Introduction to Generative AI

SALIMA LAMSIYAH, UNIVERSITY OF LUXEMBOURG

salima.lamsiyah@uni.lu

Lecture Plan

1. Introduction to Machine Learning
2. Generative AI Definition
3. Introduction to Natural Language Processing
4. Large Language Models
   - Architecture
   - Training Process
   - Usage
   - Capabilties
   - Limitations

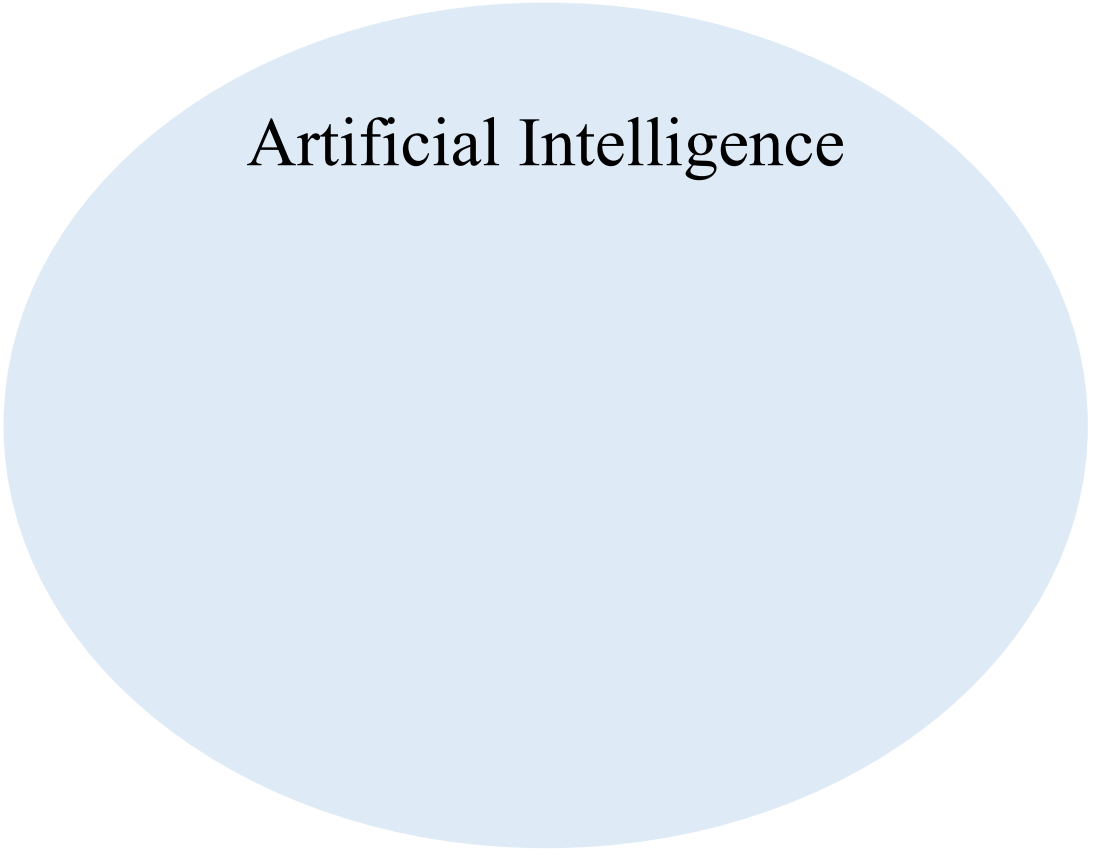
Introduction to Machine Learning

AI Terminology: Artificial Intelligence

- Artificial Intelligence or AI is a subfield of computer science that enables machines to mimic human behaviours,

- It creates intelligent systems that can perform tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, …

- AI Terminology:
  - Machine Learning
  - Deep Learning
  - Reinforcement Learning
  - Natural Language Processing
  - Generative AI
  - …

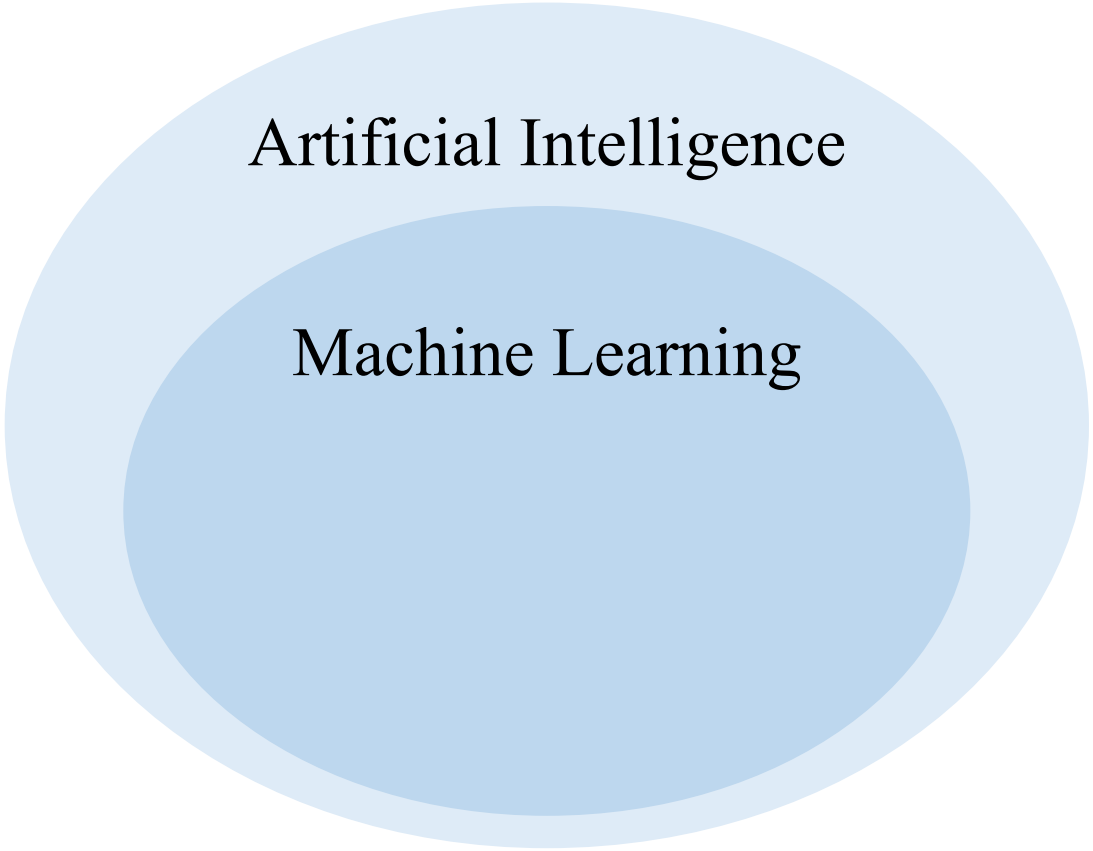


AI Terminology: Machine Learning

[Authur Samuel 1959]: Machine learning is a subfield of Artificial Intelligence that gives Computers the Ability to Learn from Data, without being explicitly programmed.

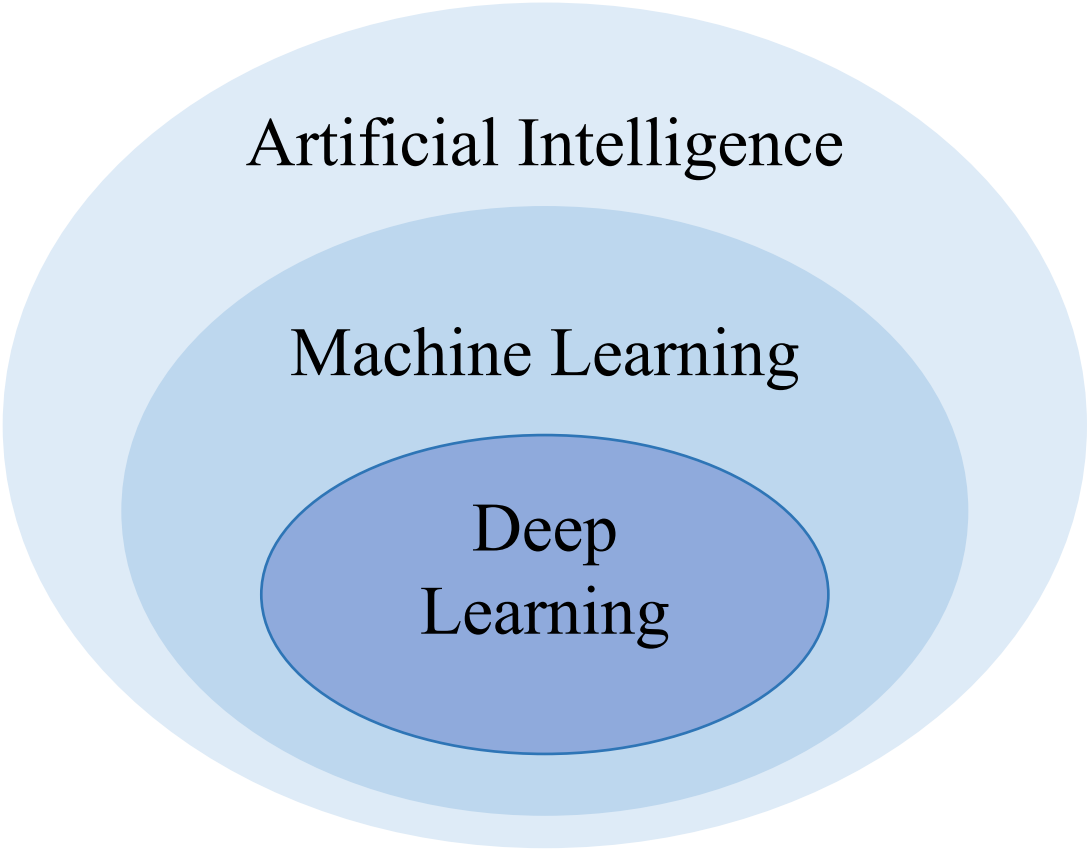
Machine Learning algorithms construct predictive models by learning from a large number of training examples.

By using machine learning, the computer can learn to recognize patterns and make decisions based on data, without requiring explicit programming instructions for every possible scenario.

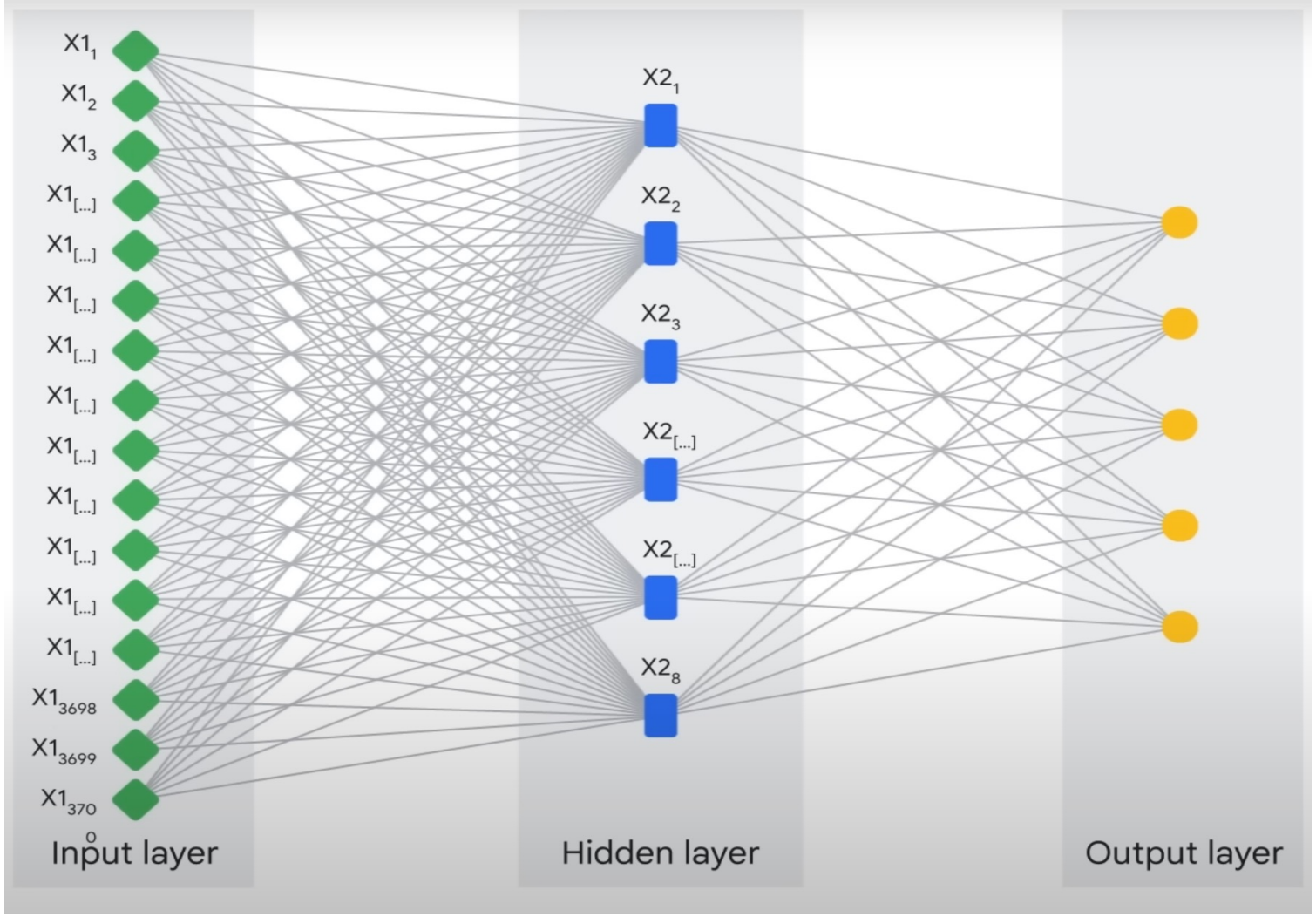


AI Terminology: Deep Learning

Deep learning (DL) is a subfield of ML that uses Artificial Neural Networks to learn complex patterns from data.



AI Terminology: Deep Learning



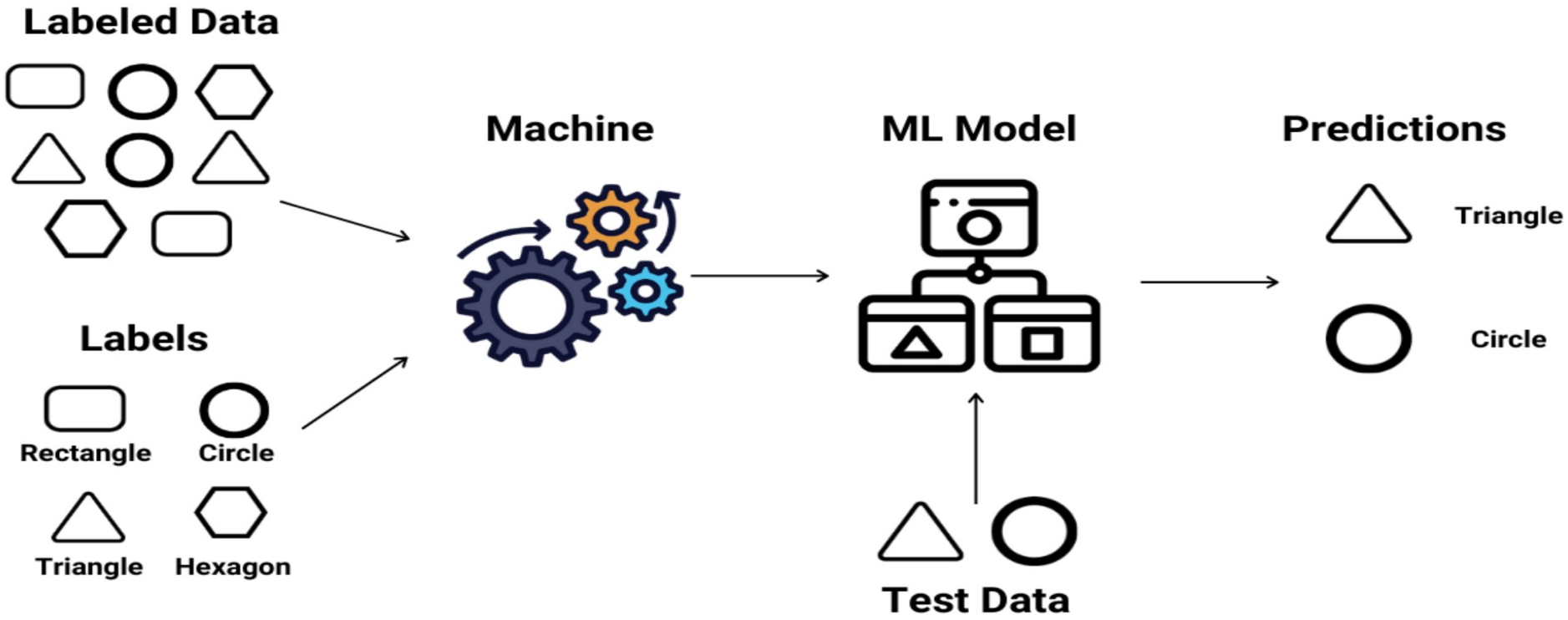
Examples of deep Learning models:

- Feed Forward Neural Networks
- Convolutional Neural Networks
- Recurrent Neural Networks
- **Transformers**
- …

Machine Learning: *Learning Paradigm*

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
3. Reinforcement Learning
4. Self-Supervised Machine Learning

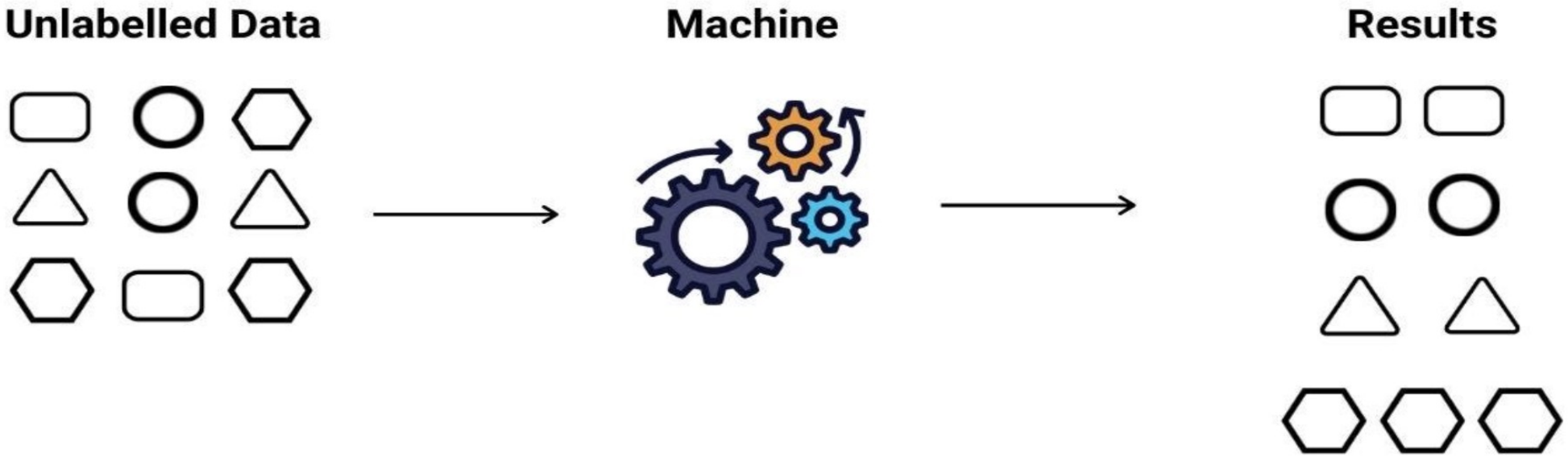
Supervised Machine Learning



Source

- **Conventional Machine Learning** (Linear Regression, Logistic Regression, Support Vector Machines, Decision Trees, Naïve Bayes, …)
- **Deep Learning models** (e.g., Feed-Forward NNs, Recurrent NNs, Convolutional NNs, …)

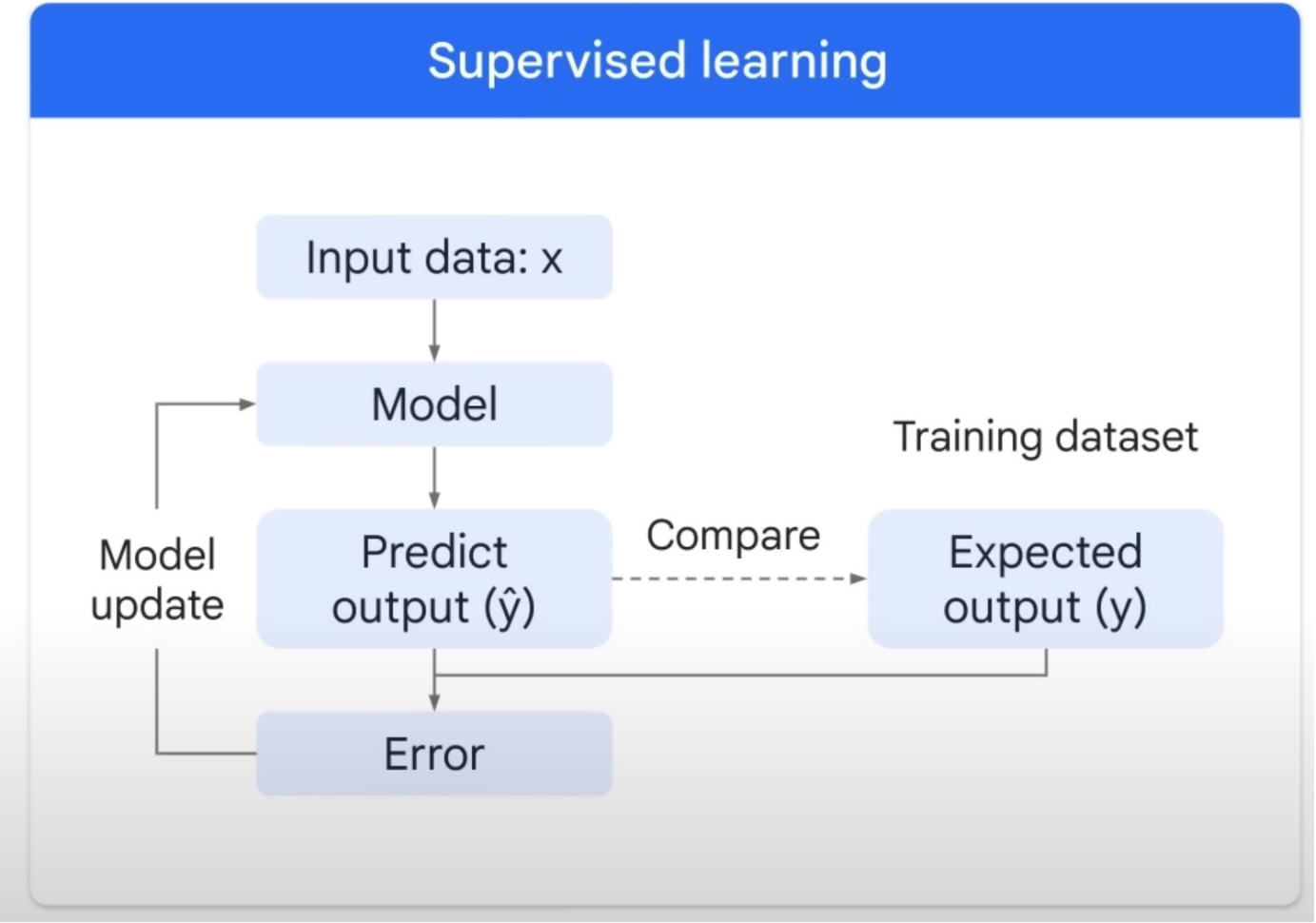
Unsupervised Machine Learning

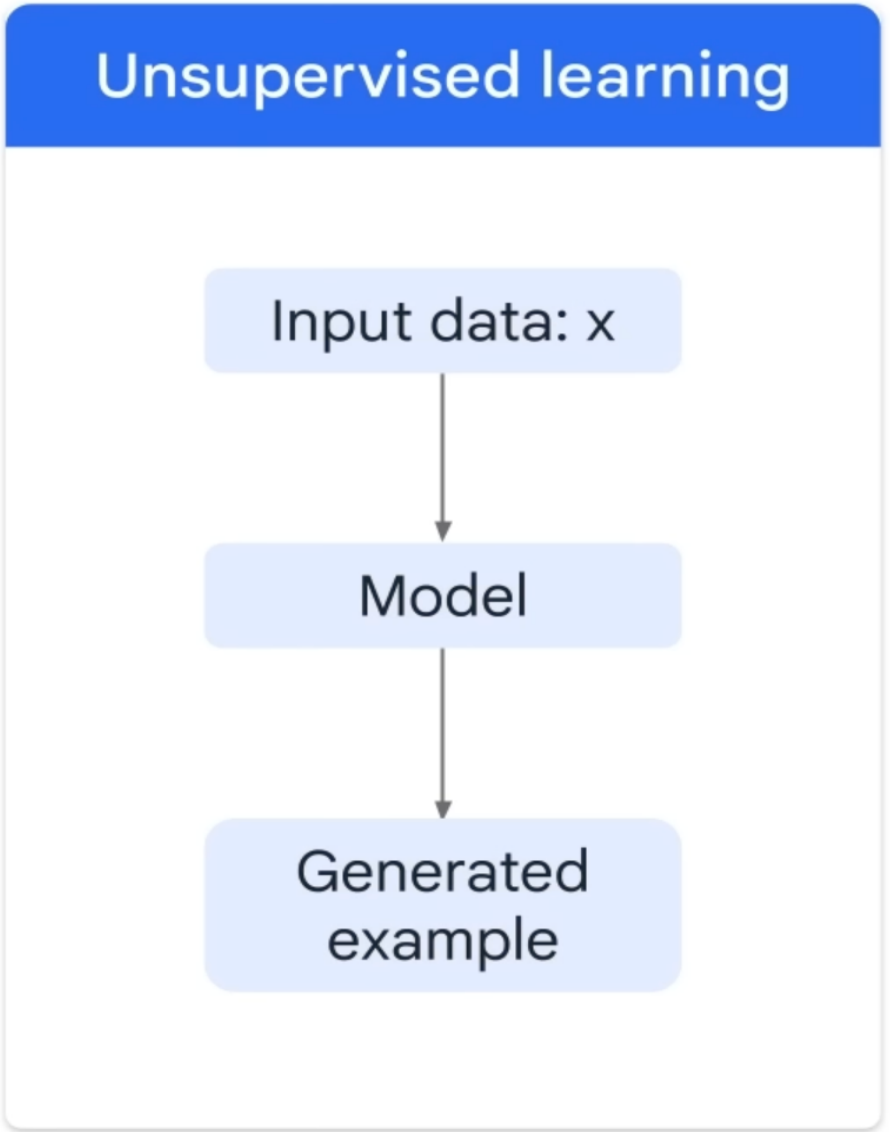


Source

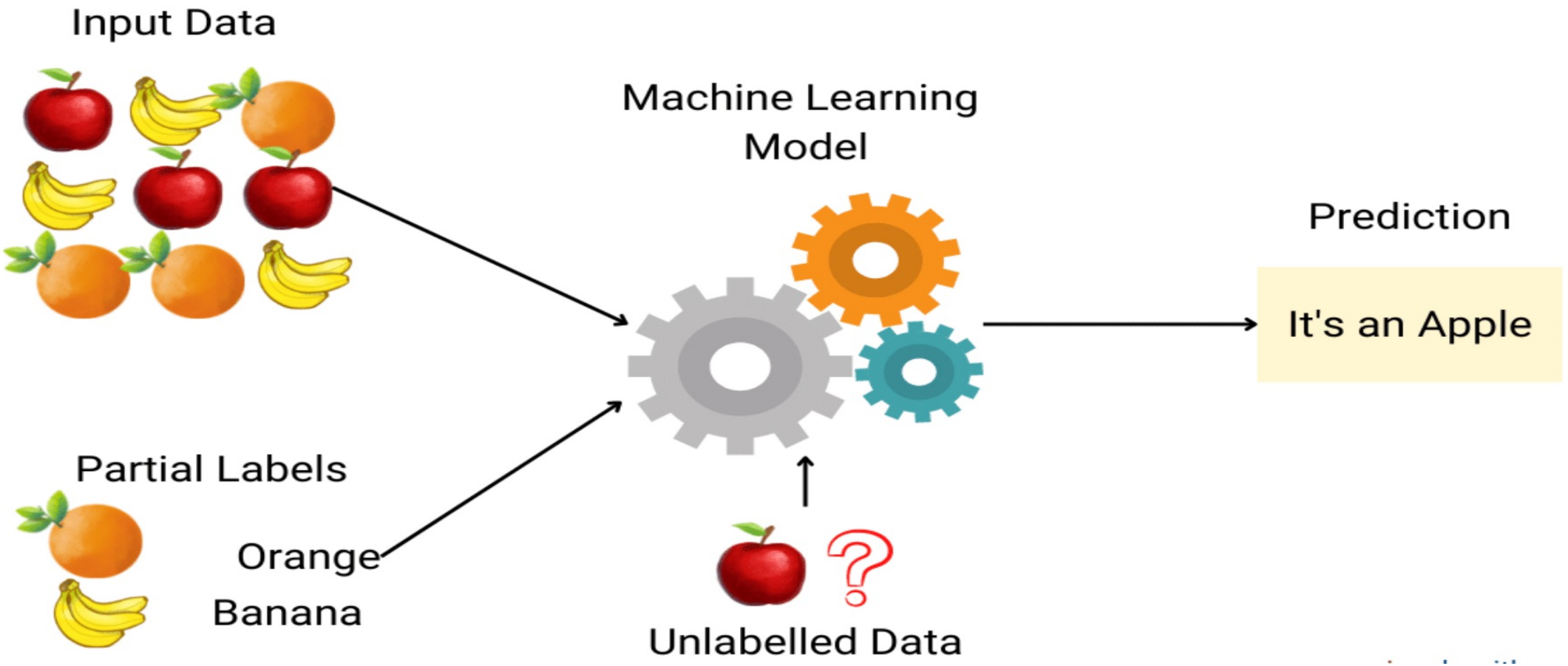
- **Clustering algorithms** (e.g., k-means clustering, hierarchical clustering),
- **Deep Learning models** (e.g., autoencoders, variational autoencoders, generative adversarial networks).

Supervised Vs Unsupervised Learning





Semi-Supervised Machine Learning

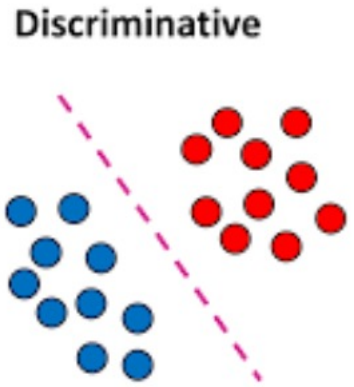


Source

Discriminative Vs Generative Machine Learning

1. *Discriminative*

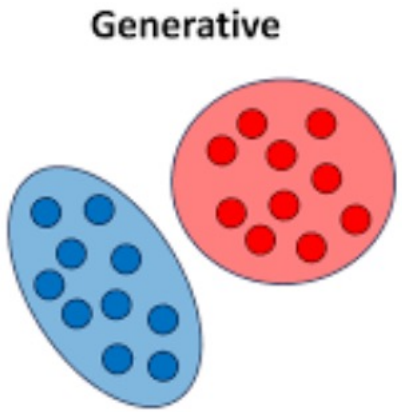
- Classify or predict
- Usually trained using labeled data
- Learns representation of features for data based on the labels using conditional probability: $\boldsymbol{P(c|d)}$



2. *Generative*

- Generates new data
- focuses on the distribution of a dataset to return a probability for a given example using Joint Probability.

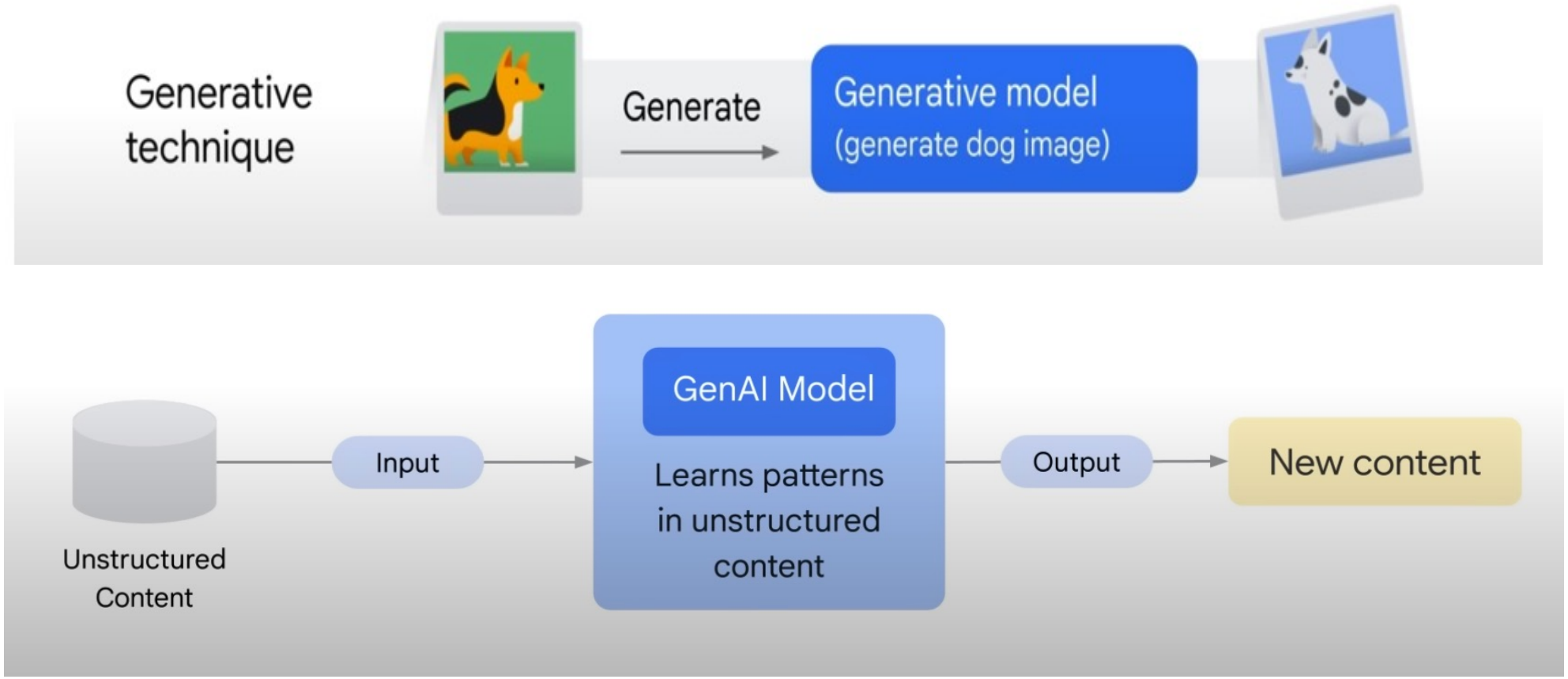
$$\boldsymbol{P(d \cap c) = P(d|c).p(c)}$$



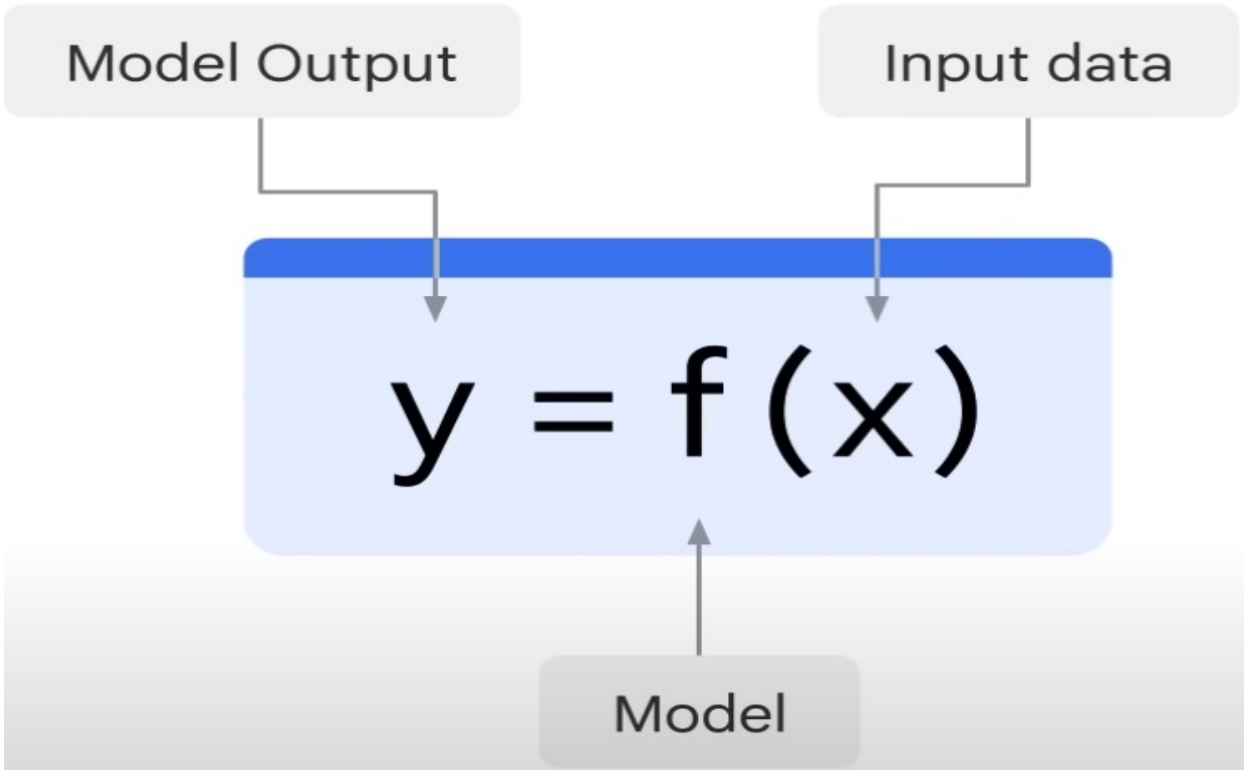
Discriminative Vs Generative Machine Learning

Source

Discriminative Vs Generative Machine Learning



Source



Not GenAI when y is a:

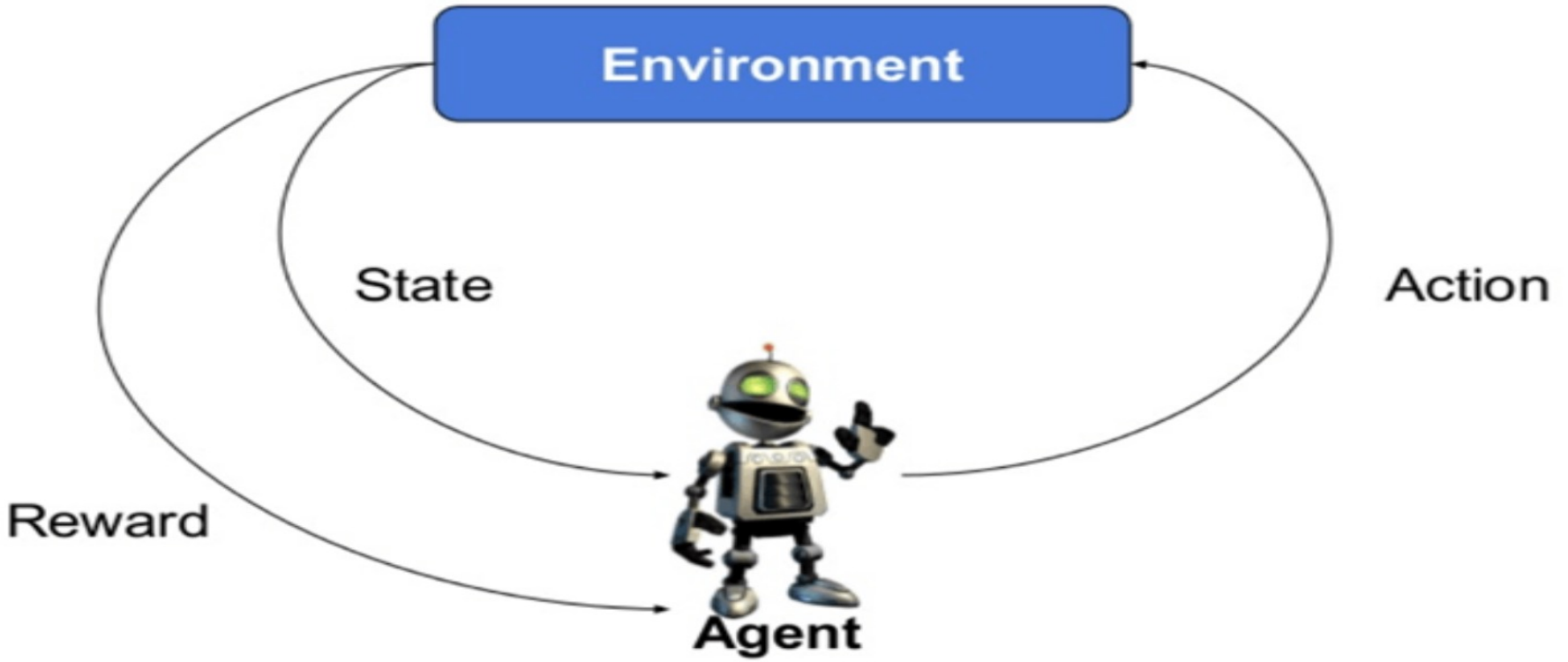
- Number
- Discrete
- Class
- Probability

Is GenAI when y is:

- Natural language
- Image
- Audio

Source

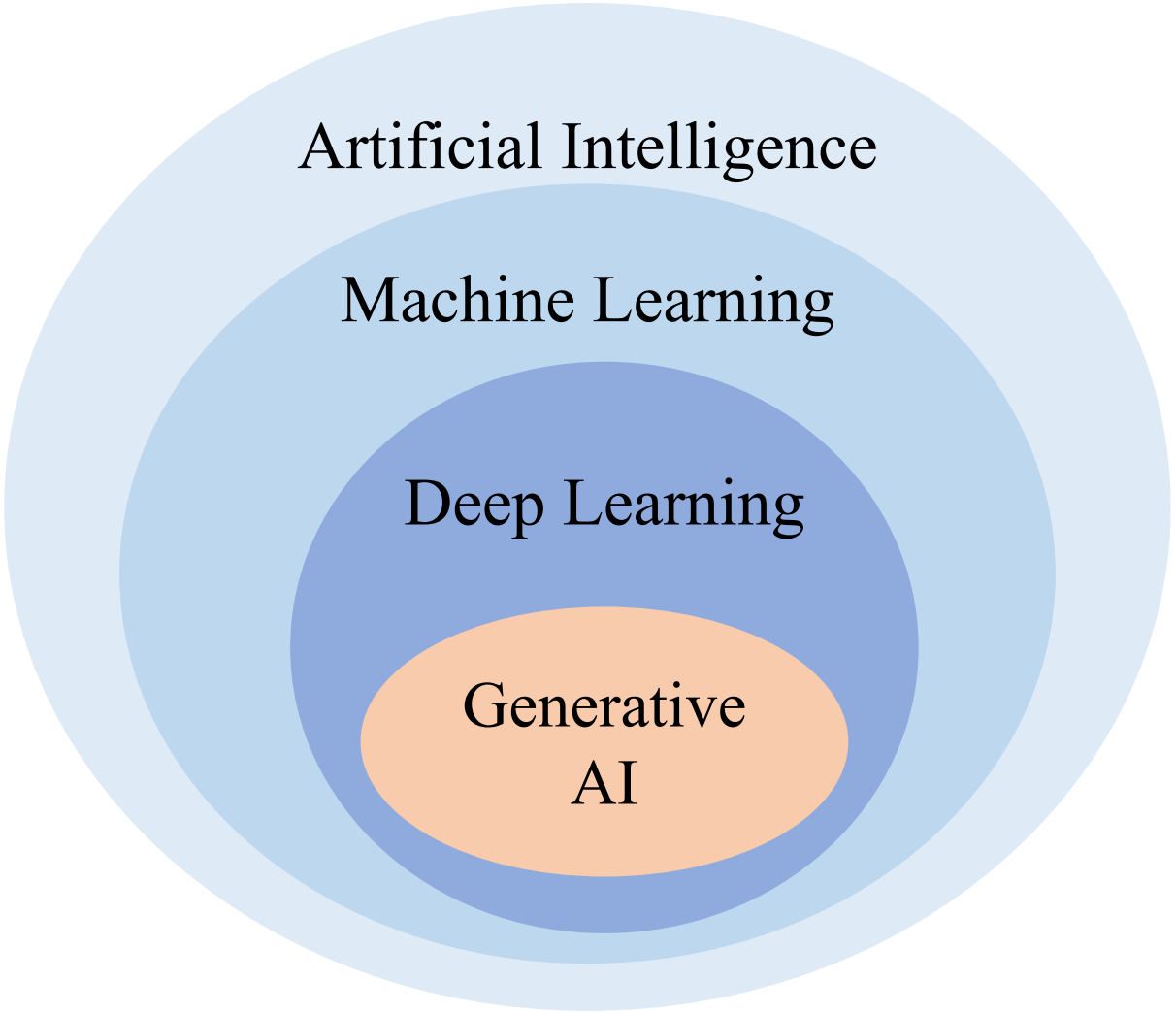
Reinforcement Learning



- Examples of RL applications include:
  - Game playing: Chess, Go, Atari, Tic-Tac-Toes games
  - Robotics, Autonomous vehicles, …

Generative AI Definition

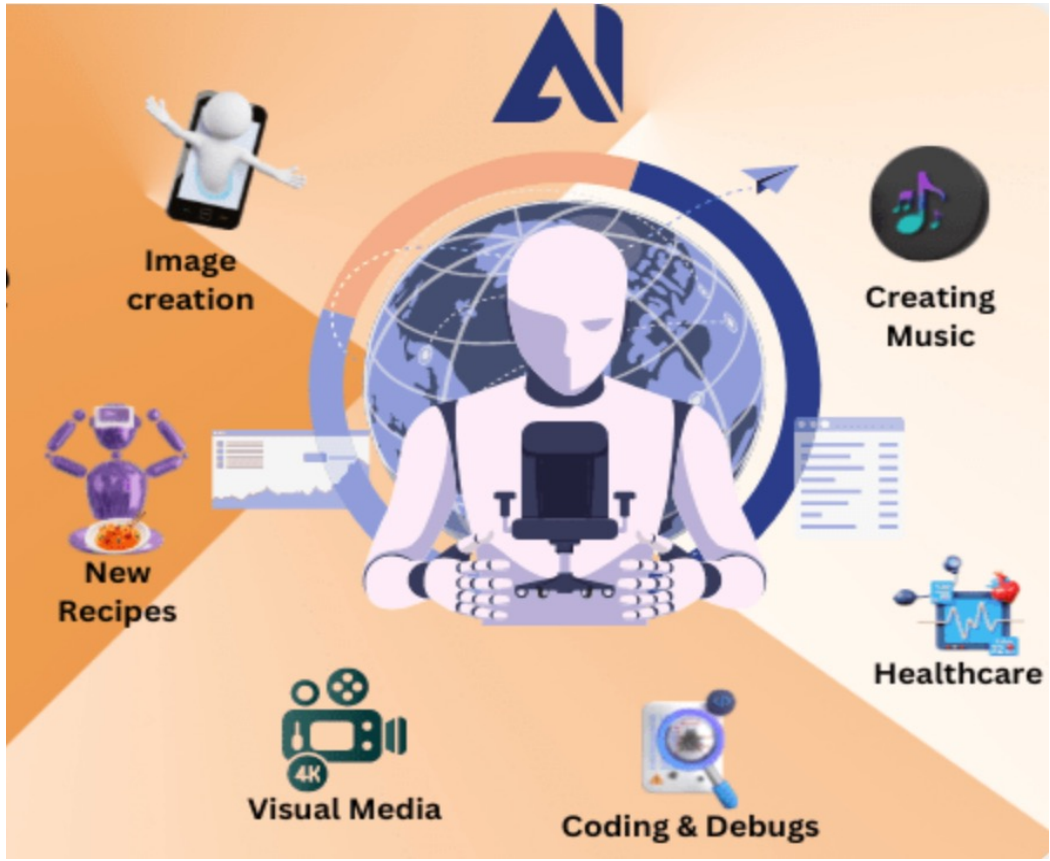
Generative AI is subset of Deep Learning



Generative AI (GenAI)

McKinsey defines the Generative AI as:

- Generative AI refers to a branch of AI that focuses on creating or generating new content, such as images, text, video, synthetic data, or other forms of media, using machine learning examples.
- It does this by learning patterns from existing data, then using this knowledge to generate new and unique outputs.
- The process of learning from existing content is called **training** and results in the creation of a **statistical model**.
- When given a prompt, GenAI uses this statistical model to predict what an expected response might be-and this generate new content.
- Recent breakthroughs in the field, such as **GPT** (Generative Pre-trained Transformer) and **Midjourney**, have significantly advanced the capabilities of GenAI.



Source

Generative AI

- **Text Generation** involves using machine learning models to generate new text based on patterns learned from existing text data. The models used for text generation can be **Markov Chains**, **Recurrent Neural Networks (RNNs),** and more recently, **Transformers**, which have revolutionized the field due to their extended attention span. Text generation has numerous applications in the realm of natural language processing, chatbots, and content creation.
  - ***Application:*** **ChatGPT**, developed by OpenAI, is a successful platform that uses Text Generation to generate human-like responses in chat conversations.

- **Image Generation** is a process of using deep learning algorithms such as **VAEs**, **GANs**, and more recently **Stable Diffusion**, to create new images that are visually similar to real-world images. Image Generation can be used for data augmentation to improve the performance of machine learning models, as well as in creating art, generating product images, and more.
  - ***Application:*** Very successful platforms such as **MidJourney** and **DALL-E** have become a popular choice for anyone seeking to generate realistic images through Image Generation techniques.

Generative AI

- **Video Generation** involves deep learning methods such as **GANs** and **Video Diffusion** to generate new videos by predicting frames based on previous frames. Video Generation can be used in various fields, such as entertainment, sports analysis, and autonomous driving.
  - ***Application:*** Platforms such as **DeepBrain** and **Synthesia** utilize Video and Speech Generation to create realistic video content, that appears as if a human was speaking on camera.

- **Data augmentation** is a process of generating new training data by applying various image transformations such as flipping, cropping, rotating, and color jittering. The goal is to increase the diversity of training data and avoid overfitting, which can lead to better performance of machine learning models.
  - ***Application***: **Synthesis AI** simplifies the process of building and optimizing machine learning models by providing a platform for creating AI models using automated machine learning techniques.

Why GenAI Now?

- Availability of large and diverse datasets
- AI models learn patters, correlations, and characteristics of large datasets
- Pre-trained state-of-the-art models

- Advancements in hardware; GPUs
- Access to cloud computing
- Open-source software, Hugging Face

- Generative Adversarial Networks (GANs)
- Transformers Architecture
- Reinforcement Learning from human feedback (RLHF)

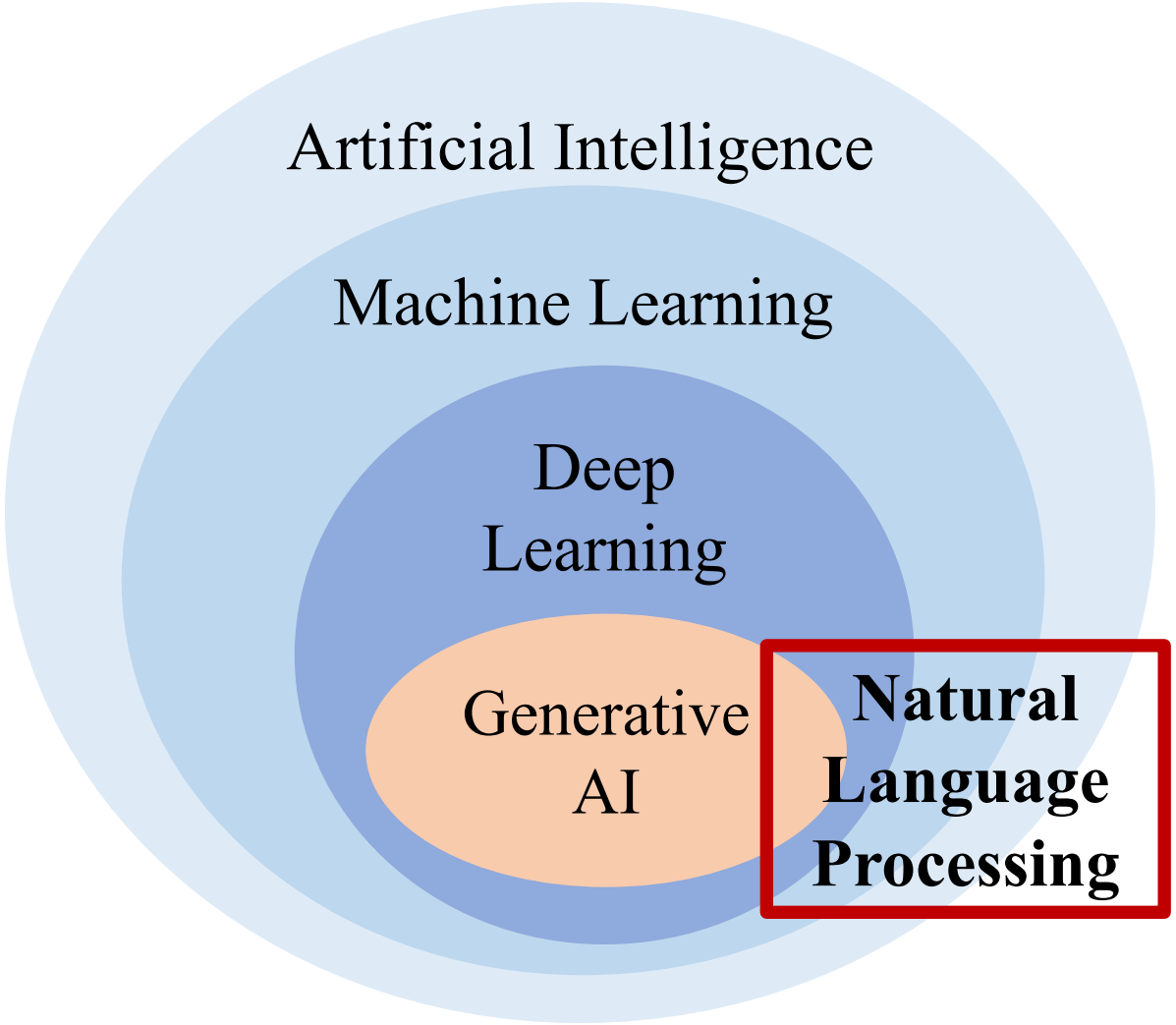
Introduction to Natural Language Processing

Natural Language Processing

- **Natural Language Processing (NLP)** is a field at the intersection of:
  - Computer Science
  - Artificial Intelligence (ML, DL, GenAI)
  - And Linguistics.

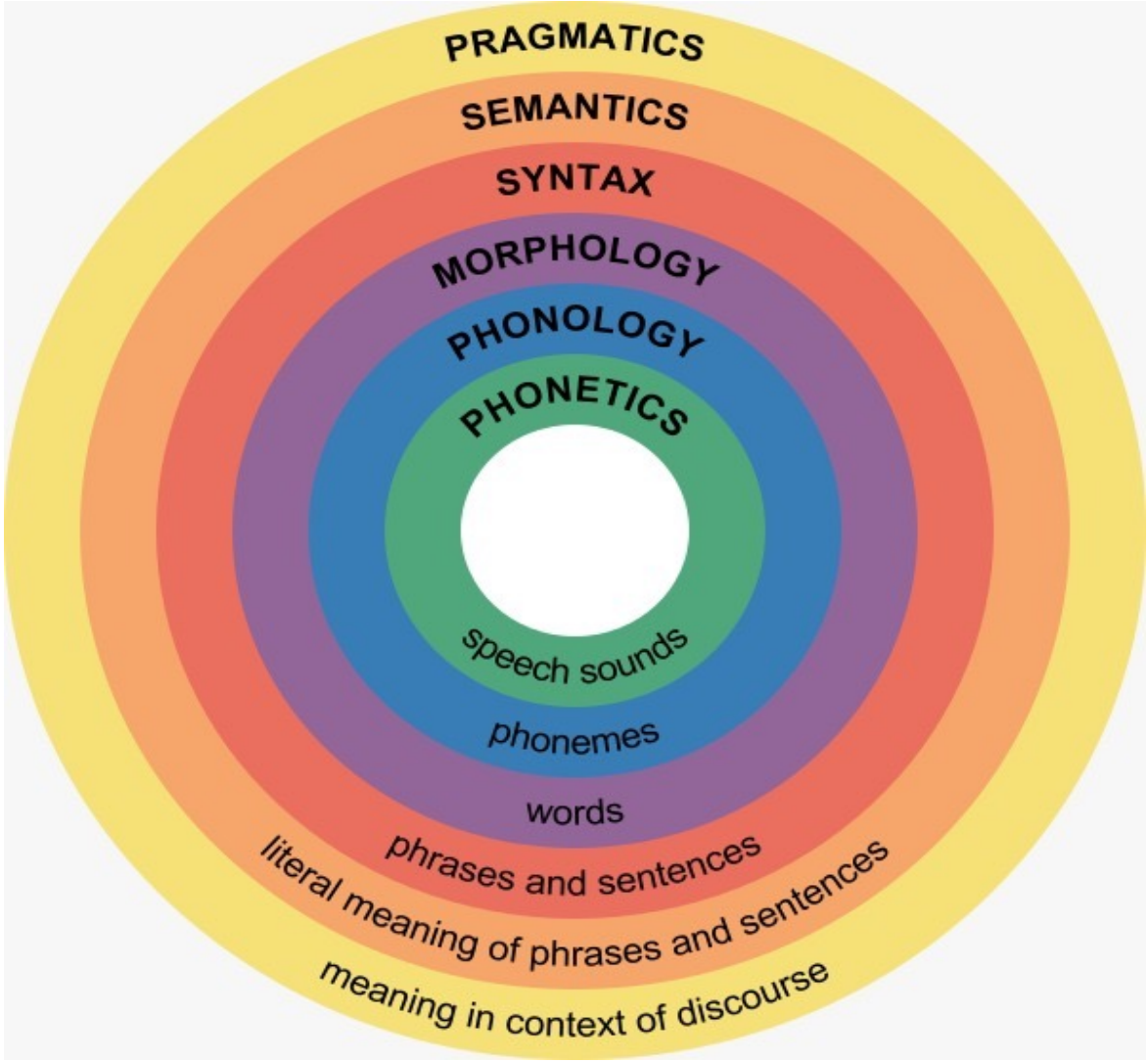
- **Goal**: for computers to process or "Understand" natural language in order to perform tasks:
  - Translation, Question Answering, Siri, Google Assistant, …

- Fully **understanding** and **representating** the **meaning** of a language is a difficult goal.
  - Perfect language understanding is AI-complete (AI-hard)



Why NLP is hard

- Language consists of **many levels of linguistic knowledge.**
- Humans fluently integrate all of these to produce and understand language
- Ideally, so would a computer!



Why is NLP hard?

- Ambiguity
- Scale
- Variation
- Expressivity
- Unknown representation

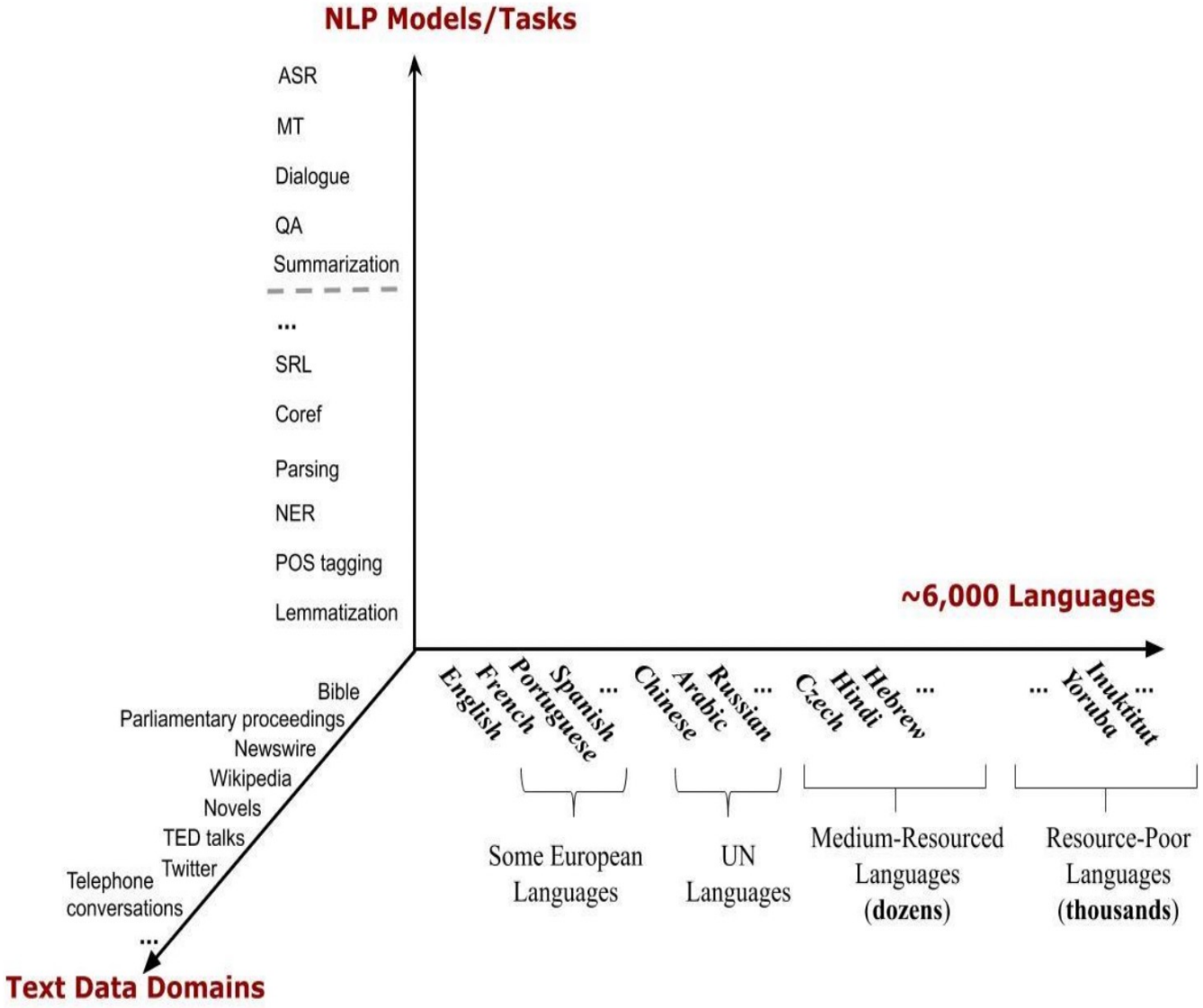
Ambiguity

- Ambiguity at multiple levels:

  - ***Lexical Ambiguity***: The fisherman go the **bank** (finance or river?)
  - ***Syntactic ambiguity*: I can see a man with a telescop**
  - ***Semantic ambiguity*: I gave a present to the children or The chicken is ready to eat**
  - ***Referential ambiguity*: Alice invited Maya for dinner but she cooked her own food (she = Alice or Maya ?)**

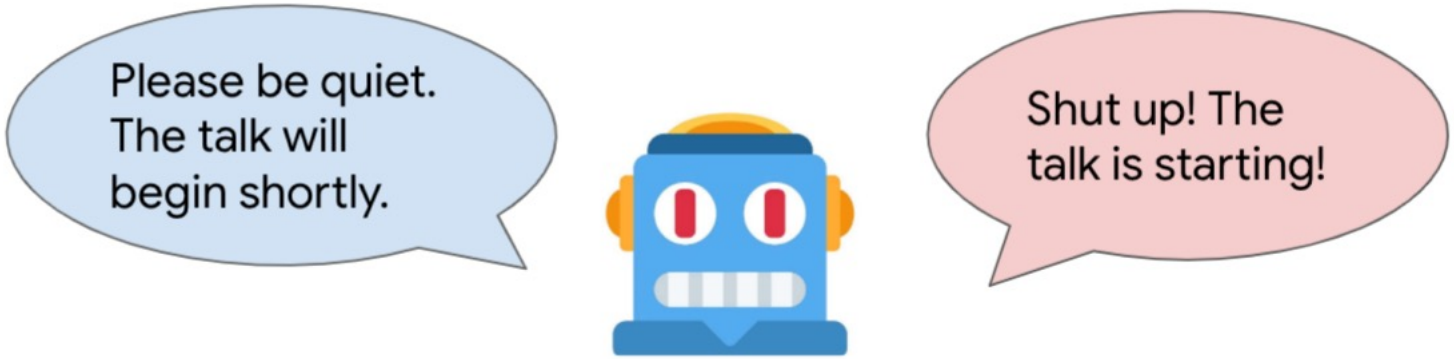
Scale and Variation

- **~7K languages**
- **Thousands of language varieties**
- **Variation of domains (news, biomedical, historical, …)**



Expressivity

- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:
  - **She gave the book to Aria** vs. **She gave Aria the book**
  - **Is that door still open?** vs. **Please close the door**



Unknown Representation

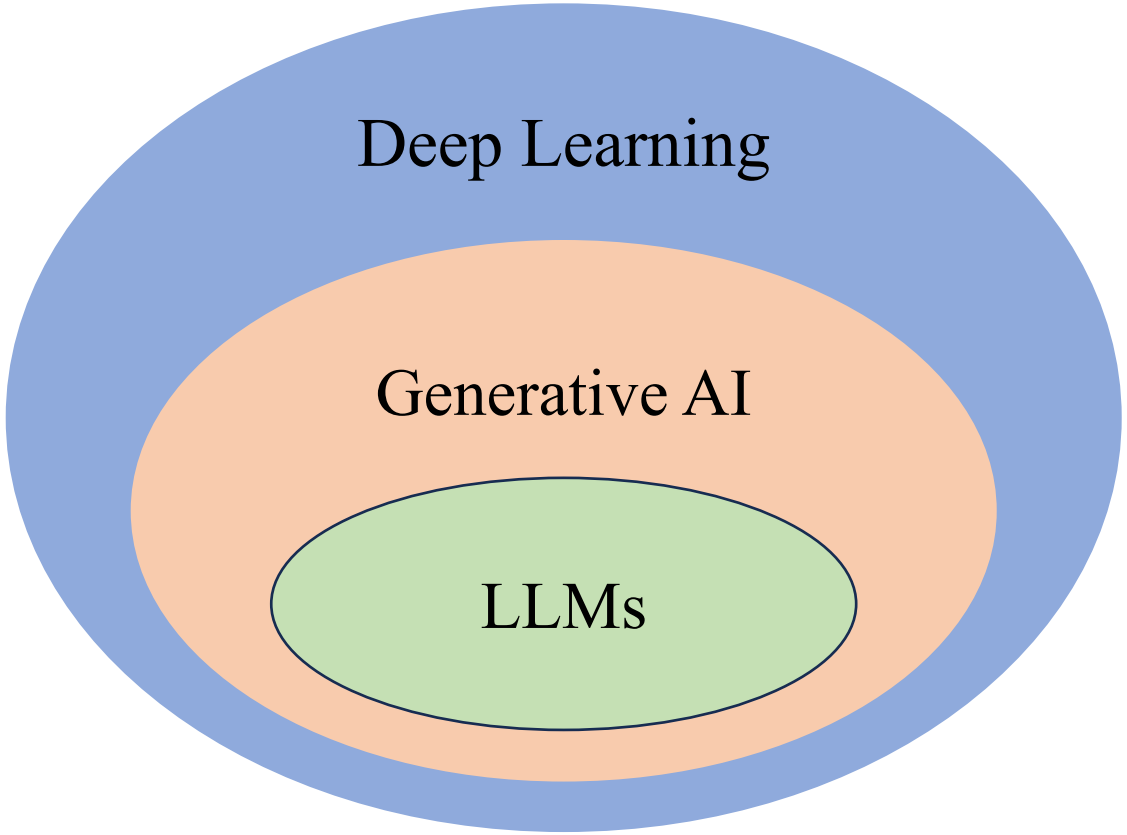
- Very difficult to capture what is the representation of the text or speech, since we don't even know how to represent the knowledge a human needs:
  - What is the "meaning" of a word or sentence?
  - How to model context?
  - Other general knowledge?

Large Language Models (LLMs)

Large Language Models (LLMs)

Large Language Models is subset of Generative AI

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



Large Language Models – Architecture

- Encoder
- Decoder
- Encoder-decoder

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

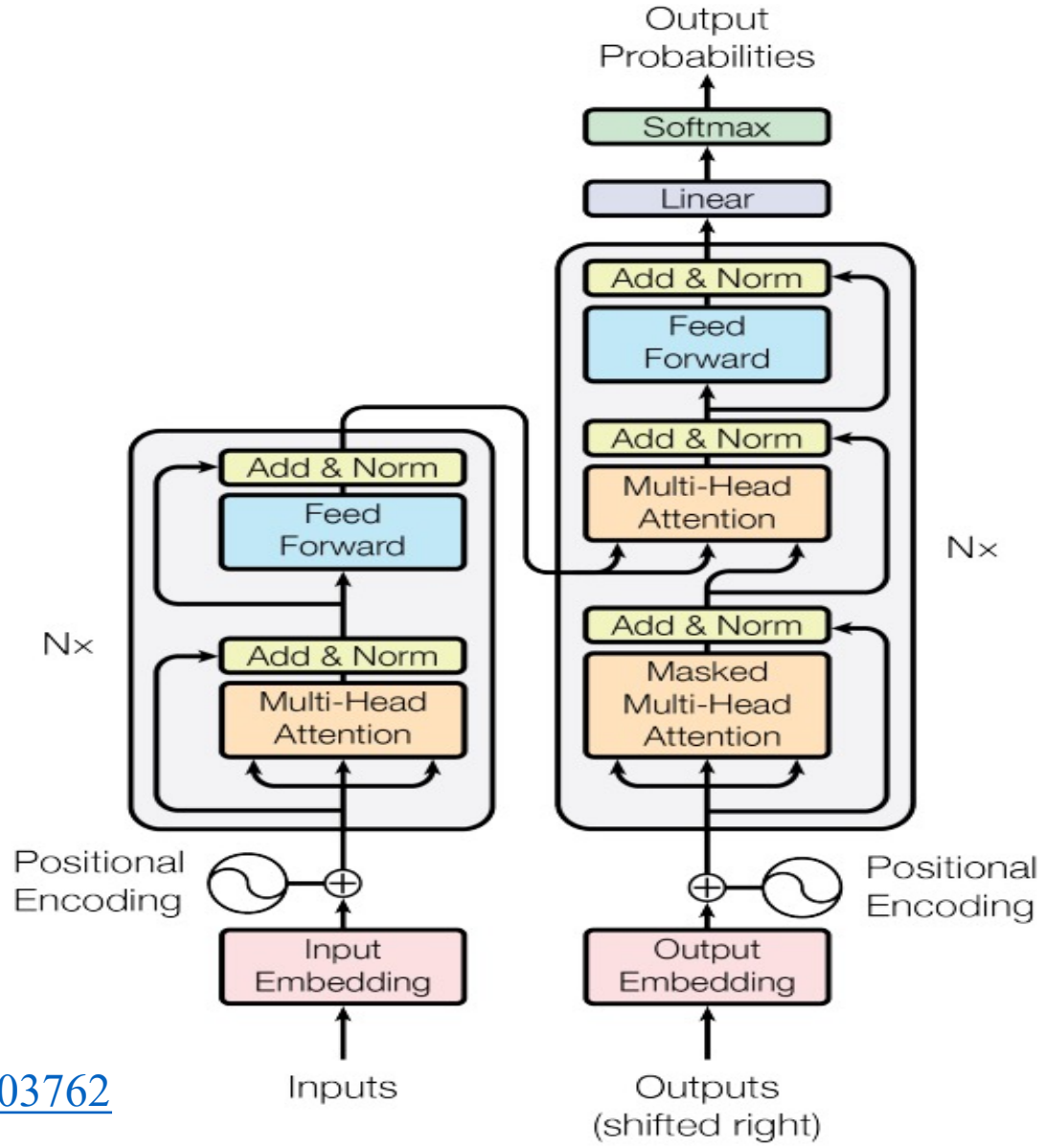
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

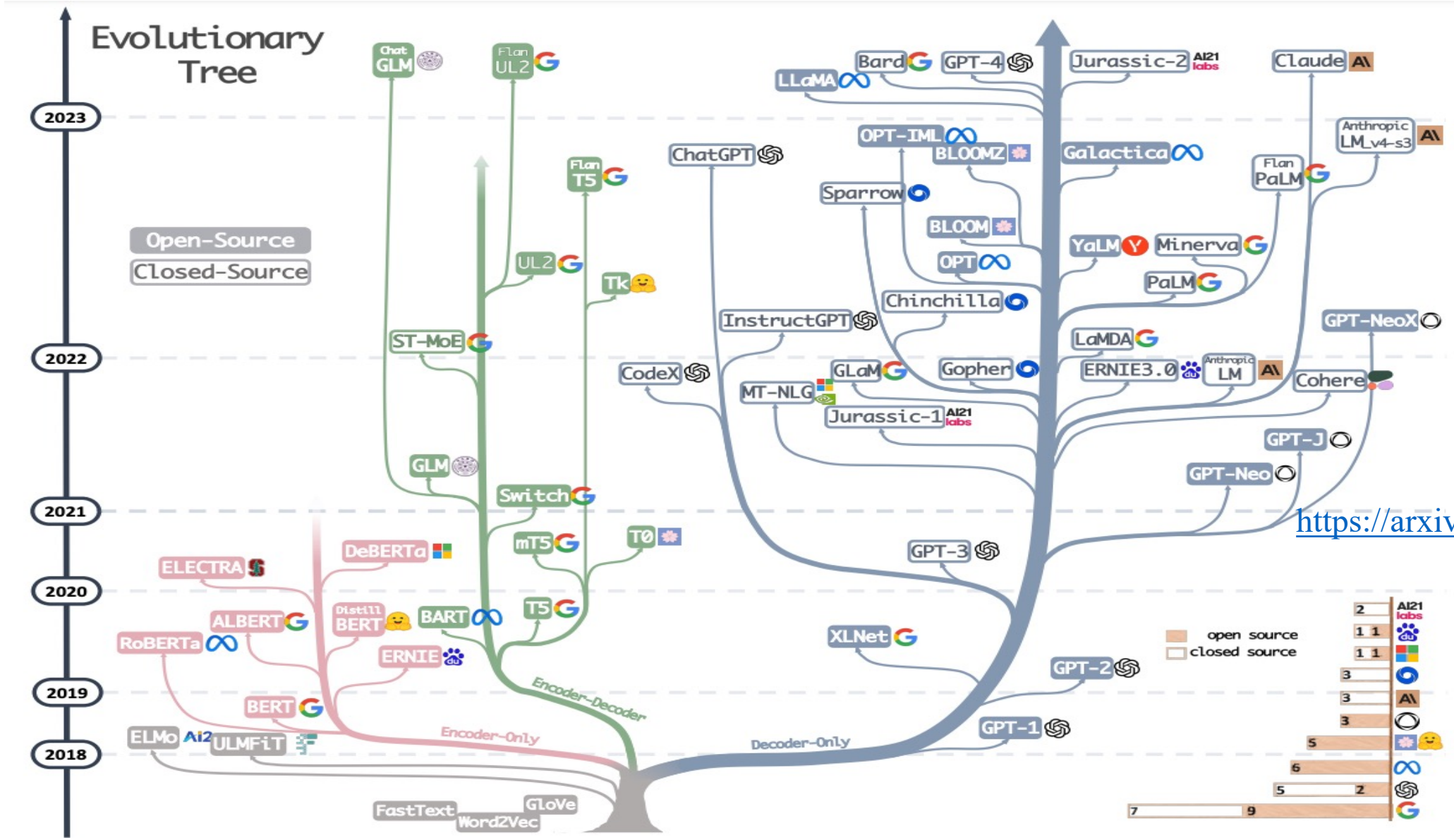
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com



https://arxiv.org/abs/1706.03762

Timeline of Language Models Evolution: 2018-2023



https://arxiv.org/abs/2304.13712

Large Language Models – Characteristic

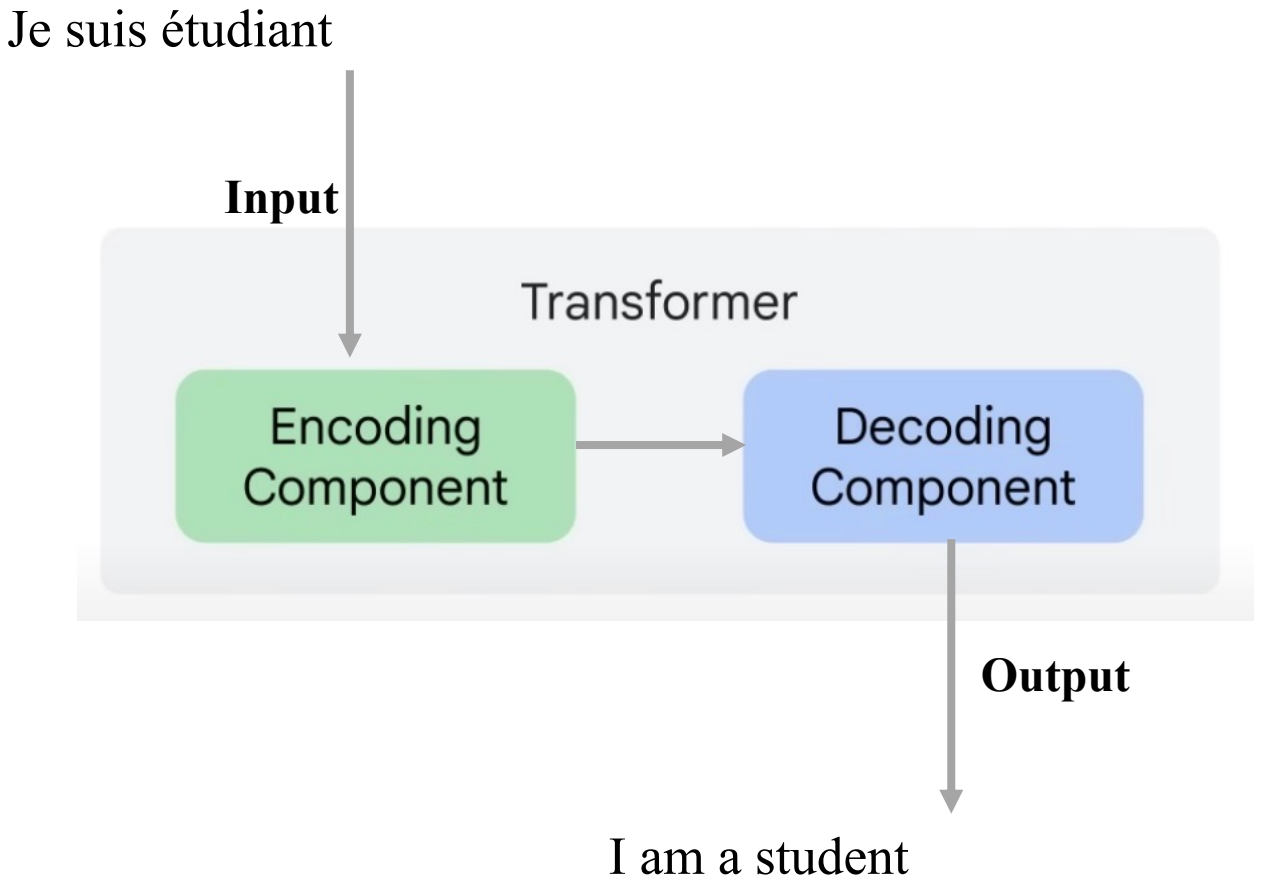
Table 1. Summary of Large Language Models.

	Characteristic		LLMs
Encoder-Decoder or Encoder-only (BERT-style)	Training: Model type: Pretrain task:	Masked Language Models Discriminative Predict masked words	ELMo [80], BERT [28], RoBERTa [65], DistilBERT [90], BioBERT [57], XLM [54], Xlnet [119], ALBERT [55], ELECTRA [24], T5 [84], GLM [123], XLM-E [20], ST-MoE [133], AlexaTM [95]
Decoder-only (GPT-style)	Training Model type: Pretrain task:	Autoregressive Language Models Generative Predict next word	GPT-3 [16], OPT [126]. PaLM [22], BLOOM [92], MT-NLG [93], GLaM [32],Gopher [83], chinchilla [41], LaMDA [102], GPT-J [107], LLaMA [103], GPT-4 [76], BloombergGPT [117]

https://arxiv.org/abs/2304.13712

How do Transformer-based LLMs Work?

A simplified version of LLM training process



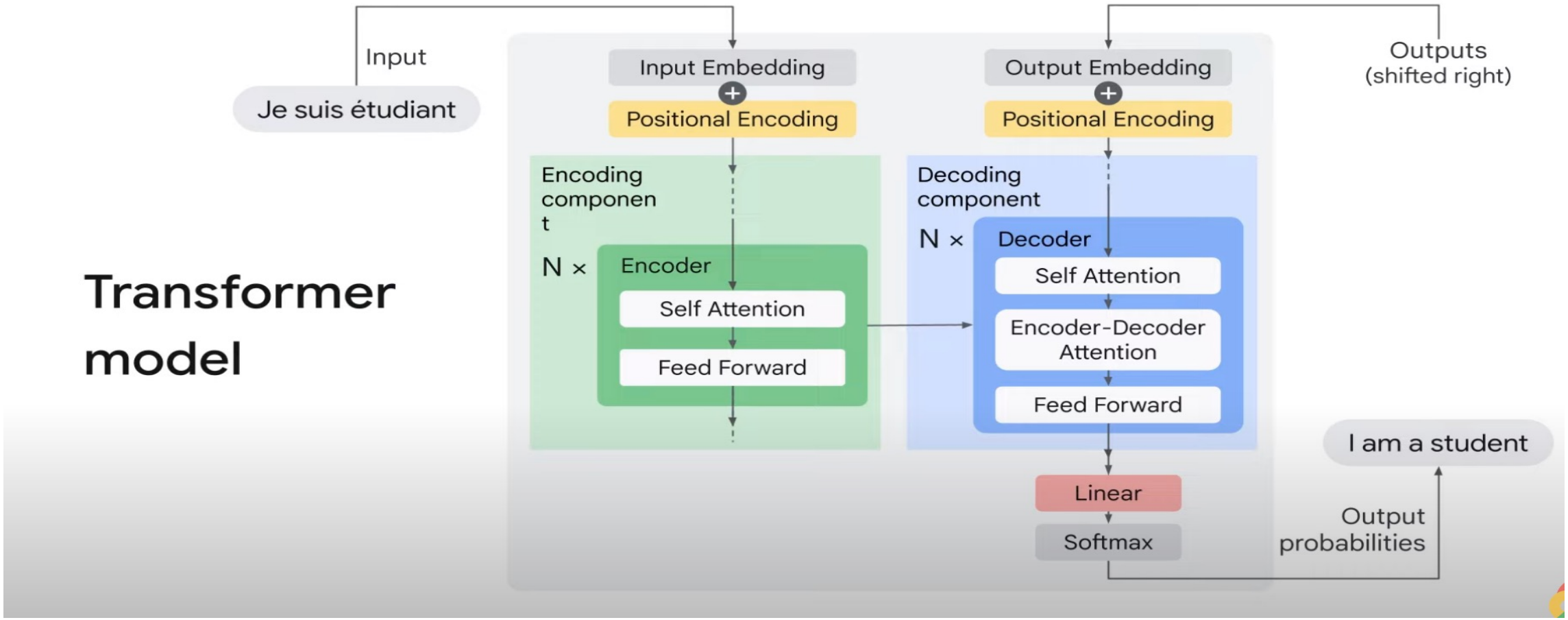
Massive advantage over RNN based encoder-decoder architecture since it allows us to:

- Take advantage of parallelization GPU/TPU.
- Process much more data in the same amount of time.
- Process all tokens at once!

https://www.youtube.com/watch?v=t45S_MwAcOw

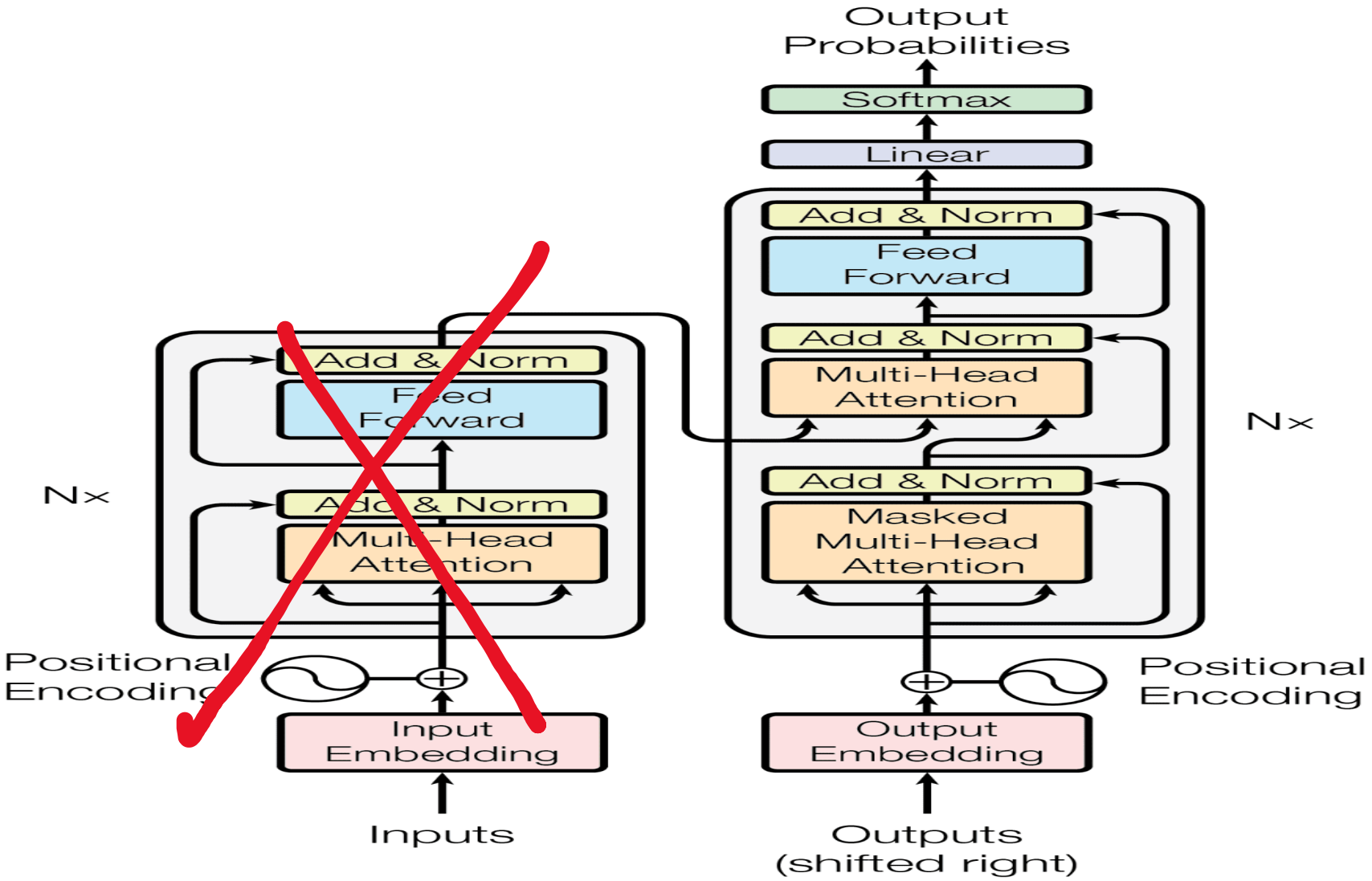
How do Transformer-based LLMs Work?

A simplified version of LLM training process



https://www.youtube.com/watch?v=t45S_MwAcOw

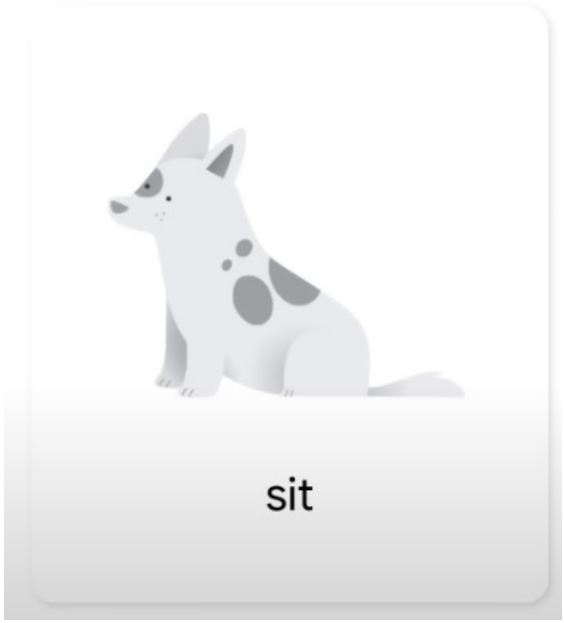
Generative Pre-trained Transformer (GPT) -- Architecture (Decoder-Only)

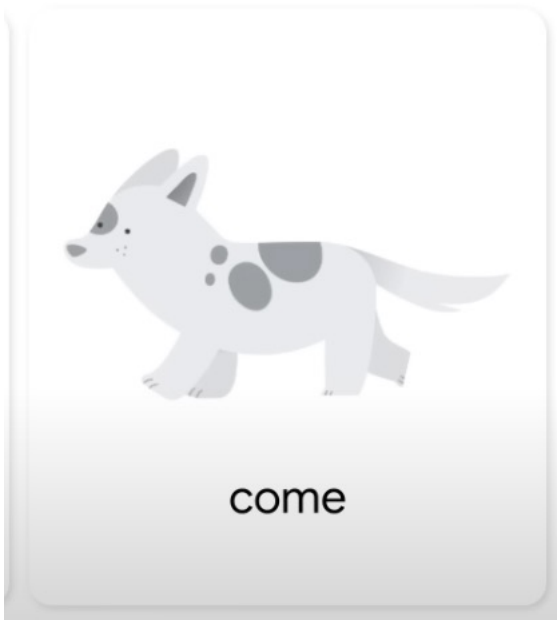


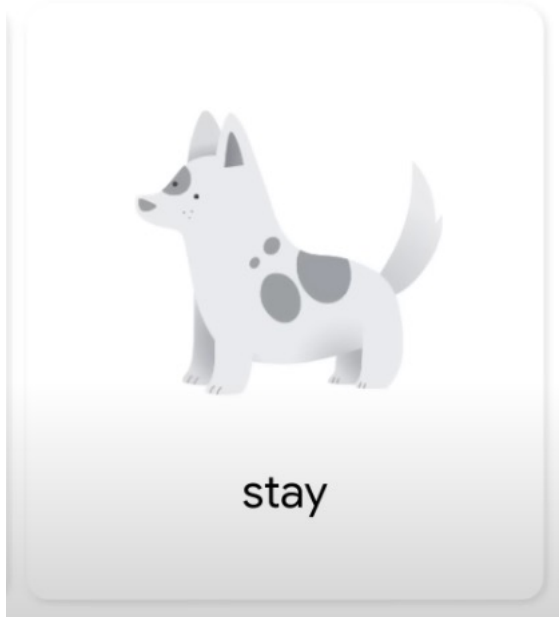
Large Language Models - Training

1. Pretraining using Self-supervised learning
2. Supervised fine-tuning (Instruction Tuning)
3. Reinforcement learning from human feedback (Alignment with human values)
   - nudging the LLM towards values you desire

Training a dog



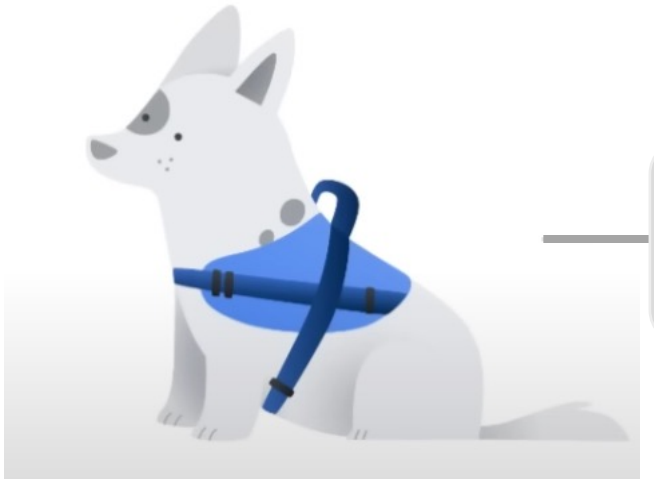




A good canine citizen

Source

Training a dog for Special Services



+ special trainings

police dog

guide dog

hunting dog

Source

Similar idea applied to Large Language Models (LLMs)

Pre-training or Self-supervised Learning

- **Model at the start:**
  - Zero knowledge about the world
  - Can't form English words (doesn’t have language skills)
- **Learning objective:** Next token prediction
- **Giant corpus of text data**
- **Often scraped from the internet "unlabeled”**
- **Self-supervised learning**
- **After training**
  - Learns language
  - Learns knowledge

Source

Two Types of Large Language Models (LLMs)

Base LLM

Predicts next word, based on text training data

Once upon a time, there was a unicorn
that lived in a magical forest with
all her unicorn friends

What is the capital of France?
What is France's largest city?
What is France's population?
What is the currency of France?

Instruction Tuned LLM

Tries to follow instructions

Fine-tune on instructions and
good attempts at following those instructions.

RLHF: Reinforcement Learning with
Human Feedback

What is the capital of France?
The capital of France is Paris.

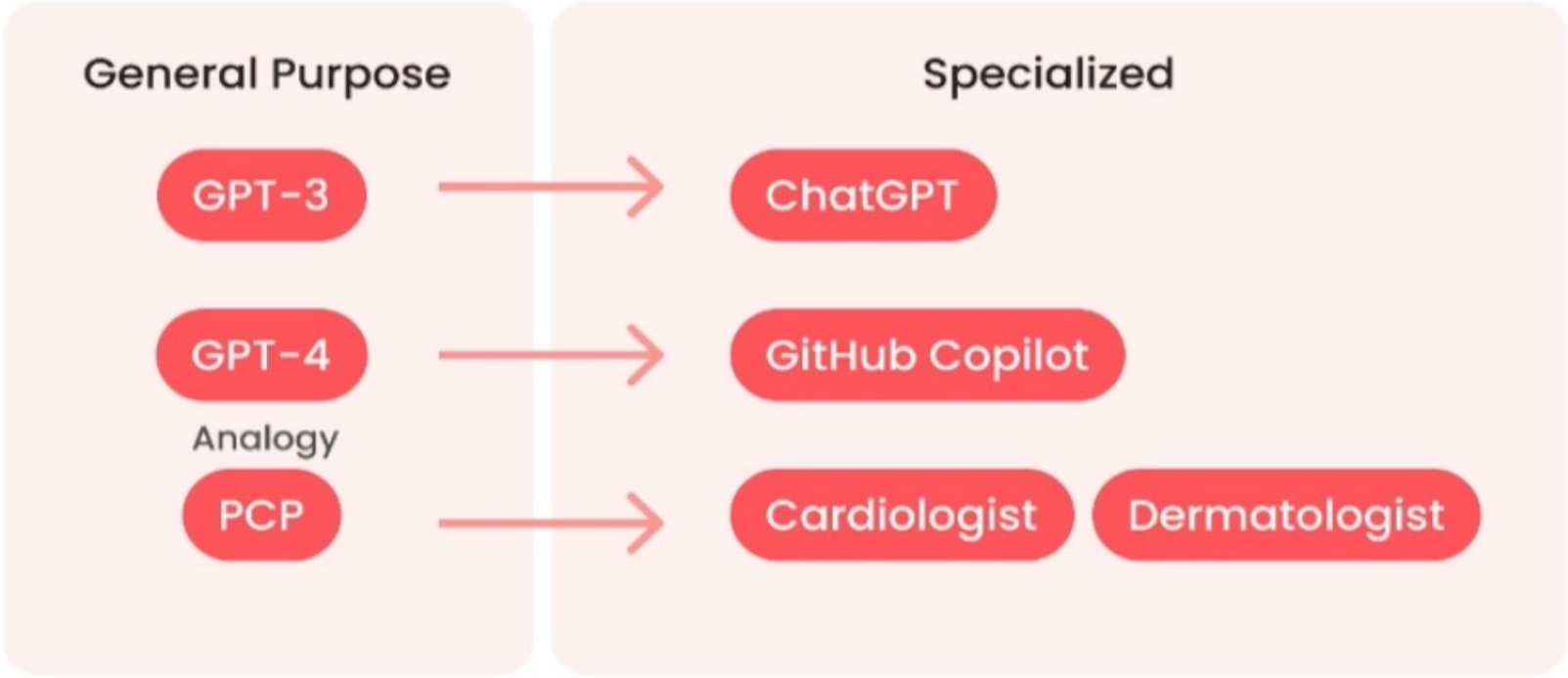
Source

How to use LLMs?

1. Fine-Tuning the LLMs (Supervised Learning)

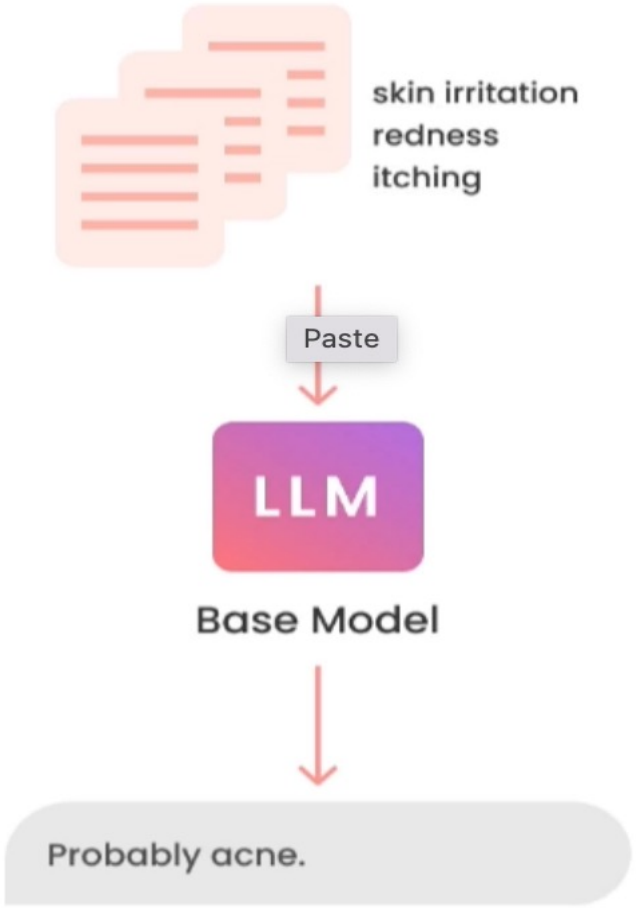
2. LLMs Prompt Engineering

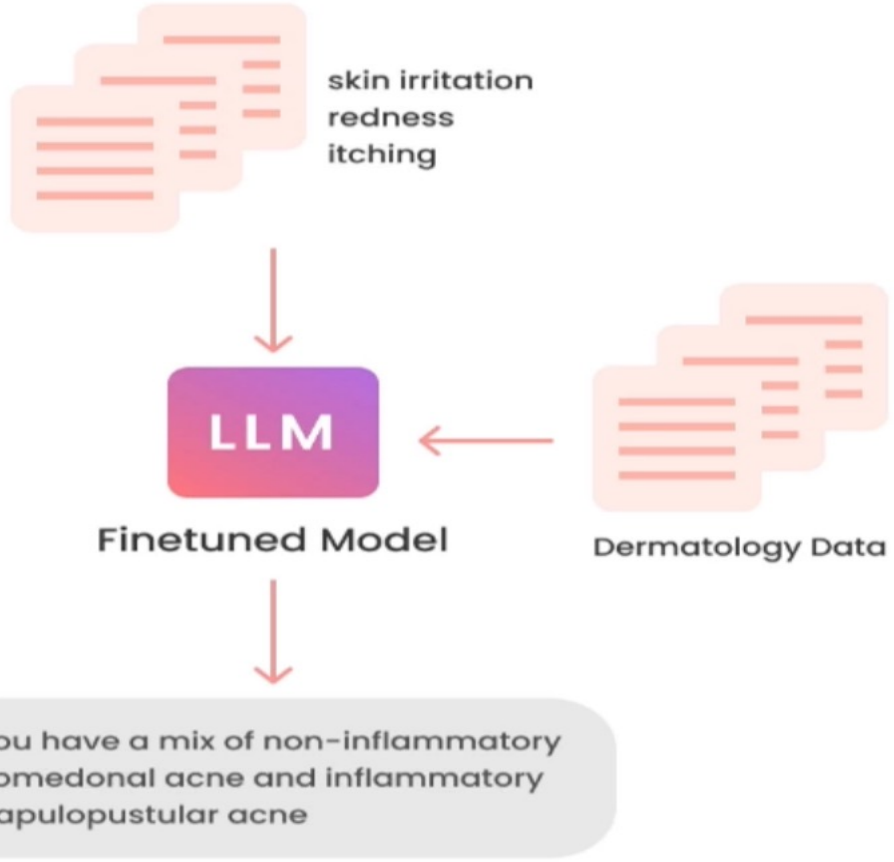
Why Fine-tuning?



Source

What does fine-tuning do for your model?





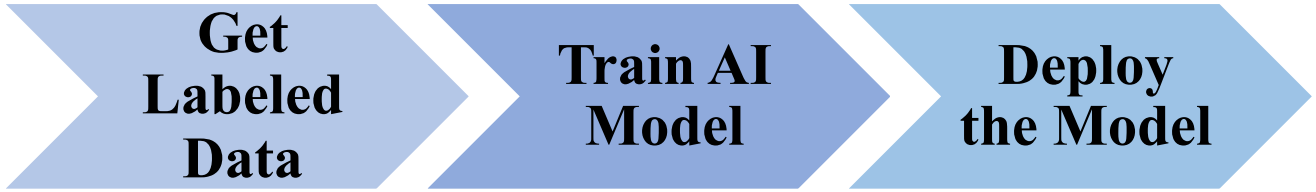
- **Steers the model to more consistent outputs**
- **Reduces hallucinations**
- **Customizes the model to a specific use case**

How to use LLMs?

1. Fine-Tuning the LLMs
2. LLMs Prompt Engineering

Prompting is revolutionizing AI Application Development

Supervised Learning (Fine-Tuning)

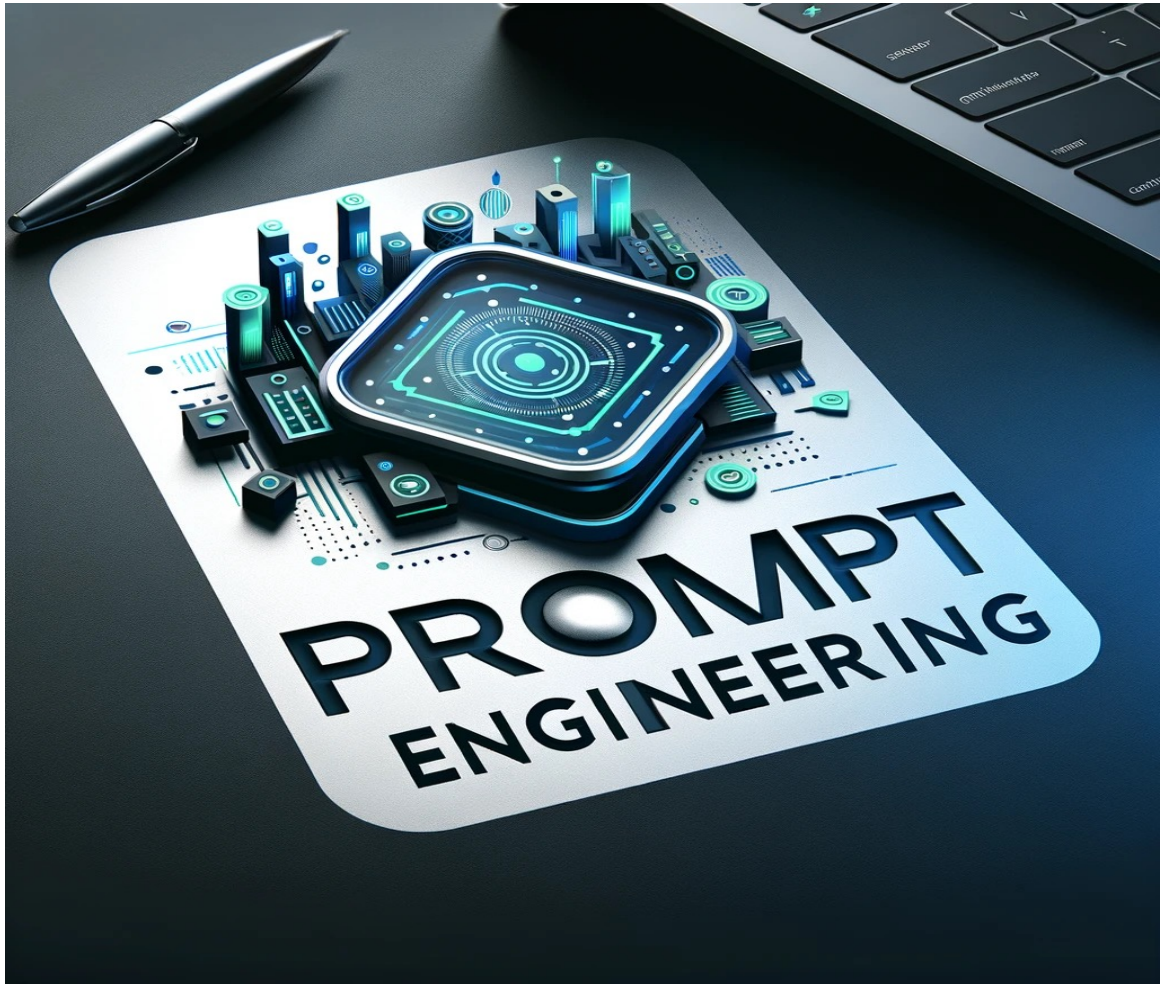


Prompt-based AI

WHAT IS PROMPT ENGINEERING

- Prompt engineering is the practice of **designing and refining specific text prompts to guide generative AI models, such as Large Language Models (LLMs), in generating desired outputs**.

- It involves crafting clear and specific instructions and allowing the model sufficient time to process information.



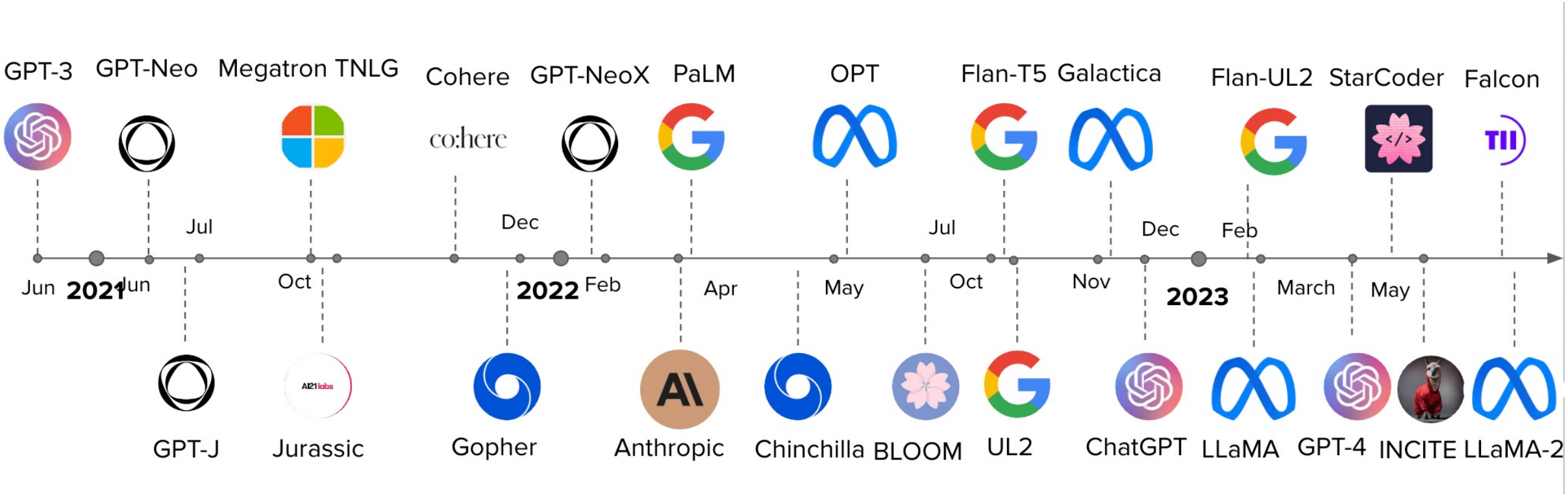
Picture generated by DALLE-3

"Prompt engineering is more about communicating than coding."

Prompting Techniques

- Many advanced prompting techniques have been designed to improve performance on complex tasks
  1. In-Context Learning (ICL)
  2. Chain-of-thought (CoT) prompting
  3. Self-Consistency
  4. Knowledge Generation Prompting
  5. ReAct
  6. …

Text-to-Text Foundation Models since GPT3



*only LLMs with >1B parameters & EN as the main training language are shown.
*Comprehensive list: https://crfm.stanford.edu/helm/v1.0/?models=1

Models Access

Open access

All model components are publicly available:

- Open source **code**
- Training **data**
  - Sources and their distribution
  - Data pre-processing and curation steps
- Model **weights**
- **Paper or blog** summarizing
  - Architecture and training details
  - Evaluation results
  - Adaptation to the model
    - Safety filters
    - Training with human feedback

Limited access

- Limited access falls somewhere in between open and closed.
- The access can be via API or through a review process of call for research proposals and then granting approved proposals limited model access.

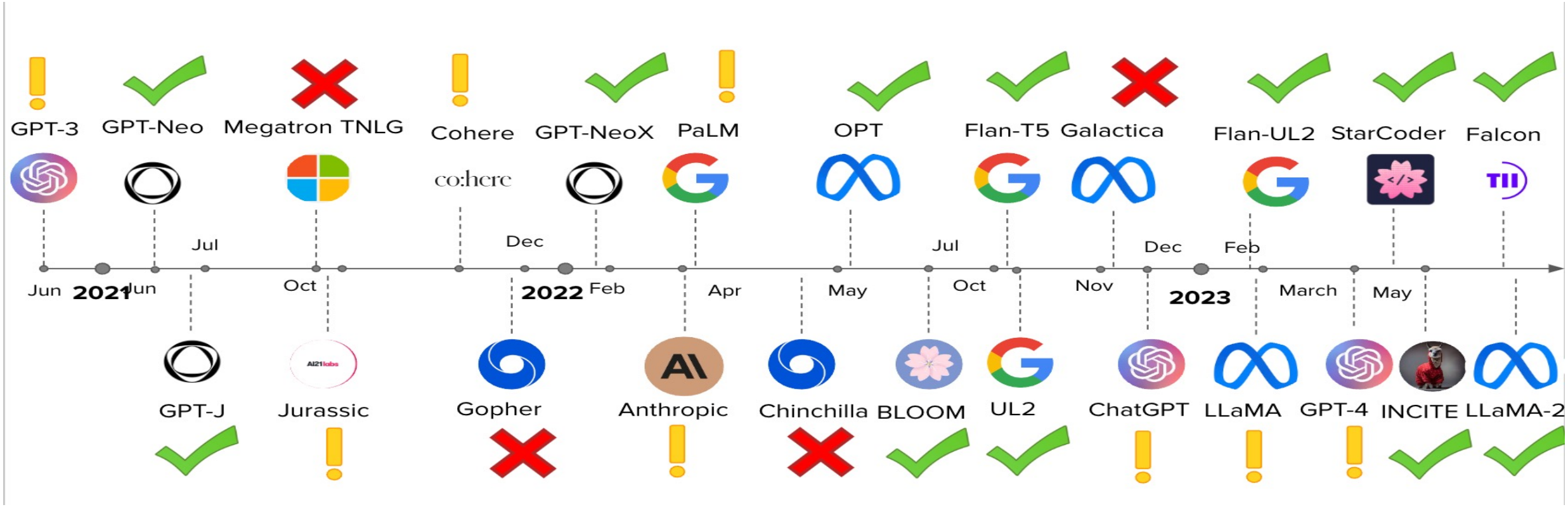
Closed access

Only research paper or blog is available and *may* include overview of

- Training data
- Architecture and training details (including infrastructure)
- Evaluation results
- Adaptation to the model
  - Safety filters
  - Training with human feedback

https://crfm.stanford.edu/helm/v1.0/?models=1

Text-to-Text Foundation Models since GPT3



*** Hugging Face has become the defacto hub for open source ML.**

https://crfm.stanford.edu/helm/v1.0/?models=1

To recap

How LLMs are Built?

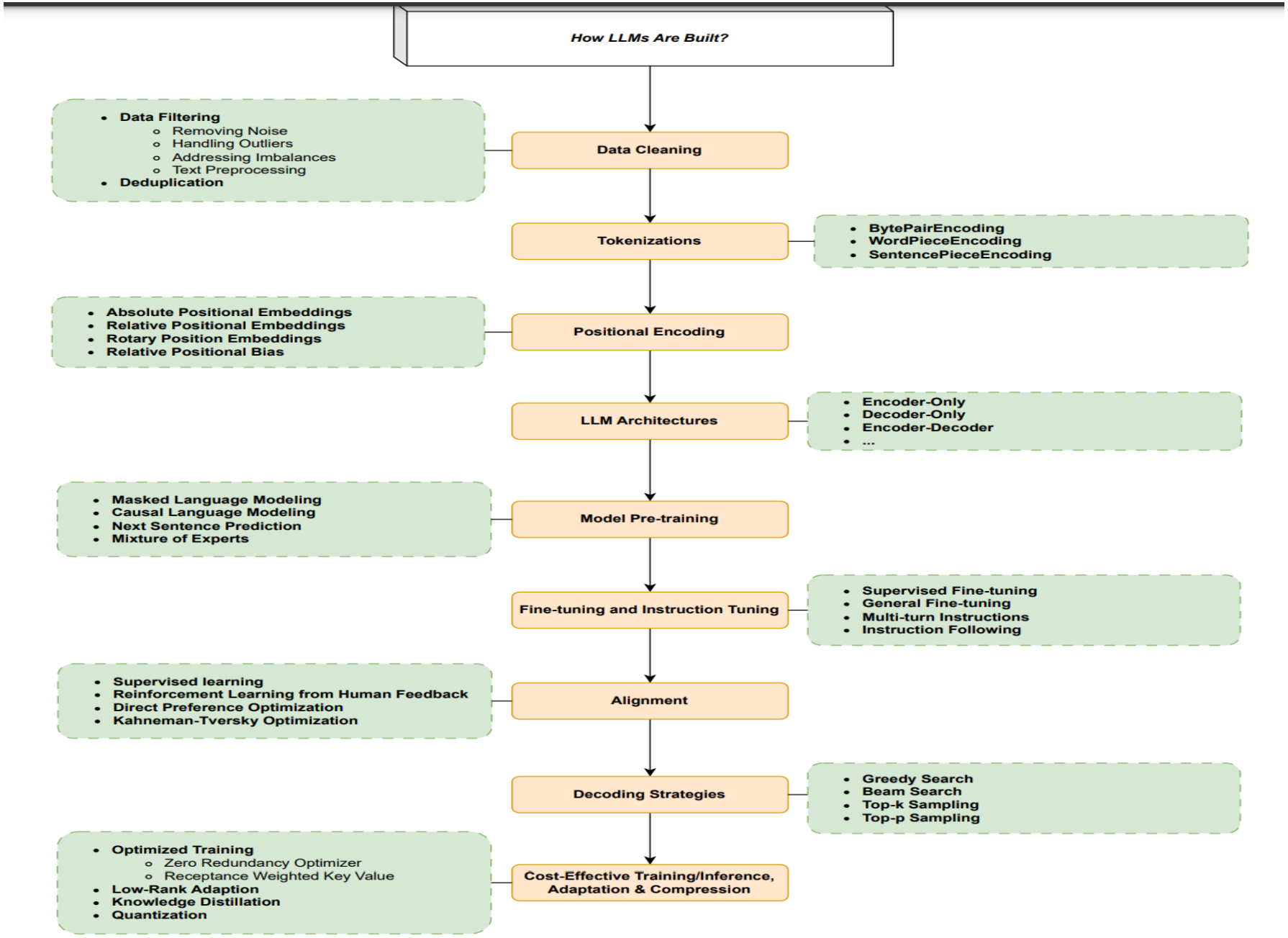


Fig. 25: This figure shows different components of LLMs.

https://arxiv.org/pdf/2402.06196

Web-based vs Software application use of LLMs

Web-based interface applications e.g. ChatGPT, Bard, or Bing Chat

Software-based applications e.g. email routing, document search

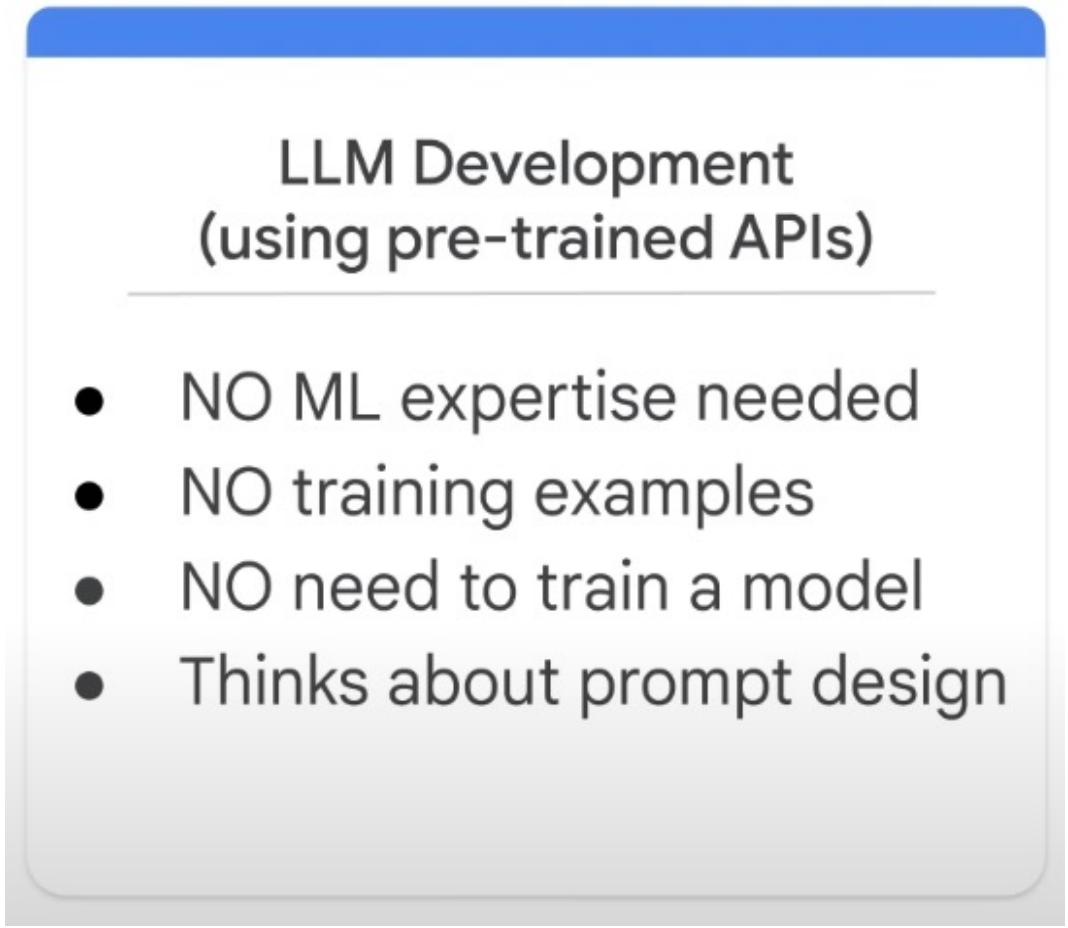
Benefits of using Large Language Models

1. A single model can be used for different tasks

2. The fine-tuning process requires minimal field data (Transfer Learning, Domain Adaptation)

3. The performance is continuously growing with more data and parameters

LLM Development vs. Traditional Development



Large Language Models – Capabilities

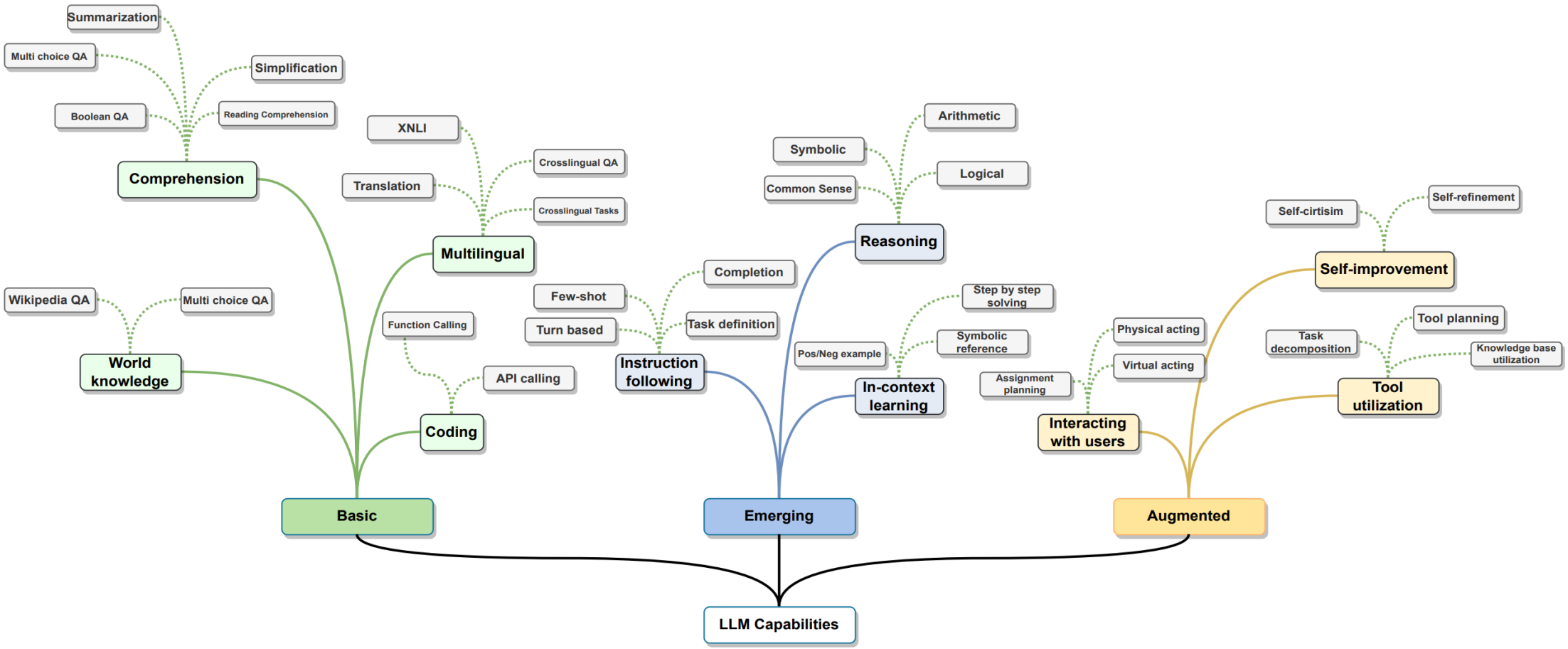


Fig. 1: LLM Capabilities.

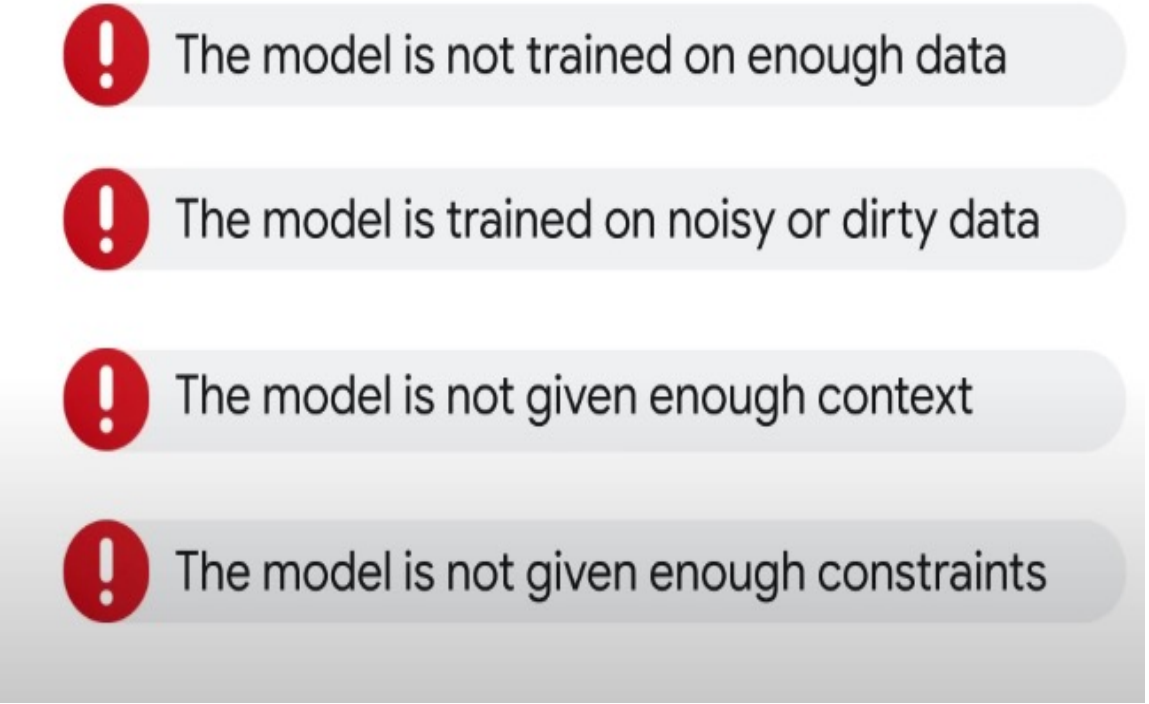
https://arxiv.org/pdf/2402.06196

Some Limitations of LLMs

- They sometimes write plausible-sounding but incorrect or nonsensical answers
- Lack of common sense
- Inability to handle complex tasks
- Limited knowledge base
- Lack of emotional intelligence
- Biases in the training data

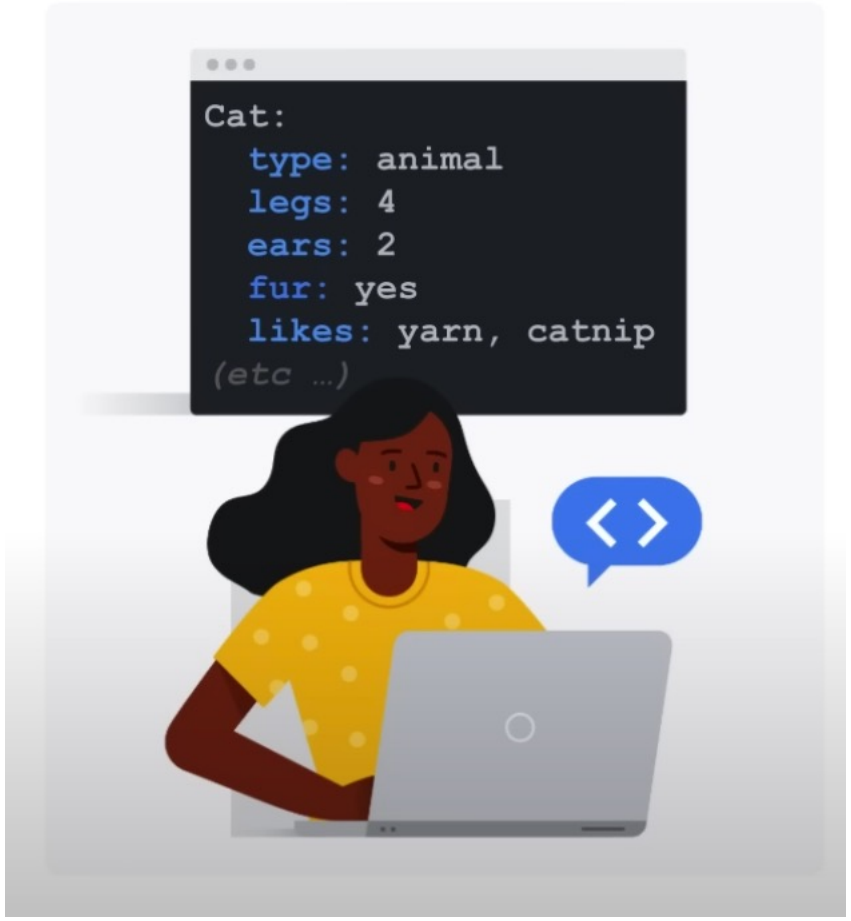
Hallucinations

Hallucination are words or phrases that are generated by the model that are often nonsensical or grammatically incorrect.

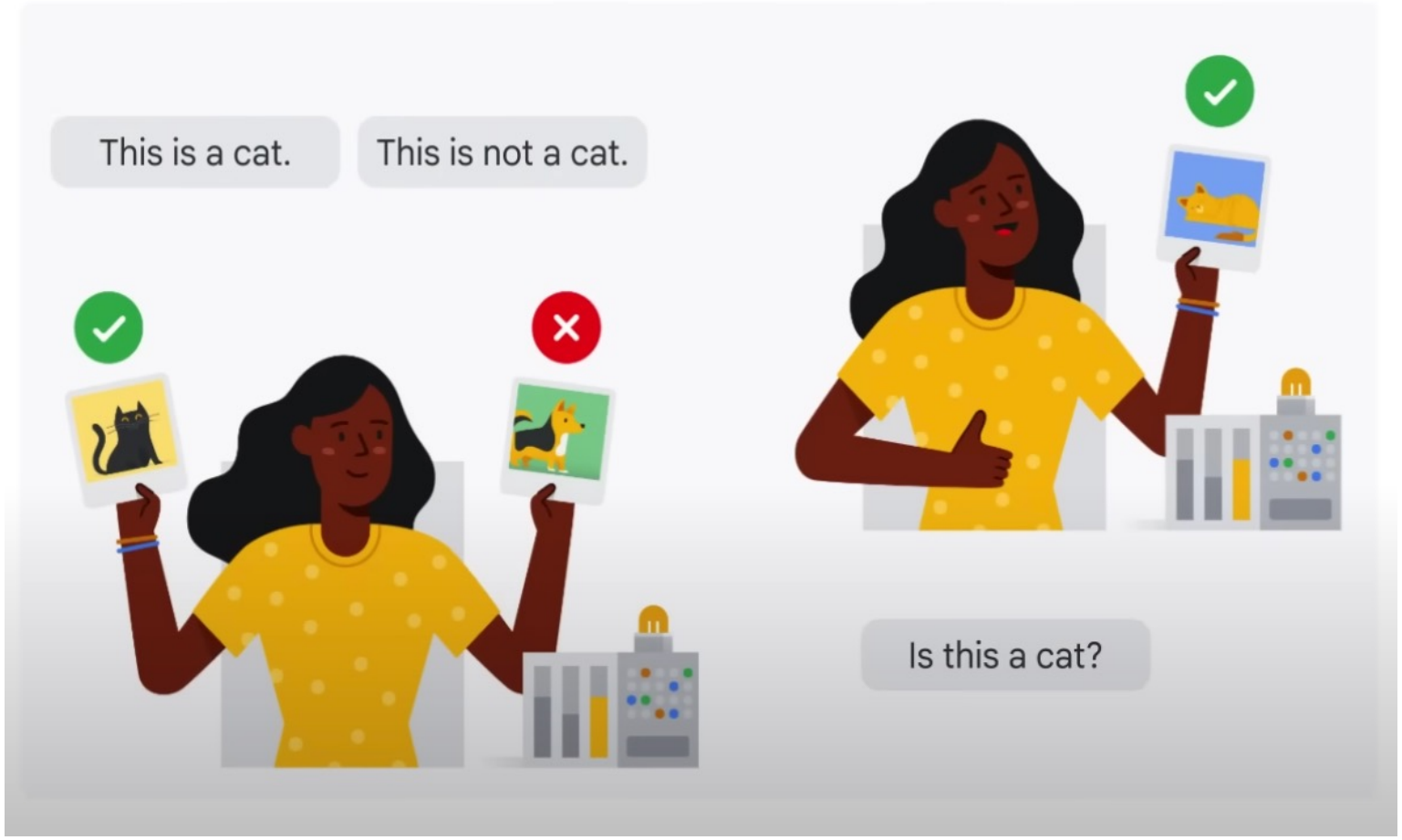


To conclude

Traditional programming

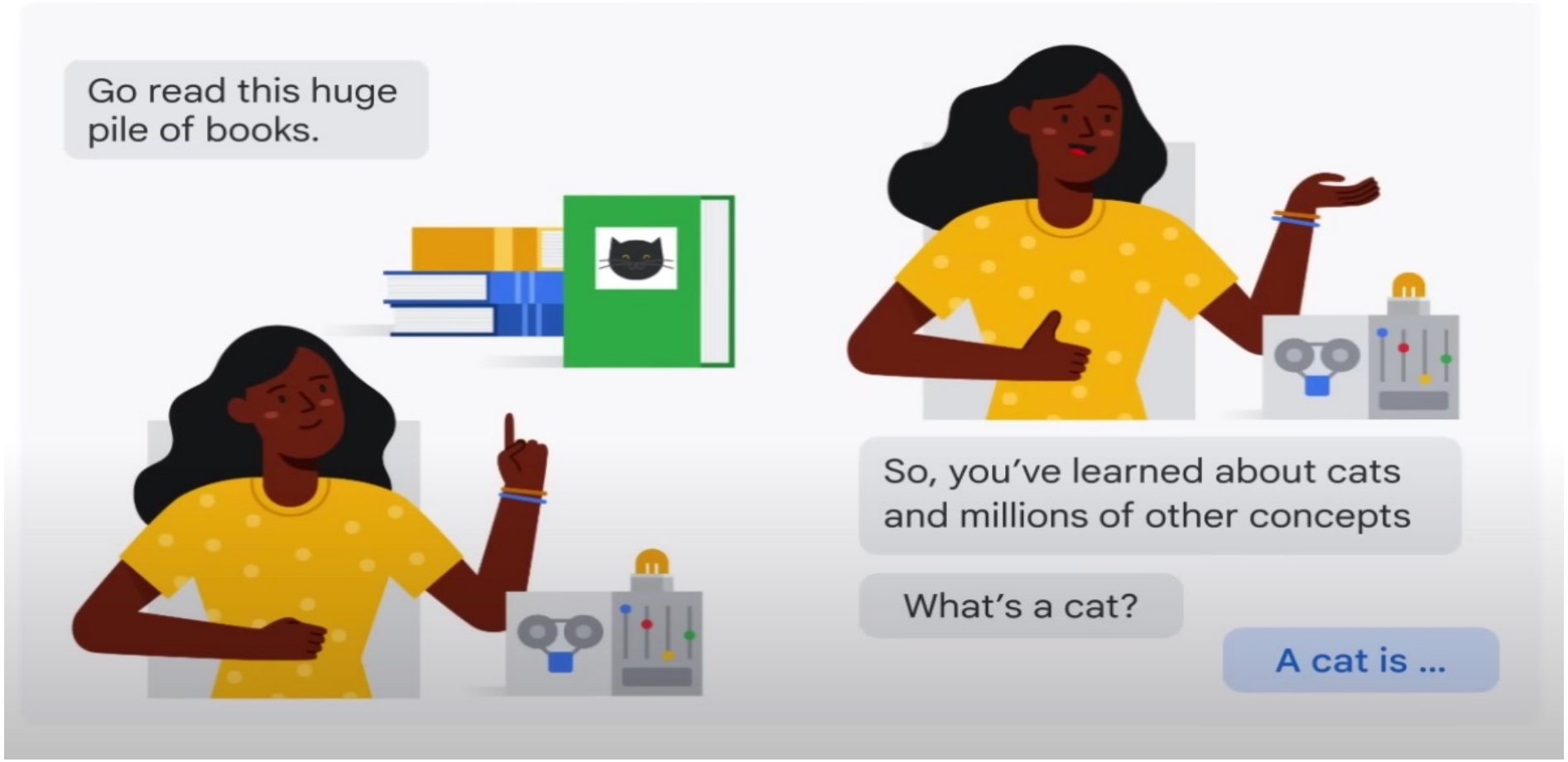


Wave of neural networks | ~2012



Source

Generative language models | LaMDA, PaLM, GPT, etc.



Source

Questions Everyone Asks

Is Generative AI a threat of an opportunity for my business?

How Exactly can I use Generative AI to gain a competitive advantage?

How can I use my data securely with Generative AI?

Reading

- https://www.cloudskillsboost.google/journeys/118/course_templates/536
- https://cloud.google.com/ai/generative-ai
- https://smlbook.org/book/sml-book-draft-latest.pdf?fbclid=IwAR2ztL1GkSuhYJHJJeWACwRFEnAtqZshuq6l-S-0Z6_MHT9o90Qzoy6eMgA
- https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/what-is-generative-ai
- GenAI studio: https://cloud.google.com/generative-ai-studio?hl=fr
- Survey on Large Language Models https://arxiv.org/pdf/2303.18223.pdf

Thank You!