Chat with PDF Document using GenAI

- Generative AI Fundamentals
- Q&A System Architecture

Generative AI

GAN - Generative Adversarial Network

VAE - Variational Autoencoder

GPT - Generative Pre-trained Transformer

LSTM - Long Short-Term Memory

AR - Autoregressive Model

AI - Artificial Intelligence

SVM - Support Vector Machine

Machine Learning

EDA - Exploratory Data Analysis

ML - Machine Learning

RL - Reinforcement Learning

MSE - Mean Squared Error

DL - Deep Learning

Data Science

NLP - Natural Language Processing

PCA - Principal Component Analysis

AUC - Area Under the Curve

ROC - Receiver Operating Characteristic

NER - Named Entity Recognition

RNN - Recurrent Neural Network

Natural Language Processing

CNN - Convolutional Neural Network

POS - Part of Speech

BERT - Bidirectional Encoder Representations from Transformers

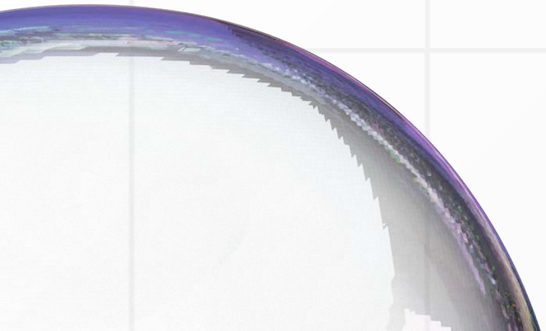
IoU - Intersection over Union

AR - Augmented Reality

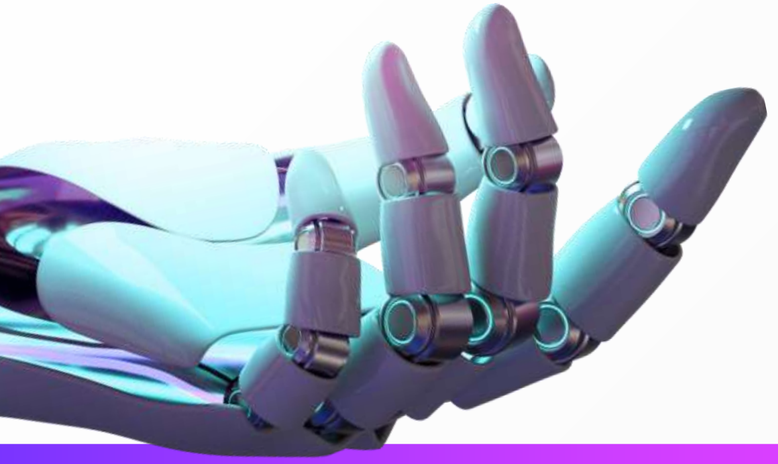
GAN - Generative Adversarial Network

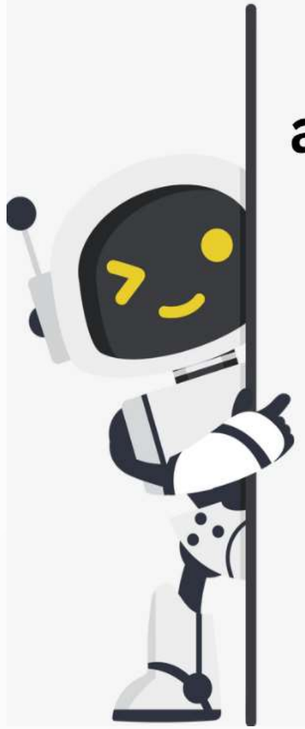
Computer Vision

OCR - Optical Character Recognition

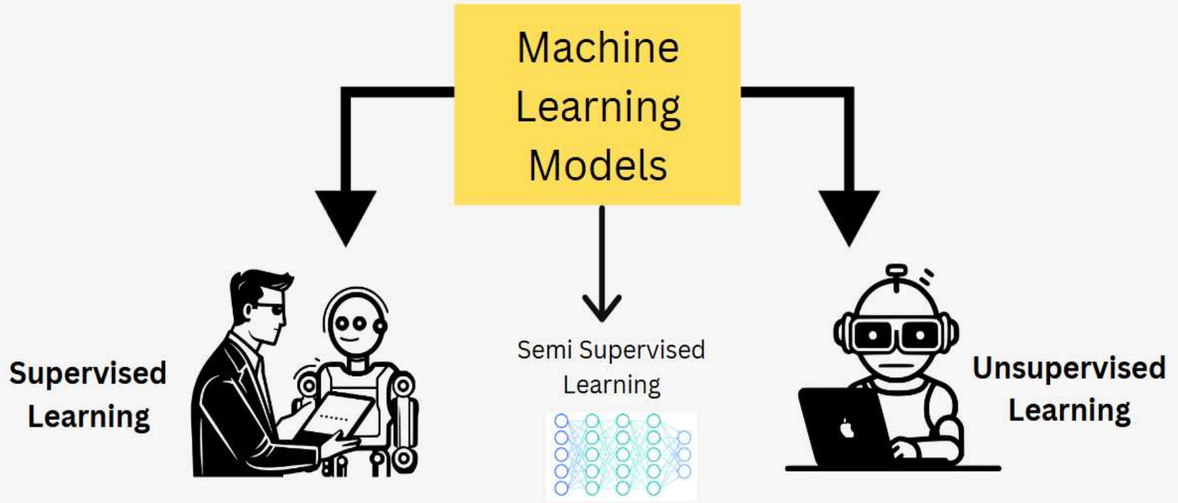


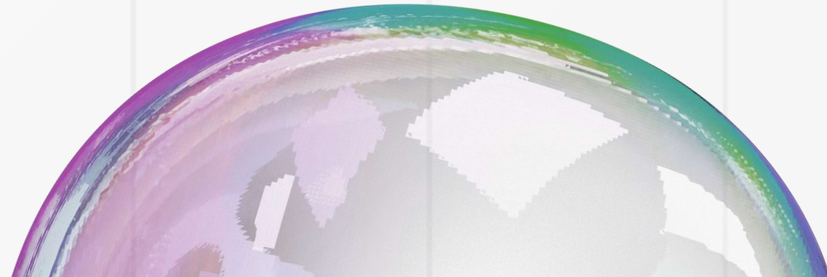
Machine Learning

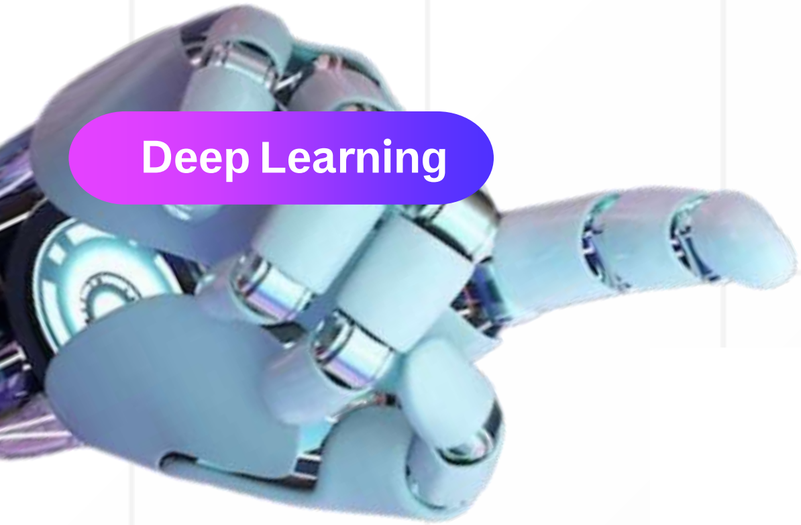




ML is a subfield of AI and gives computers the ability to learn without explicitly programming.

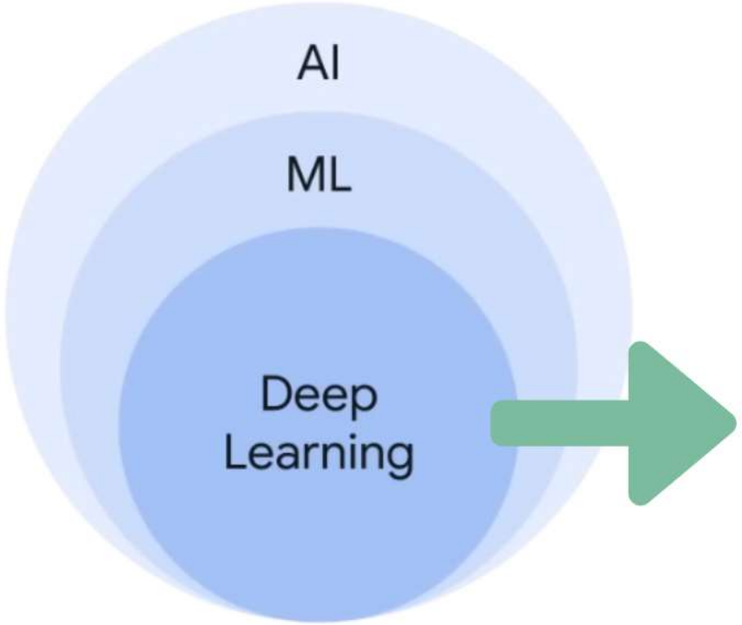


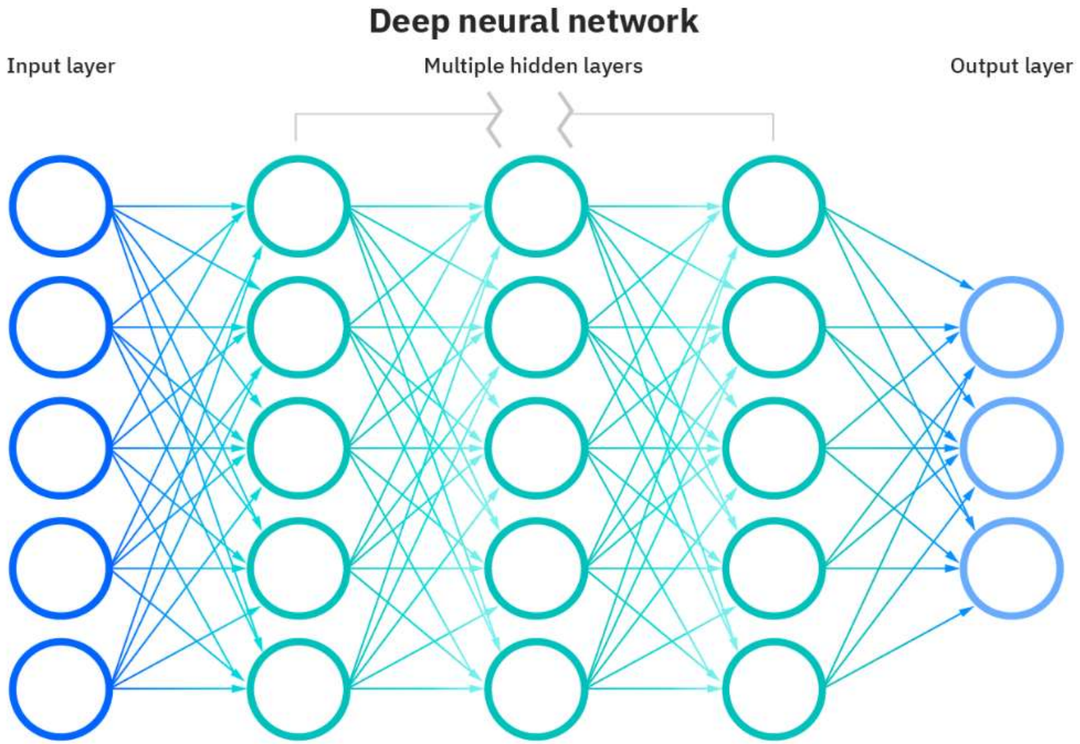


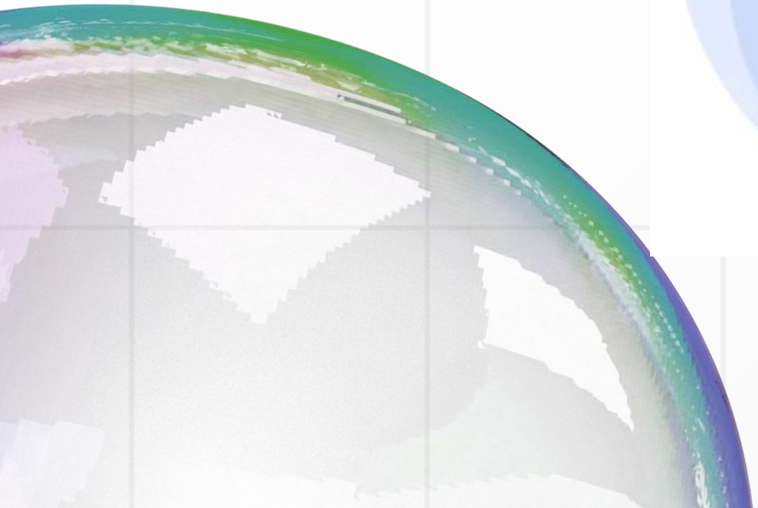


Deep Learning

Deep Learning

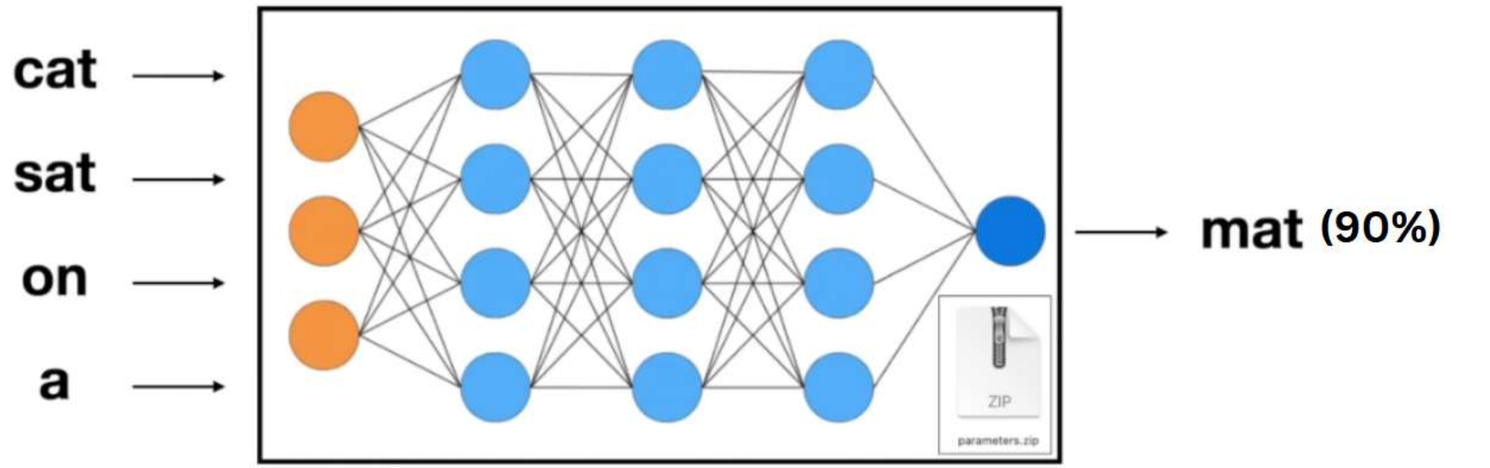






Neural Network

Predicts the next word in the sequence.

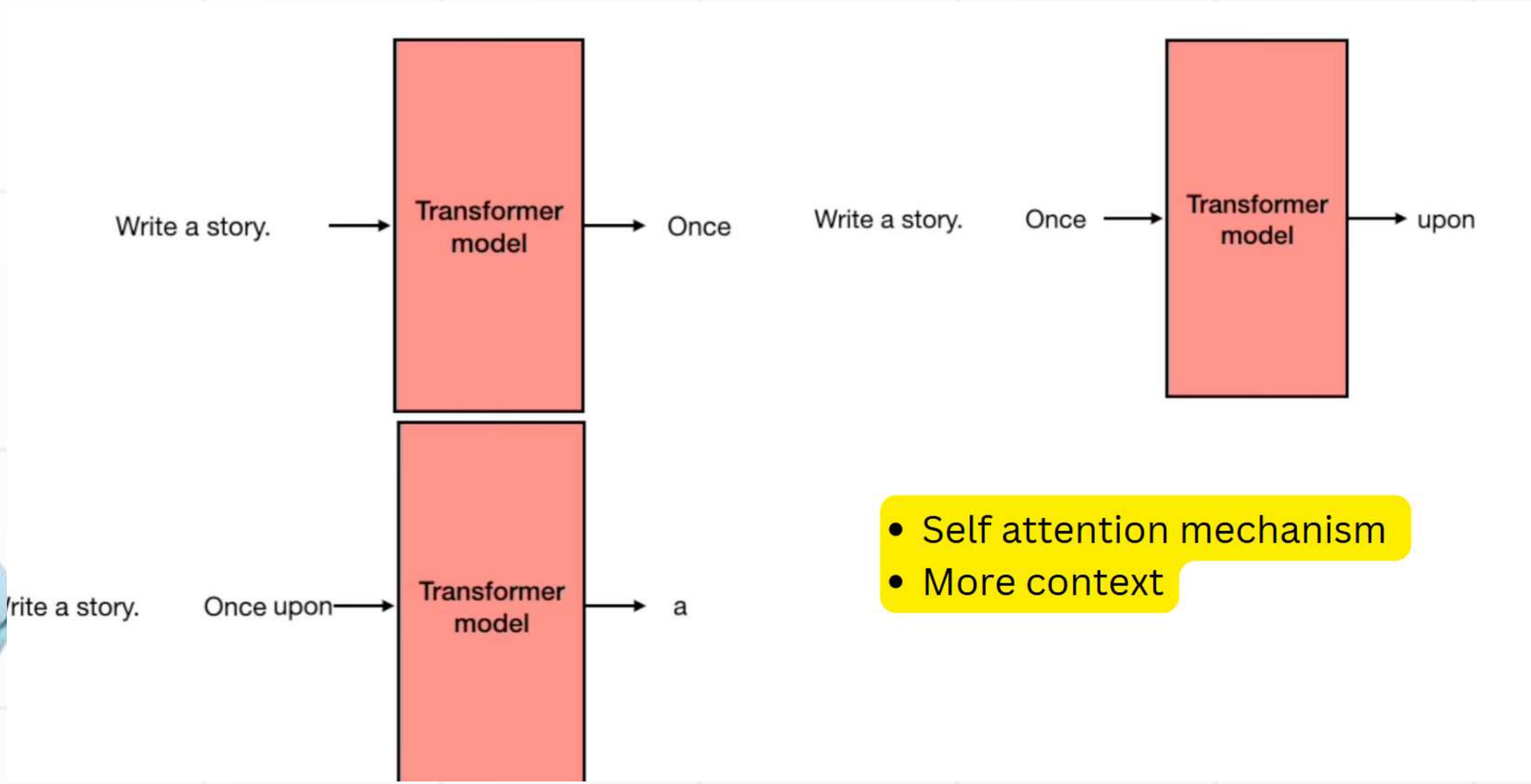


e.g. context of 4 words

predict next word

Transformer

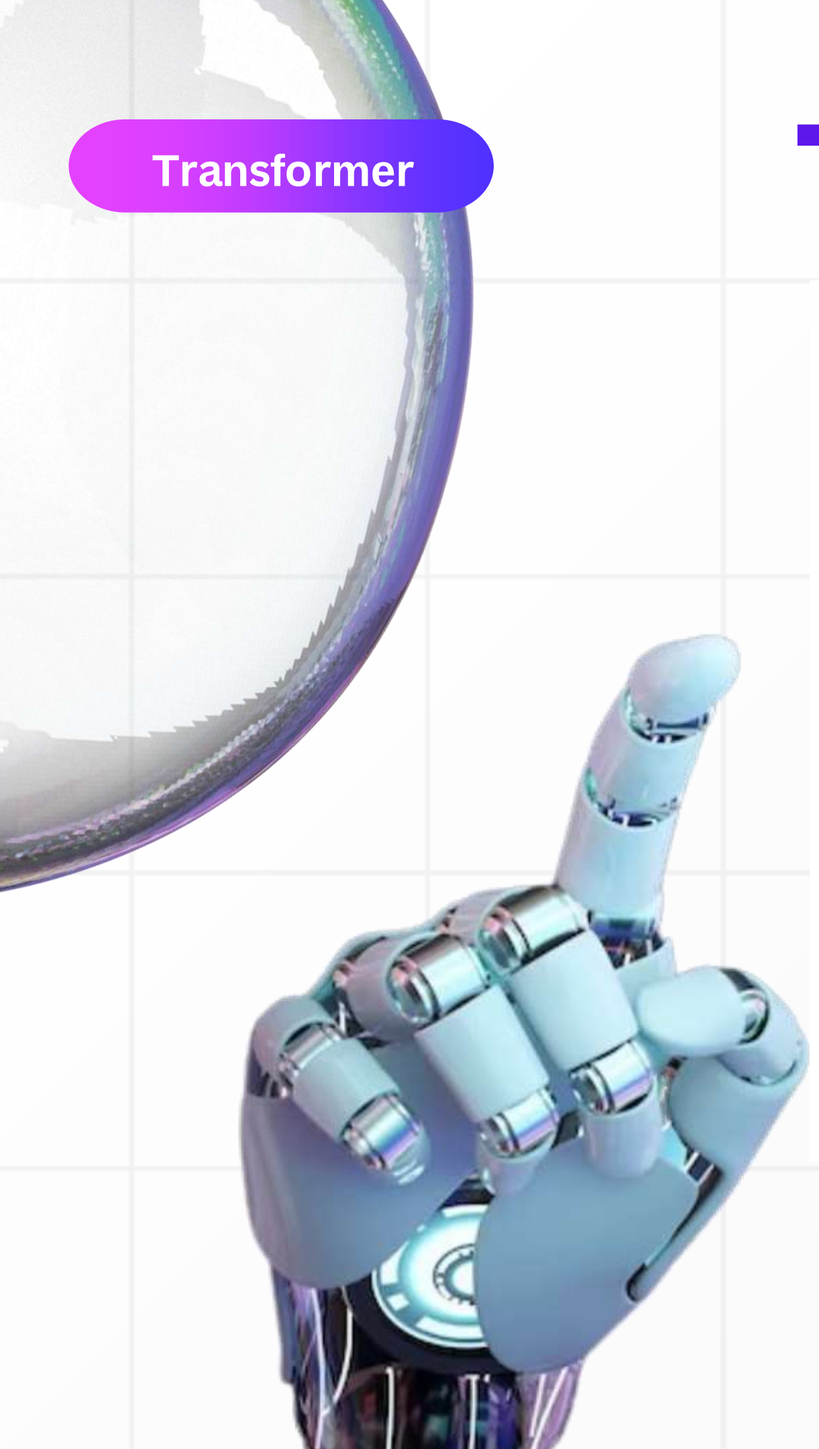
Transformer Model



- Self attention mechanism
- More context

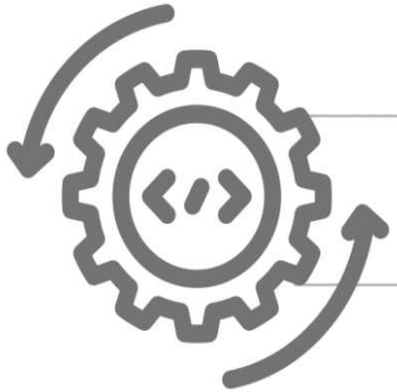
Transformer

Transformer Model



My				→	name
My	name			→	is
My	name	is		→	Sylvain
My	name	is	Sylvain	→	.

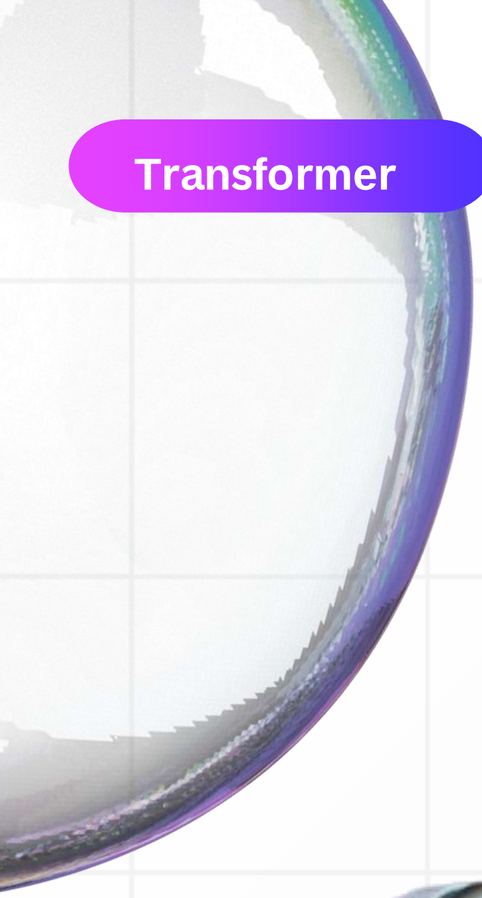
Transformer Model

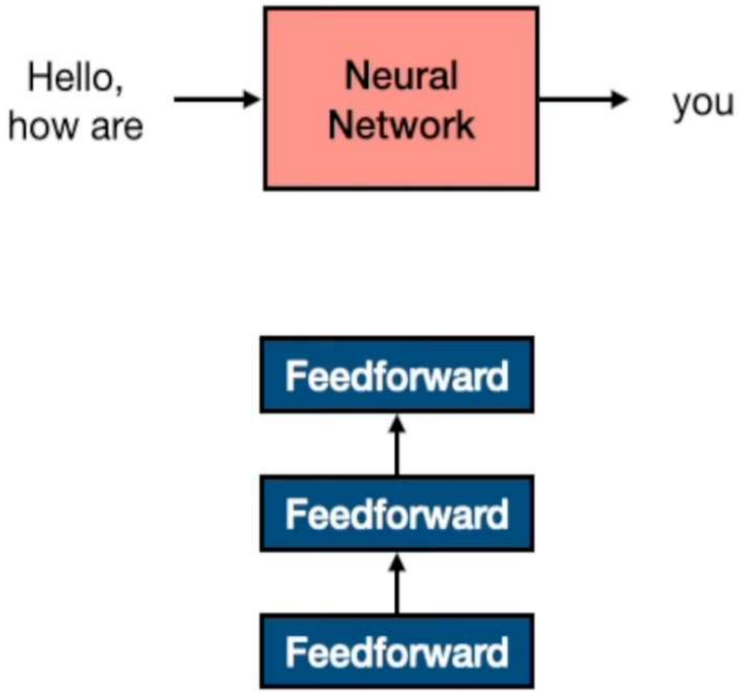


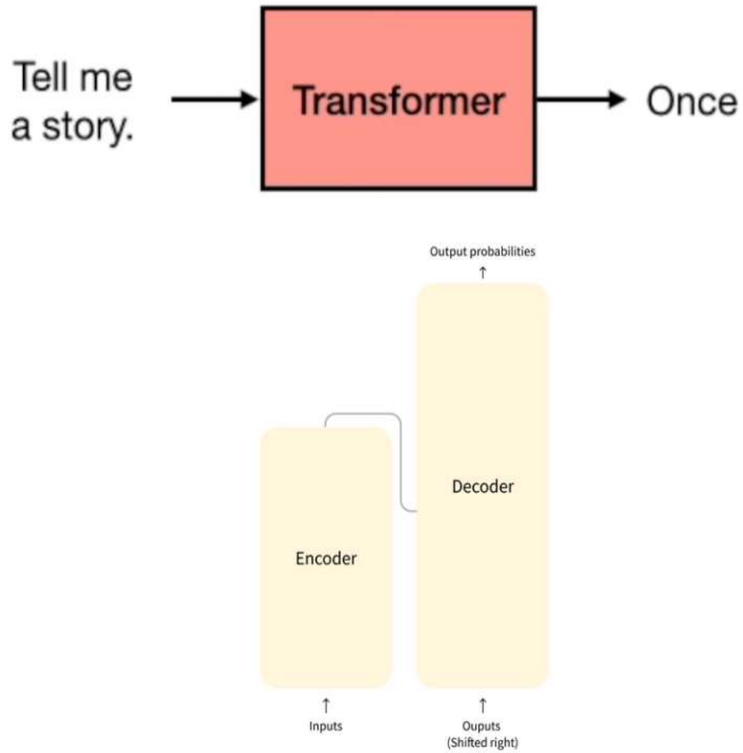
with almost 99% accuracy

Transformer

Neural Network vs Transformer Model

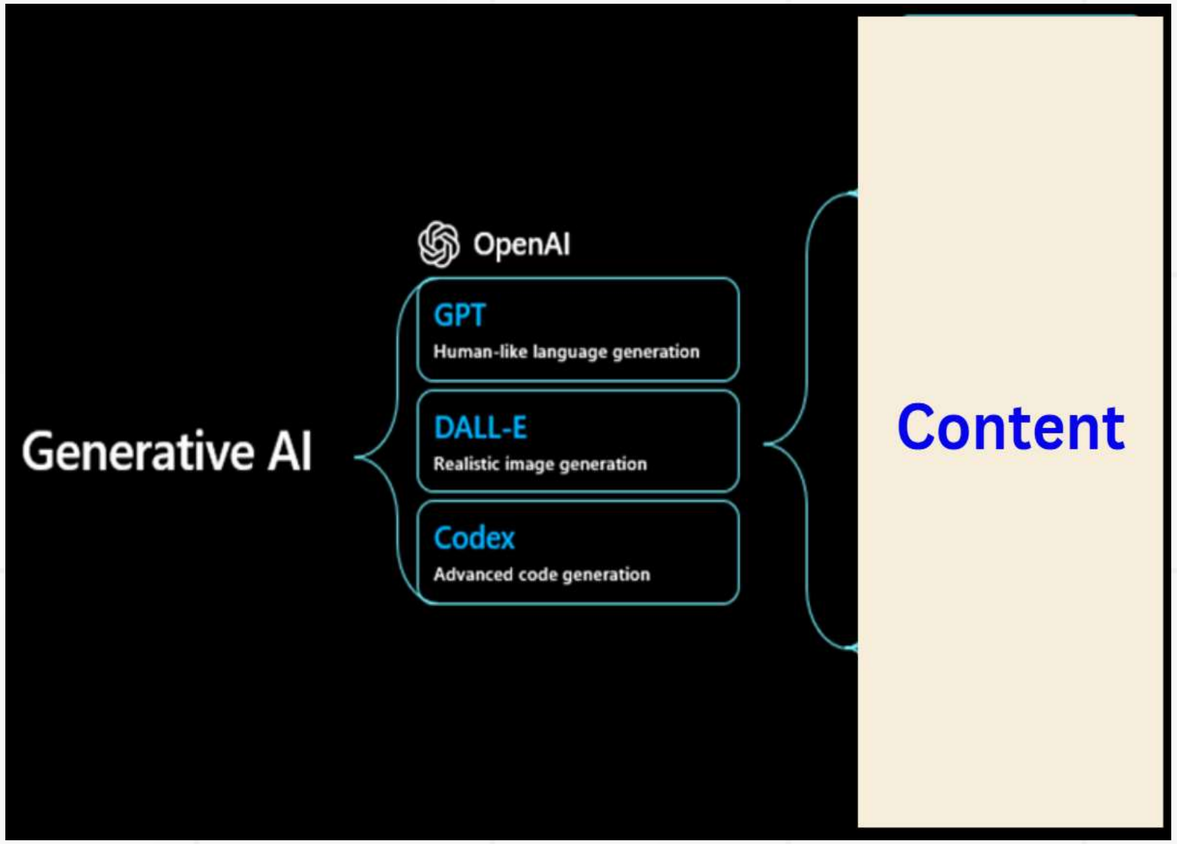






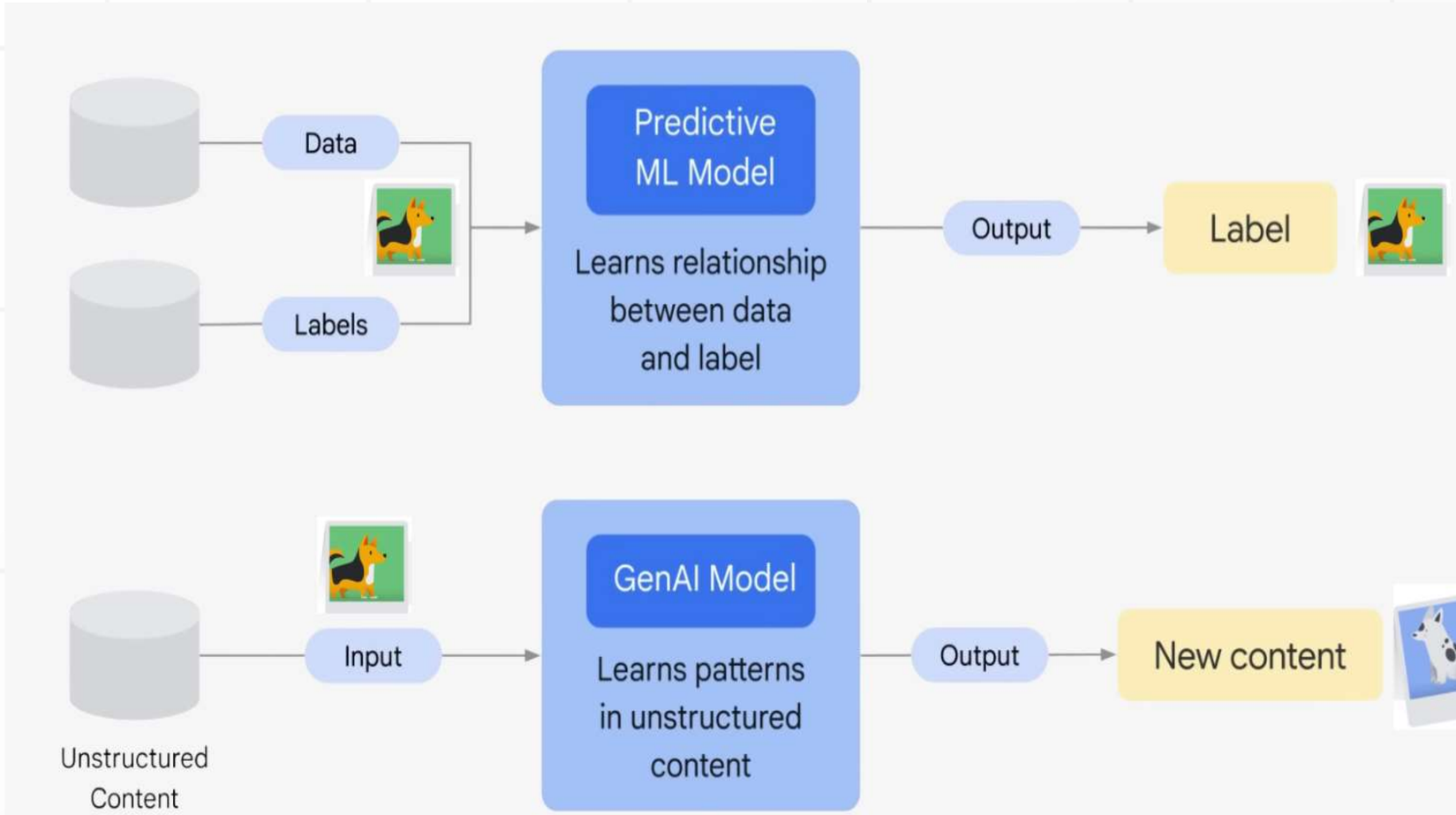
Generative AI

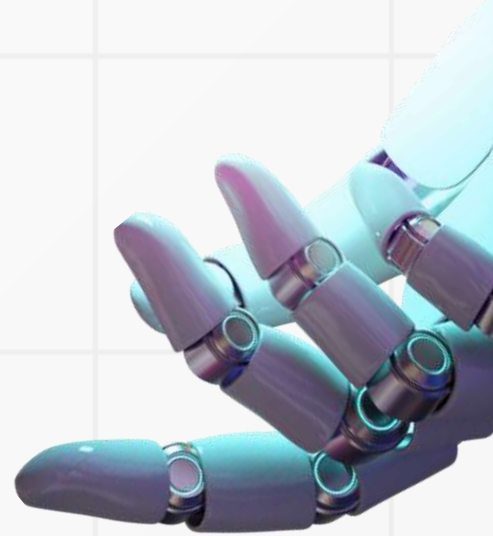
Generative AI is a type of artificial intelligence technology that can produce various types of content, including text, imagery, audio and synthetic data.

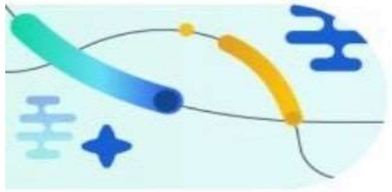


Transformer

Traditional Model vs Generative ai models







Evolution of AI Architecture:
Traditional ML to Generative AI

Traditional ML

Data Pre-Processing
Cleaning and preparing data for analysis

Feature Engineering
Extracting important features from data

Training & Tuning
Training models on data and adjusting parameters for optimal performance

Deployment & Monitoring
Implementing models in real-world applications and monitoring their performance

Generative AI

Data Pre-Processing
Cleaning and preparing data for analysis

Prompt Engineering/Fine Tuning
Designing effective prompts to guide AI in generating desired outputs

Foundational/Fine-Tuned LLM
Using foundation and Fine-tuned language models for sophisticated content generation

Deployment & Monitoring
Implementing models in real-world applications and monitoring their performance

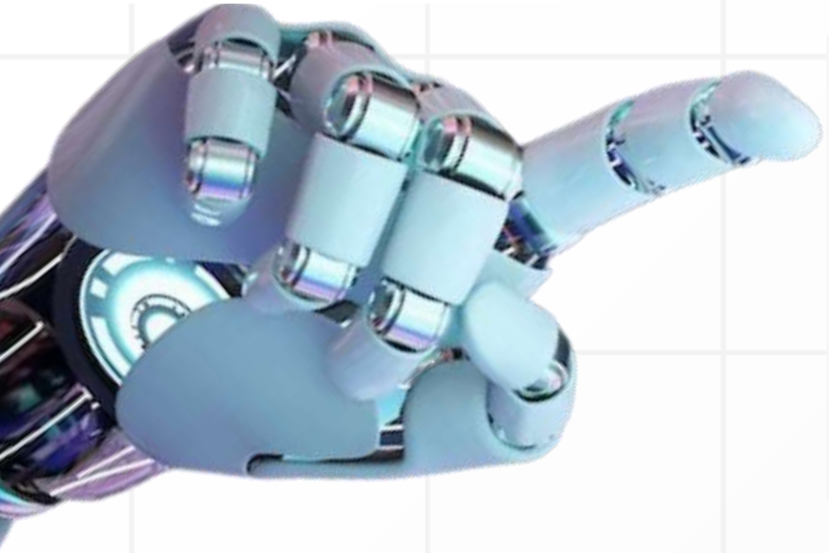
Tech Stack for Traditional ML

- **ML Frameworks:** Keras, Theano
- **ML API's & SDK:** IBM Watson
- **Database:** SQL Server, Oracle
- **ML Ops:** Docker, Jenkins

Tech Stack for Generative AI

- **Gen AI Orchestration:** Langchain, llamaindex
- **LLM Models:** OpenAI, Anthropic
- **Vector Database:** Pinecone, Weaviate
- **LLM Ops:** Prompt Layer, Helicone

Large Language Models



Gemini

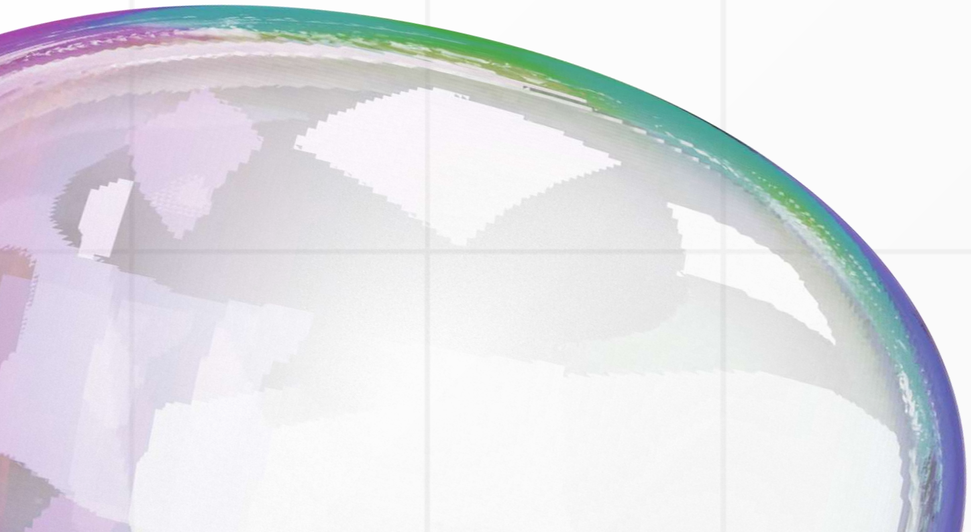
ChatGPT

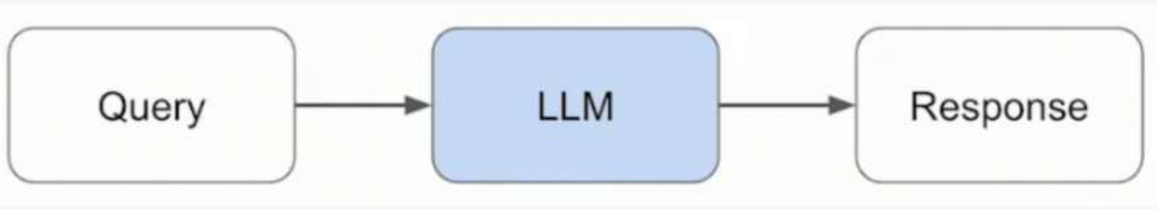
DALL·E 2

Midjourney

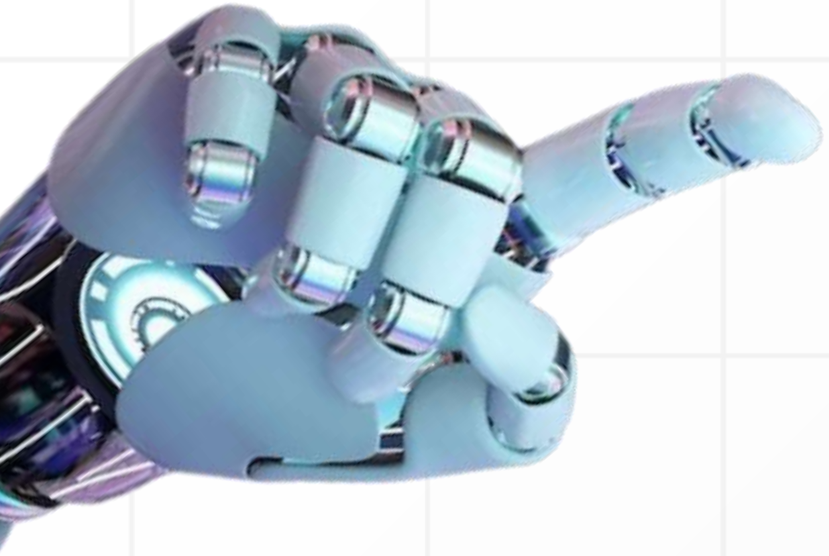
Stable Diffusion

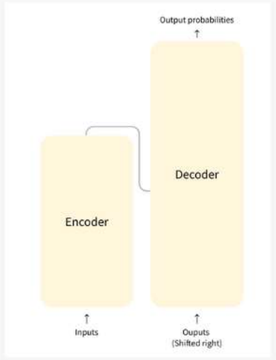
Generative AI applications are built using LLMs

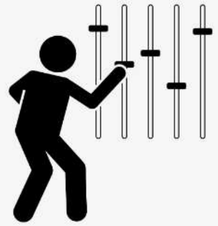




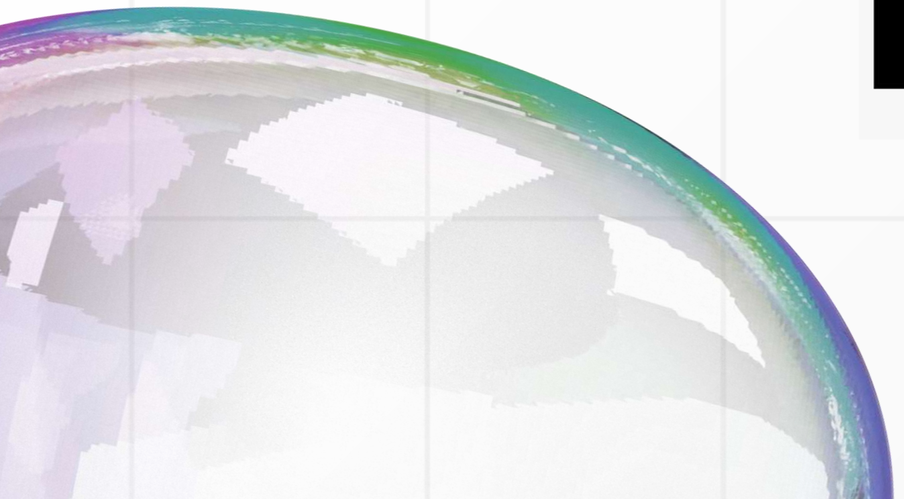
LLM Training







Data + **Architecture** + **Training** = **LLM**



Custom LLM Training

every ~year

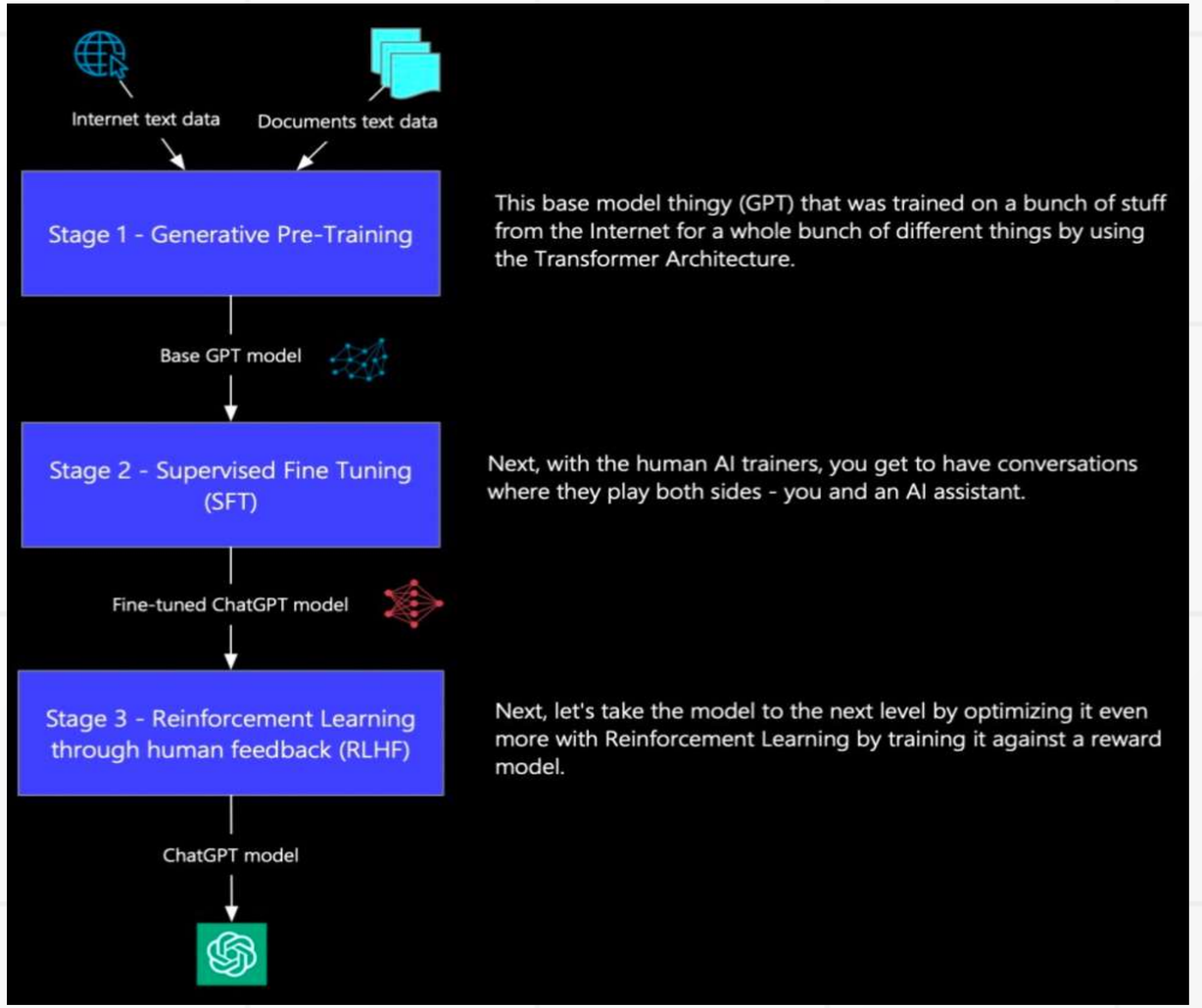
Stage 1: Pretraining

1. Download ~10TB of text.
2. Get a cluster of ~6,000 GPUs.
3. Compress the text into a neural network, pay ~$2M, wait ~12 days.
4. Obtain **base model**.

every ~week

Stage 2: Finetuning

1. Write labeling instructions
2. Hire people (or use scale.ai!), collect 100K high quality ideal Q&A responses, and/or comparisons.
3. Finetune base model on this data, wait ~1 day.
4. Obtain **assistant model.**
5. Run a lot of evaluations.
6. Deploy.
7. Monitor, collect misbehaviors, go to step 1.

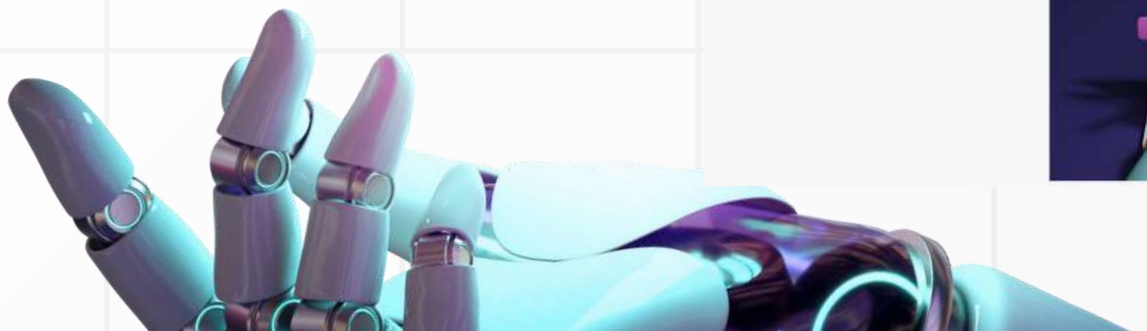


LLM Hallucination

- Missing Context
- Not tailored outputs
- Specialised vocabulary

Model	Accuracy	Hallucination Rate	Answer Rate
GPT 4	97.0 %	3.0 %	100.0 %
GPT 3.5	96.5 %	3.5 %	99.6 %
Llama 2 70B	94.9 %	5.1 %	99.9 %
Llama 2 7B	94.4 %	5.6 %	99.6 %
Llama 2 13B	94.1 %	5.9 %	99.8 %
Cohere-Chat	92.5 %	7.5 %	98.0 %
Cohere	91.5 %	8.5 %	99.8 %
Anthropic Claude 2	91.5 %	8.5 %	99.3 %
Mistral 7B	90.6 %	9.4 %	98.7 %
Google Palm	87.9 %	12.1 %	92.4 %
Google Palm-Chat	72.8 %	27.2 %	88.8 %

- **Fine tuning**
- **Prompt Engineering**
- **Retrieval Augmented Generation (RAG)**



Transformer

LLM Use case

Input Data

Books and Literature of multiple Languages,

Online content - websites, news, blogs

Social Media, Online chats and discussions

Wikipedia

Training

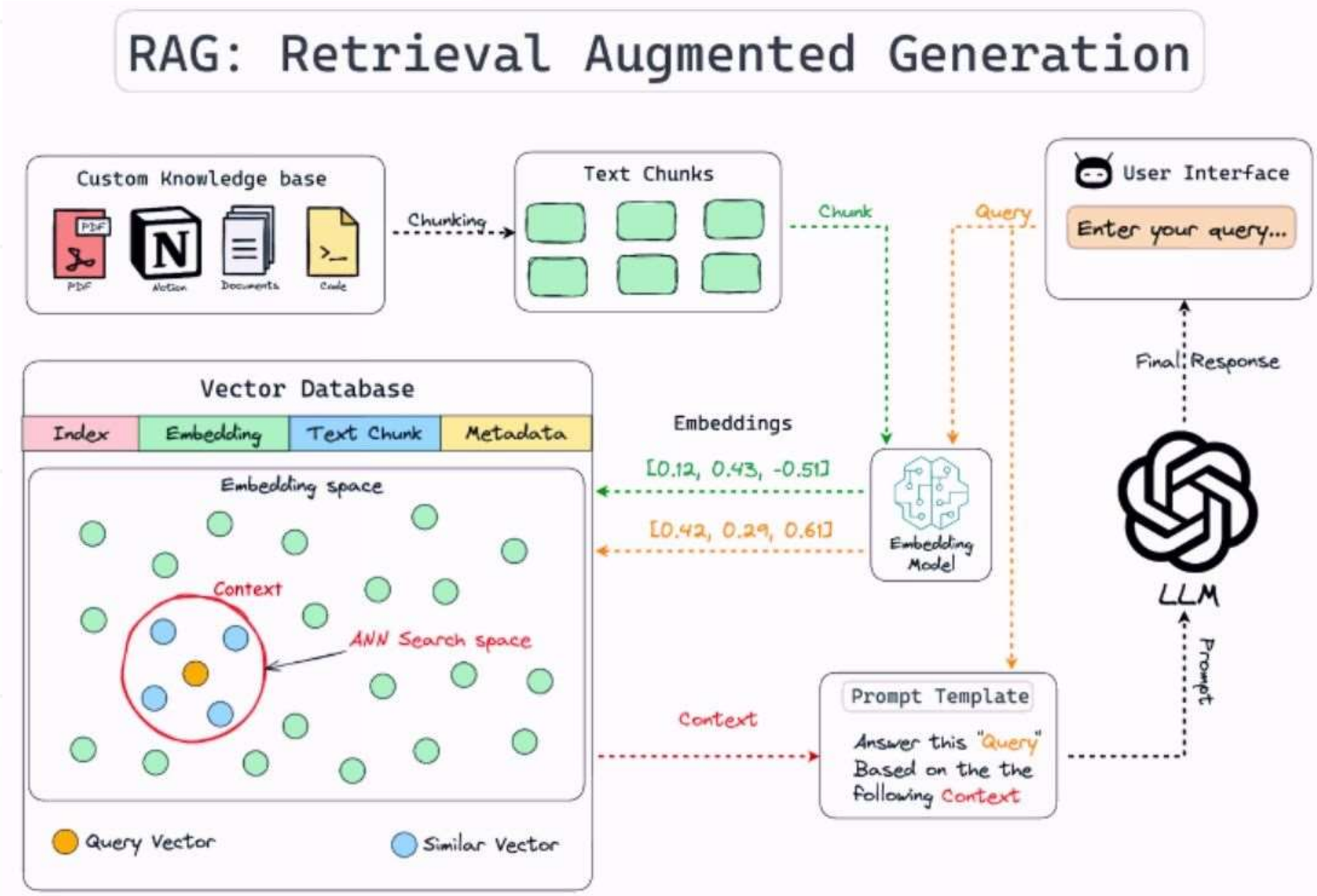
Large Language Model
Bard/ChatGPT

Adaptation

Tasks

- Creative Writing
- Answering Questions
- Text Summarization
- Language Translation
- Sentiment Analysis
- Interactive Conversation
- Information Retrieval

Building RAG



Project Architecture