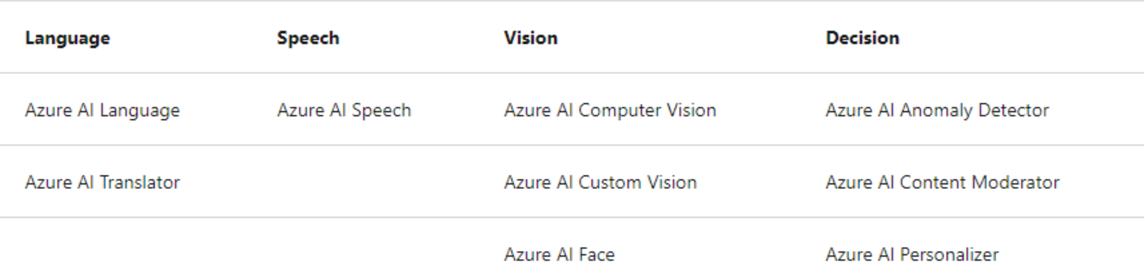
Create and consume Azure AI services

Saturday, June 22, 2024 3:50 PM

# Introduction: Azure AI services

Azure AI services are cloud-based services that encapsulate AI capabilities. Think of AI services as a set of individual services that you can use as building blocks to compose sophisticated applications

Ai services includes a wide range of invidividual services across multiple categories, as shown in the following table:



Language	Speech	Vision	Decision
Azure AI Language	Azure AI Speech	Azure AI Computer Vision	Azure AI Anomaly Detector
Azure AI Translator		Azure AI Custom Vision	Azure AI Content Moderator
		Azure AI Face	Azure AI Personalizer

Some common AI scenarios:

- Azure AI Document Intelligence - An optical character recognition (OCR) solution that can extract semantic meaning from forms, such as invoices, receipts, and others.
- Azure AI Immersive Reader - A reading solution that supports people of all ages and abilities.
- Azure Cognitive Search - A cloud-scale search solution that uses AI services to extract insights from data and documents.
- Azure OpenAI - An Azure Cognitive Service that provides access to the capabilities of OpenAI GPT-4.

# Provision an Azure AI services resource

Azure AI services include a wide range of AI capabilities for this you need to create appropriate resources in an Azure subscription to define an endpoint

## Provisioning options:

You can choose between the following provisioning options:

### Multi-service resource

You can create a single resource that enables you to use the Azure AI Language, Azure AI Vision, Azure AI Speech, and other services. This approach enables you to manage a single set of access credentials to consume multiple services at a single endpoint and a single point of billing

### Single-service resource

You provision each AI service individually , for example this approach enables you to create AI Language and AI Vision resources in your Azure subscription. Which means you would have to use separate endpoints for each service. It also enables you to manage billing separately for each service. Single-Service resources generally offer a free tier (with usage restrictions), making them a good choice to try out a service before using it in a production application

### Training and prediction resources

Some AI services require separate resources for model training and prediction. This enables you to manage billing for training custom models separately from model consumption. And in most cases enables you to use a dedicated service-specific resource to train a model, but a generic AI services resource to make the model available to applications for inferencing

# Identify endpoints and keys

When you provision an Azure AI services service resource in your Azure subscription, you are defining an endpoint through which the service can be consumed by an application.

To consume the service through the endpoint, applications require the following information:

- The endpoint URI. This is the HTTP address at which the REST interface for the service can be accessed. Most AI services software development kits (SDKs) use the endpoint URI to initiate a connection to the endpoint.
- A subscription key. Access to the endpoint is restricted based on a subscription key. Client applications must provide a valid key to consume the service. When you provision an AI services resource, two keys are created - applications can use either key. You can also regenerate the keys as required to control access to your resource.
- The resource location. When you provision a resource in Azure, you generally assign it to a location, which determines the Azure data center in which the resource is defined. While most SDKs use the endpoint URI to connect to the service, some require the location.

From <https://learn.microsoft.com/en-us/training/modules/create-manage-ai-services/3-identify-keys-endpoints>

# Use a REST API

Azure AI services provide REST application programming interfaces (APIs) that client applications can use to consume services. In most cases, service functions can be called by submitting data in JSON format over an HTTP request, which may be a POST, PUT, or GET request depending on the specific function being called. The results of the function are returned to the client as an HTTP response, often with JSON contents that encapsulate the output data from the function

From <https://learn.microsoft.com/en-us/training/modules/create-manage-ai-services/4-use-rest>

# Use SDKS

Software development kits (SDKs) for common programming languages abstract the REST interfaces for most AI services. SDK availability varies by individual AI services, but for most services there's an SDK for languages such as:

- Microsoft C# (.NET Core)
- Python
- JavaScript (Node.js)
- Go
- Java

Each SDK includes packages that you can install in order to use service-specific libraries in your code, and online documentation to help you determine the appropriate classes, methods, and parameters used to work with the service.

From <https://learn.microsoft.com/en-us/training/modules/create-manage-ai-services/5-use-sdk>