HOW TO BUILD A CHATBOT

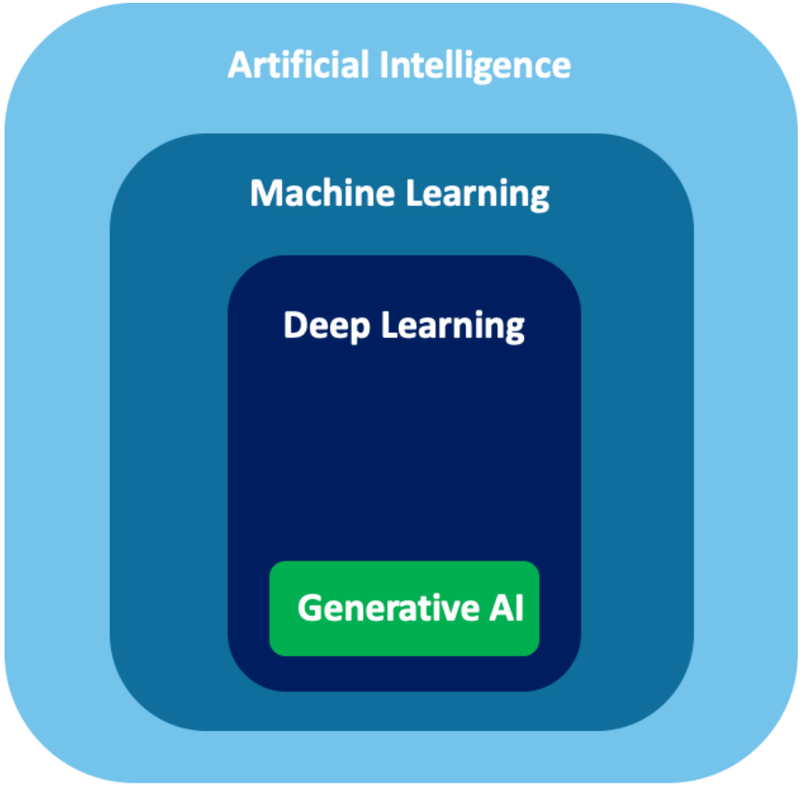
Session 1 - Introduction to LLMs

Author: Martin Kovacs

SESSION 1 AGENDA

1. Introduction to GenerativeAI
2. Exploring Large Language Models
3. Prompt Engineering
4. Deployment and Interaction with LLMs

INTRODUCTION TO GENERATIVE AI



1956

Artificial Intelligence – field in computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

1997

Machine Learning– subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

2017

Deep Learning– a machine learning technique in which deep neural networks are used to process data and make decisions.

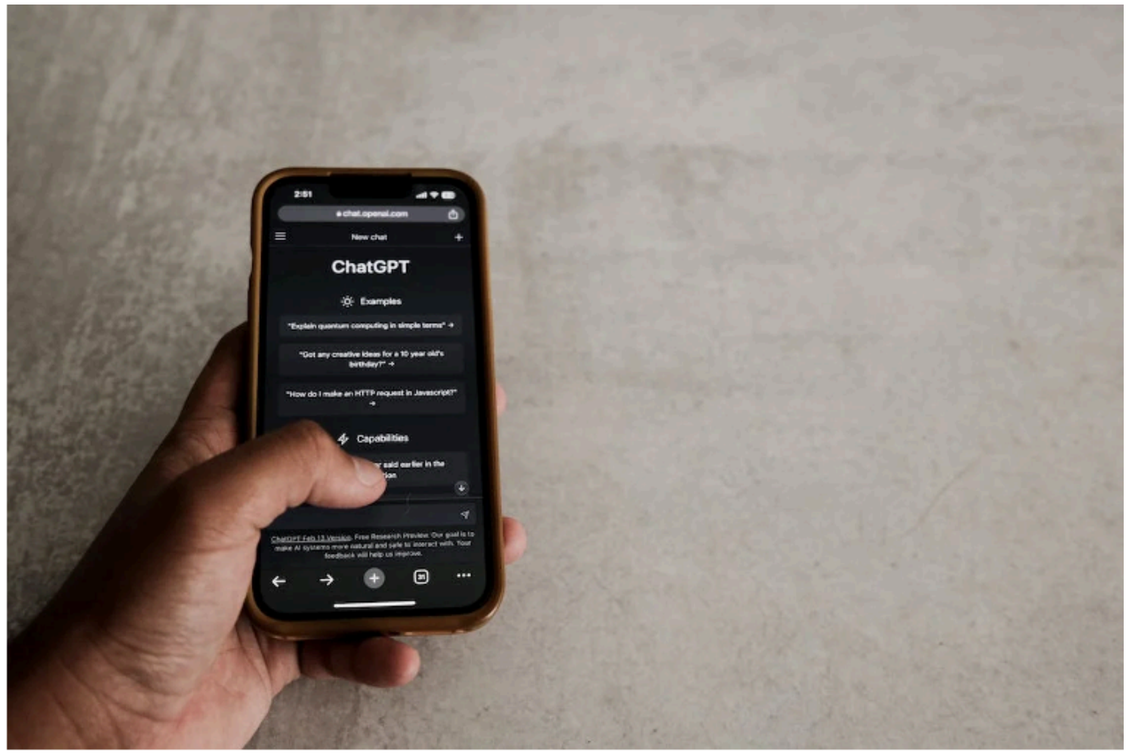
2021

Generative AI– create written or visual content given prompts.

INTRODUCTION TO GENERATIVE AI

"Generative AI is a type of artificial intelligence that creates new content—such as text, images, music, or code—by learning patterns from existing data and generating original outputs based on that knowledge.“

– Answered by ChatGPT



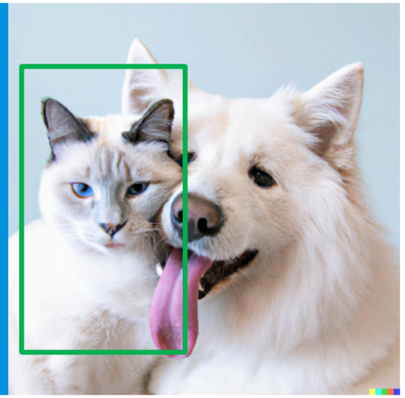
INTRODUCTION TO GENERATIVE AI

Discriminative AI vs Generative AI.

Discriminative AI

"Is there a cat in the image?"

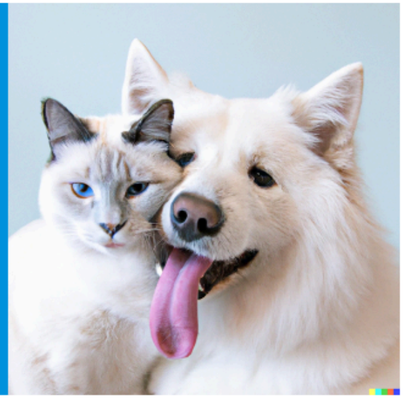
e.g. Image classification / object detection



Generative AI

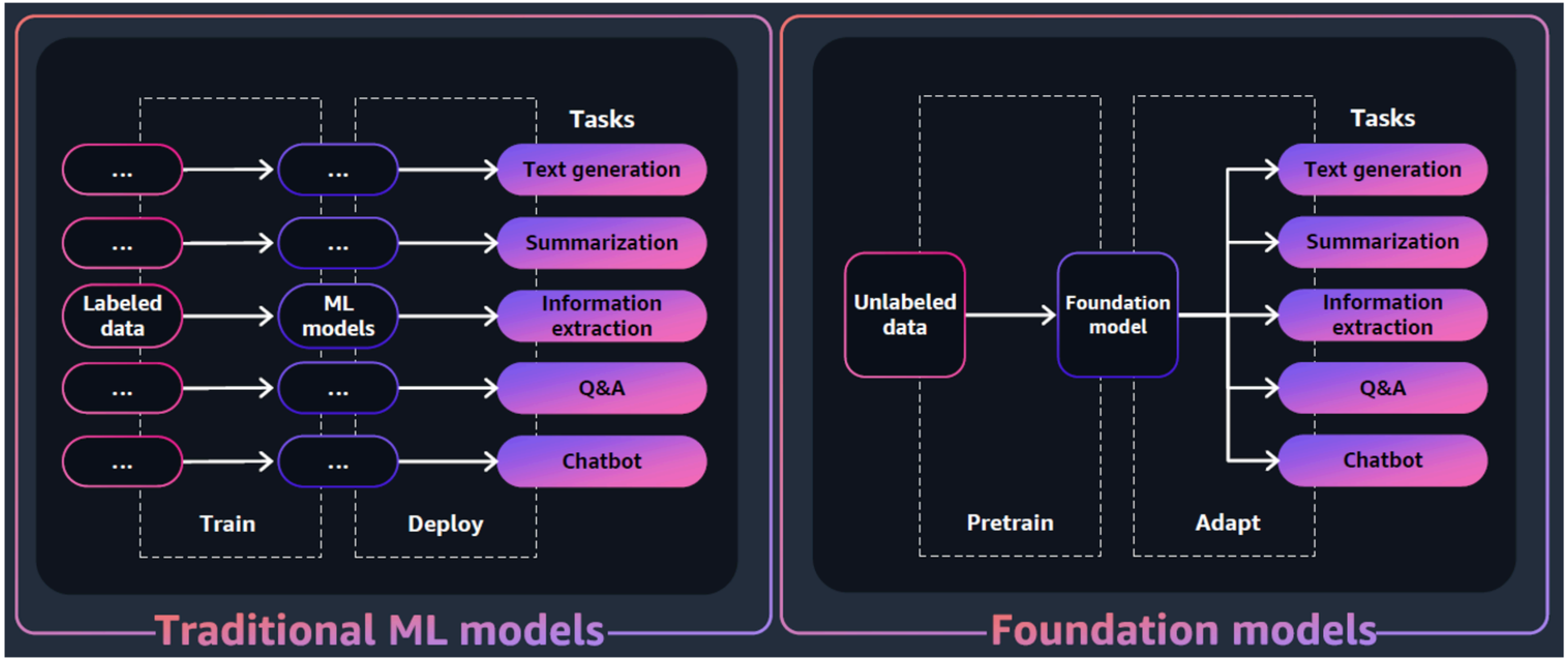
"Draw an image of a dog with its tongue out hugging a white siamese cat."

e.g. Image generation



INTRODUCTION TO GENERATIVE AI

Traditional models vs Foundation models

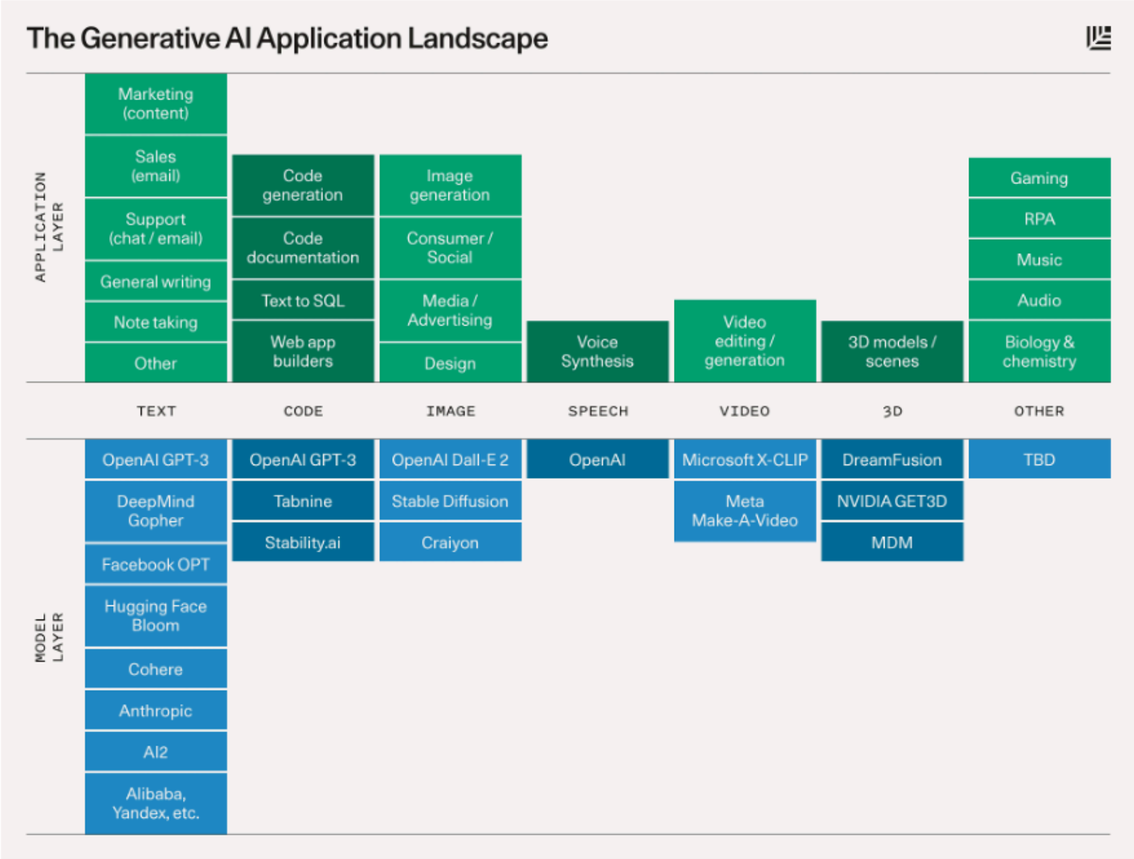


[2]

INTRODUCTION TO GENERATIVE AI

Input & Output Data

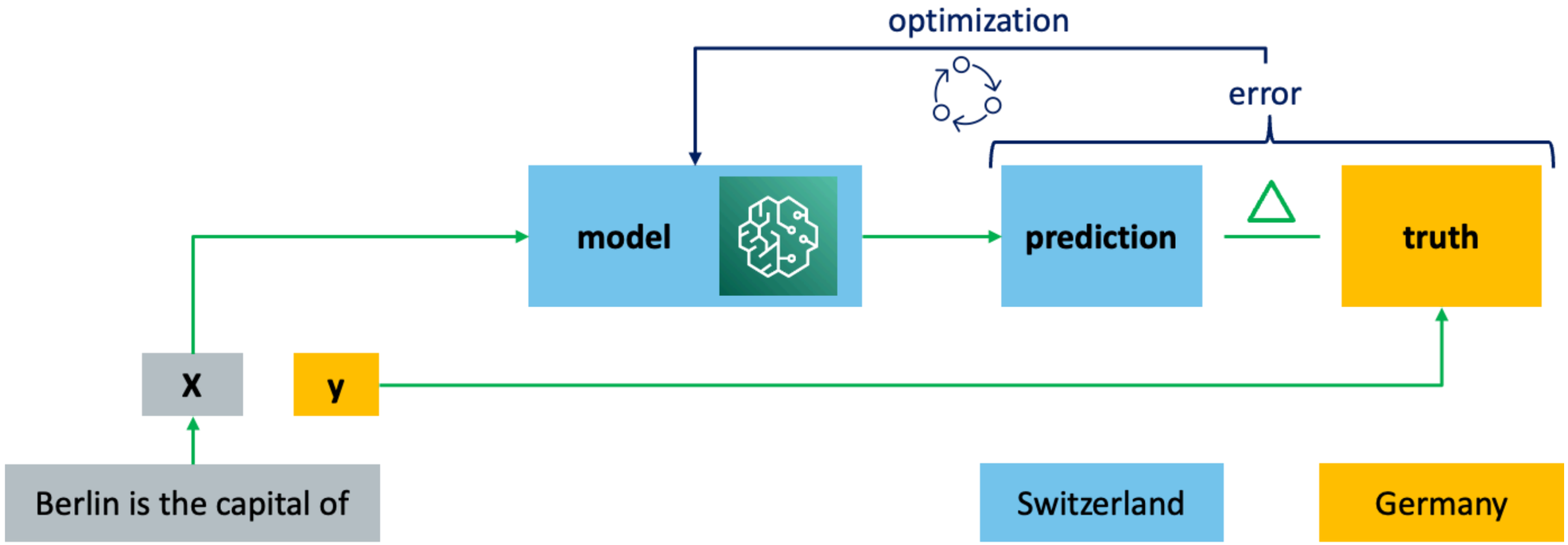
- Text
- Images
- Speech
- Video
- 3D models
- Multimodal (e.g. image and text)



Types of Foundation Models and their applications [1]

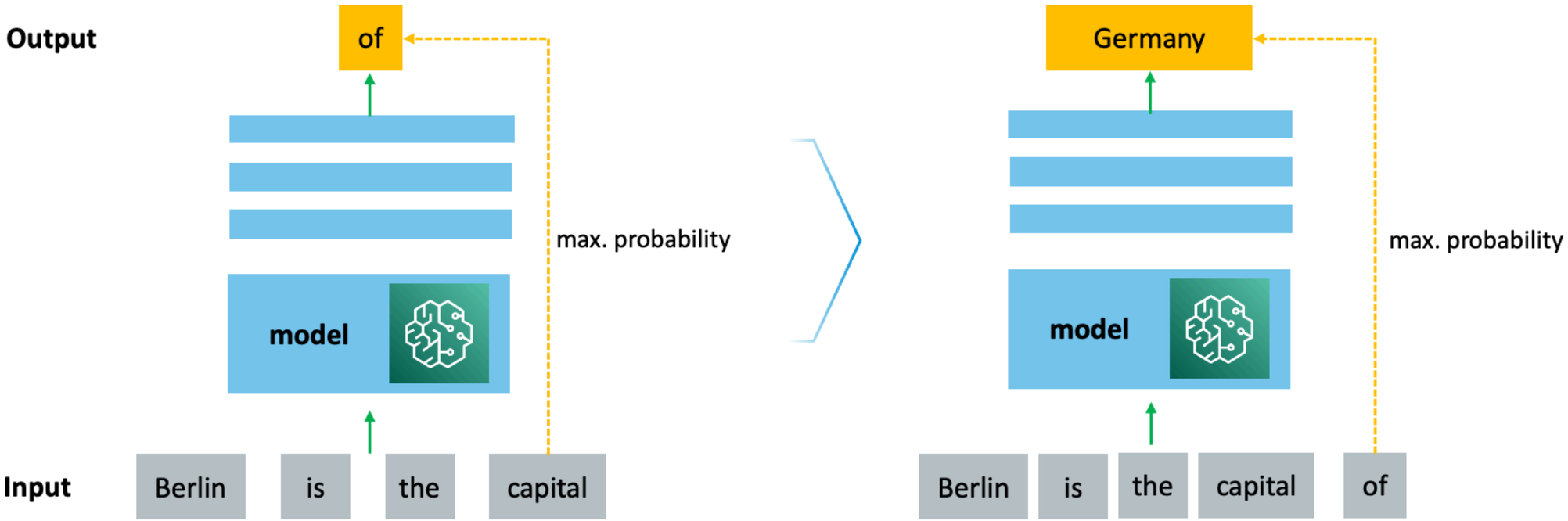
EXPLORING LARGE LANGUAGE MODELS

How LLMs are trained.



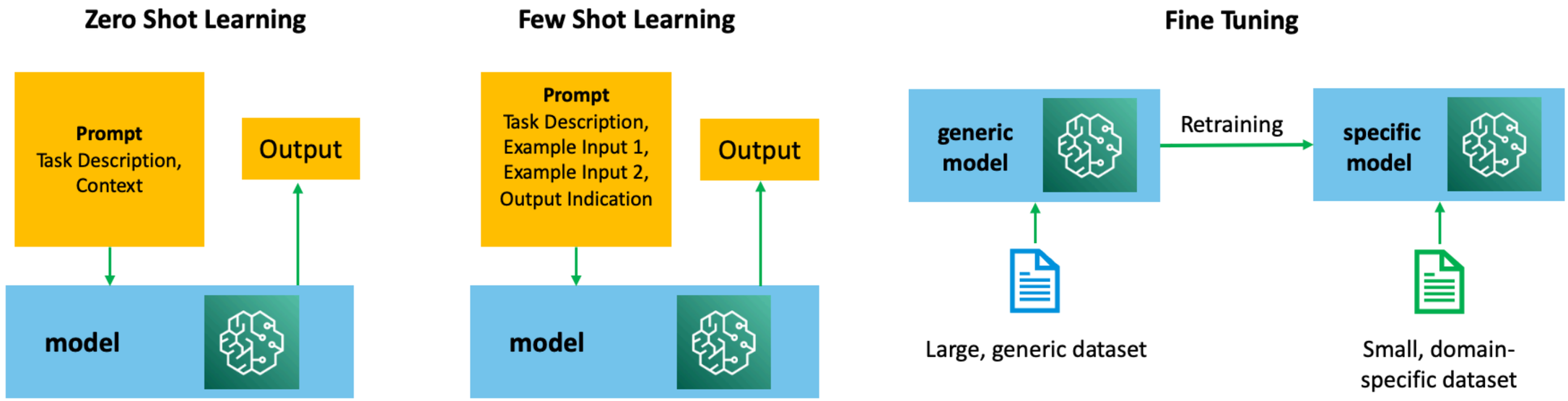
EXPLORING LARGE LANGUAGE MODELS

How LLMs generate (predict) outputs.



EXPLORING LARGE LANGUAGE MODELS

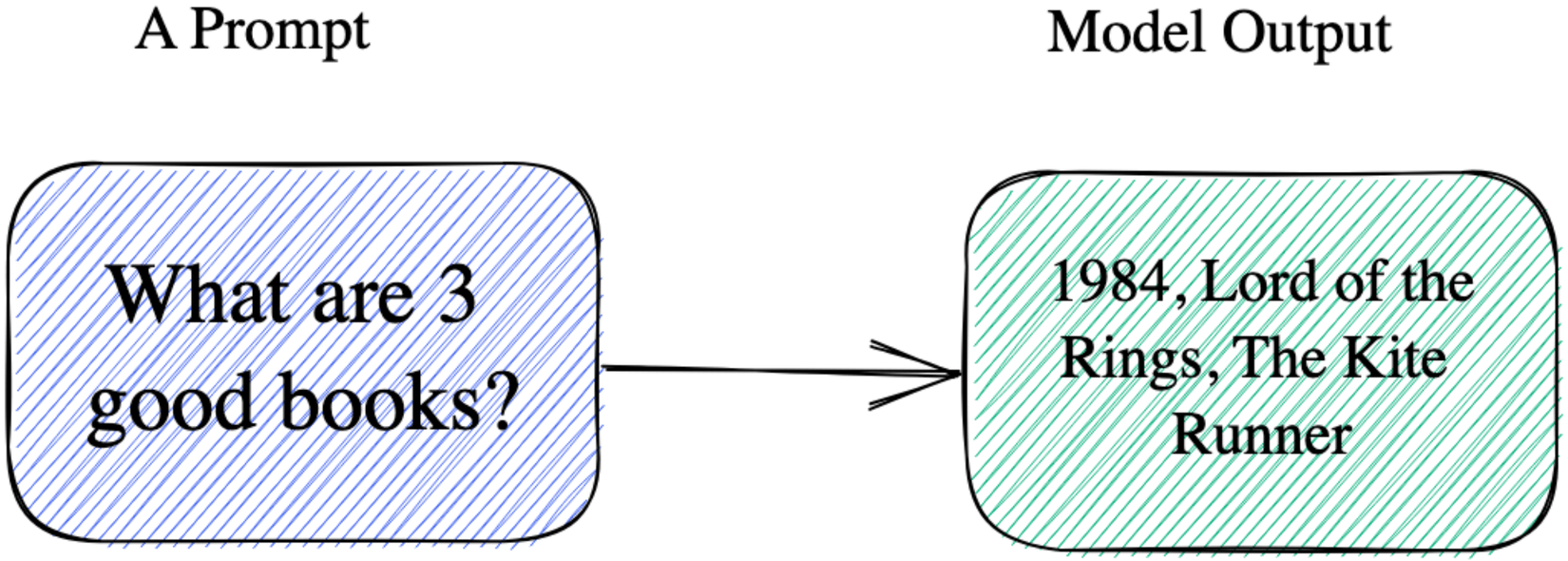
How LLMs can be used.



“A prompt for a LLM is a piece of text that is used to instruct a model to generate a specific output.”

PROMPT ENGINEERING

How to instruct LLMs?



[3]

“A prompt for a LLM is a piece of text that is used to instruct a model to generate a specific output.”

PROMPT ENGINEERING

Building blocks of a prompt:

Role / Context:

- This defines the desired persona or perspective of the AI response.
- Context information about the situation or task

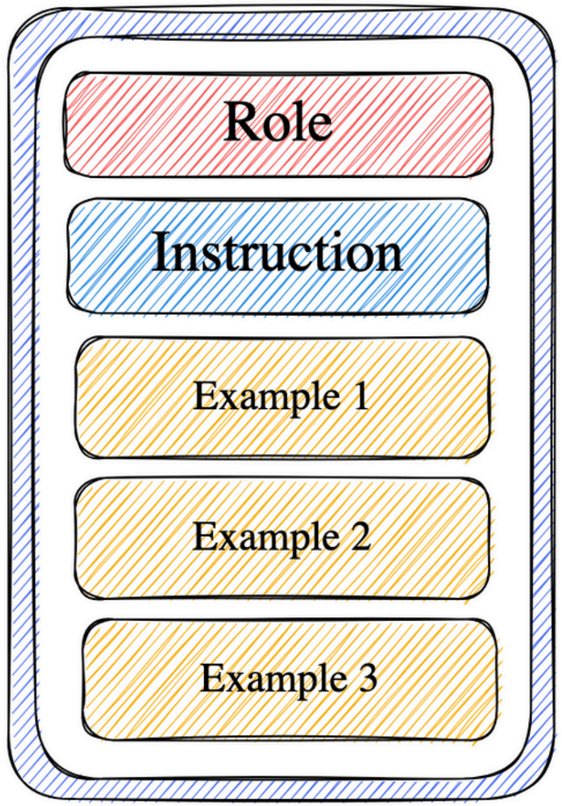
Instruction:

- This specifies the task or goal you want the AI to accomplish.

Examples:

- These provide additional context or guidance to help the AI understand the desired output.

A Combined Techniques Prompt



[3]

PROMPT ENGINEERING

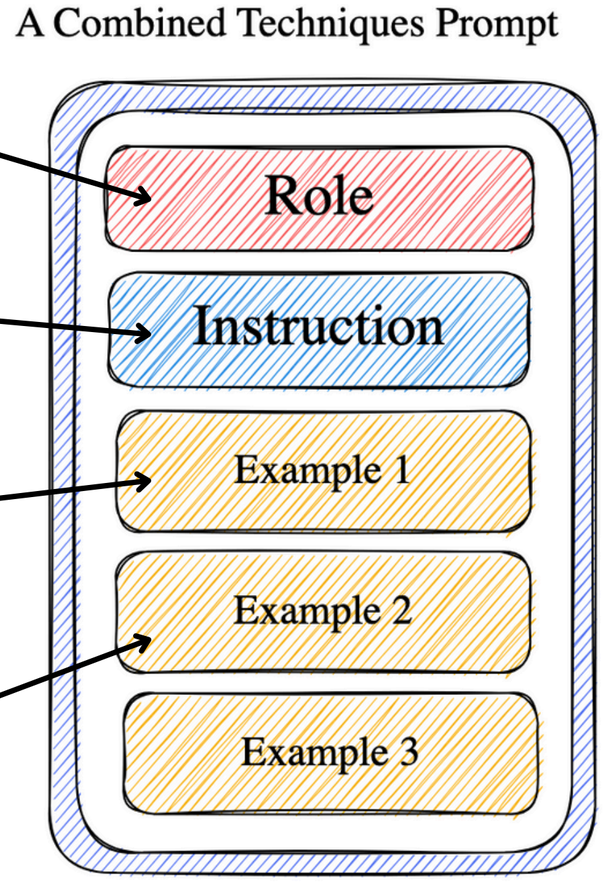
A simple prompt example.

Act as a knowledgeable AI tutor specializing in machine learning, tailored to the curriculum of a university of applied science.

Please explain in machine learning related topic in a easy understandable way

Q: Explain the difference between supervised and unsupervised learning.
A: Supervised learning uses labeled data to train a model to make predictions, while unsupervised learning ...

Q: What are the advantages and disadvantages of using decision trees for classification?
A: Decision trees are easy to interpret and can handle both numerical and categorical data. However, they can be prone to overfitting and ...

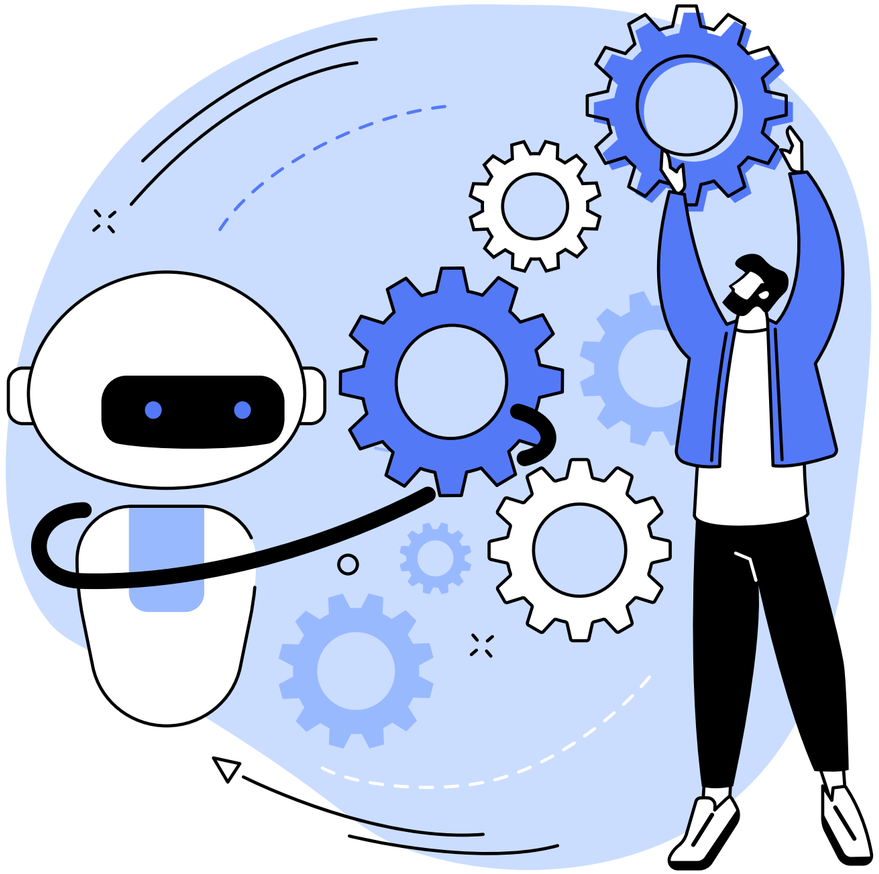


[3]

PROMPT ENGINEERING

Basic use cases of LLMs.

- Summarize text
- Condense text
- Generate code
- Explain code
- Bug fixing

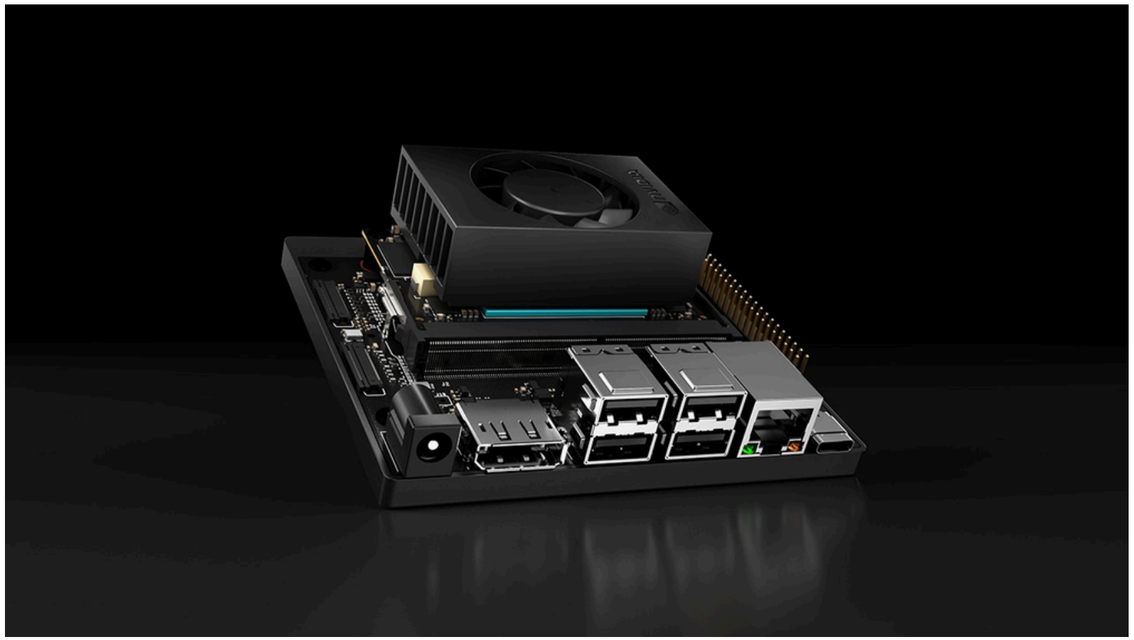


- Brainstorming
- Creative writing
- Data analysis
- Entity Extraction
- Sentiment analysis

DEPLOYMENT AND INTERACTION WITH LLMS

Different ways to deploy LLMs:

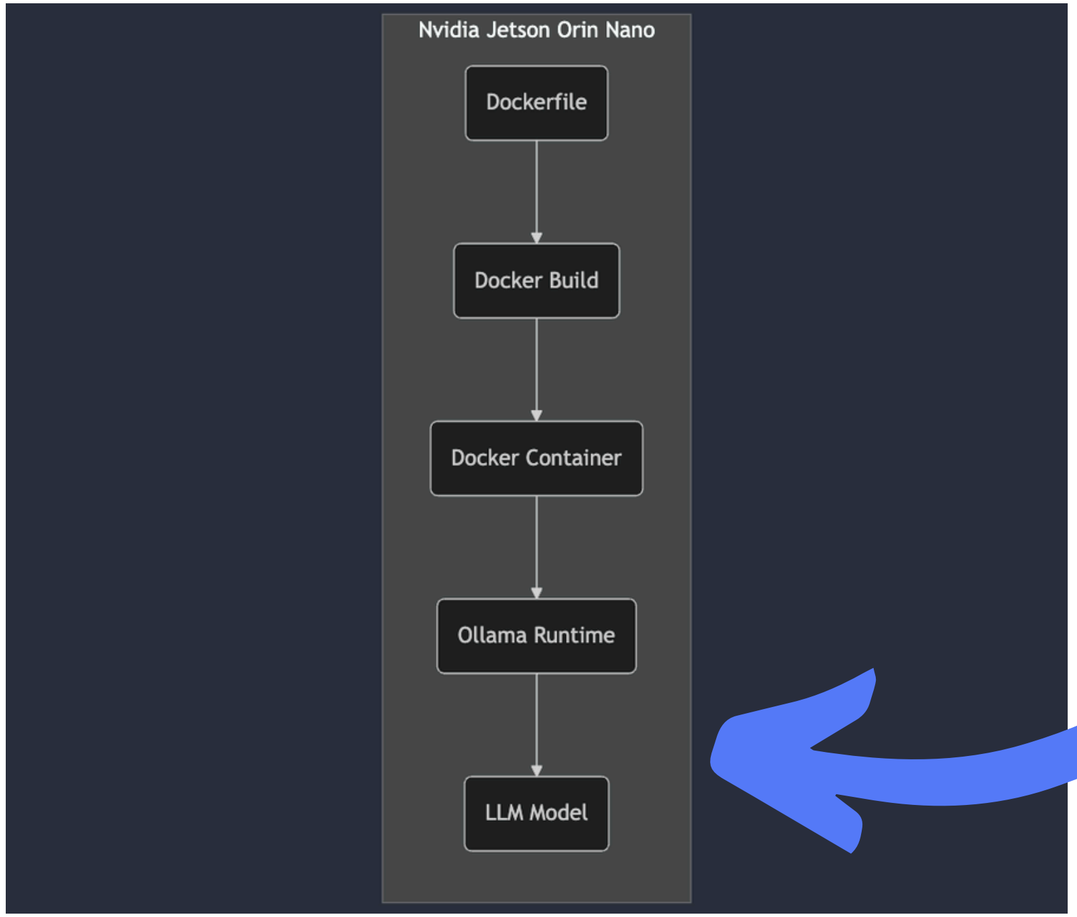
- Cloud (e.g. AWS, Azure, GCP)
- On-premise (e.g. local infrastructure)
- **Edge devices (e.g. Nvidia Jetson)**



[4]

DEPLOYMENT AND INTERACTION WITH LLMS

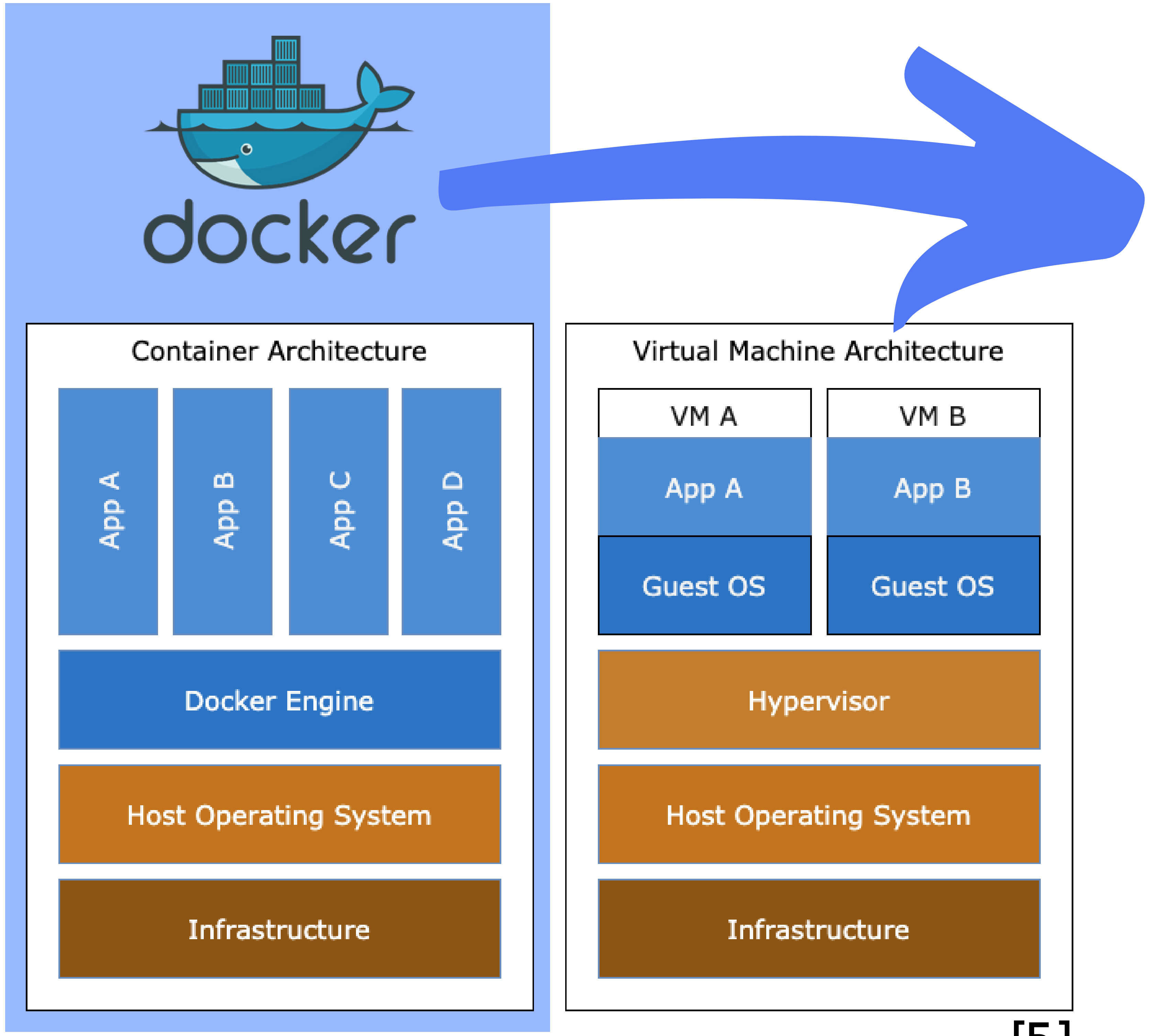
Deploy LLM with Ollama & Docker on Nvidia Jetson Orin Nano



Ollama is an open-source project that serves as a powerful and user-friendly platform for running LLMs on your local machine.

DEPLOYMENT AND INTERACTION WITH LLMS

Docker Excursion



[5]

```
nvidia@jao-60:/$ jetson-containers run $(autotag ollama) ollama run mistral
Namespace(packages=['ollama'], prefer=['local', 'registry', 'build'], disable=[''], user='dustynv
', output='/tmp/autotag', quiet=False, verbose=False)
-- L4T_VERSION=36.2.0  JETPACK_VERSION=6.0  CUDA_VERSION=12.2
-- Finding compatible container image for ['ollama']
cu122/ollama:r36.2.0
+ docker run --runtime nvidia -it --rm --network host --volume /tmp/argus_socket:/tmp/argus_socke
t --volume /etc/enctune.conf:/etc/enctune.conf --volume /etc/nv_tegra_release:/etc/nv_tegra_relea
se --volume /tmp/nv_jetson_model:/tmp/nv_jetson_model --volume /var/run/dbus:/var/run/dbus --volu
me /var/run/avahi-daemon/socket:/var/run/avahi-daemon/socket --volume /var/run/docker.sock:/var/r
un/docker.sock --volume /mnt/NVME/jetson-containers/dev/data:/data --device /dev/snd --device /de
v/bus/usb --device /dev/video0 --device /dev/video1 cu122/ollama:r36.2.0 ollama run mistral
pulling manifest
pulling e8a35b5937a5...  73%                          | 3.0 GB/4.1 GB   34 MB/s
```

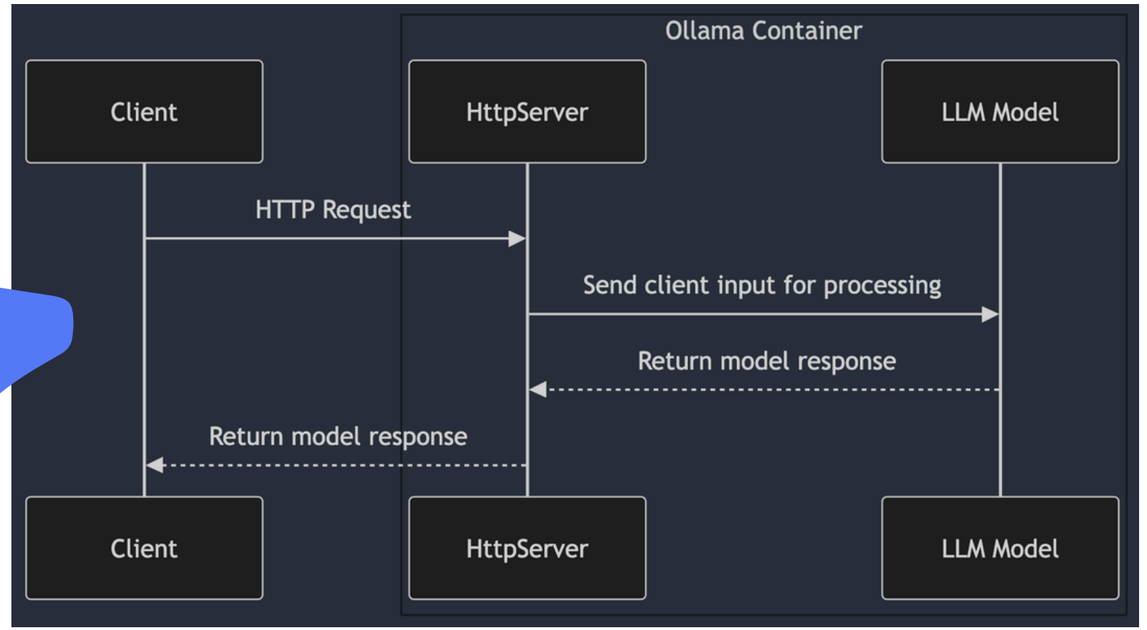
DEPLOYMENT AND INTERACTION WITH LLMS

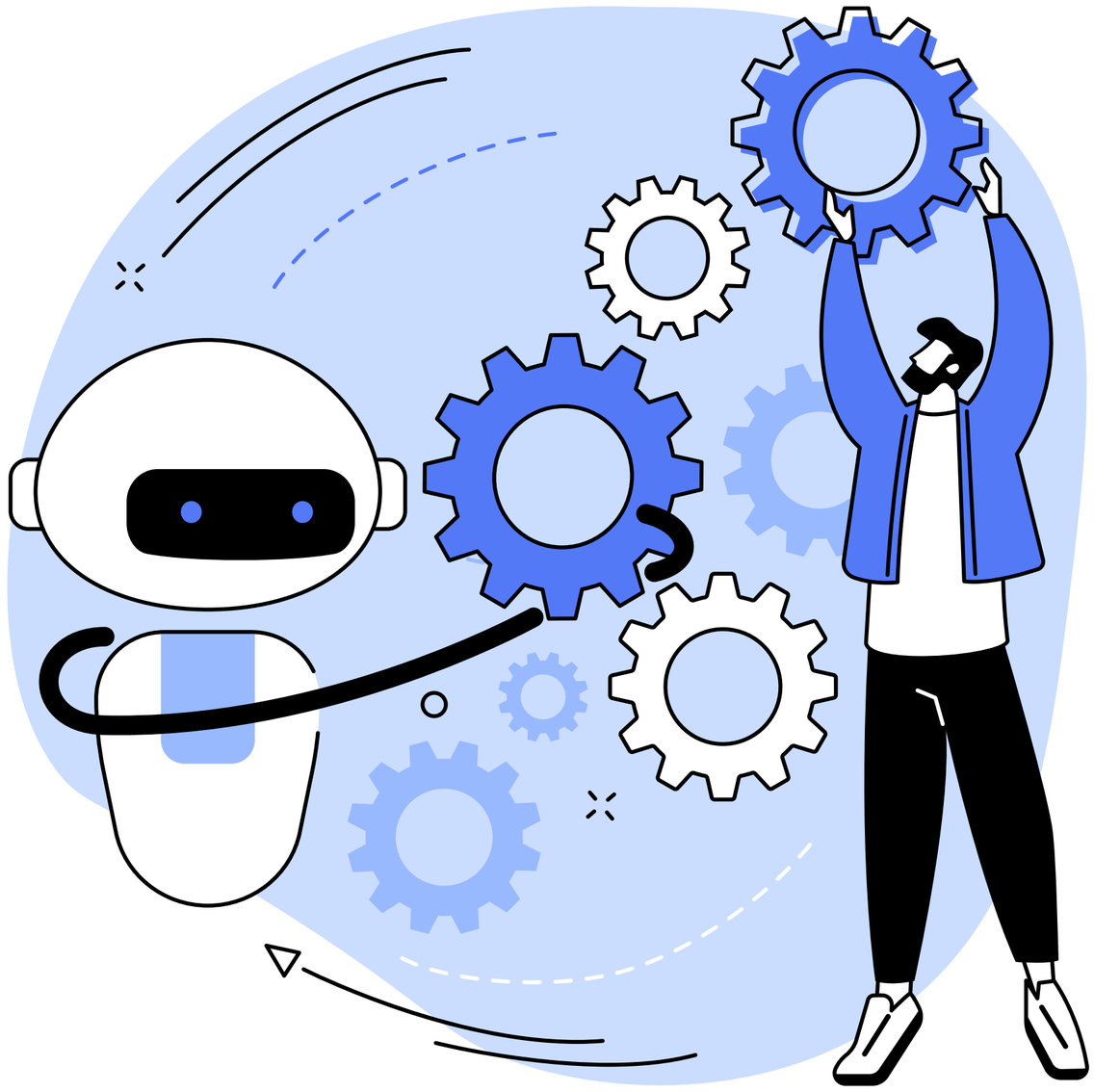
```python
# Simple HTTP Request via requests

# Define the URL of the deployed LLM
url = "http://localhost:11434/api/generate"

# Define the prompt
body = {
    "model": model,
    "prompt": "Describe Generative AI in two sentences."
}

# Send the POST request
response = requests.post(url, json=body)
```





IT’S YOUR TURN

Sources:

[1]: Generative AI and Innovation Management - investigating the impact of generative AI on creativity and innovation in organizations. - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Generative-AI-application-landscape-Huang-Grady-2022_fig5_370761753 [accessed 14 Sept 2024]

[2]: Generative AI and Innovation Management - investigating the impact of generative AI on creativity and innovation in organizations. - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Generative-AI-application-landscape-Huang-Grady-2022_fig5_370761753 [accessed 14 Sept 2024]

[3]: https://learnprompting.org/de/docs/basics/prompting

[4]: https://www.antratek.de/media/opti_image/avif/catalog/product/cache/0c2253ca5cb32d2cfd90eba2caa6b5a5/n/v/nvidia-jetson-orin-nano-developer-kit-2c50-d.avif

[5]: https://bitovi.github.io/academy/static/img/docker/1-what-is-docker/docker-arch.png