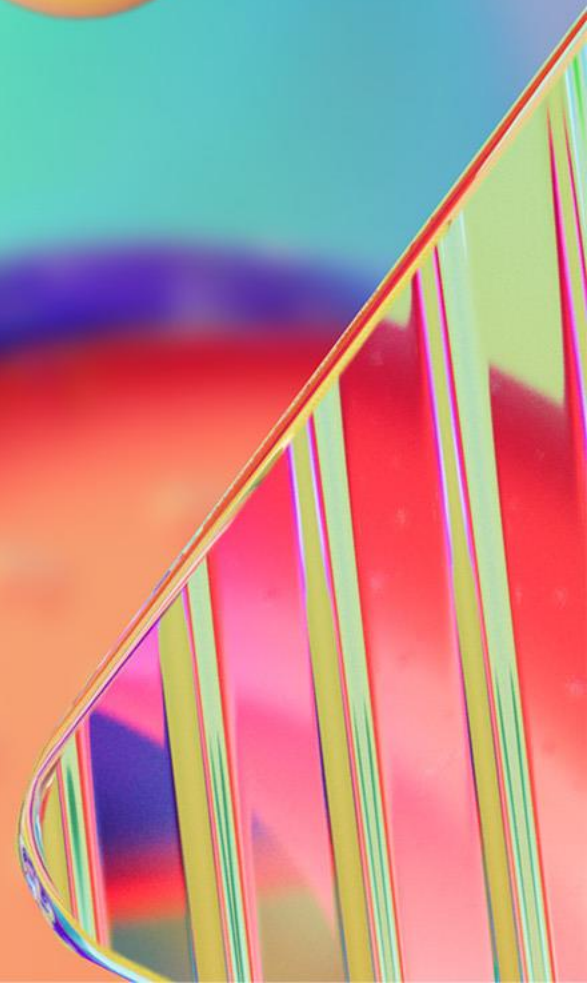
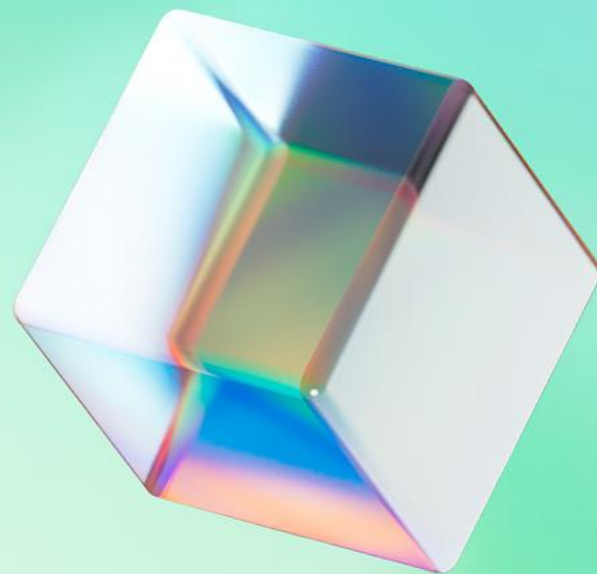
Learn how to build an app using vision capabilities of gpt-4o

Build proof of concept and explore prompts



Demo

Toei Yanyong



Announcement

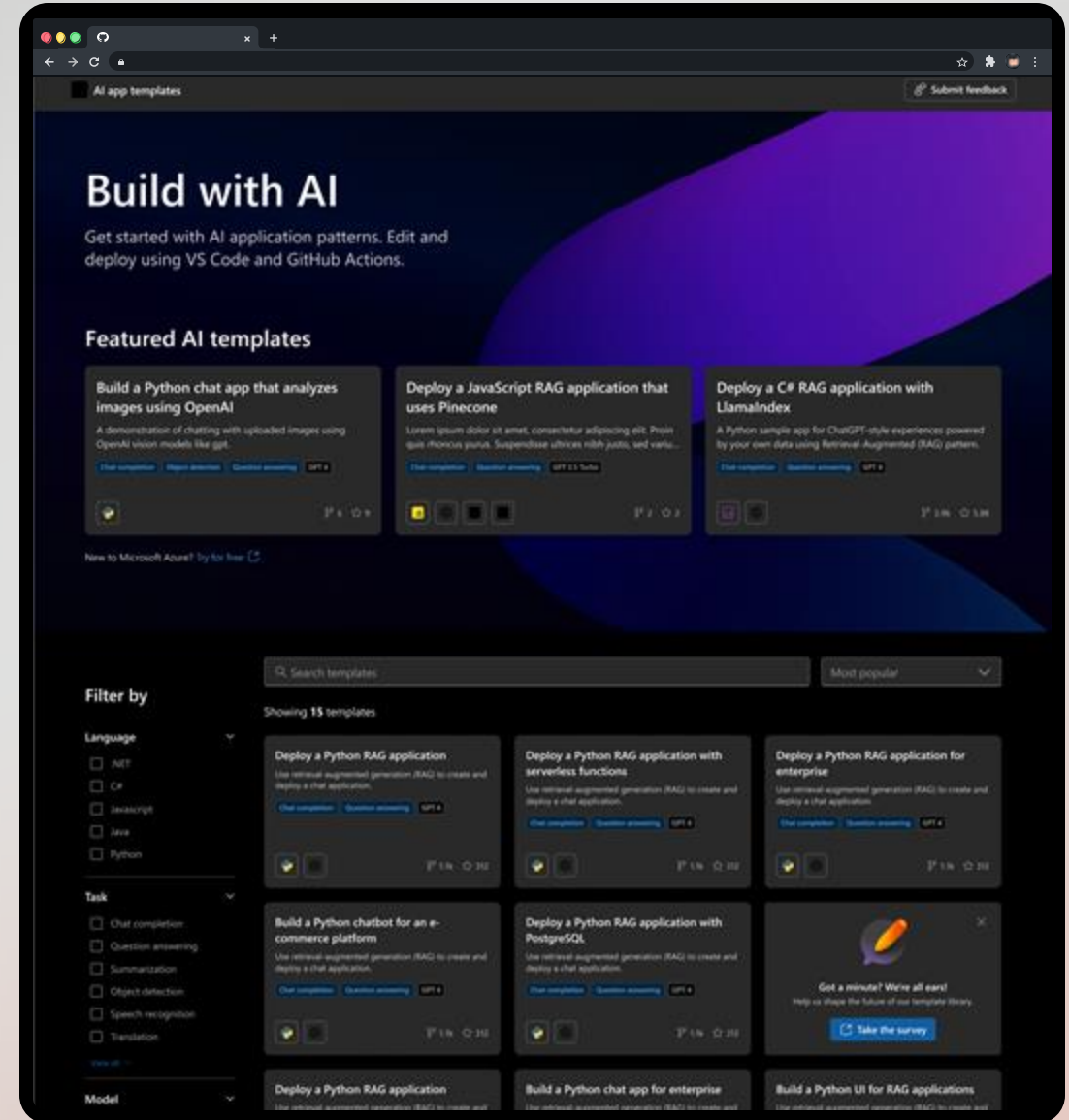
Azure AI application templates gallery

Quick start your AI development

AI 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.



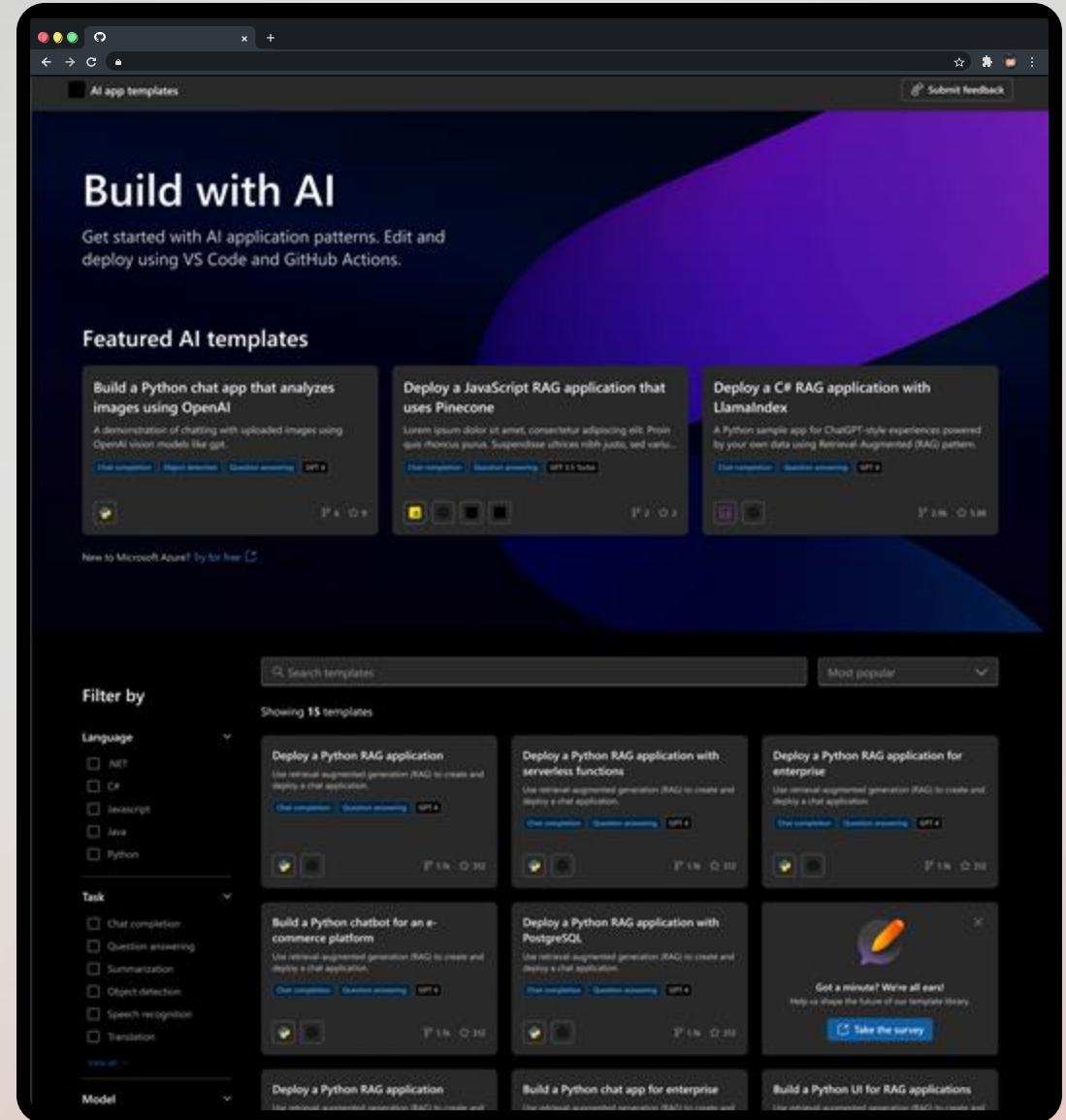
<http://aka.ms/aiapps>

Workflow Deployment

```
azd init -t azure-search-openai-javascript
```

```
azd login
```

```
azd up // main.bicep
```



<http://aka.ms/aiapps>

Building an AI App



Exploration

Experiment with LLMs in a Playground & proof of concept

Gather example data

Prompt Engineering

Define what successful output looks like



Development

Design user interactions

Integrate systems needed (vectorDB)

Prompts in source control

Integrate LLM into app



Evaluations & CICD

Setup Prompt evaluations & CICD

Logging & Monitoring

System to gather customer ratings

Automated testing



Pilot

Experiment with LLMs

Get real usage from customers

Customer direct feedback

Measure business outcomes



Production

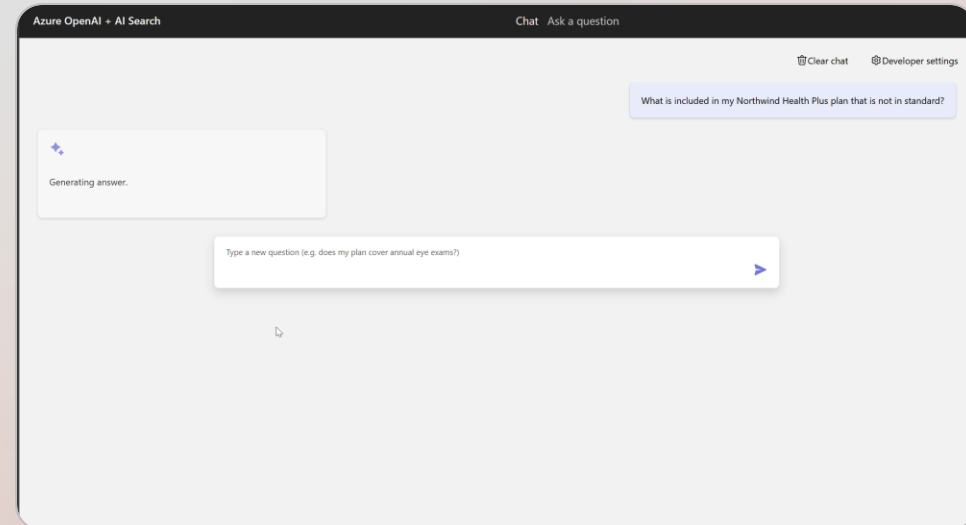
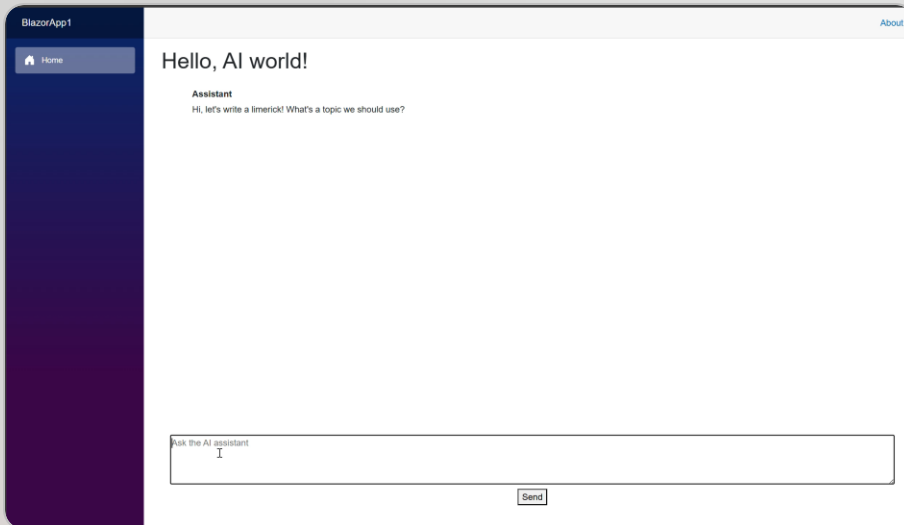
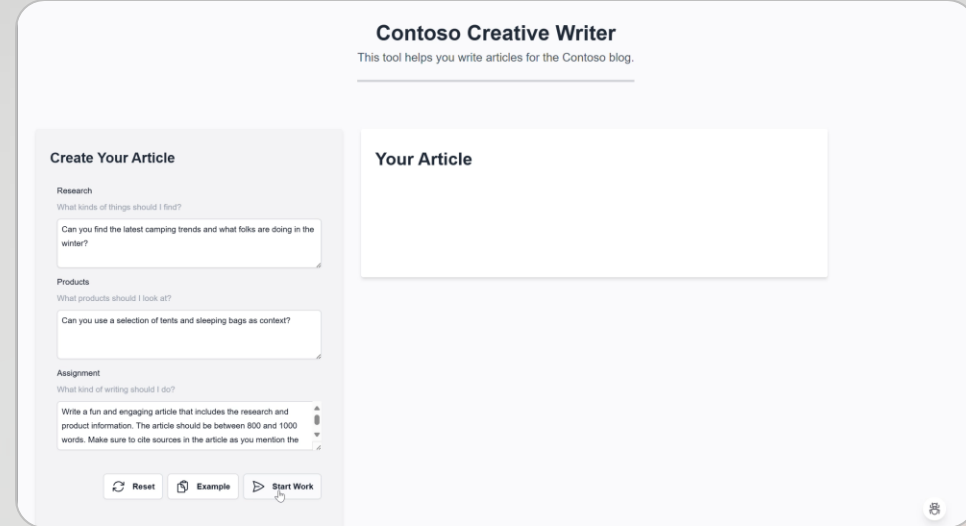
GDPR

Cost Optimization

Additional RAI Safeguards

Token load balancer

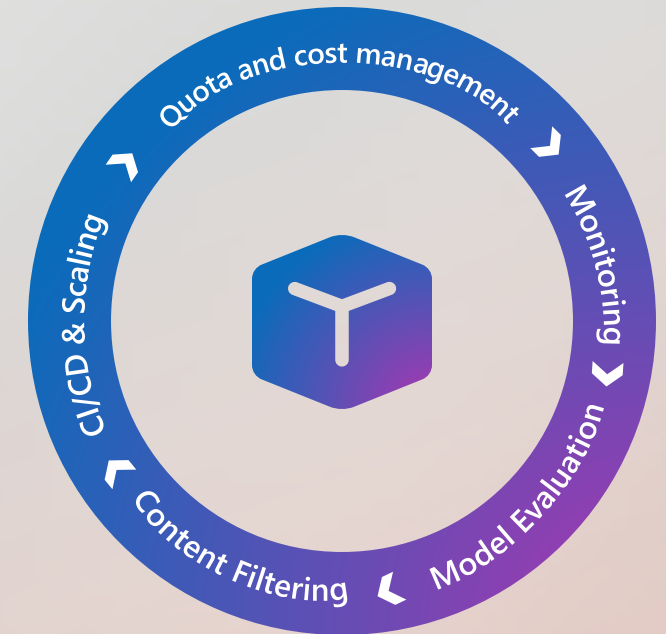
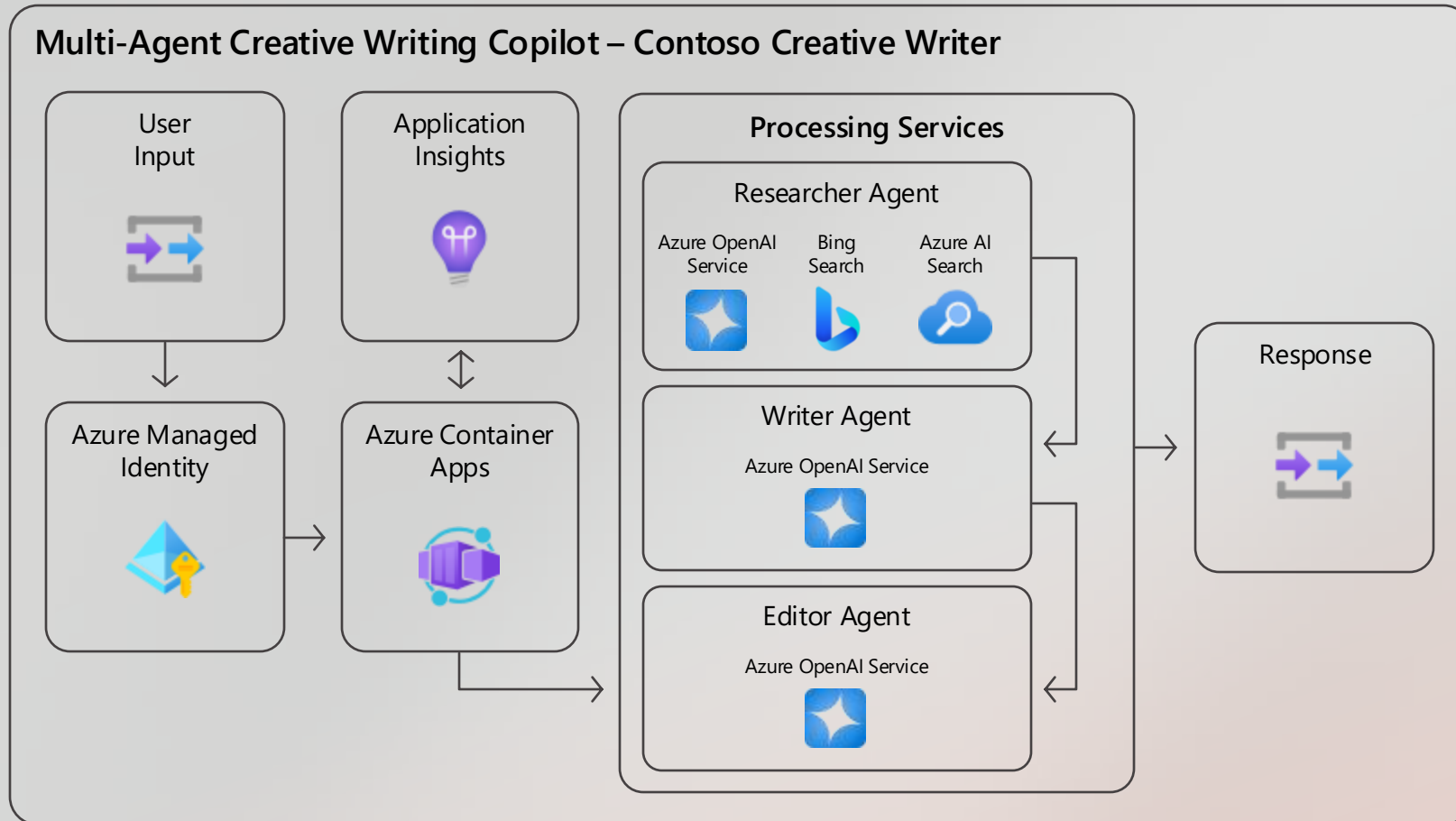
Operationalizing AI Apps



aka.ms/aiapps

Operationalizing AI Apps: Demo Scenario

AI-powered Content Generation – Contoso Creative Writer



Operationalization Loop



azd init -t contoso-creative-writer

<https://aka.ms/ai-apps-docs/multiagents-qs>

Demo Scenario

Operationalizing AI Apps in Azure



AI Governance
with AI Hubs



AI Foundry
Playgrounds




AI
Evaluations



CI/CD of AI
Apps



AI App
Monitoring,
Tracing &
Experiments



Demo – Operationalizing AI Apps

Toei Yanyong

Demo: Governance with AI Hubs

ai-hub-ay6ykg13d772i

- [+ New project](#)
- [Refresh](#)
- [Delete project](#)
- [Reset view](#)

[Filter](#) [Columns](#)

Name	Hub	Description
ai-project-experiments	ai-hub-ay6ykg13d772i	
ai-project-enterprise-chat	ai-hub-ay6ykg13d772i	
ai-project-ay6ykg13d772i	ai-hub-ay6ykg13d772i	Project for the AI Content

Description

Ignite AI Hub - Provisioned from an AI App Template

Hub properties

Name	ai-hub-ay6ykg13d772i	Location	swedencentral
Subscription	MCAPS-Hybrid-REQ-66283-2023-dgartne	Resource Group	rg-ignite-cc-demo2

- [Manage in Azure Portal](#)
- [Get API endpoints and keys](#)
- [View subscription quota](#)

[Delete hub](#)

Users 3

[View all](#) →

- ai-hub-ay6ykg13d772i Azure AI Administrator
- Dan Gartner (You) Dan.Gartner@microsoft.com Owner
- Groups and application permissions + 1 Manage in Azure Portal

[+ New user](#)

- Management center**
- All hubs + projects
 - Quota
 - Hub (ai-hub-ay6ykg13d772i)
 - Overview**
 - Users
 - Models + endpoints
 - Connected resources
 - Compute
- Project (ai-project-ay6ykg13d772i)
- Overview
 - Users
 - Models + endpoints
 - Connected resources
- [Go to project](#)

Demo: Developer Playgrounds

ai-hub-ay6ykg13d772i

[+ New project](#) [Refresh](#) [Delete project](#) [Reset view](#)

[Filter](#) [Columns](#)

Name	Hub	Description
ai-project-experiments	ai-hub-ay6ykg13d772i	
ai-project-enterprise-chat	ai-hub-ay6ykg13d772i	
ai-project-ay6ykg13d772i	ai-hub-ay6ykg13d772i	Project for the AI Content

Description

Ignite AI Hub - Provisioned from an AI App Template

Hub properties

Name ai-hub-ay6ykg13d772i	Location swedencentral
Subscription MCAPS-Hybrid-REQ-66283-2023-dgartne	Resource Group rg-ignite-cc-demo2

[Manage in Azure Portal](#)
[Get API endpoints and keys](#)
[View subscription quota](#)

[Delete hub](#)

Users 3

[View all](#)

- ai-hub-ay6ykg13d772i Azure AI Administrator
- Dan Gartner (You) Dan.Gartner@microsoft.com Owner
- Groups and application permissions + 1 [Manage in Azure Portal](#)

[+ New user](#)

Management center

- All hubs + projects
- Quota
- Hub (ai-hub-ay6ykg13d772i)
- Overview**
- Users
- Models + endpoints
- Connected resources
- Compute

Project (ai-project-ay6ykg1...)

- Overview
- Users
- Models + endpoints
- Connected resources

[Go to 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?

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.

 Reset

 Example

 Start Work

Your Article



Model Evaluation

Assess and compare AI application performance

Help

Automated evaluations Manual evaluations Evaluator library

Evaluate the quality and safety of your generative AI applications with industry standard metrics to compare and choose the best version based on your need. [Learn more about metrics.](#)

+ New evaluation

Refresh

Cancel

Delete

Compare

View options

Default

Show all runs

Switch to dashboard view

Evaluation process

Search

Filter

Columns

Evaluations	Status	Created on ↓	Groundedness	Relevance	Retrieval score	Coherence	Similarity	Fluency
evaluate-api-multi-modal-eval-dataset-cdd383-	Completed	Nov 14, 2024 7:09 PM	--	--	--	--	--	--
Remote Evaluation	Completed	Nov 14, 2024 7:09 PM	5	5	--	4.67	--	4.33
jolly_cheese_fy3zt824	Completed	Nov 14, 2024 7:08 PM	5	5	--	4.67	--	4.33
evaluation_sad_pear_qvjljynwhp	Completed	Nov 14, 2024 6:57 PM	--	--	--	--	--	--
evaluate-api-multi-modal-eval-dataset-6744e71	Completed	Nov 13, 2024 3:12 PM	--	--	--	--	--	--
Remote Evaluation	Completed	Nov 13, 2024 3:12 PM	4.33	5	--	5	--	4.33
patient_mango_wp8wrhx4	Completed	Nov 13, 2024 3:11 PM	4.33	4.67	--	5	--	4
wheat_ant_inh46h0s	Completed	Nov 13, 2024 2:59 PM	4.33	4.67	--	4.33	--	4.33
evaluation_tough_head_w6qs3pv3f7	Completed	Nov 13, 2024 2:03 PM	5	5	--	5	--	5

Announcement

Azure AI evaluation and experimentation

Scale AI applications using CI/CD workflows

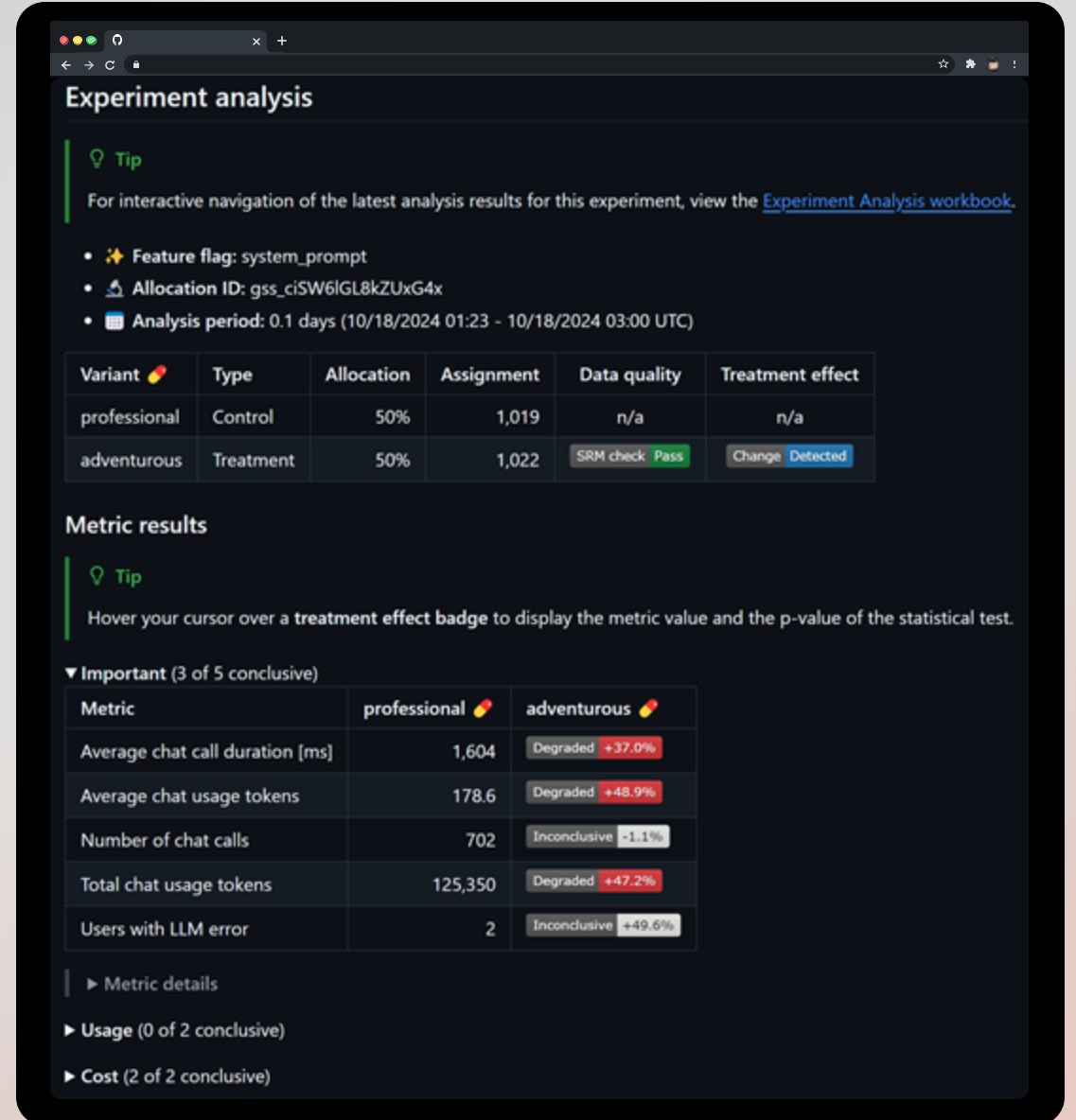
Streamline workflows with GitHub Actions for evaluation (in public preview) and A/B experimentation (in private preview)

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>



The screenshot shows the 'Experiment analysis' dashboard in a dark theme. It includes a tip about the 'Experiment Analysis workbook', a list of experiment details (Feature flag: system_prompt, Allocation ID: gss_ciSW6iGL8kZUxG4x, Analysis period: 0.1 days), a table of variants, and a 'Metric results' section with a table of key metrics.

Variant	Type	Allocation	Assignment	Data quality	Treatment effect
professional	Control	50%	1,019	n/a	n/a
adventurous	Treatment	50%	1,022	SRM check: Pass	Change Detected

Metric	professional	adventurous
Average chat call duration [ms]	1,604	Degraded +37.0%
Average chat usage tokens	178.6	Degraded +48.9%
Number of chat calls	702	Inconclusive -1.1%
Total chat usage tokens	125,350	Degraded +47.2%
Users with LLM error	2	Inconclusive +49.6%

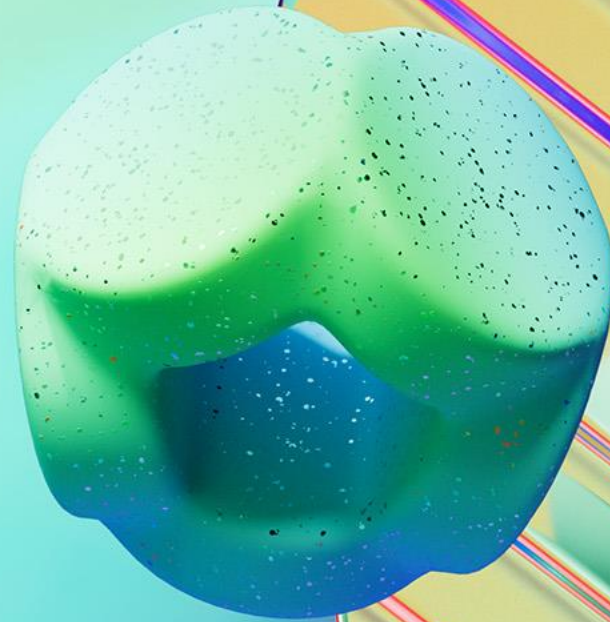
Azure AI evaluation and experimentation

Scale AI applications using CI/CD workflows

The screenshot shows a GitHub Actions workflow run for 'Evaluate #4' in the repository 'gartdan / contoso-writer-ignite'. The workflow is in a 'Success' state, manually triggered 2 days ago. The total duration is 3m 32s, and the billable time is 4m. There is 1 artifact. The workflow file 'evaluate.yml' is shown, triggered on 'workflow_dispatch'. A job named 'evaluate' completed successfully in 3m 21s. Below the job details, the 'evaluate summary' section displays 'Promptflow Evaluation Results' in a table.

	research_context	gpt_relevance	gpt_fluency	gpt_coherence	gpt_groundedness
0	Can you find the latest camping trends and what folks are doing in the winter?	1	5	5	5
1	Can you find the latest trends in hiking shoes?	5	5	5	5
2	Find information about the best snow camping spots in the world	5	5	5	5

Get started today!

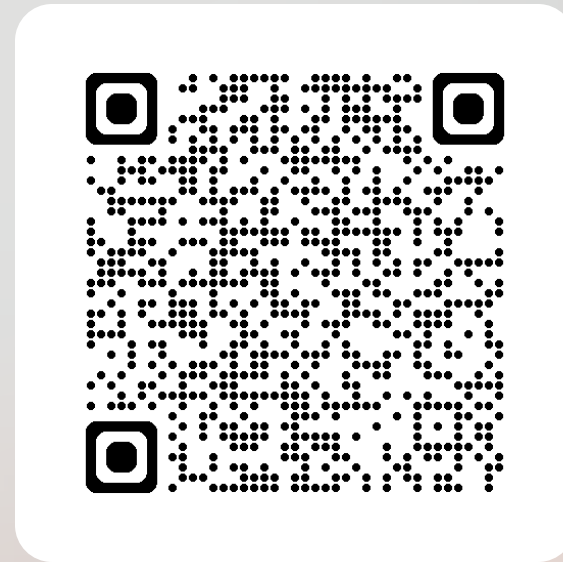


AI App Template Gallery



aka.ms/aiapps

GitHub Copilot for Azure



aka.ms/GetGitHubCopilotForAzure

How did we do?

Tell us your thoughts
about our sessions.



