Multimodal Prompting

Generative AI

Module 1 – Lesson 5

Today’s activities

- Introduction to multimodal applications
- Multimodal LLMs
- Prompting MLLMs
- Multimodal use cases

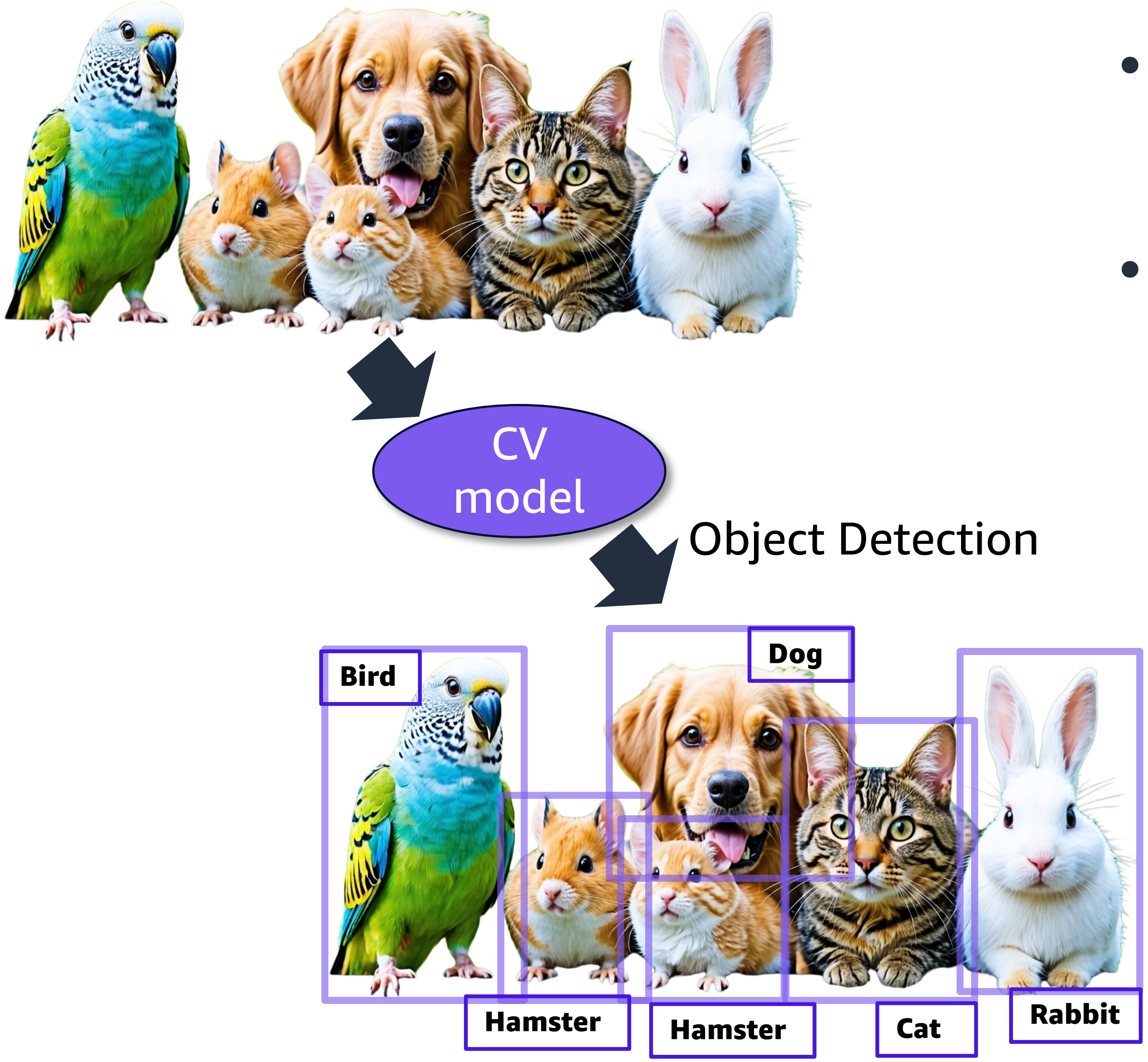
Introduction to multimodal applications

Your marketing company has been hired by a pet shop:

- They need to create individual flyers about each pet based on their images and bio

- How can you do that with traditional, single modality models?

We can consider an image-only model



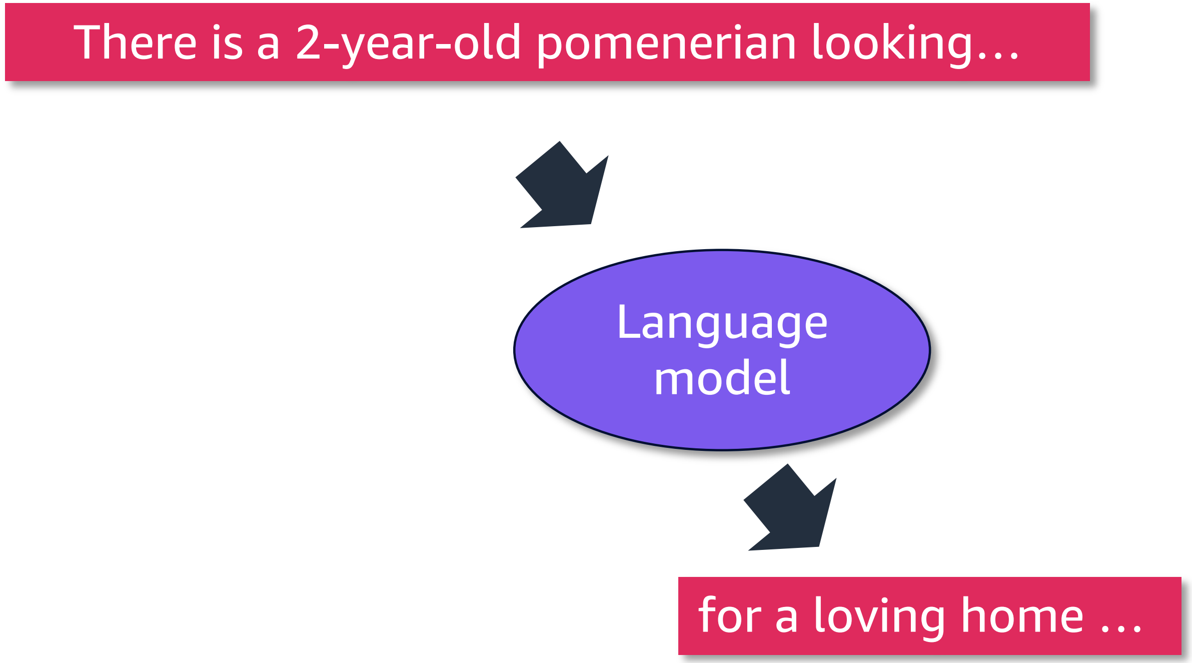
CV model

Object Detection

- Learns information only from images
  - Computer vision (CV)
- Can accomplish basic image understanding tasks
  - Classifying images for our campaigns depending on the theme needed
    - Ex. Cat class, dog class
  - Object detection
  - Semantic segmentation
  - It won’t accept the input search query or won’t be able to generate text

We can consider a text-only model

- Learns information only from text
- Can accomplish basic text understanding tasks
  - Generating text based in a campaign description
- Won’t be able to generate a description of the pets from images



What does “multimodal” mean?

- Humans are naturally multimodal in the way we interact with the world!
- Perceive the world using multiple senses:
  - Vision, hearing, smell, taste and touch
- Engage in non-verbal communication
  - Gestures
  - Facial expressions
  - Body language
  - Eye contact
  - Appearance

Why multimodal?

- Generative AI shifted from **prediction** to **interaction**
- **Multimodality** is a way to boost AI performance to interact with humans to solve real world problems

Data modalities

- **Image Data**:
  - The most versatile format for model inputs
  - There’s much more visual data than text data
    - Phones and webcams constantly take pictures and videos today
  - It can be used to represent:
    - Text
    - Tabular data
    - Audio
    - And to some extent, videos

Data modalities

- **Text Data**:
  - Text is a powerful mode for model outputs
  - A model that can understand/generate text can be used for many tasks:
    - Summarization
    - Translation
    - Reasoning
    - question answering
    - etc.

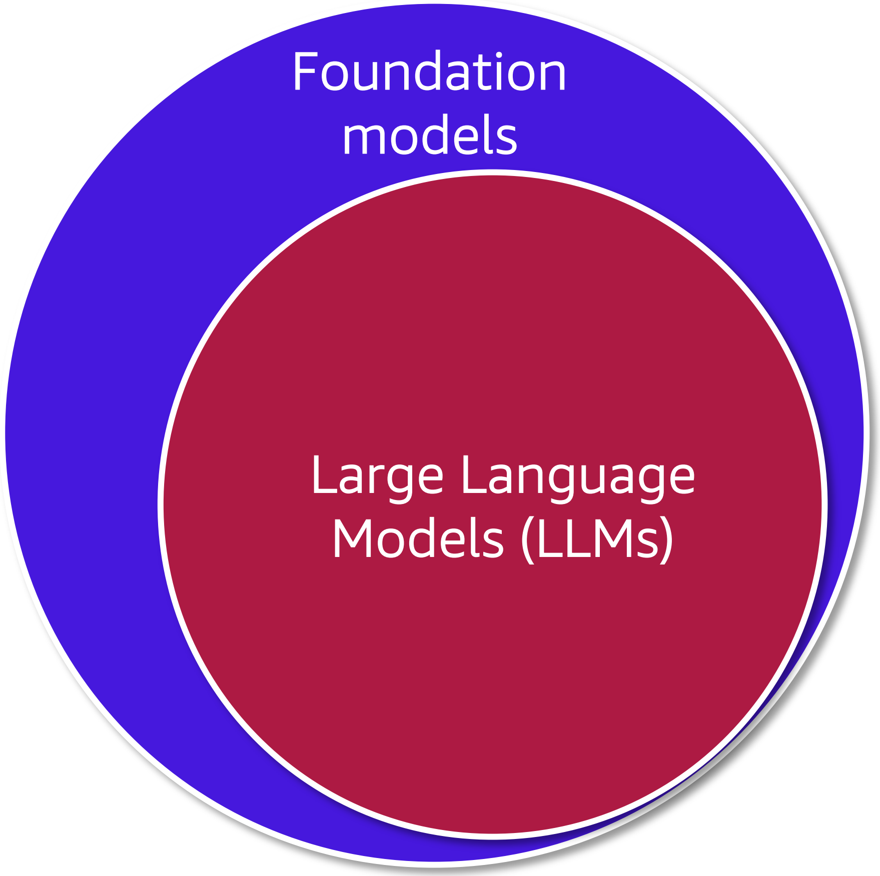
Other data modalities

- Video
- Audio
- Haptic data
- Electrical signals

- In this course, we will focus mainly on **text** and **image** data.

Multimodal LLMs

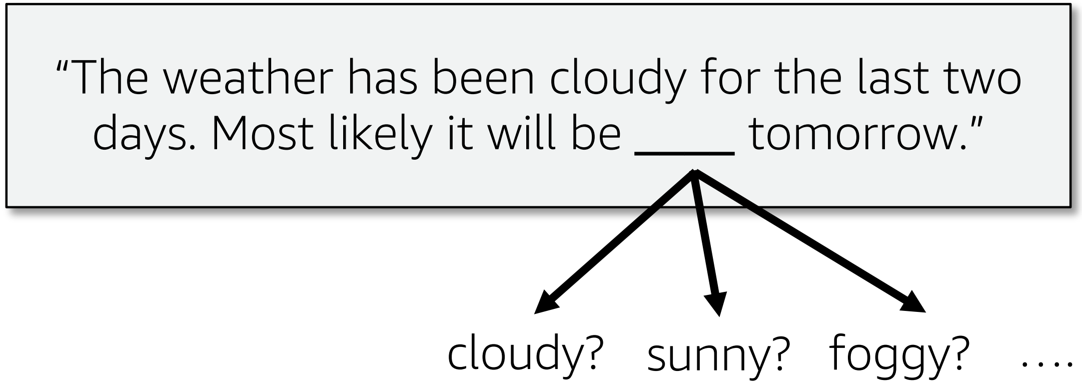
Review: Large Language Models (LLMs)



Foundation models

Large Language Models (LLMs)

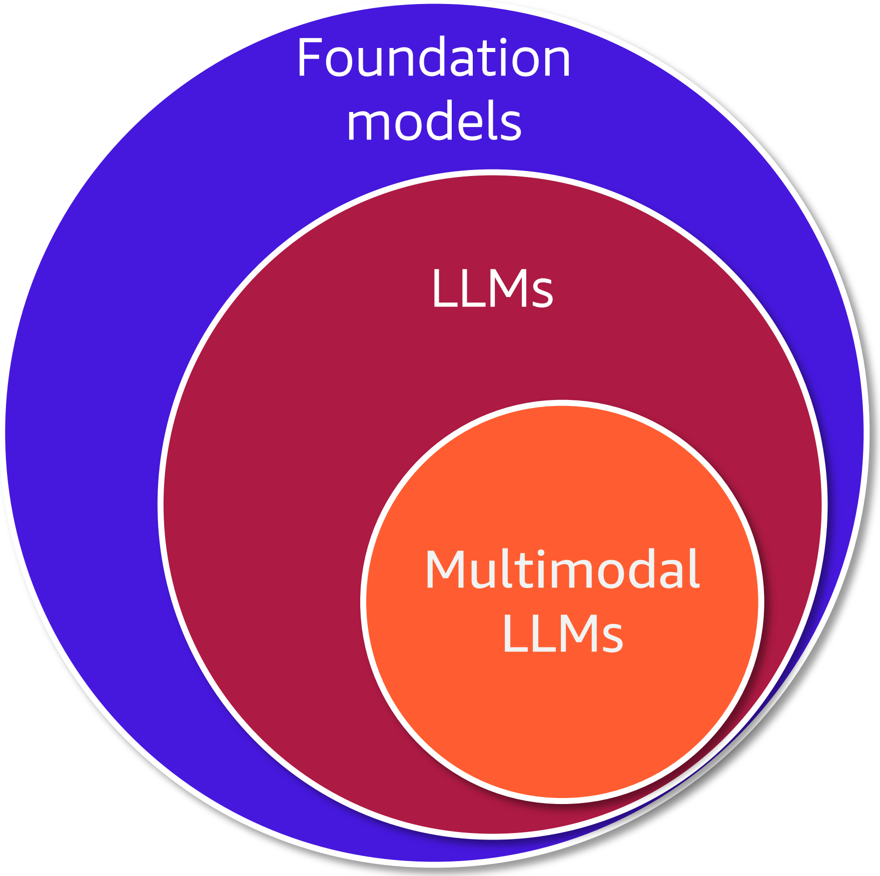
- Foundation models trained on **text**
- Large ML models that learn the **probabilities of words** being used in certain contexts
- **Training task:** Learn to predict the missing word in a text sequence



"The weather has been cloudy for the last two days. Most likely it will be _____ tomorrow."

cloudy? sunny? foggy? ….

Multimodal LLMs (MLLMs)



- Large language models trained on **multiple modalities**

- MLLMs typically use encoders and adapters to Equip LLMs with cross modal capabilities
  - **Vision encoder**
  - **Video encoder**
  - **Audio encoder**

Prompting MLLMs

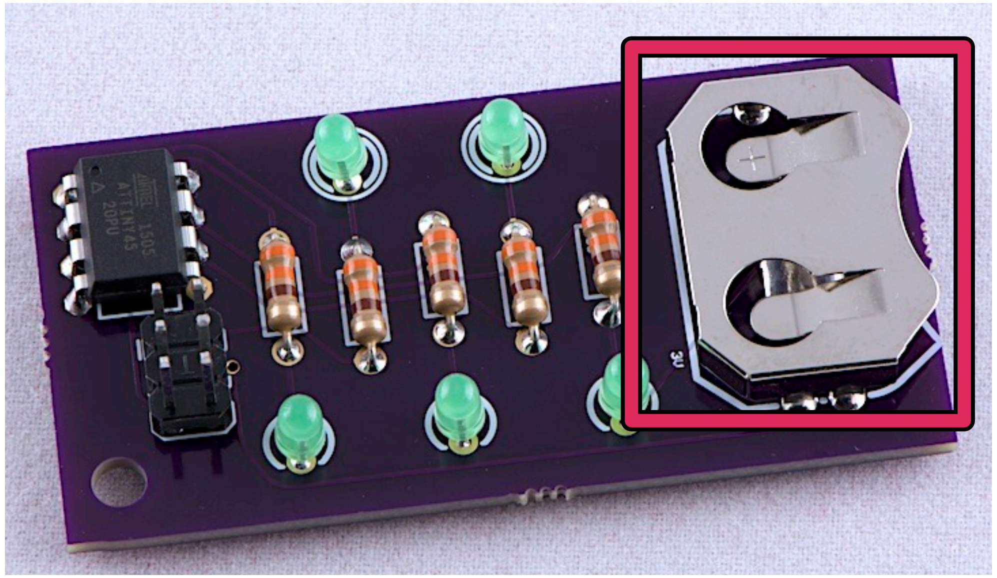
Prompting MLLMs

- **Text** prompts:
    - Follow best prompting strategies discussed in previous lessons
- **Image** prompts:
    - **Input format**: Most MLLMs use base64-encoded format
    - **Image size**: Adhere to the image size limitations (e.g. <5MB)
    - **Multiple images**: Most MLLMs can only analyze a limited number of images
    - **Image format**: Follow the image format specified for the MLLM (e.g. jpg, png, etc.)
    - **Image clarity**: Avoid blurry images
    - **Image placement**: In most cases, it works better when images come before text
    - **Image resolution**: Be within the image resolution limits of the MLLM

Multimodal use cases

Visual question answering

- Instead of relying only on text for the context, you can give the model both text and images
- **Examples**:
  - Generate text descriptions of images
  - Query using both text and images
    - Image analysis using text prompts



What is the purpose of the highlighted part in the circuit board?

Text-based image retrieval

- Image search matters not only for search engines but also for enterprises to be able to search through all their internal images and documents
- **Examples**:
  - Given a text query, find images whose captions/metadata are closest to this text query
  - Given a text query, find all images whose embeddings are closest to this embedding

Find chairs in stock

Can bring images with closest embeddings to the text

In-stock #: 235

In-stock #: 15

Using also image metadata

Deep image similarity retrieval

- Given an image, find similar images
- **Examples**:
  - Retrieving similar images for Amazon products
  - Identifying other product from the manufacturer

Next lesson

- This lesson introduced multimodal models and applications

Thank you!