Microsoft

Information Assistant Accelerator

Architecture document

Prepared for WW Public Sector
11/28/2023

Prepared by
WW Public Sector Solutions

Contents

Executive Summary

Overview

The Information Assistant industry accelerator showcases integration between Azure and OpenAI's large language models. It leverages Azure AI Search for data retrieval and ChatGPT-style Q&A interactions. Using the Retrieval Augmented Generation (RAG) design pattern with Azure Open AI's GPT models, it provides a natural language interaction to discover relevant responses to user queries. Azure AI Search simplifies data ingestion, transformation, indexing, and multilingual translation.

Information Assistant is provided to customers and to partners as a Microsoft Industry Accelerator at no charge to accelerate their "time to value" on Azure.

Challenge

Create an "assistant" that will enable users to "**Have a conversation with their private/domain specific data**" by leveraging the power of a large language model where the responses are:

Current – Responses can be based on the latest "up to date" information

Relevant – The responses must leverage their own *private/domain specific* data

Controlled – Provide the ability to adjust the response from precise to creative

Referenced – Respond using references/citations

Personalized – Respond with your role/persona in mind

Explainable – Providing details on the "thought process" that was used

Introduction

Purpose

This document provides an architectural overview of the accelerator with a focus on how it has been assembled and can be deployed.

The purpose of the architecture document is to provide an **architectural overview** of the industry sales accelerator. The document focuses on how the accelerator has been assembled and can be deployed. It aims to provide a comprehensive understanding of the internal structure, cross-cutting concepts, and fundamental decisions of the architecture.

Audience

This document is targeted primarily at Individuals working with Public Sector and Defense customers who would like to deploy the industry sales accelerator. Roles such as the ATU (Industry Execs, Account Technology Strategists), STU (Azure Specialists), CSU (Cloud Solution Architects), GBB (Global Black Belts) and NTO (National Technology Officers) and NSO (National Security Officers) will find this document particularly relevant.
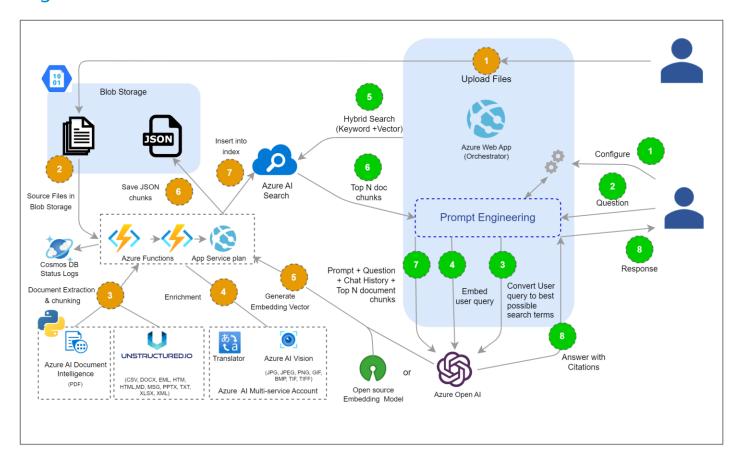
This document is intended for individuals working with the Public Sector and other industries who are interested in deploying the industry sales accelerator. The document is particularly relevant to roles such as Industry Advisors (IBB), Account Technology Strategists (ATU), Azure Specialists (STU), Technical Specialists (CSU), Global Black Belts (GBB), Cloud Solution Architect (CSA) and National Technology Officers (NTO). It provides information on the internal structure, cross-cutting concepts, and fundamental decisions of architecture.

Solution Architecture

The IA Accelerator contains several features, many of which have their own documentation.

- Examples of custom Retrieval Augmented Generation (RAG), Prompt Engineering, and Document Pre-Processing.
- Azure AI Search Integration to include text search of both text documents and images.

- Customization and Personalization to enable enhanced AI interaction.

High Level Architecture

1   Solution Deployment

1.1   Pre-Requisites

To find the latest pre-requisites for the Information Assistant Accelerator, view our [pre-requisites on GitHub](#).

1.2   Deployment

A detailed set of deployment instructions and scripts can be found on our GitHub Project at [PubSec-Info-Assistant/docs/deployment/deployment.md at main · microsoft/PubSec-Info-Assistant (github.com)](#)

Solution Estimator and Licensing

The proposed cost estimation is prepared based Sandbox environment; estimation may vary from customer to customer. [PubSec-Info-Assistant/docs/deployment/deployment.md at main · microsoft/PubSec-Info-Assistant (github.com)](#)

Appendix A: Azure Configuration

Executive summary

The following Azure data centers are used in your deployment:

Location	Number of resources
global	1
westeurope	21

The following components are used in your deployment:

Trend	Category	Component Type	Number
	Container	Media	1
	ETL (Extract, Transform, Load)	Log Analytic Workspace	1
	NoSQL Database	Cosmos DB	1
	Organizational Unit	Management Group	1
	Organizational Unit	Resource Group	1
	Organizational Unit	Subscription	6
	Security	Vault	1
	Serverless Compute	App Service plan	3
	Serverless Compute	Function	6
	Serverless Compute	Web Site	3

	Storage	CORS Rule	1
	Storage	Storage account	2
	Virtual Network	Network Interface IP Configuration	9

App Services

infoasst-enrichmentweb-xxxxx

⚙ Settings

Web Site State	Running
Authentication	Off
Always On	True
Load Balancing	LeastRequests
Nb of workers	1
Enabled	True
Location	West Europe
Outbound Ip Addresses	*******
Repository Site Name	infoasst-enrichmentweb-xxxxx
Scm Site Stopped	False
Use 32 Bits Worker Process	False

App Service Plan

Service Plan Name	infoasst-enrichmentasp-xxxxx
Operating System	Linux
Geographical Location	West Europe
Maximum Number Of Workers	30
Number of Sites	0
Status	Ready
Sku Name	P1v3
Sku Tier	PremiumV3
Sku Family	Pv3
Sku Capacity	1

Sku Size	P1v3

Capacity	Maximum capacity: '5', Default capacity: '1', Minimum capacity: '1'
Rules	Array with two scaling rules
Rule 1: Increase	If CPU percentage > 60% in the last 5 minutes, increase count by 1, cooldown: 5 minutes
Rule 2: Decrease	If CPU percentage < 20% in the last 10 minutes, decrease count by 1, cooldown: 15 minutes
Metric	CpuPercentage
Time Grain:	PT1M
Statistic	Average
Time Window	'PT5M' for Rule 1, 'PT10M' for Rule 2
Time Aggregation	Average
Operator	'GreaterThan' for Rule 1, 'LessThan' for Rule 2
Threshold	60 for Rule 1, 20 for Rule 2
Direction	'Increase' for Rule 1, 'Decrease' for Rule 2
Cooldown	'PT5M' for Rule 1, 'PT15M' for Rule 2

Quotas

FILE SYSTEM STORAGE

Review the file system usage across your App Service plans and apps.

Threshold	350.0 GB
Usage	2.6 GB

Site diagnostics

The site diagnostics options have been configured as follows.

Request Tracing Enabled	True
Detailed logging enabled	True
Remote Debugging Enabled	False

App Settings

The following application settings have been used for the Web Site.

Key	Value

EMBEDDINGS_QUEUE	Hidden
LOG_LEVEL	Hidden
DEQUEUE_MESSAGE_BATCH_SIZE	Hidden
AZURE_BLOB_STORAGE_ACCOUNT	Hidden
AZURE_BLOB_STORAGE_CONTAINER	Hidden
AZURE_BLOB_STORAGE_UPLOAD_CONTAINER	Hidden
AZURE_BLOB_STORAGE_ENDPOINT	Hidden
COSMOSDB_URL	Hidden
COSMOSDB_LOG_DATABASE_NAME	Hidden
COSMOSDB_LOG_CONTAINER_NAME	Hidden
COSMOSDB_TAGS_DATABASE_NAME	Hidden
COSMOSDB_TAGS_CONTAINER_NAME	Hidden
MAX_EMBEDDING_REQUEUE_COUNT	Hidden
EMBEDDING_REQUEUE_BACKOFF	Hidden
AZURE_OPENAI_SERVICE	Hidden
AZURE_OPENAI_EMBEDDING_DEPLOYMENT_NAME	Hidden
AZURE_SEARCH_INDEX	Hidden
AZURE_SEARCH_SERVICE	Hidden
TARGET_EMBEDDINGS_MODEL	Hidden
EMBEDDING_VECTOR_SIZE	Hidden
AZURE_SEARCH_SERVICE_ENDPOINT	Hidden
WEBSITES_CONTAINER_START_TIME_LIMIT	Hidden
SCM_DO_BUILD_DURING_DEPLOYMENT	Hidden
ENABLE_ORYX_BUILD	Hidden
APPLICATIONINSIGHTS_CONNECTION_STRING	Hidden
AZURE_KEY_VAULT_ENDPOINT	Hidden
WEBSITE_HTTPLOGGING_RETENTION_DAYS	Hidden
COSMOSDB_KEY	Hidden
AZURE_SEARCH_SERVICE_KEY	Hidden
BLOB_CONNECTION_STRING	Hidden
AZURE_OPENAI_SERVICE_KEY	Hidden
AZURE_BLOB_STORAGE_KEY	Hidden

📄 Default documents

The following default documents are defined.

- Default.htm
- Default.html
- Default.asp
- index.htm
- index.html

- iisstart.htm
- default.aspx
- index.php
- hostingstart.html

🌐 Host Names

The following Host Names are defined.

- infoasst-enrichmentweb-xxxxx.azurewebsites.net
- infoasst-enrichmentweb-xxxxx.scm.azurewebsites.net

🏷️ Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	

infoasst-web-xxxxx

⚙️ Settings

Web Site State	Running
Authentication	On
Always On	True
Load Balancing	LeastRequests
Nb of workers	1
Enabled	True
Location	West Europe
Outbound Ip Addresses	*******
Repository Site Name	infoasst-web-xxxxx
Scm Site Stopped	False
Use 32 Bits Worker Process	False

App Service Plan

Service Plan Name	infoasst-asp-xxxxx
Operating System	Linux
Geographical Location	West Europe
Maximum Number Of Workers	10
Number of Sites	0
Status	Ready

Sku Name	S1
Sku Tier	Standard
Sku Family	S
Sku Capacity	3
Sku Size	S1

None

Quotas

Review the file system usage across your App Service plans and apps.

Threshold	350.0 GB
Usage	98.7 MB

Site diagnostics

The site diagnostics options have been configured as follows.

Request Tracing Enabled	True
Detailed logging enabled	True
Remote Debugging Enabled	False

App Settings

The following application settings have been used for the Web Site.

Key	Value
AZURE_BLOB_STORAGE_ACCOUNT	Hidden
AZURE_BLOB_STORAGE_ENDPOINT	Hidden
AZURE_BLOB_STORAGE_CONTAINER	Hidden
AZURE_BLOB_STORAGE_UPLOAD_CONTAINER	Hidden
AZURE_OPENAI_SERVICE	Hidden
AZURE_OPENAI_RESOURCE_GROUP	Hidden
AZURE_SEARCH_INDEX	Hidden
AZURE_SEARCH_SERVICE	Hidden
AZURE_SEARCH_SERVICE_ENDPOINT	Hidden
AZURE_OPENAI_CHATGPT_DEPLOYMENT	Hidden
AZURE_OPENAI_CHATGPT_MODEL_NAME	Hidden

AZURE_OPENAI_CHATGPT_MODEL_VERSION	Hidden
USE_AZURE_OPENAI_EMBEDDINGS	Hidden
EMBEDDING_DEPLOYMENT_NAME	Hidden
AZURE_OPENAI_EMBEDDINGS_MODEL_NAME	Hidden
AZURE_OPENAI_EMBEDDINGS_MODEL_VERSION	Hidden
APPINSIGHTS_INSTRUMENTATIONKEY	Hidden
COSMOSDB_URL	Hidden
COSMOSDB_LOG_DATABASE_NAME	Hidden
COSMOSDB_LOG_CONTAINER_NAME	Hidden
COSMOSDB_TAGS_DATABASE_NAME	Hidden
COSMOSDB_TAGS_CONTAINER_NAME	Hidden
QUERY_TERM_LANGUAGE	Hidden
AZURE_CLIENT_ID	Hidden
AZURE_TENANT_ID	Hidden
AZURE_SUBSCRIPTION_ID	Hidden
IS_GOV_CLOUD_DEPLOYMENT	Hidden
CHAT_WARNING_BANNER_TEXT	Hidden
TARGET_EMBEDDINGS_MODEL	Hidden
ENRICHMENT_APPSERVICE_NAME	Hidden
APPLICATION_TITLE	Hidden
SCM_DO_BUILD_DURING_DEPLOYMENT	Hidden
ENABLE_ORYX_BUILD	Hidden
APPLICATIONINSIGHTS_CONNECTION_STRING	Hidden
AZURE_KEY_VAULT_ENDPOINT	Hidden
WEBSITE_HTTPLOGGING_RETENTION_DAYS	Hidden
AZURE_SEARCH_SERVICE_KEY	Hidden
AZURE_OPENAI_SERVICE_KEY	Hidden
AZURE_BLOB_STORAGE_KEY	Hidden
COSMOSDB_KEY	Hidden
AZURE_CLIENT_SECRET	Hidden

🖹 Default documents

The following default documents are defined.

- Default.htm
- Default.html
- Default.asp
- index.htm
- index.html
- iisstart.htm
- default.aspx

- index.php
- hostingstart.html

🌐 Host Names

The following Host Names are defined.

- infoasst-web-xxxxx.azurewebsites.net
- infoasst-web-xxxxx.scm.azurewebsites.net

🏷 Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	
azd-service-name	backend

infoasst-func-xxxxx

⚙ Settings

Web Site State	Running
Authentication	Off
Always On	True
Load Balancing	LeastRequests
Nb of workers	1
Enabled	True
Location	West Europe
Outbound Ip Addresses	*******
Repository Site Name	infoasst-func-xxxxx
Scm Site Stopped	False
Use 32 Bits Worker Process	False

App Service Plan

Service Plan Name	infoasst-func-asp-xxxxx
Operating System	Linux
Geographical Location	West Europe
Maximum Number Of Workers	10
Number of Sites	0
Status	Ready
Sku Name	S2

Sku Tier	Standard
Sku Family	S
Sku Capacity	2
Sku Size	S2

Capacity	Maximum capacity: '10', Default capacity: '2', Minimum capacity: '2'
Rules	Array with two scaling rules
Rule 1: Increase	If CPU percentage > 60% in the last 5 minutes, increase count by 2, cooldown: 5 minutes
Rule 2: Decrease	If CPU percentage < 40% in the last 5 minutes, decrease count by 2, cooldown: 2 minutes
Metric	CpuPercentage
Time Grain:	PT1M
Statistic	Average
Time Window	'PT5M' for Rule 1, 'PT5M' for Rule 2
Time Aggregation	Average
Operator	'GreaterThan' for Rule 1, 'LessThan' for Rule 2
Threshold	60 for Rule 1, 40 for Rule 2
Direction	'Increase' for Rule 1, 'Decrease' for Rule 2
Cooldown	'PT5M' for Rule 1, 'PT2M' for Rule 2

Quotas

Review the file system usage across your App Service plans and apps.

Threshold	350.0 GB
Usage	69.0 KB

Site diagnostics

The site diagnostics options have been configured as follows.

Request Tracing Enabled	False
Detailed logging enabled	False
Remote Debugging Enabled	False

⚙ App Settings

The following application settings have been used for the Web Site.

Key	Value
AzureWebJobsStorage	Hidden
WEBSITE_CONTENTAZUREFILECONNECTIONSTRING	Hidden
WEBSITE_CONTENTSHARE	Hidden
FUNCTIONS_WORKER_RUNTIME	Hidden
FUNCTIONS_EXTENSION_VERSION	Hidden
WEBSITE_NODE_DEFAULT_VERSION	Hidden
APPLICATIONINSIGHTS_CONNECTION_STRING	Hidden
APPINSIGHTS_INSTRUMENTATIONKEY	Hidden
BLOB_STORAGE_ACCOUNT	Hidden
BLOB_STORAGE_ACCOUNT_ENDPOINT	Hidden
BLOB_STORAGE_ACCOUNT_UPLOAD_CONTAINER_NAME	Hidden
BLOB_STORAGE_ACCOUNT_OUTPUT_CONTAINER_NAME	Hidden
BLOB_STORAGE_ACCOUNT_LOG_CONTAINER_NAME	Hidden
CHUNK_TARGET_SIZE	Hidden
TARGET_PAGES	Hidden
FR_API_VERSION	Hidden
AZURE_FORM_RECOGNIZER_ENDPOINT	Hidden
COSMOSDB_URL	Hidden
COSMOSDB_LOG_DATABASE_NAME	Hidden
COSMOSDB_LOG_CONTAINER_NAME	Hidden
COSMOSDB_TAGS_DATABASE_NAME	Hidden
COSMOSDB_TAGS_CONTAINER_NAME	Hidden
PDF_SUBMIT_QUEUE	Hidden
PDF_POLLING_QUEUE	Hidden
NON_PDF_SUBMIT_QUEUE	Hidden
MEDIA_SUBMIT_QUEUE	Hidden
TEXT_ENRICHMENT_QUEUE	Hidden
IMAGE_ENRICHMENT_QUEUE	Hidden
MAX_SECONDS_HIDE_ON_UPLOAD	Hidden
MAX_SUBMIT_REQUEUE_COUNT	Hidden
POLL_QUEUE_SUBMIT_BACKOFF	Hidden
PDF_SUBMIT_QUEUE_BACKOFF	Hidden
MAX_POLLING_REQUEUE_COUNT	Hidden
SUBMIT_REQUEUE_HIDE_SECONDS	Hidden
POLLING_BACKOFF	Hidden
MAX_READ_ATTEMPTS	Hidden

ENRICHMENT_ENDPOINT	Hidden
ENRICHMENT_NAME	Hidden
ENRICHMENT_LOCATION	Hidden
TARGET_TRANSLATION_LANGUAGE	Hidden
MAX_ENRICHMENT_REQUEUE_COUNT	Hidden
ENRICHMENT_BACKOFF	Hidden
ENABLE_DEV_CODE	Hidden
EMBEDDINGS_QUEUE	Hidden
AZURE_SEARCH_SERVICE_ENDPOINT	Hidden
AZURE_SEARCH_INDEX	Hidden
AZURE_BLOB_STORAGE_KEY	Hidden
AZURE_FORM_RECOGNIZER_KEY	Hidden
BLOB_CONNECTION_STRING	Hidden
COSMOSDB_KEY	Hidden
ENRICHMENT_KEY	Hidden
AZURE_SEARCH_SERVICE_KEY	Hidden

📄 Default documents

The following default documents are defined.

- Default.htm
- Default.html
- Default.asp
- index.htm
- index.html
- iisstart.htm
- default.aspx
- index.php

🌐 Host Names

The following Host Names are defined.

- infoasst-func-xxxxx.azurewebsites.net
- infoasst-func-xxxxx.scm.azurewebsites.net

🏷️ Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	

Key Vaults

infoasst-kv-xxxxx

⚙ Settings

Name	infoasst-kv-xxxxx
Vault Uri	https://infoasst-kv-xxxxx.vault.azure.net/
Sku (Pricing tier)	Standard
Directory ID	*******
Soft-delete	Enabled
Purge protection	Disabled
Enabled for Deployment	False
Enabled for Template Deployment	True
Location	westeurope

Storage Account

infoasststorexxxxx

⚙ Settings

The storage infoasststorexxxxx has the following settings:

Resource group	infoasst-demo_xxxxx
Status	Available
Location	westeurope
Replication	Standard Locally Redundant Storage
Account kind	StorageV2
Created date	2023-11-13 22:53:16.00 PM
SKU	Standard_LRS
Provisioning state	Succeeded
Blob EndPoint	https://infoasststorexxxxx.blob.core.windows.net/
File EndPoint	https://infoasststorexxxxx.file.core.windows.net/
Queue EndPoint	https://infoasststorexxxxx.queue.core.windows.net/
Table EndPoint	https://infoasststorexxxxx.table.core.windows.net/
Geo Primary Region	westeurope

Blob Service Cors Rules

Allowed Origins	Allowed methods	Allowed headers	Exposed headers	Max age
• *	• GET • PUT • OPTIONS • POST • PATCH • HEAD	• *	• *	86400

🏷 Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	

infoasststoremediaxxxxx

⚙ Settings

The storage infoasststoremediaxxxxx has the following settings:

Resource group	infoasst-demo_xxxxx
Status	Available
Location	westeurope
Replication	Standard Locally Redundant Storage
Account kind	StorageV2
Created date	2023-11-10 19:43:13.00 PM
SKU	Standard_LRS
Provisioning state	Succeeded
Blob EndPoint	https://infoasststoremediaxxxxx.blob.core.windows.net/
File EndPoint	https://infoasststoremediaxxxxx.file.core.windows.net/
Queue EndPoint	https://infoasststoremediaxxxxx.queue.core.windows.net/
Table EndPoint	https://infoasststoremediaxxxxx.table.core.windows.net/
Geo Primary Region	westeurope

🏷 Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	

Media Services

infoasstmediasvcxxxxx

⚙ Settings

Id	*******
Name	infoasstmediasvcxxxxx
Type	Microsoft.Media/mediaservices
Location	West Europe
Public Network Access	Disabled
Storage Authentication	System

🏷 Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	

Cosmos DB

infoasst-cosmos-xxxxx

⚙ Settings

Status	Online
Location	West Europe
Endpoint/URI	Core SQL URI: https://infoasst-cosmos-xxxxx.documents.azure.com:443/
Read Locations	• West Europe
Write Locations	• West Europe
Consistency Policy	• Default Consistency Level : Session • Max Interval In Seconds : 5 • Max Staleness Prefix : 100

🏷 Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	

Log Analytics Workspaces

infoasst-la-xxxxx

⚙ Settings

Resource group	infoasst-demo_xxxxx
Provisioning State	Succeeded
Location	westeurope
Workspace ID	*******

🏷 Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	

Resource Groups

infoasst-demo_xxxxx

📦 Resources

Name	Location	Type
infoasst-fr-xxxxx	westeurope	Microsoft.CognitiveServices/accounts
infoasst-enrichmentweb-xxxxx	westeurope	Microsoft.Web/sites
infoasst-func-asp-xxxxx-Autoscale	westeurope	Microsoft.Insights/autoscalesettings
infoasst-web-xxxxx	westeurope	Microsoft.Web/sites
infoasst-asp-xxxxx	westeurope	Microsoft.Web/serverFarms
infoasst-cosmos-xxxxx	westeurope	Microsoft.DocumentDB/databaseAccounts
infoasst-func-xxxxx	westeurope	Microsoft.Web/sites
infoasststoremediaxxxxx	westeurope	Microsoft.Storage/storageAccounts
infoasstmediasvcxxxxx	westeurope	Microsoft.Media/mediaservices
infoasst-search-xxxxx	westeurope	Microsoft.Search/searchServices
infoasst-ai-xxxxx	westeurope	Microsoft.Insights/components
infoasst-func-asp-xxxxx	westeurope	Microsoft.Web/serverFarms
infoasststorexxxxx	westeurope	Microsoft.Storage/storageAccounts
Failure Anomalies - infoasst-ai-xxxxx	global	microsoft.alertsmanagement/smartDetectorAlertRules
infoasst-lw-xxxxx	westeurope	Microsoft.Insights/workbooktemplates
infoasstvi-xxxxx	westeurope	Microsoft.VideoIndexer/accounts
infoasst-la-xxxxx	westeurope	Microsoft.OperationalInsights/workspaces
infoasst-enrichmentasp-xxxxx	westeurope	Microsoft.Web/serverFarms
infoasst-kv-xxxxx	westeurope	Microsoft.KeyVault/vaults
infoasst-enrichmentasp-xxxxx	westeurope	Microsoft.Insights/autoscalesettings
infoasst-enrichment-cog-xxxxx	westeurope	Microsoft.CognitiveServices/accounts
infoasstmediasvcxxxxx/default	westeurope	Microsoft.Media/mediaservices/streamingEndpoints

🏷️ Tags

Tag Key	Tag Value
ProjectName	Information Assistant
BuildNumber	

Microsoft

Appendix B: Resources

Topics	Resource
Learning	Microsoft AI School Get started with Azure OpenAI Service - Training | Microsoft Learn Quickstart - Get started using GPT-35-Turbo and GPT-4 with Azure OpenAI Service - Azure OpenAI Service | Microsoft Learn Azure OpenAI Service – Advanced Language Models | Microsoft Azure Hybrid search - Azure AI Search | Microsoft Learn Explore computer vision in Microsoft Azure - Learn | Microsoft Docs https://docs.microsoft.com/en-us/learn/paths/process-classify-images-with-azure-cognitive-vision-services/Evaluate text with Azure Cognitive Language Services
Certifications	Microsoft Certified: Azure AI Fundamentals Microsoft Certified: Azure AI Engineer Associate Microsoft Certified: Azure Data Scientist Associate
Reference Architecture	Azure reference Architectures
Responsible AI resources	Microsoft's Responsible AI principles Our approach to AI Azure OpenAI Transparency Note and Code of Conduct