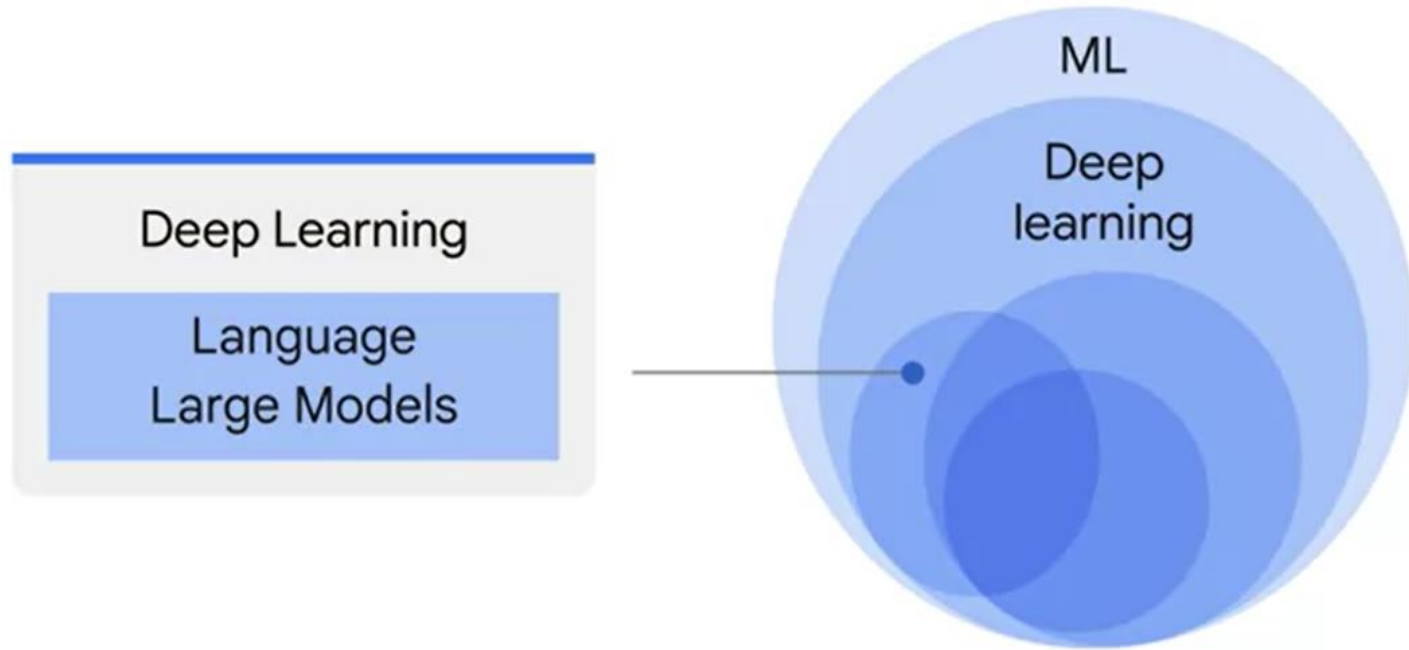
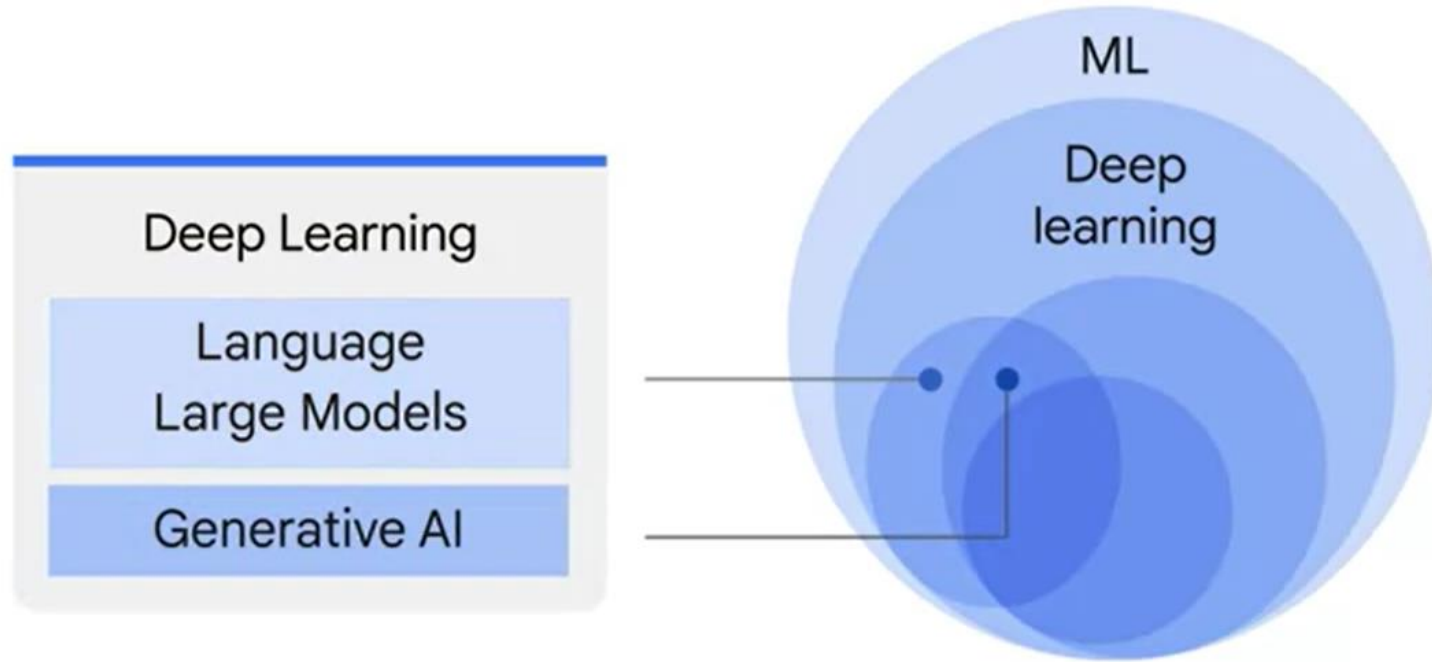
LLM Models

Large Language Models (LLMs) are a subset of Deep Learning



Large Language Models (LLMs) also intersects with Generative AI



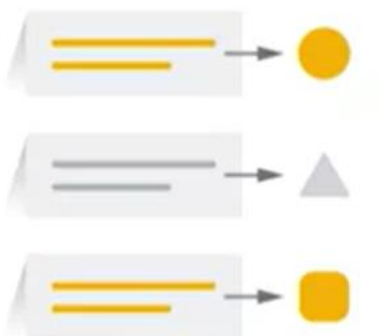
Generative AI can generate:

- Text
- Video
- Audio
- Code
- Other synthetic data

What are LLM Models?

Large, general-purpose language models can be pre-trained and then fine-tuned for specific purposes

Large language models are trained to solve common language problems, like...



Text classification

Question answering

Document summarization

Text generation

...then be tailored to solve specific problems in different fields, like...

Retail

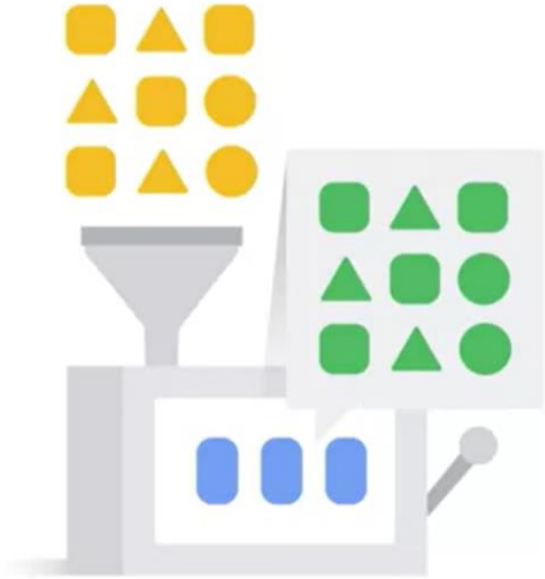
Finance

Entertainment

Trained with a relatively small size of field datasets

LLM models have three common features:

- **Large**
  - Large training dataset
  - Large number of parameters
- **General purpose**
  - Commonality of human languages
  - Resource restriction
- **Pre-trained and fine-tuned**



Benefits of using LLM models:

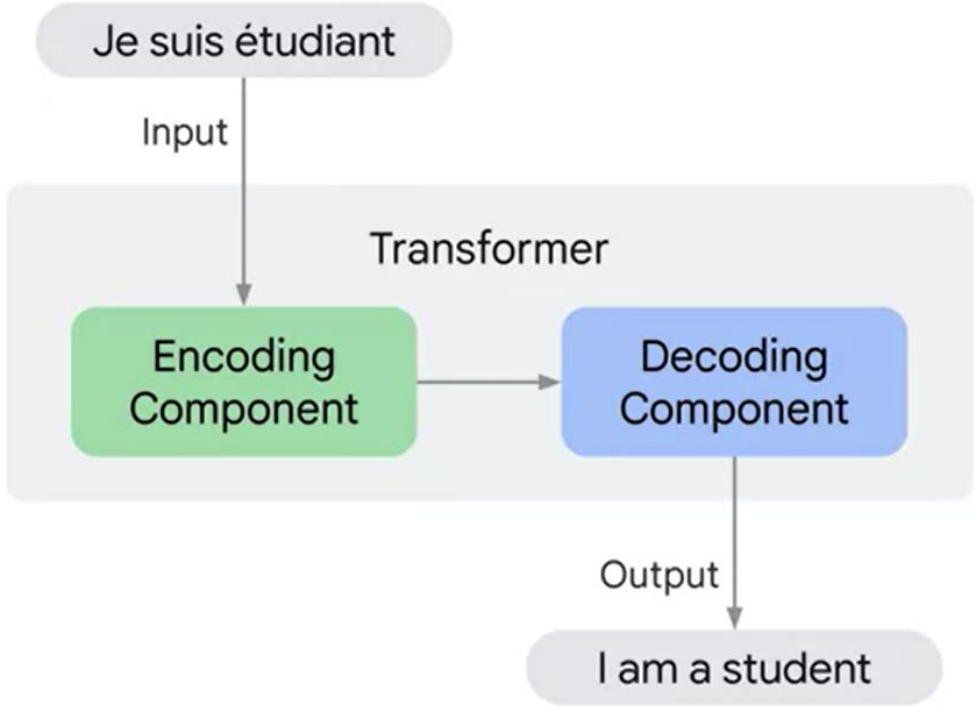
- A single model can be used in different tasks
- The fine-tuned process requires minimum data
- The performance is continuously growing with more data and parameters

Example of LLM models:

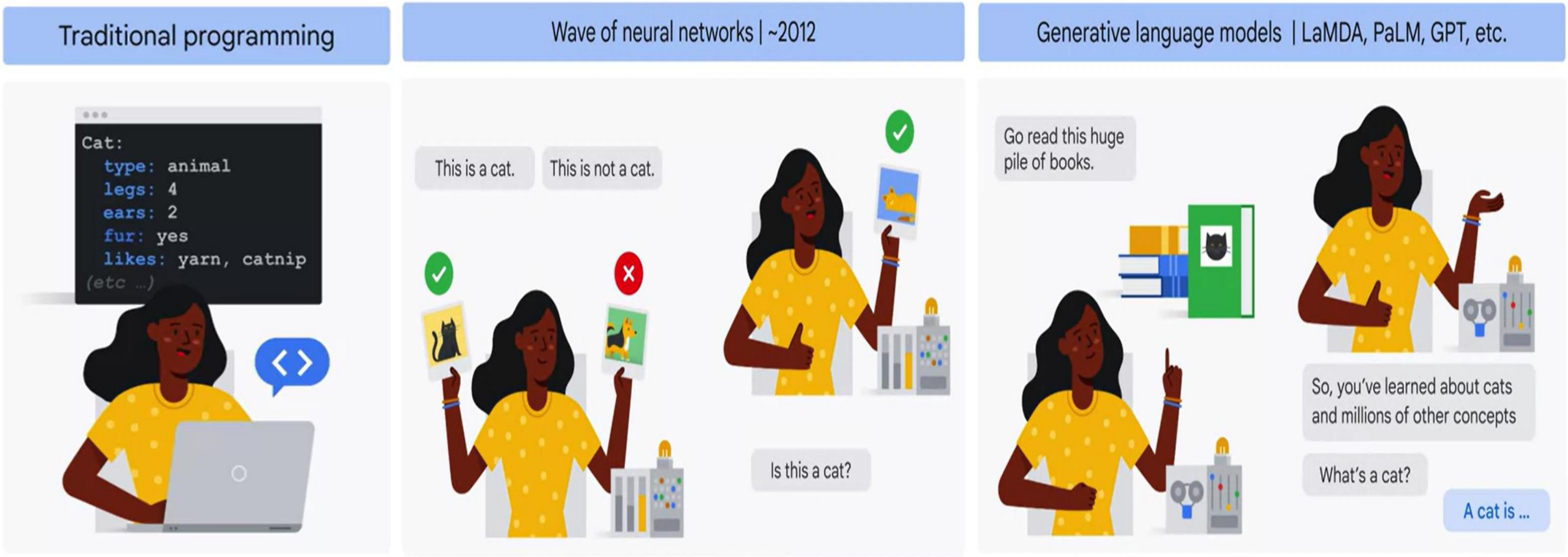
Pathways Language Model (PaLM)

- Has 540 billion parameters.
- Leverages the new Pathway system.
- Orchestrates distributed computation for accelerators.

PaLM is transformer model:



Different between traditional, NN, and GenAI model:



LLM Development vs. Traditional Development

LLM Development (using pre-trained APIs)

- NO ML expertise needed
- NO training examples
- NO need to train a model
- Thinks about prompt design

Traditional ML Development

- YES ML expertise needed
- YES training examples
- YES need to train a model
- YES compute time + + hardware
- Thinks about minimizing a loss function

What are Prompts and Prompt 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.

There are 3 main kinds of LLM models:

- Generic language model- These predicts next word based on language in the training data
- Instruction Tuned- Trained to predict a response to the instruction given in the input
- Dialog Tuned- Trained to have a dialog by predicting the next response

Dialog-Tuned Model:

Dialog-tuned models are a special case of instruction tuned where requests are typically framed as questions to a chat bot.

Dialog tuning is a further specialization of instruction tuning that is expected to be in the context of a longer back and forth conversation, and typically works better with natural question-like phrasings.

Tuning:

Tuning

The process of adapting a model to a new domain or set of custom use cases by training the model on new data. For example, we may collect training data and “tune” the LLM specifically for the legal or medical domain.

Thank you!