
PLURALSIGHT

Intro to AI/ML in Azure

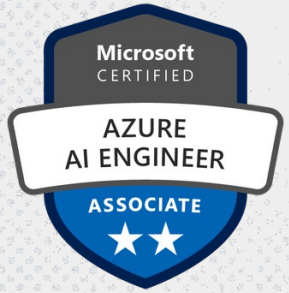
Welcome!

Tarek Atwan
Instructor, Pluralsight

Objectives

At the end of this course, you will be able to:

- Have a good understanding of Azure **Machine Learning Workspace** and **Azure AI Services**
- Be ready to continue your journey and explore taking **DP-100** and **AI-102** Certification



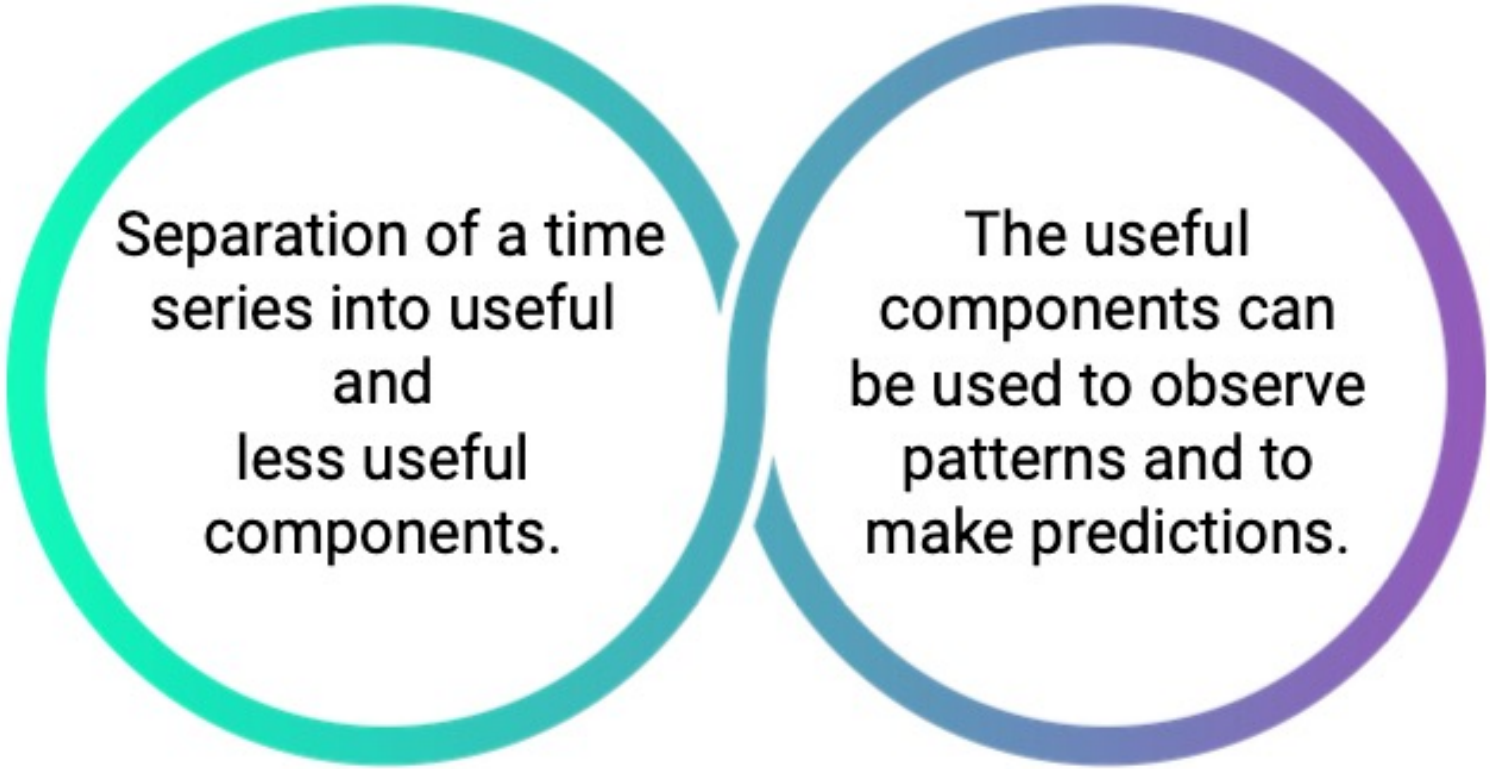
Microsoft Certified: Azure AI Engineer Associate

Microsoft Certified: Azure Data Scientist Associate

Time Series Automated ML


When analyzing time series, finding seasonal patterns is just one part of the job.

Time Series Decomposition



Time Series Decomposition

Components

01 **Level:** What is the average value of the series?

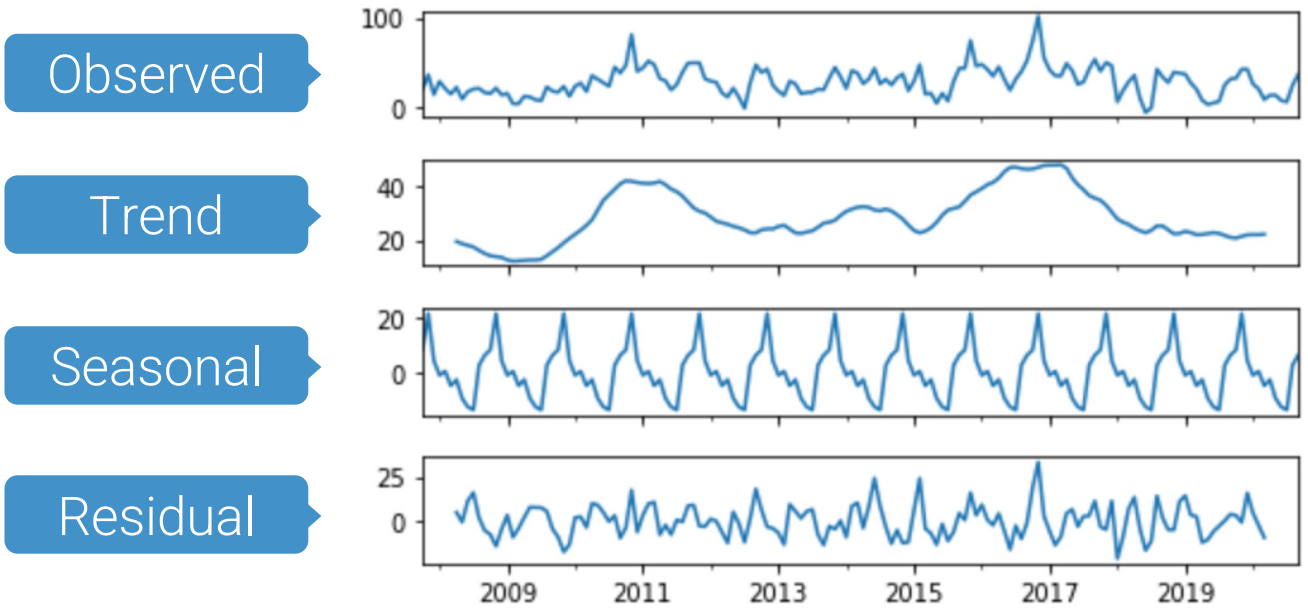
02 **Trend:** Is there an overall direction of movement?

03 **Periodicity:** Do patterns occur in cycles?

04 **Residual:** How much noise exists in the data?

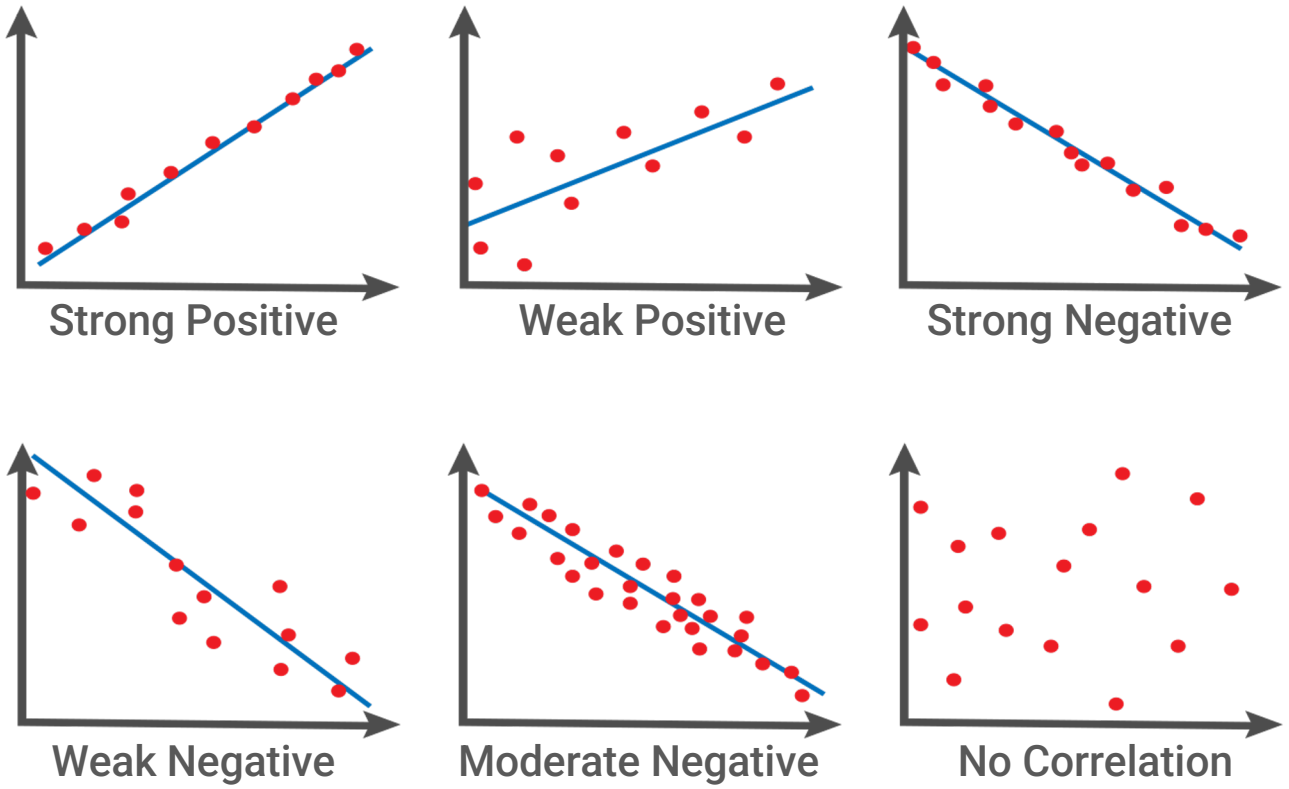
Identifying Patterned Relationships and Correlation

Another important task is identifying any relationships between time series patterns and determining if these relationships are predictable.

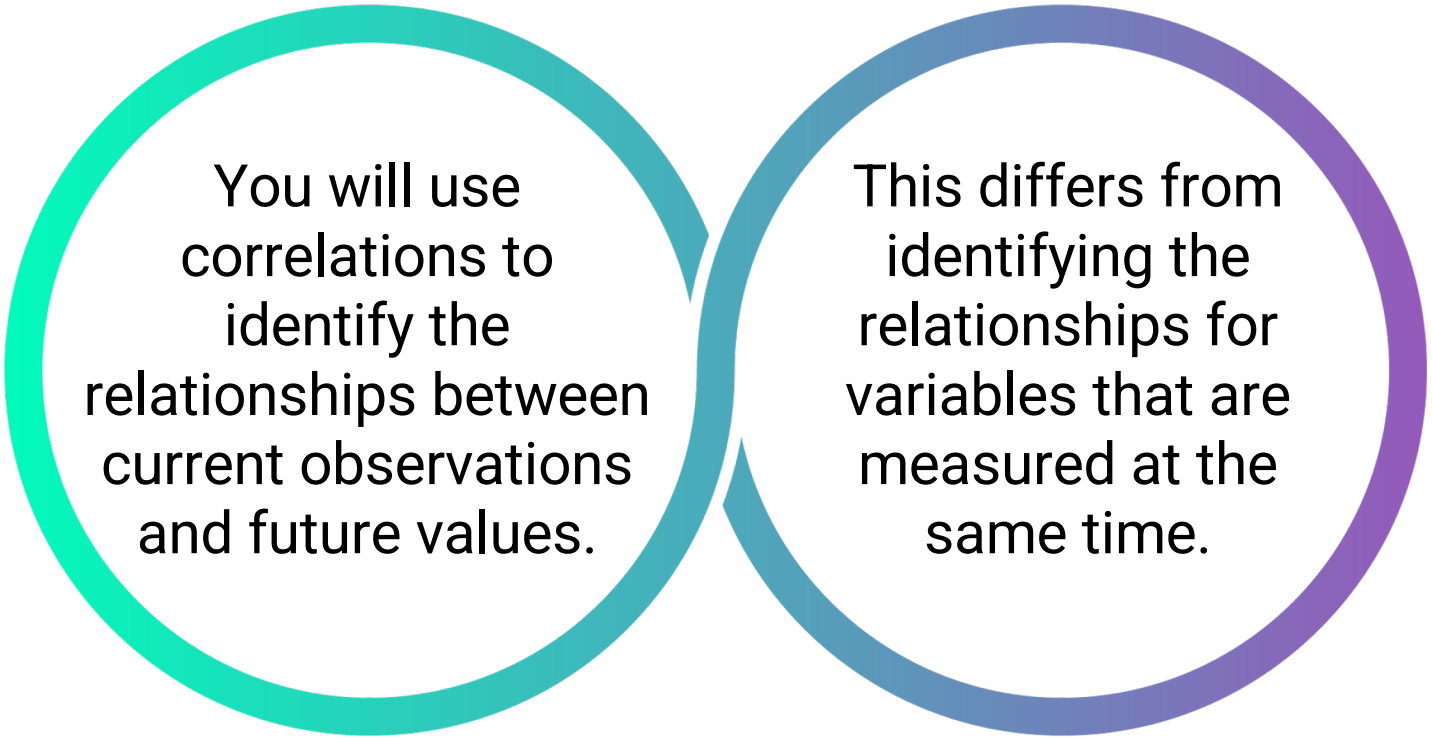


In statistics, a **correlation** defines the relationship between two or more variables.

Comparison of Correlation Relationships



In time Series



Introduction to Time Series Forecasting

The financial world makes great use of time series forecasting because of its origin in mathematics and statistics.

People sometimes refer to time series forecasting as a statistical tool, but there's a lot of overlap between statistical tools and machine learning models. Both can solve similar problems.

As you learned previously, time series analysis involves analyzing time series data to identify meaningful patterns in the data.

Time series forecasting involves using a model that is based on historical data to predict future values in the time series.

Time series forecasting can prove difficult.

External factors, such as holidays, breaking news, and special events, can impact the usual behavior of the patterns.

Additionally, it can be challenging to select the best statistical technique for analysis.

Automating Time Series Forecasting

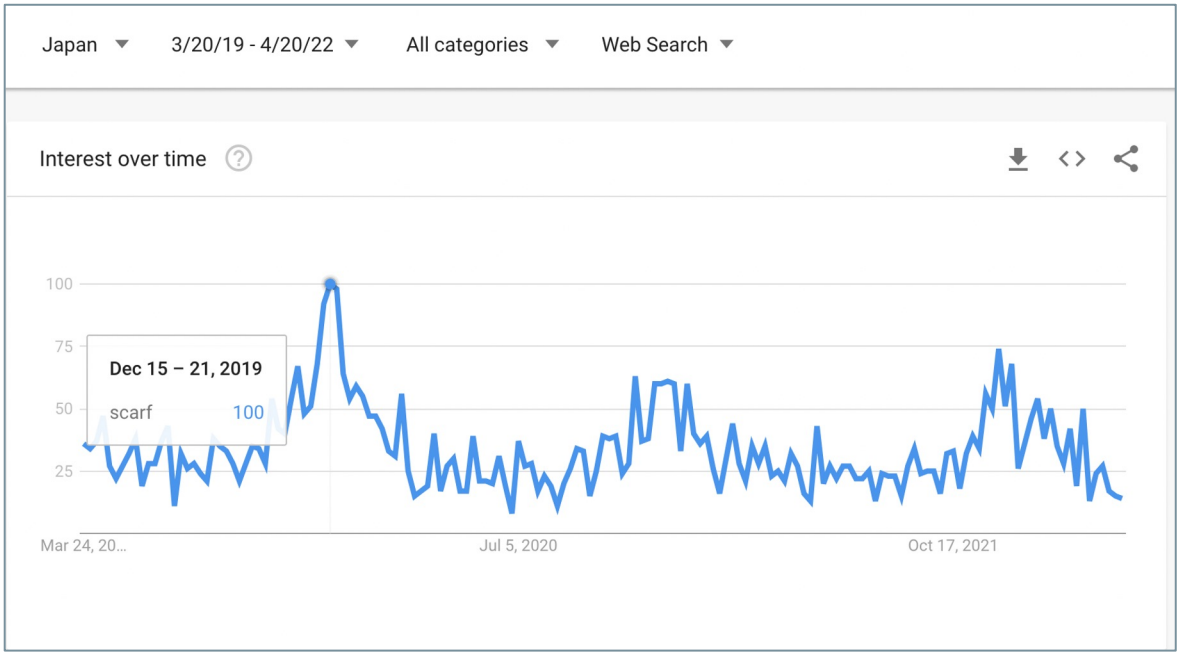
We've been asked to find the best season to sell scarves in Japan and to forecast the demand for scarves for one year.

We don't know anything about the scarf market in Japan, so we've obtained some data from Google Trends to figure out the optimal selling season.

Automating Time Series Forecasting

Using our time analysis skills, we identified that people in Japan have more interest in scarves in the winter months because of the weather.

So, it might be optimal to start marketing campaigns in August and start selling scarves by October.



External factors,
like weather or holidays,
impact a time series and
sometimes **introduce noise**.

Noise is information in the data that is essentially meaningless and distracts from the overall trend, or signal, of the data.

Noise acts as another independent variable in our models, but we can use some statistical techniques to manually reduce it.

Correlations can be helpful,
but they don’t provide enough
information to infer the relationship
between two variables.

GenAI

What is Generative AI?

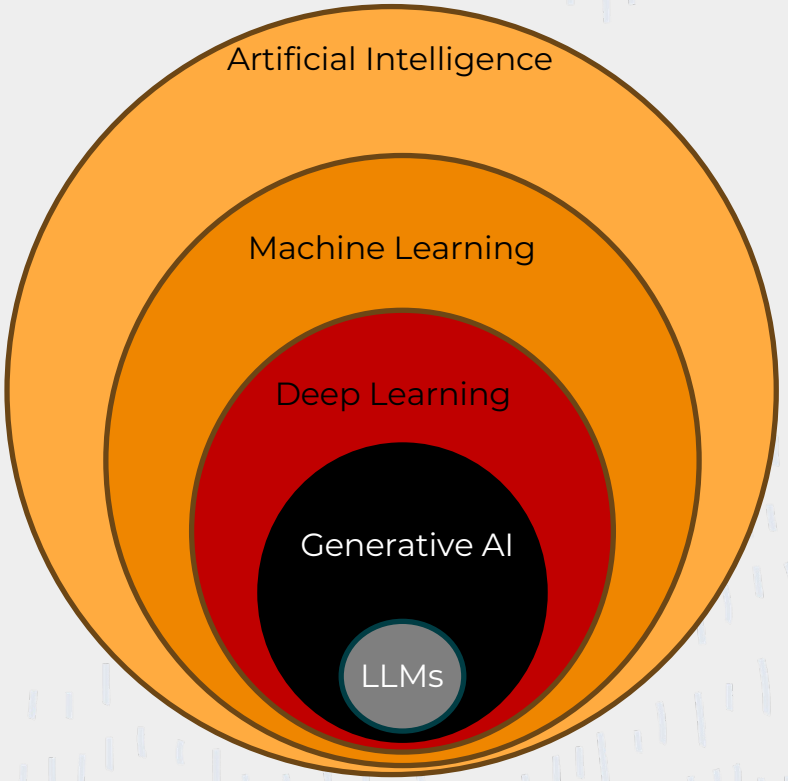
AI refers to the broad concept of machines or computers performing tasks that typically require human intelligence. This includes reasoning, learning, problem-solving, perception, language understanding, etc.

ML is a subset of AI focused on the idea that machines can learn from data, identify patterns, and make decisions with minimal human intervention

DL is a subset of ML that uses neural networks with many layers (deep networks) to model complex patterns in data.

Generative AI refers to a class of AI, often realized through DL, that focuses on generating new content or data that is similar to but distinct from the training data.

LLMs are a type of deep learning model designed to understand, generate, and interact with human language at a large scale. They are trained on vast amounts of text data.

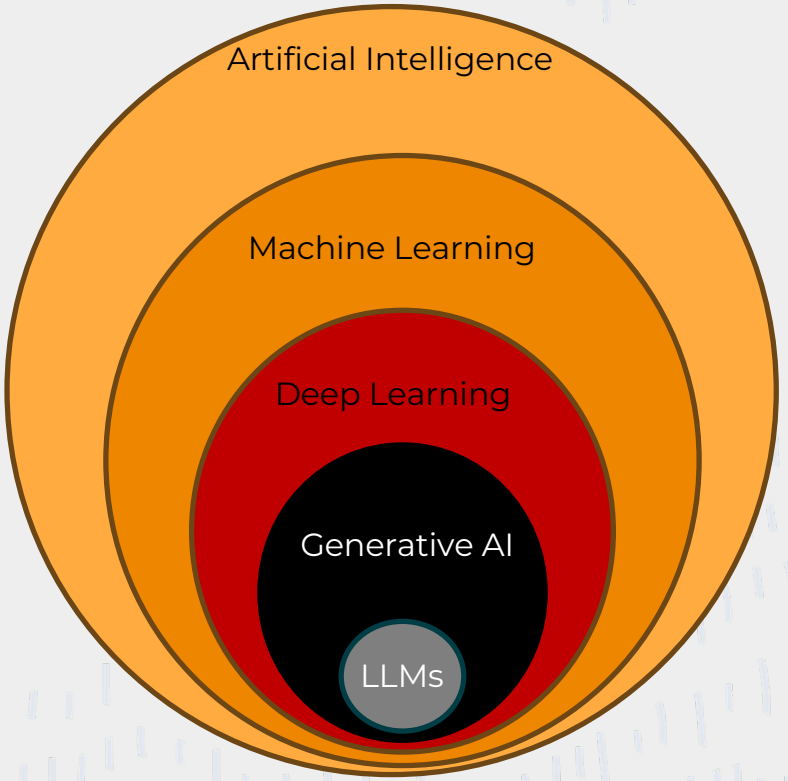


What is Generative AI?

Generative AI refers to a subset of artificial intelligence where the primary goal is to create or generate new data that is similar but not identical to the training data. It's about models that can learn from existing data to **generate** new, **unseen** data or patterns that maintain a **statistical resemblance** to the original dataset

These models are capable of understanding and replicating complex data distributions, allowing them to **produce** highly realistic and diverse outputs.

It evolved from ML/DL and is essential in fields like **content creation** and **data augmentation**. It includes models like **Generative Adversarial Networks** (GANs) and **Large Language Models** (LLMs)



Discriminative AI vs Generative AI

Discriminative AI

- Discriminative models learn the **conditional probability distribution P(Y|X)** They focus on understanding the **boundary** between different classes in the data, essentially distinguishing between different types of data inputs.

Generative AI

- Generative models are designed to learn the **joint probability distribution P(X,Y)** of inputs X and outputs Y Their goal is to **understand and replicate** the way data is generated, enabling them to produce new data instances that are similar to the training data.

Discriminative AI vs Generative AI (Examples)

Discriminative AI

1. **Convolutional Neural Networks (CNNs)**: Used for image classification.
2. **Recurrent Neural Networks (RNNs)**: Common in speech recognition and natural language processing.
3. **Support Vector Machines (SVMs), Logistic Regression, etc.**: Traditional ML algorithms for classification tasks.

Generative AI

1. **Generative Adversarial Networks (GANs)**: Used for generating realistic images, artworks, etc.
2. **Variational Autoencoders (VAEs)**: Often used in image generation and denoising.
3. **Language Models like GPT (Generative Pre-trained Transformer)**: Used for generating coherent and contextually relevant text.

Discriminative AI vs Generative AI (Objective)

Discriminative AI

A discriminative AI and its algorithms can be used to:

- Differentiate
- Classify
- Identify Patterns
- And Draw Conclusions
- Example: Email spam filters
- They are best applied to classification tasks.

Generative AI

A generative AI can generate new content/output as:

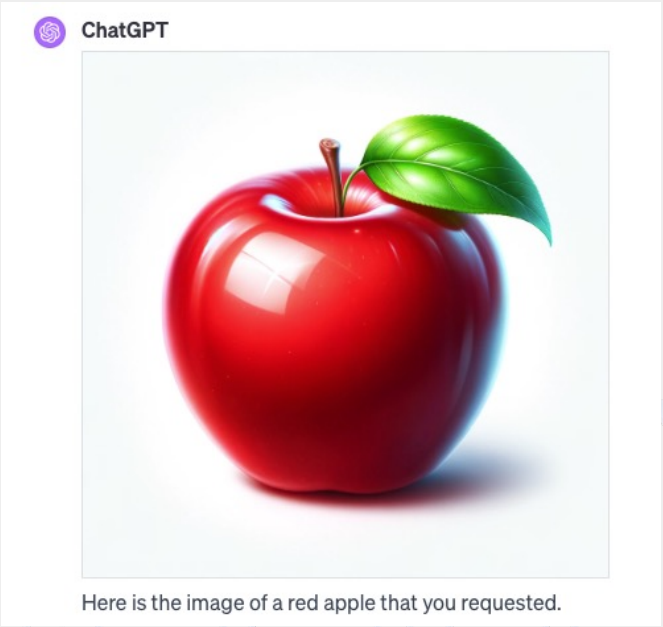
- Text
- Images
- Audio
- Video
- Code
- And new data

Discriminative AI vs Generative AI

Discriminative AI

Is the image an Orange or an Apple?

Generative AI



I want an image of a Red Apple.

Generative AI for Automation

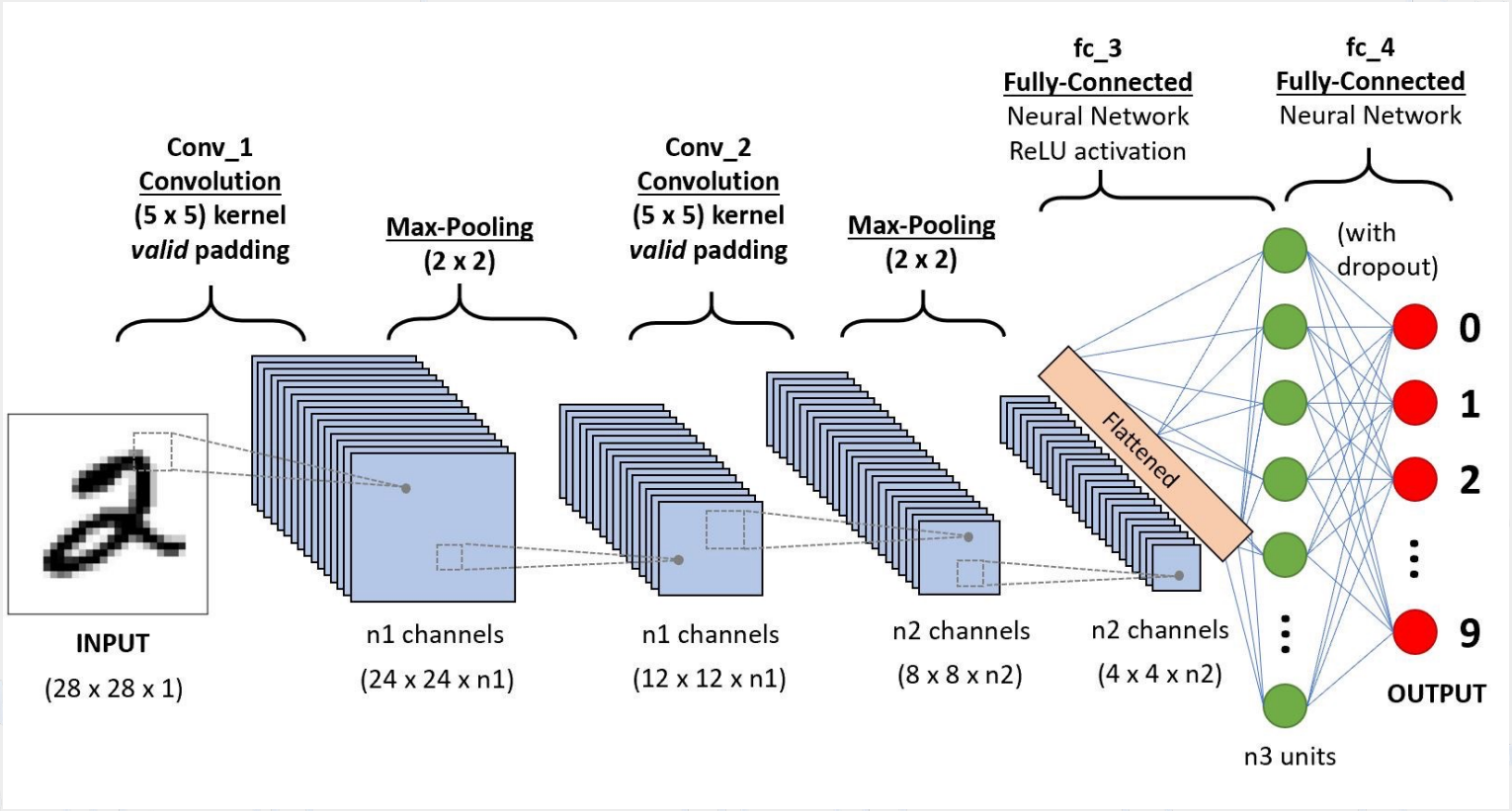
- Generative AI will not replace your Job or automate your Job
- Rather, Generative AI automates tasks
- A Job will involve a large number of tasks that can be automated
- Not every task can be fully automated. An analysis needs to be done on which tasks are good candidates for Generative AI automation
- Think of Generative AI as a Co-Pilot (Your assistant)

Generative AI Automation vs Augmentation

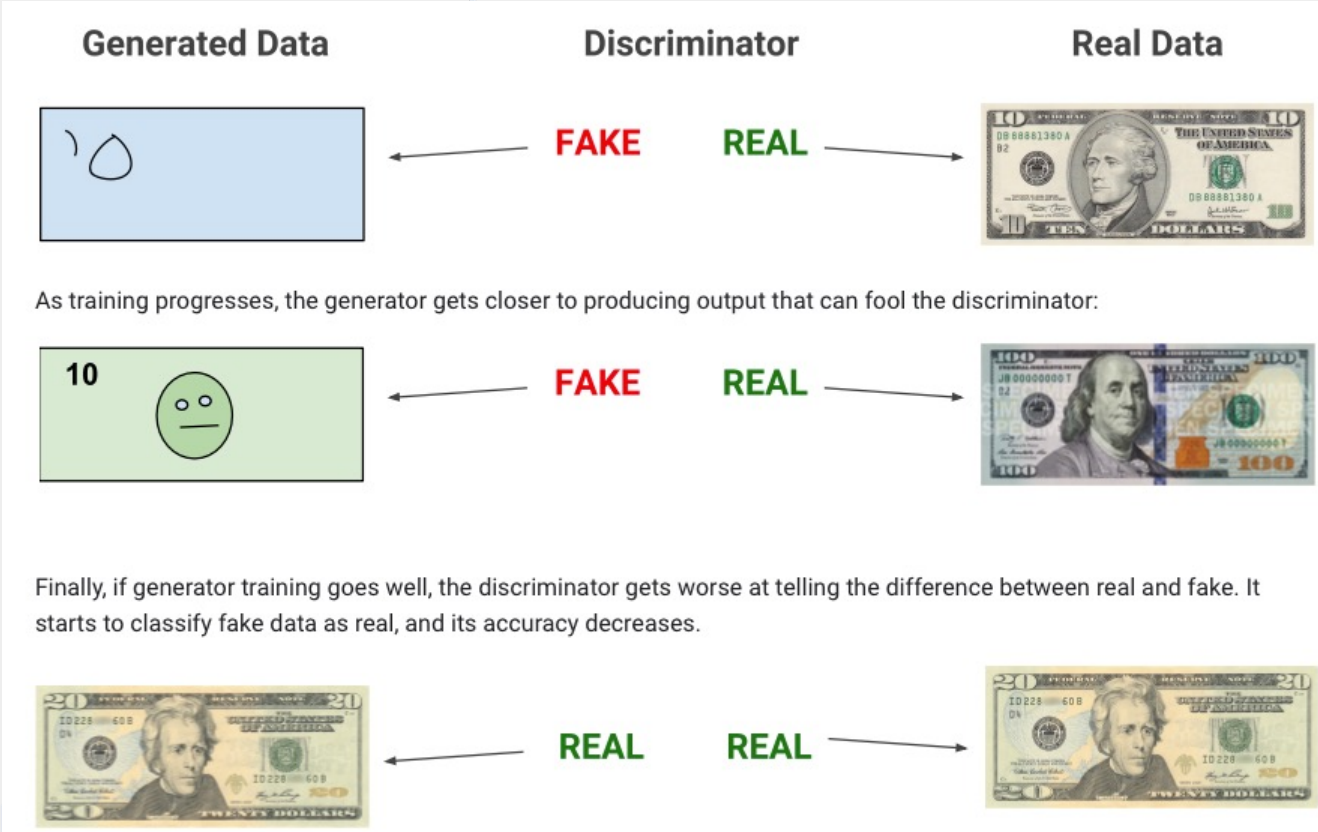
- In some tasks and businesses, you will start with augmentation, and then move toward automation.

- **Augmentation**: generative AI is used to augment (help/support/enhance) human capabilities. It enhances the quality and efficiency of tasks performed by humans.
- **Automation**: generative AI to **fully automate** certain tasks or processes. The AI system takes over the entire function, performing it from start to finish without the need for human intervention

CNN Architecture

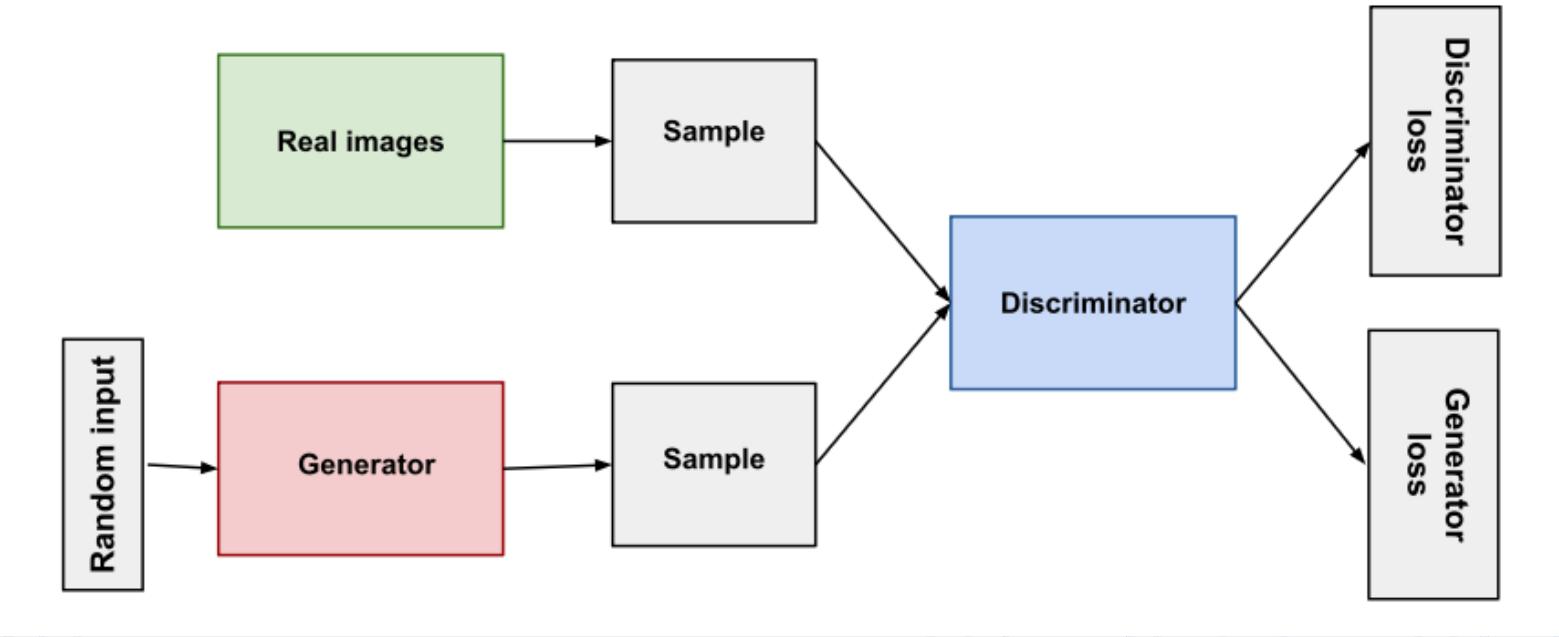


GANs Architecture



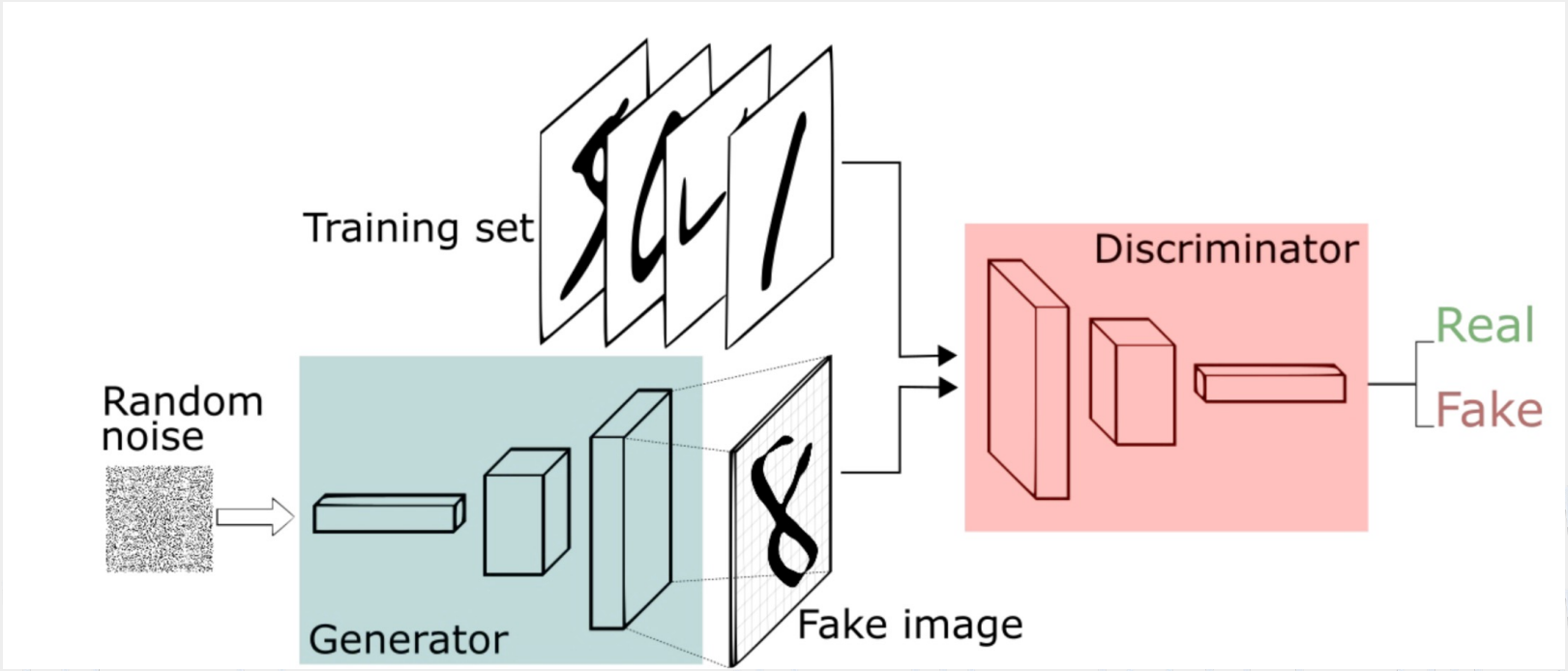
Source: Google Developer

GANs Architecture



GAN Demo - DCGAN

https://github.com/tatwan/generative_ai_class/blob/main/Activities/Class%20Activity/Ex_4_GANs_demo/GANs_DCGAN.ipynb



Core Generative AI
Other Key Concepts

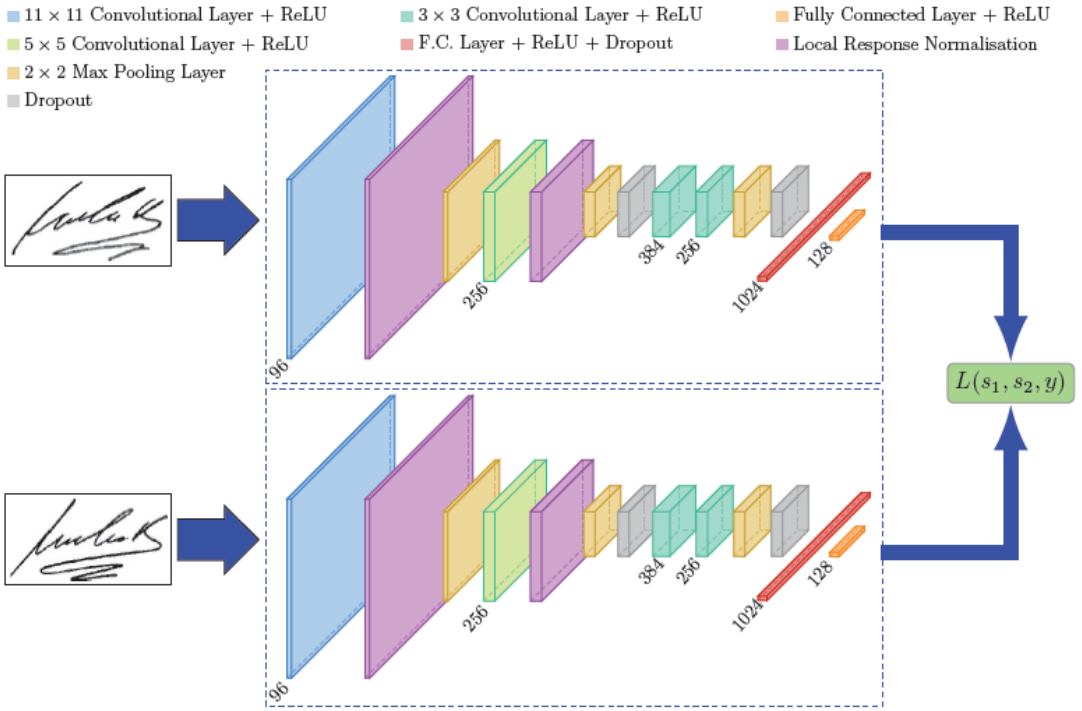
One-Shot Learning

- **Definition**: The ability of a model to learn from a single example or a few examples. In generative AI, this means being able to generate new data that is similar to a given example with minimal training data.
- **Application**: Useful in situations where large datasets are not available, such as rare disease diagnosis in healthcare.

One-Shot Learning

Siamese neural network

One Shot Image Recognition



Zero-Shot Learning

- **Definition**: The ability of a model to understand and perform tasks it has not explicitly been trained on. In generative AI, this might involve generating content or solving problems in domains not covered in the training data.
- **Application**: Enables more flexible and versatile AI systems, like a language model generating text in a genre it was not specifically trained on.

Zero-Shot Learning

Contrastive Language-Image Pretraining (CLIP)

- Developed by OpenAI
- Primarily a **transformer-based** model that has been widely used for zero-shot learning.
- CLIP consists of **two models**: a **text transformer** for encoding text embeddings and a **vision transformer (ViT)** for encoding image embeddings

Few-Shot Learning

- **Definition**: Similar to one-shot learning, but the model learns from a small number of examples rather than just one.
- **Application**: Useful in personalized AI applications, where the model adapts to individual preferences or needs with limited data.

Fine-Tuning

- **Definition**: Adjusting a pre-trained model on a new, typically smaller, dataset. This allows the model to specialize in a specific task or domain while leveraging the knowledge it gained during its initial training.
- **Application**: Common in adapting large language models to specific industries or niches, like legal or medical language.

Transfer Learning

- **Definition**: The process of taking a model trained on one task and applying it to a different, but related, task. This is often done by fine-tuning a pre-trained model.
- **Application**: Widely used to apply large-scale models trained on general tasks to more specific tasks, such as using a model trained on general images to identify specific types of objects.

Foundation Models

- **Definition**: Foundation models are a class of large-scale models that are pre-trained on extensive and diverse datasets, often unsupervised or self-supervised. They have a wide range of capabilities and can be adapted to various tasks.
- **Characteristics**:
  - **Scale**: They are typically very large, both in terms of the size of the model (number of parameters) and the dataset used for training.
  - **Generalizability**: These models are designed to be general-purpose, meaning they can be fine-tuned or adapted to perform a wide variety of tasks, often with state-of-the-art performance.
  - **Examples**: Models like GPT-3, BERT, and other large language models are considered foundation models.
- **Usage**: Foundation models serve as a starting point for further task-specific training (fine-tuning) or for developing new models and applications.

Large Language Models (LLMs)

Large Language Models (LLMs) are a type of artificial intelligence model specifically designed to understand, generate, and interact with human language at a large scale. They have become a significant focus in the field of Natural Language Processing (NLP) due to their remarkable ability to handle a wide range of language-related tasks.

Large Language Models (LLMs)

1. **Outdated Knowledge**: **LLMs** are unable to provide real-time or recent data as they rely solely on the training data provided to them.
2. **Inability to act**: **LLMs** cannot perform actions or interact with external systems, limiting their functionality. For example, they cannot initiate web searches, query databases in real-time, or use a calculator for multiplying numbers.
3. **Lack of context and additional information**: **LLMs** may struggle to understand and incorporate context from previous prompts or conversations. They may not remember previously mentioned details or fail to provide additional relevant information beyond the given prompt.
4. **Complexity and Learning Curve**: Developing applications using large language models often requires a deep understanding of AI concepts, complex algorithms, and APIs. This can pose a challenge for developers who may not have expertise in these areas.
5. **Hallucinations**: **LLMs** have a lot of general knowledge about the world implicit in their weights. However, they may have an insufficient understanding about certain subjects, and generate responses that are not factually correct or coherent. For example, they might produce information that does not exist or provide inaccurate details.
6. **Bias and Discrimination**: Depending on the data they were trained on, LLMs can exhibit biases, which can be of religious, ideological, political, and other nature.

Prompt Design vs Engineering

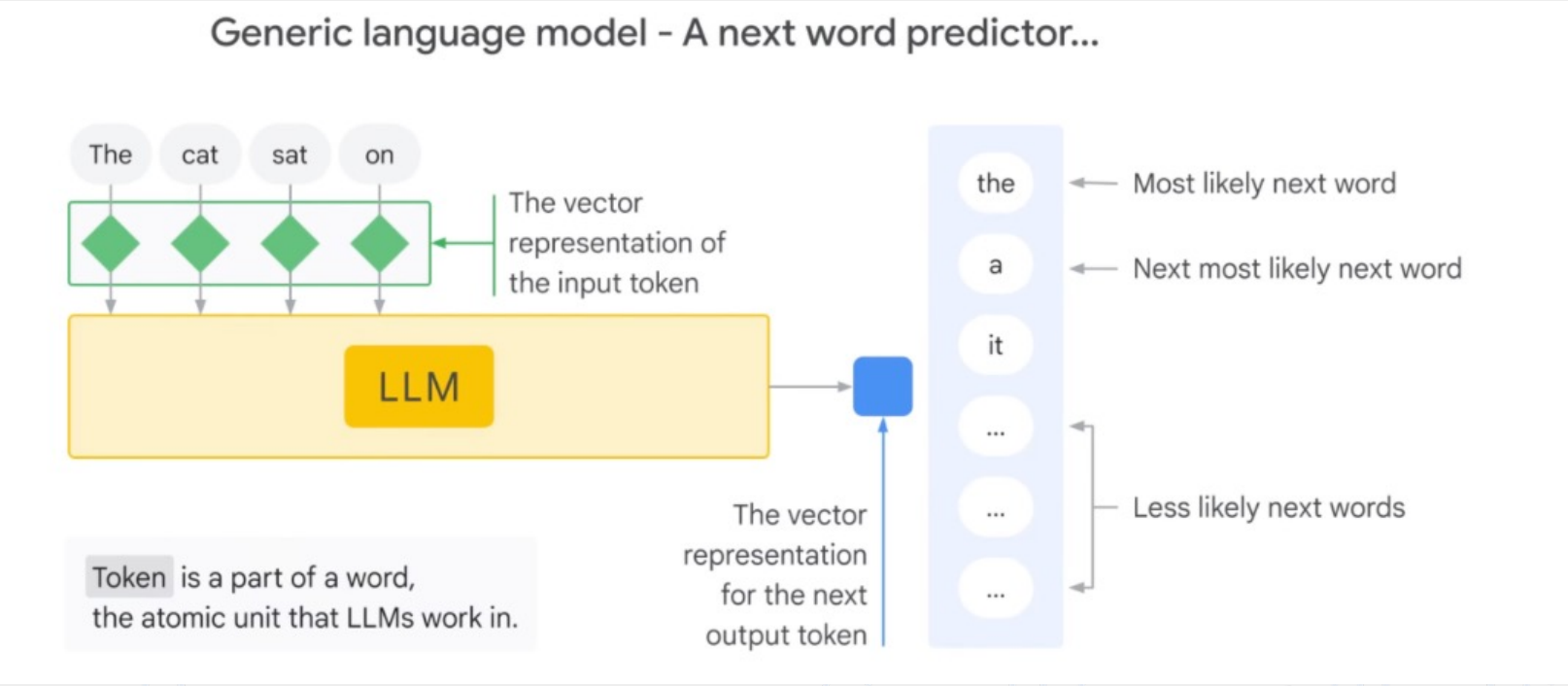
Prompt Design

Prompts involve instructions and context passed to a language model to achieve a desired task.

Prompt Engineering

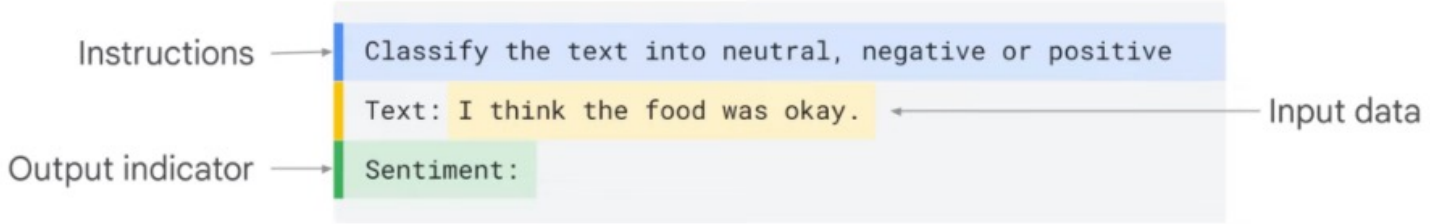
Prompt engineering is the practice of developing and optimizing prompts to efficiently use language models for a variety of applications.

LLM Design



LLM Design

Elements of the Prompt



Advanced Prompt Engineering

Summary of Prompting Techniques

Prompting Technique	When to use?
Zero-shot prompting	Effective on various basic tasks but not as effective for advanced tasks
Few-shot in-context learning	Add high-quality demonstrations to steer the model better
Chain-of-thought (CoT)	Applies chain of thoughts to elicit models to use reasoning steps before responding
ReAct	Enables LLMs to leverage external tools and knowledge to improve performance and reduce hallucination
Prompt chaining	Chains several prompts to achieve complex tasks that involve different subtasks
Tree of thoughts	Encourages exploration over thoughts that serve as intermediate steps for general problem solving with language models
Retrieval Augmented Generation (RAG)	Helps to leverage external knowledge to optimize prompt context and improve output quality and reliability

Retrieval Augmented Generation (RAG)

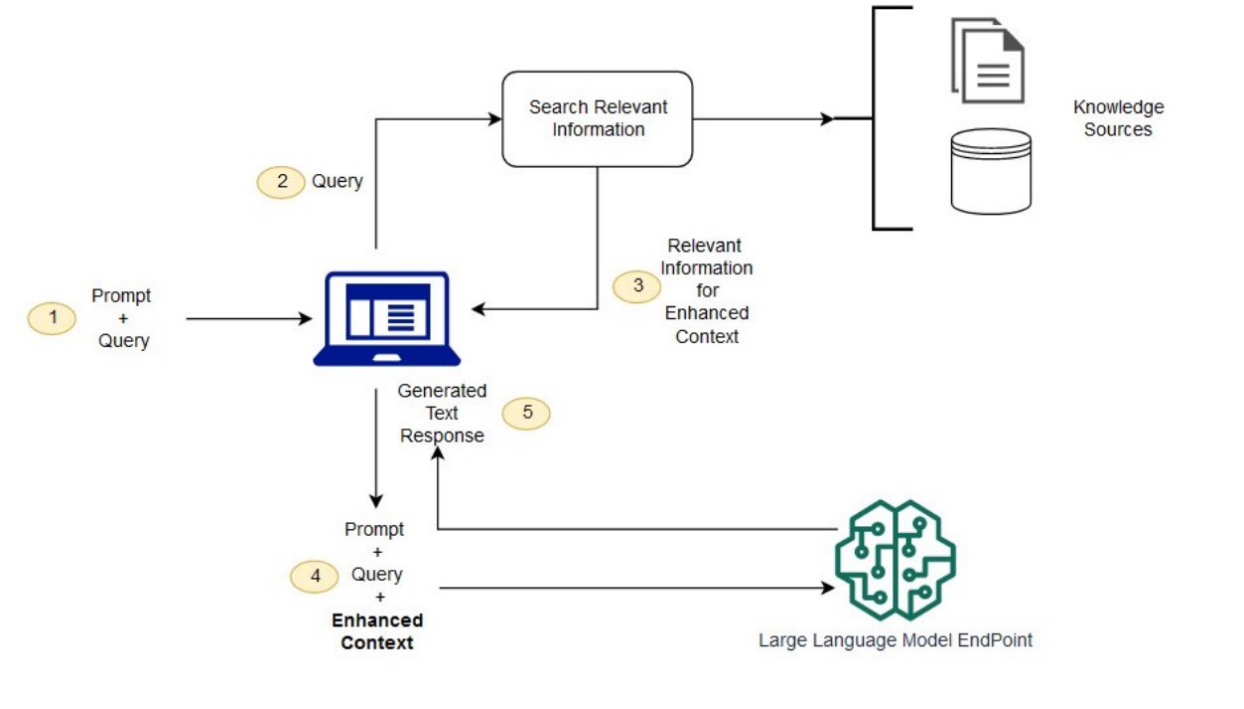
RAG: Gives LLM access to external data sources

Retrieval-augmented generation (**RAG**) is an advanced artificial intelligence technique that combines **information retrieval** with **text generation**. It enhances the accuracy and reliability of generative AI models by allowing them to retrieve relevant information from external sources and incorporate it into generated text.

The **RAG** architecture integrates a **neural retriever** and a **neural generator**. The retriever is used to fetch relevant context or information from a large corpus of data (like a database or a collection of documents), and the generator then uses this retrieved information to construct a response or output.

Retrieval Augmented Generation (RAG)

RAG: Gives LLM access to external data sources



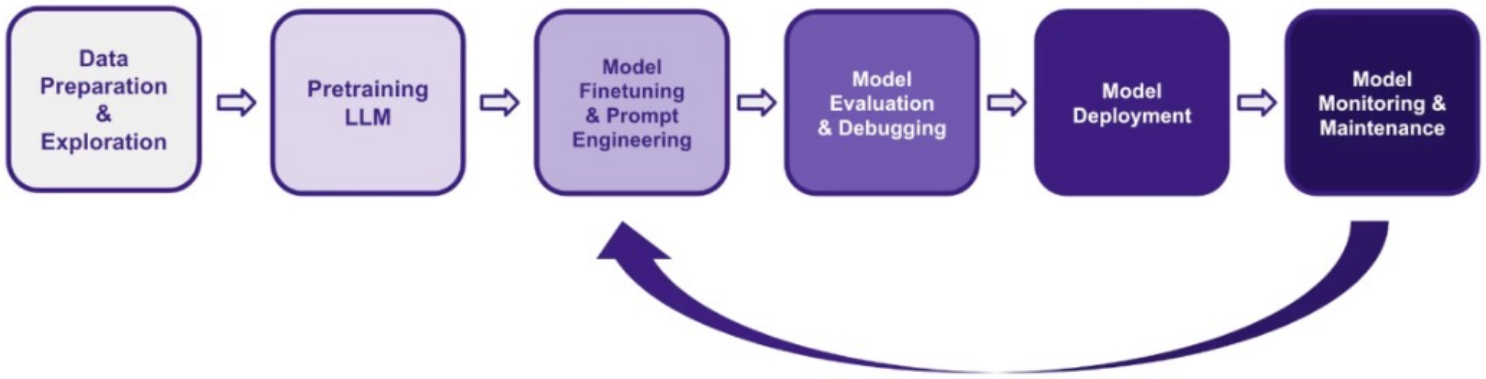
LLMOps

LLMOps, or Large Language Model Operations, is a set of tools and best practices used to manage the lifecycle of large language models (LLMs) and LLM-powered applications, including development, deployment, and maintenance.

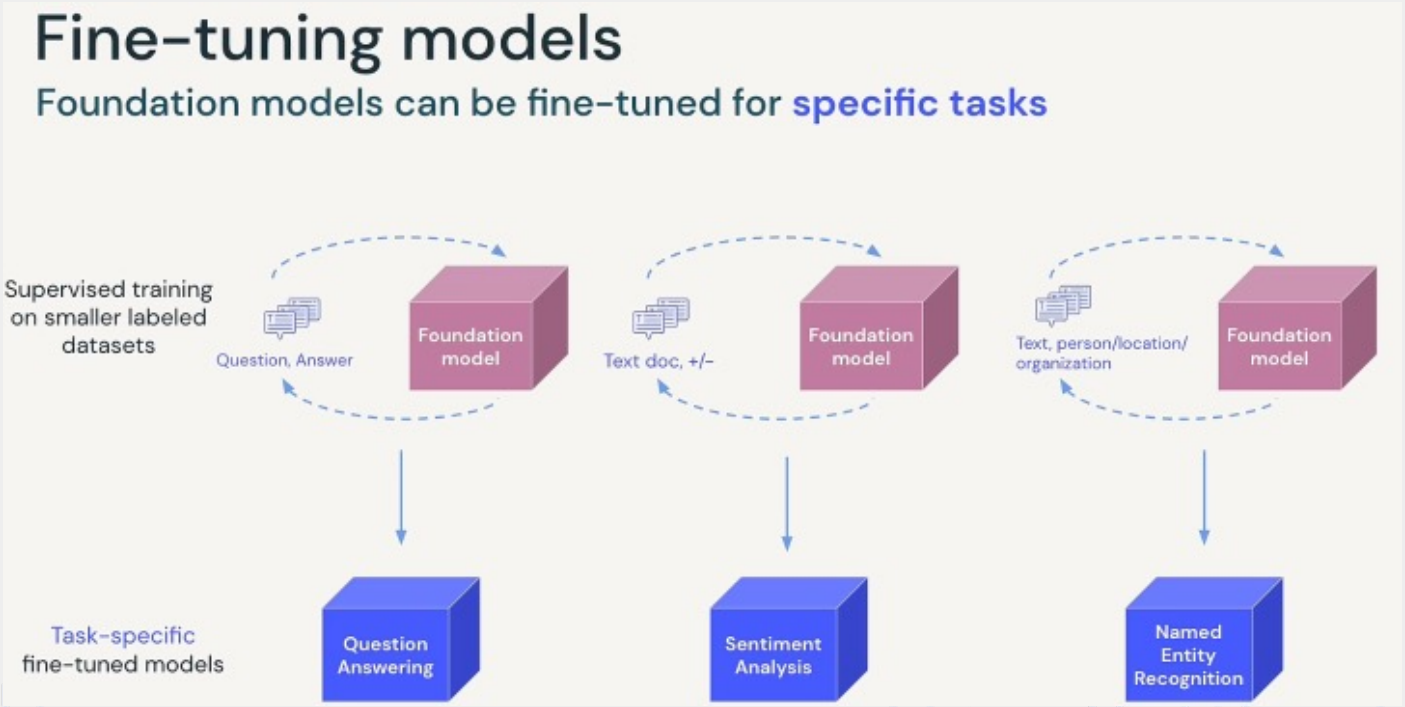
It focuses on the **operational** capabilities and **infrastructure** required to **fine-tune** existing **foundational models** and deploy these refined models as part of a product. LLMOps incorporates prompt management, LLM chaining, monitoring, and observability techniques not typically found in conventional MLOps. It is similar to MLOps but is specifically tailored to the unique requirements of LLMs and LLM-powered applications

LLMOps

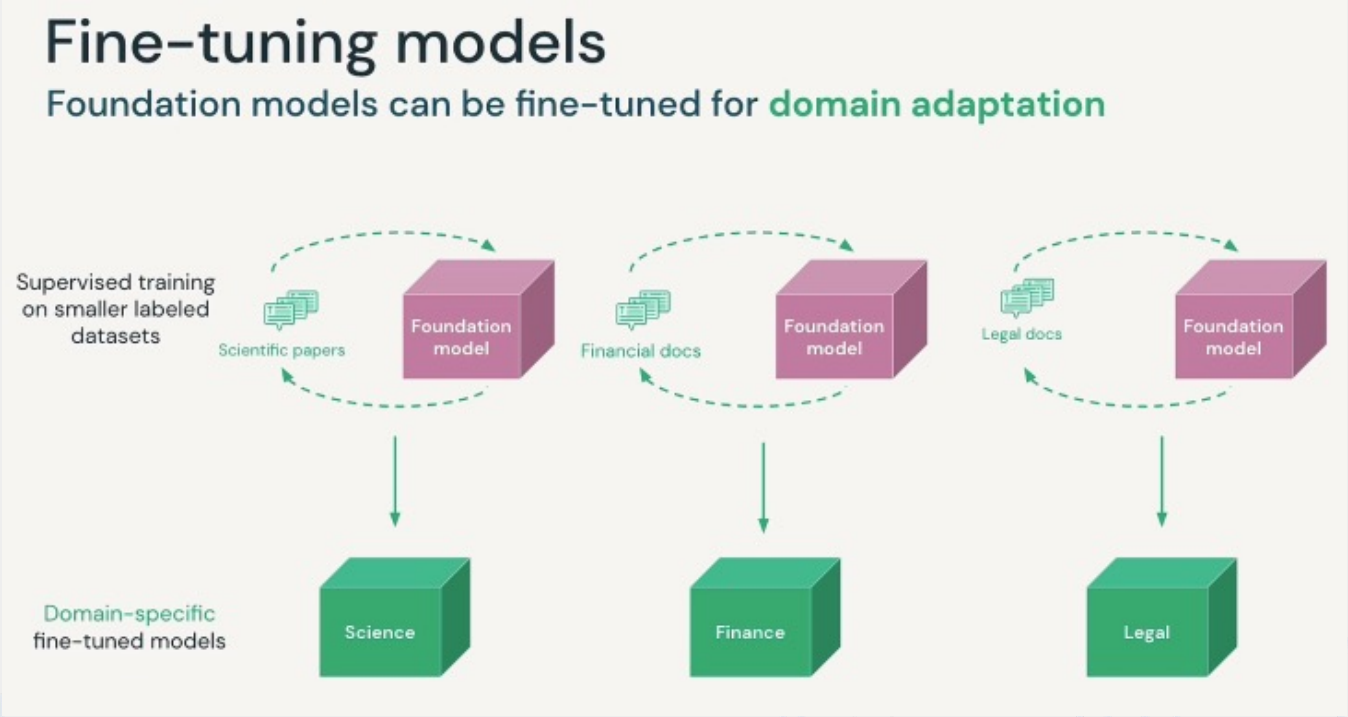
Overview of LLMOps Life Cycle



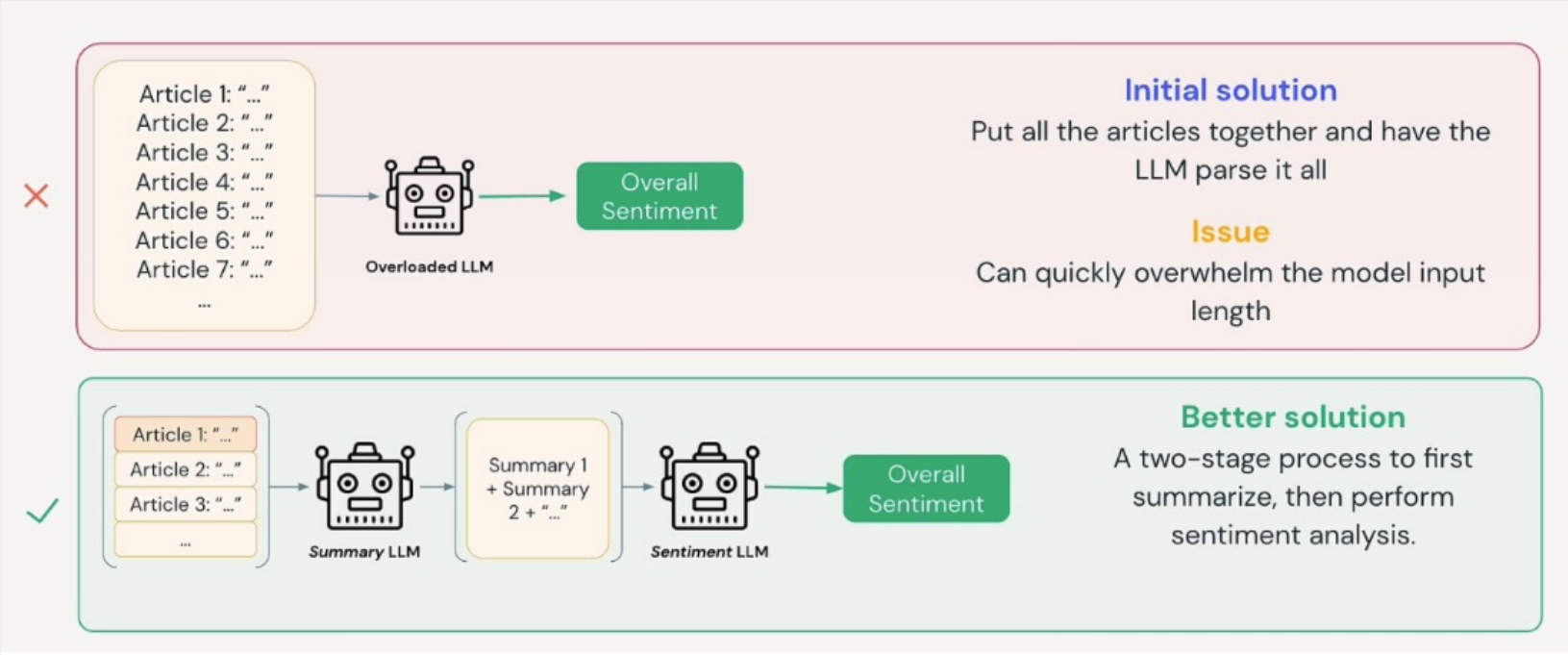
LLM Fune Tuning



LLM Fune Tuning



Mixing LLM Flavors in a Workflow



Prompt Injection

Prompt Injection

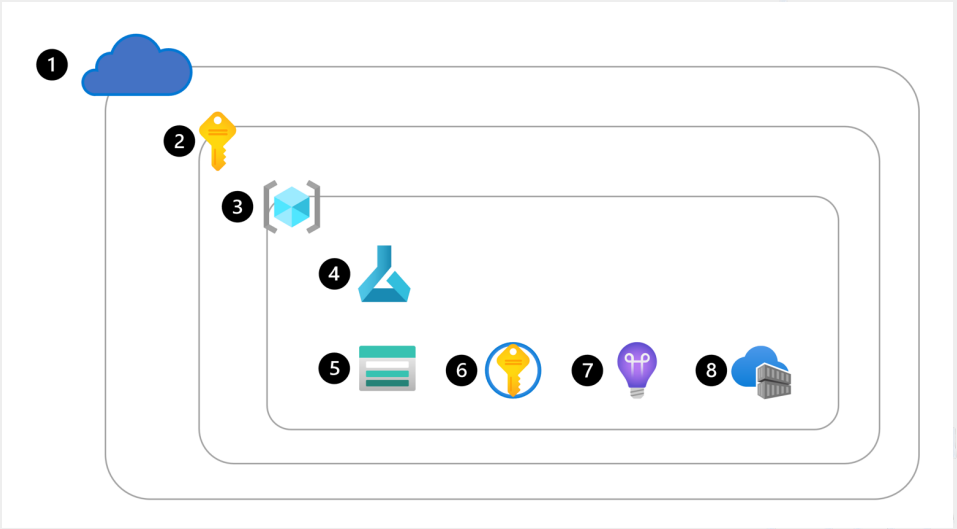
- **Definition:** Inserting a specific instruction or prompt within the input text to **manipulate the normal behavior** of LLMs.
- Other prompt injection cases:
  - Generating malicious code
  - Instructing agent to give wrong information
  - Revealing confidential information

ML Workspace (Studio)

https://azure.microsoft.com/en-us/products

Azure ML Services

- When a **workspace** is provisioned, Azure will automatically create other Azure resources within the same resource group to support the workspace:
- **Azure Storage Account**: To store files and notebooks used in the workspace, and to store metadata of jobs and models.
- **Azure Key Vault**: To securely manage secrets such as authentication keys and credentials used by the workspace.
- **Application Insights**: To monitor predictive services in the workspace.
- **Azure Container Registry**: Created when needed to store images for Azure Machine Learning environments.



Azure ML Compute Instances

A compute instance is a fully managed cloud-based workstation optimized for your machine learning development environment.

- **Compute instances**: Similar to a virtual machine in the cloud, managed by the workspace. Ideal to use as a development environment to run (Jupyter) notebooks.
- **Compute clusters**: On-demand clusters of CPU or GPU compute nodes in the cloud, managed by the workspace. Ideal to use for production workloads as they automatically scale to your needs.
- **Kubernetes clusters**: Allows you to create or attach an Azure Kubernetes Service (AKS) cluster. Ideal to deploy trained machine learning models in production scenarios.
- **Attached computes**: Allows you to attach other Azure compute resources to the workspace, like Azure Databricks or Synapse Spark pools.
- **Serverless compute**: A fully managed, on-demand compute you can use for training jobs.

Datastores

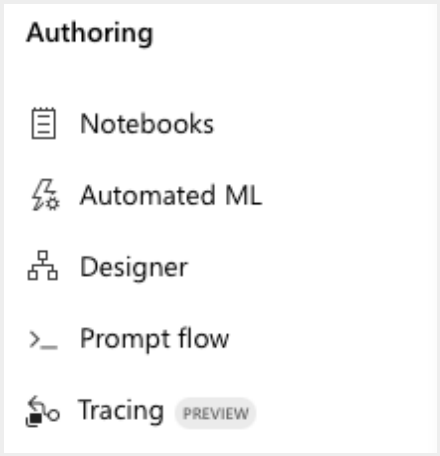
The workspace doesn't store any data itself. Instead, all data is stored in **datastores**, which are references to Azure data services.

- workspaceartifactstore: Connects to the azureml container of the Azure Storage account created with the workspace. Used to store compute and experiment logs when running jobs.
- workspaceworkingdirectory: Connects to the file share of the Azure Storage account created with the workspace used by the **Notebooks** section of the studio. Whenever you upload files or folders to access from a compute instance, it's uploaded to this file share.
- workspaceblobstore: Connects to the Blob Storage of the Azure Storage account created with the workspace. Specifically the azureml-blobstore-... container. Set as the default datastore, which means that whenever you create a data asset and upload data, it's stored in this container.
- workspacefilestore: Connects to the file share of the Azure Storage account created with the workspace. Specifically the azureml-filestore-... file share.

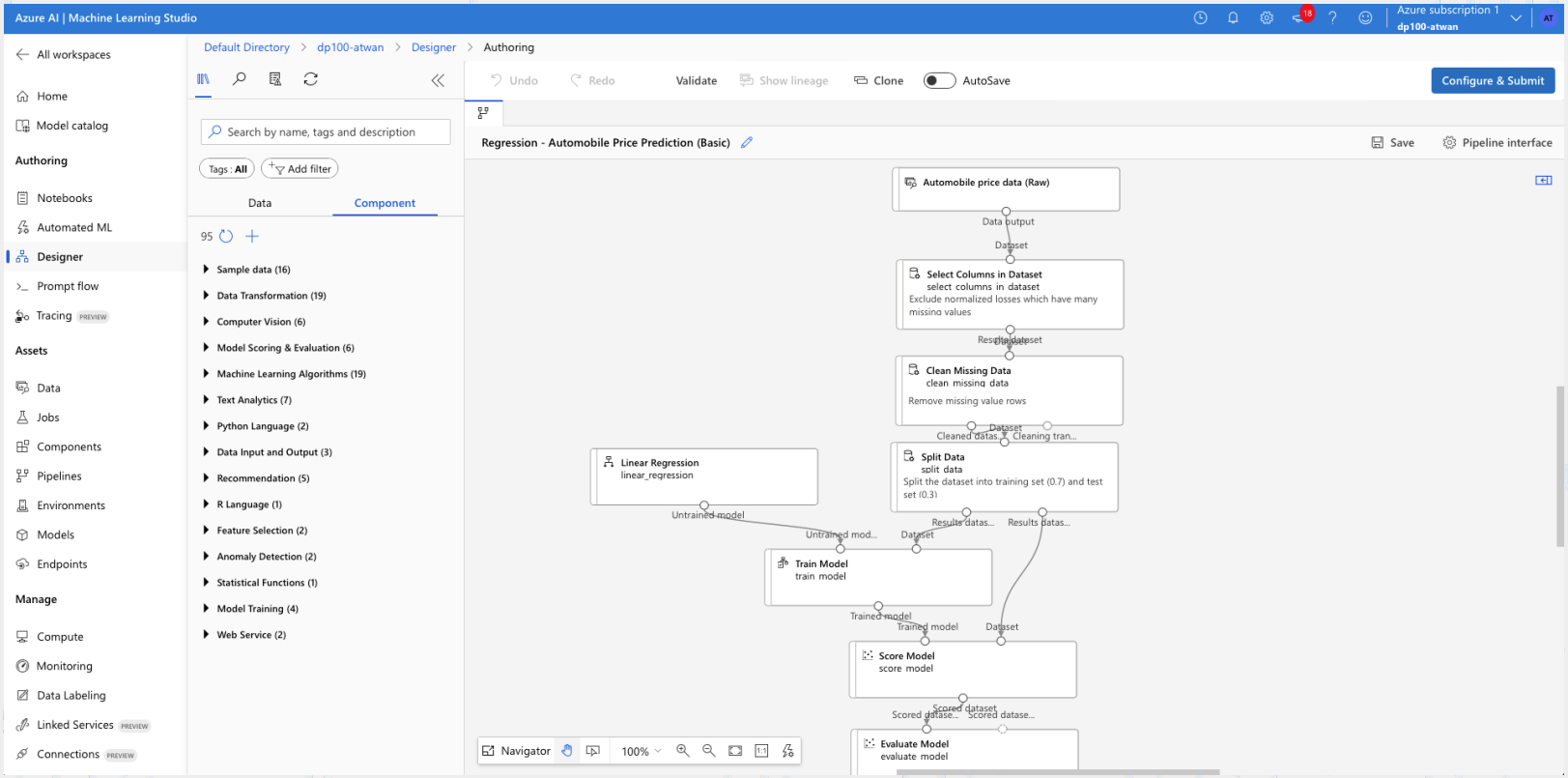
ML Assets

- Models
- Environments
- Data
- Components

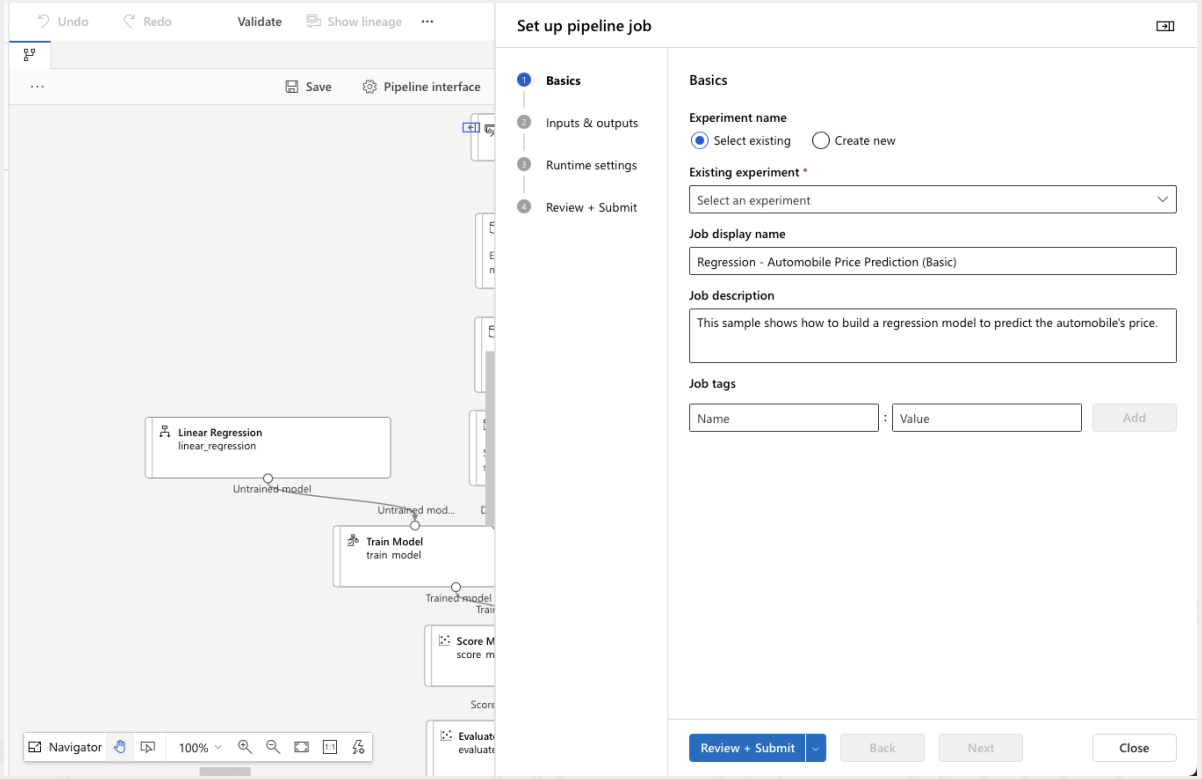
Authoring within ML Workspace



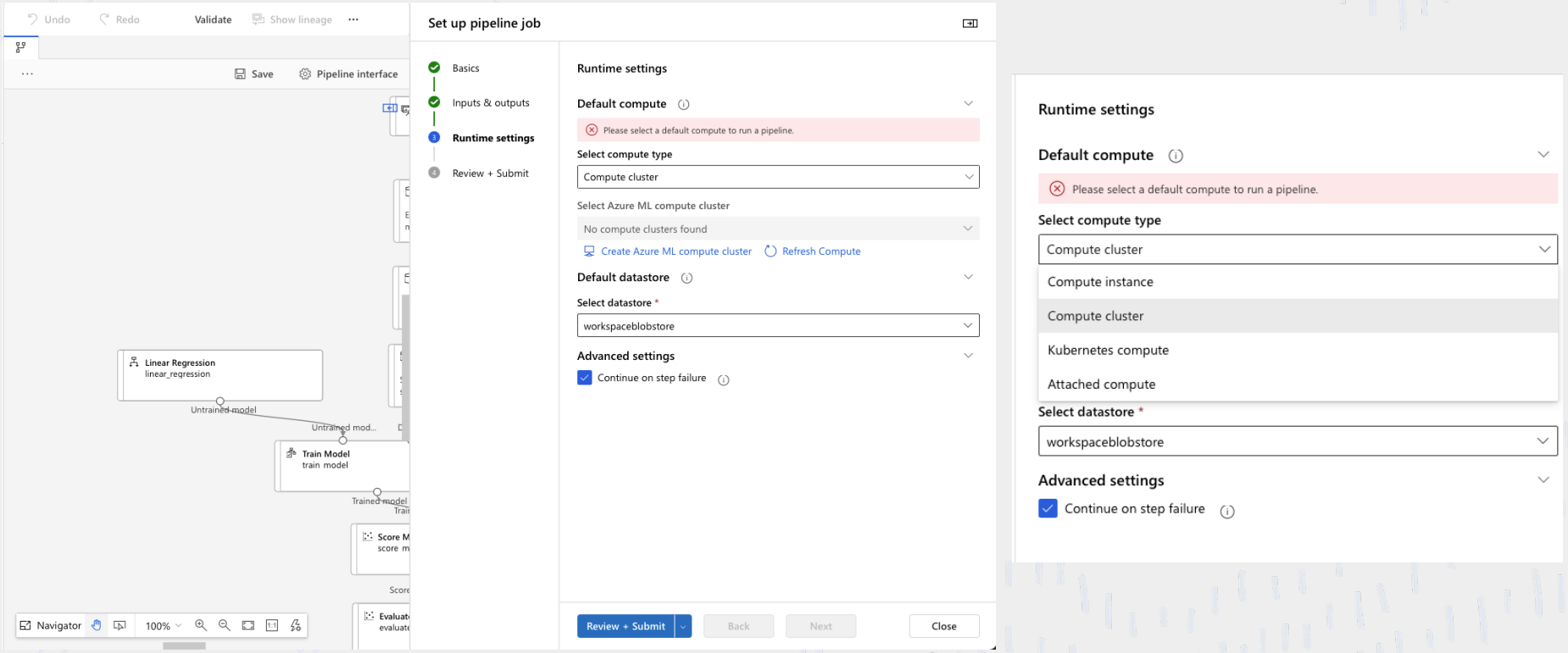
Designer



Designer

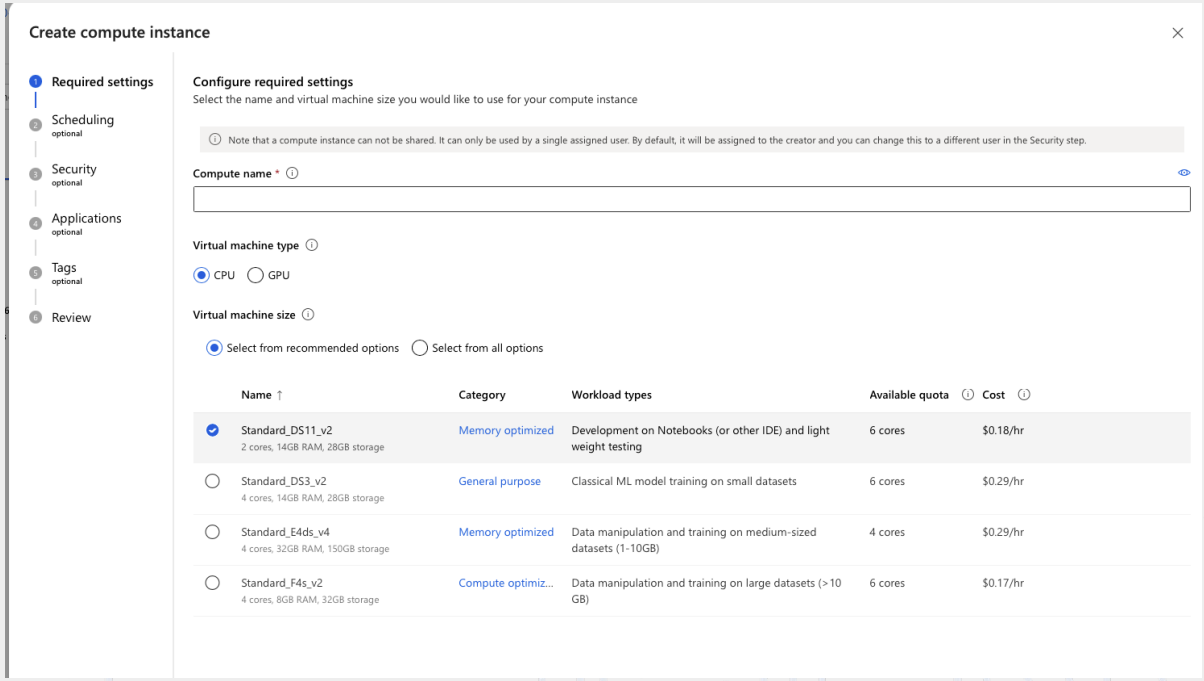


Designer – A Compute Cluster



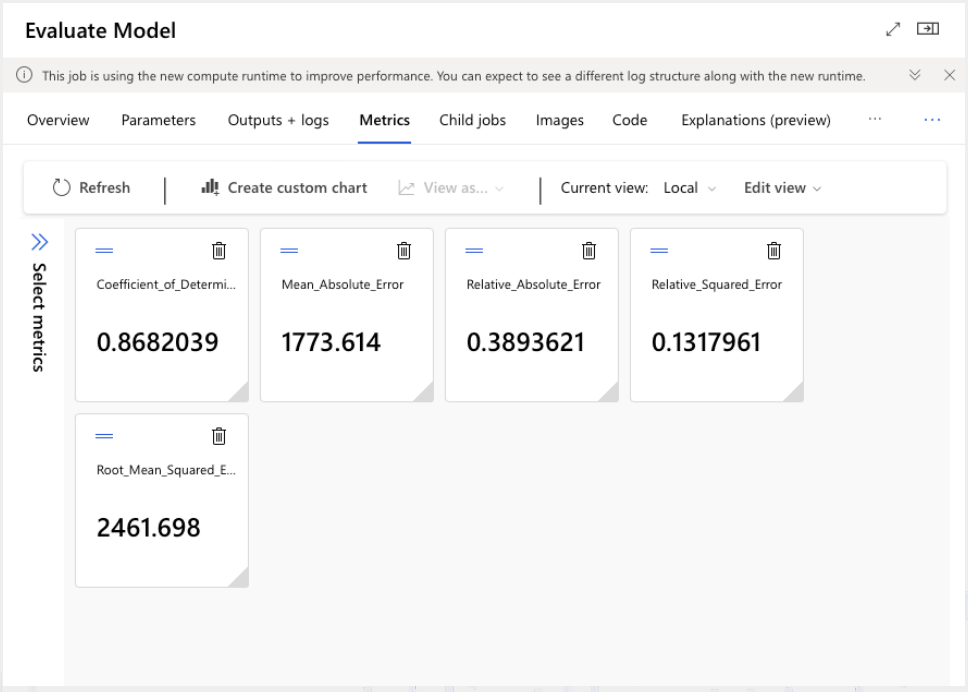
What is a Computer Instance in ML Workspace

- A compute instance is a fully managed cloud-based workstation optimized for your machine learning development environment.



What is a Computer Instance in ML Workspace

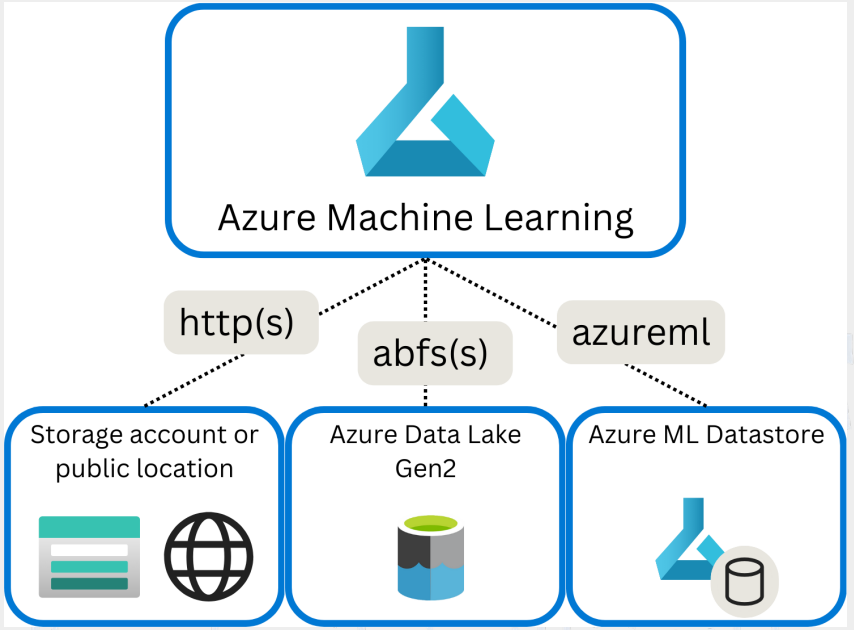
- A compute instance is a fully managed cloud-based workstation optimized for your machine learning development environment.



Data in Azure ML Workspace

To find and access data in Azure Machine Learning, you can use **Uniform Resource Identifiers** (**URIs**).

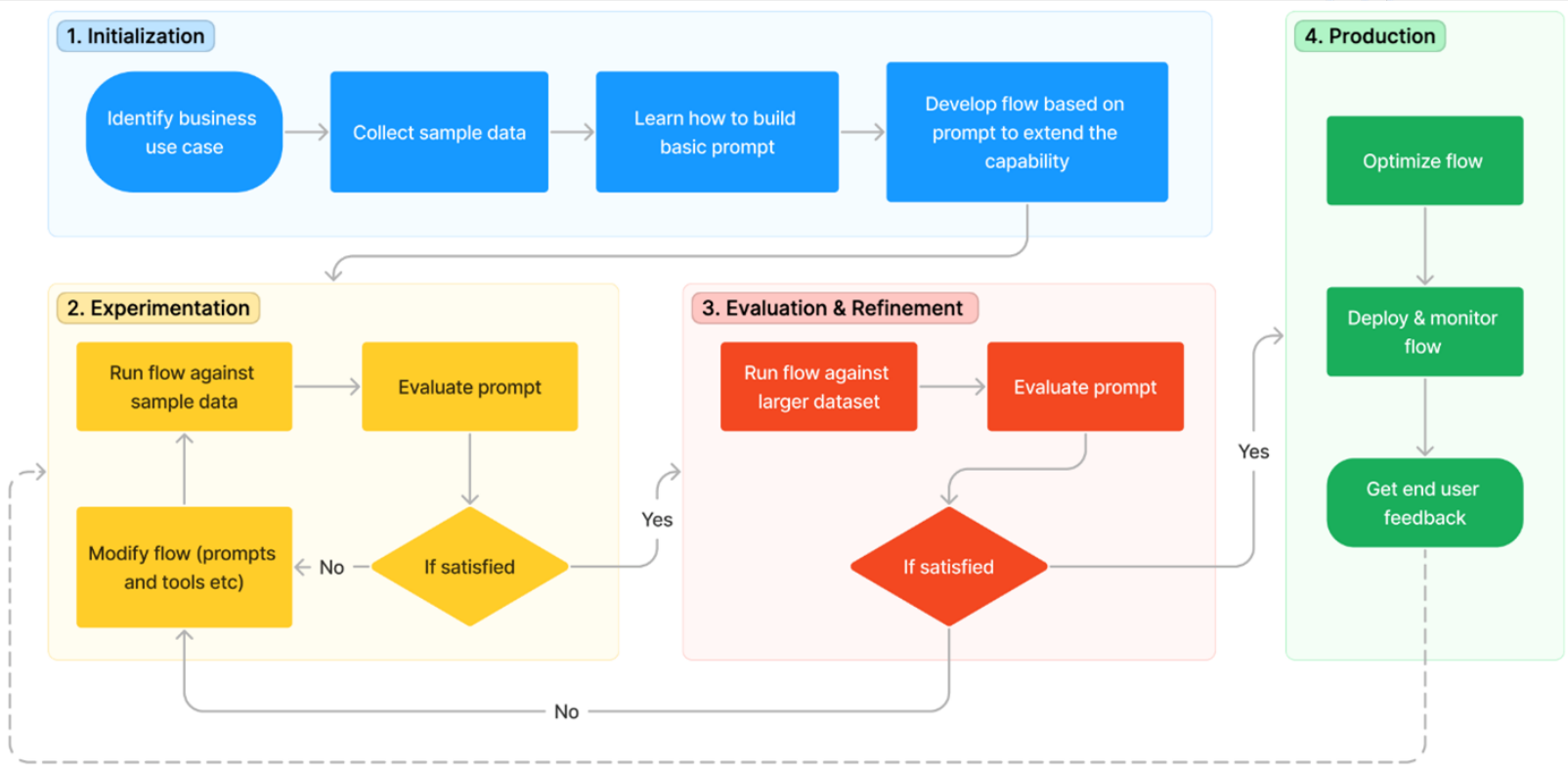
- http(s): Use for data stores publicly or privately in an Azure Blob Storage or publicly available http(s) location.
- abfs(s): Use for data stores in an Azure Data Lake Storage Gen 2.
- azureml: Use for data stored in a datastore.



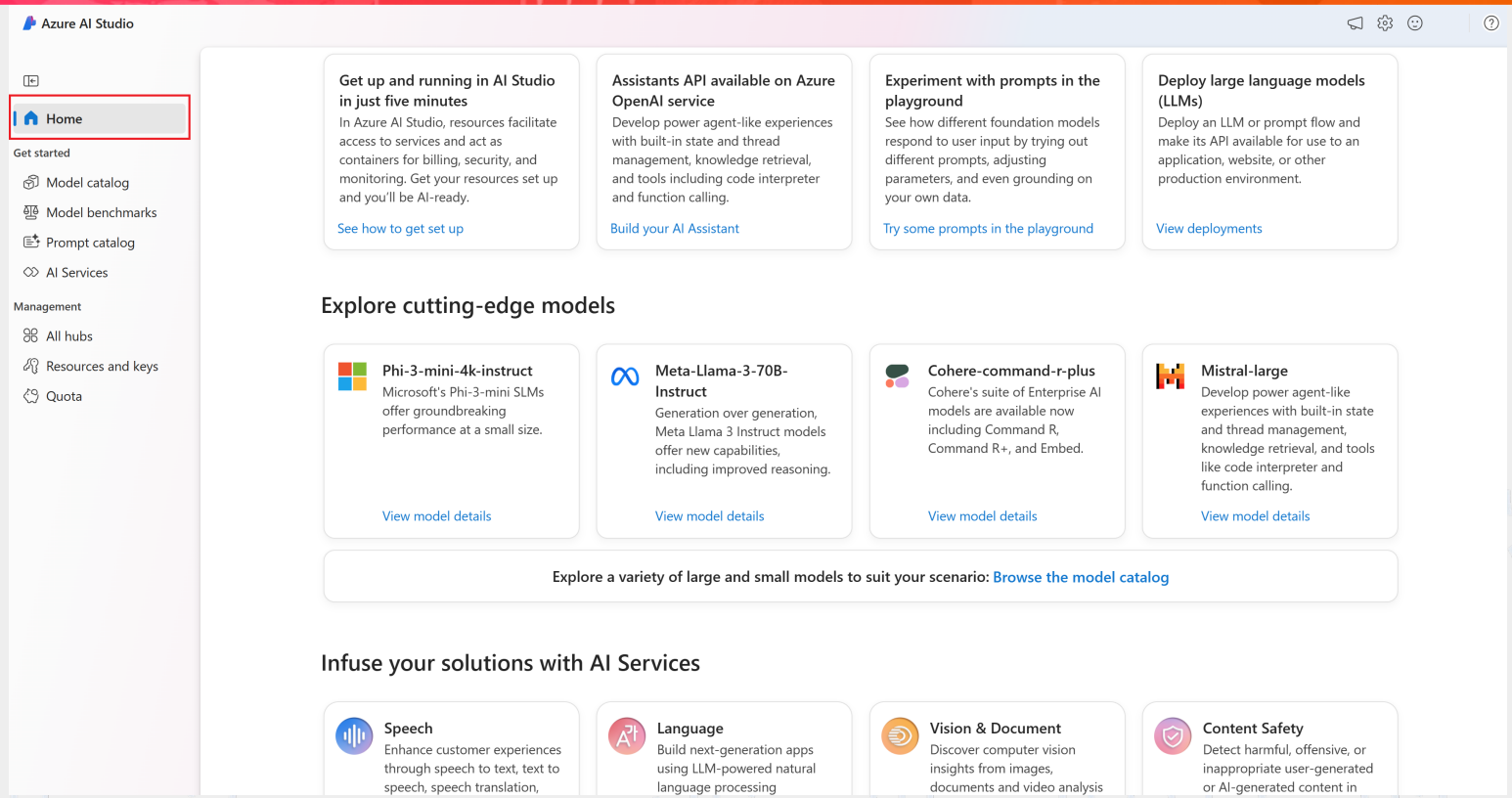
AI Studio

https://azure.microsoft.com/en-us/products

ML Studio – Prompt Flow



AI Studio

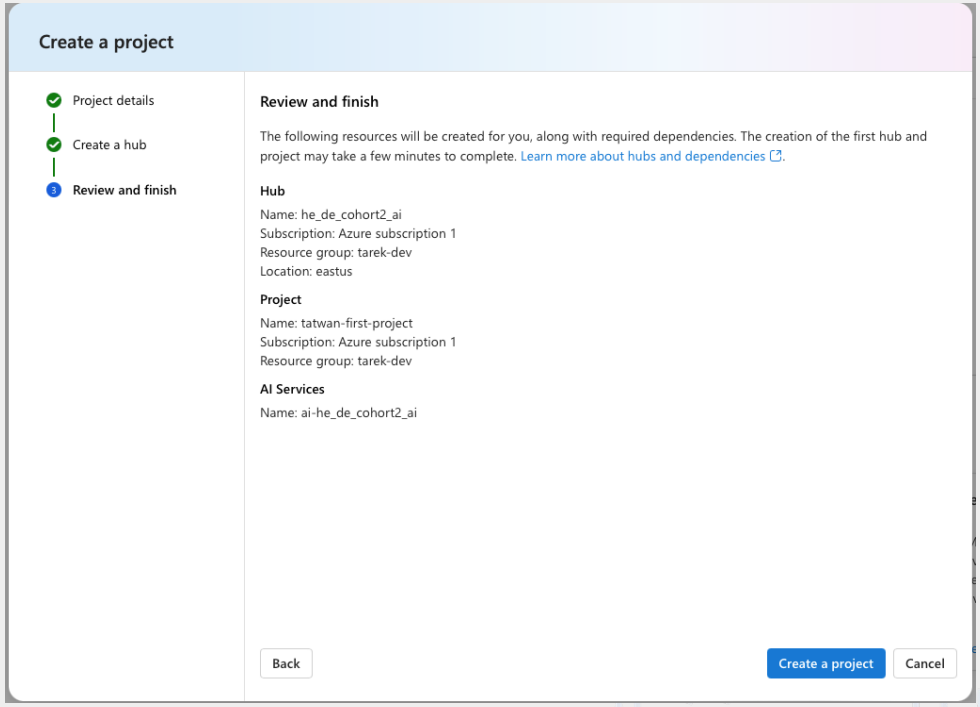


What is AI Studio for

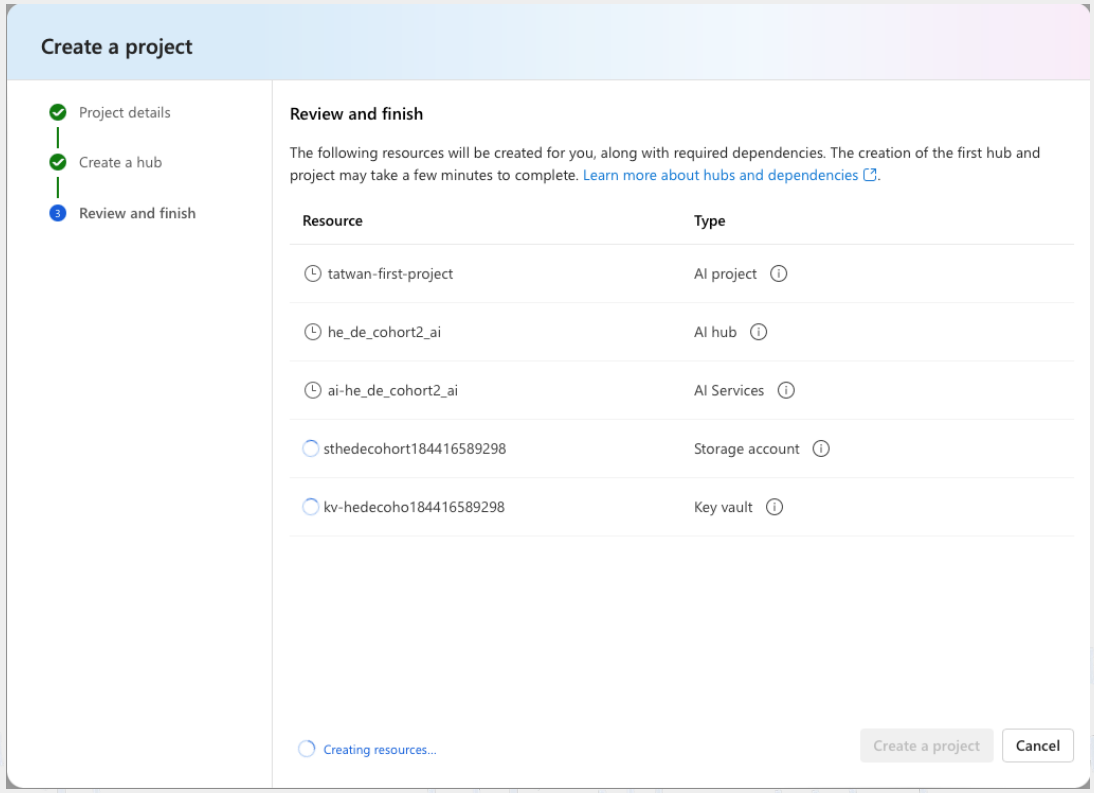
AI Studio is designed for developers to:

- Build **generative AI** applications on an enterprise-grade platform.
- Directly from the studio you can interact with a project code-first via Visual Studio Code (Web).
- AI Studio is a trusted and inclusive platform that empowers developers of all abilities and preferences to innovate with AI and shape the future.
- Seamlessly **explore**, **build**, **test**, and **deploy** using cutting-edge AI tools and ML models, grounded in responsible AI practices.
- Build together as one team. Your AI Studio hub provides enterprise-grade security, and a collaborative environment with shared files and connections to pretrained models, data and compute.
- Organize your way. Your AI Studio project helps you save state, allowing you iterate from first idea, to first prototype, and then first production deployment. Also easily invite others to collaborate along this journey.

AI Project and AI Hub



AI Project and AI Hub



Thank you!

If you have any additional questions, please ask!