Introduction to Generative AI

Focusing on NLP & Tamil

Abinaya Mahendiran
Aug 30, 2023

Abinaya Mahendiran

- CTO, Nunnari Labs,

- Program Manager, IITM

- M.Tech IT, International Institute of Information and Technology Bangalore

- Mentor at TeamEpic, Springboard, Topmate.io

- Volunteer at AI Tamil Nadu, WTM, Data Conversations, GHCI, WAI, Women Who Code

- Interests: Building NLP/NLU/NLG/MLOps/Gen AI systems, Open source, Applied Research

https://abinayam02.github.io

@freakynut

https://medium.com/@abinayamahendiran

https://www.linkedin.com/in/abinayamahendiran/

https://topmate.io/abinaya_mahendiran

Agenda

What is Generative AI?

Types of Models

Traditional Approach

Transfer Learning

Types of Fine-tuning
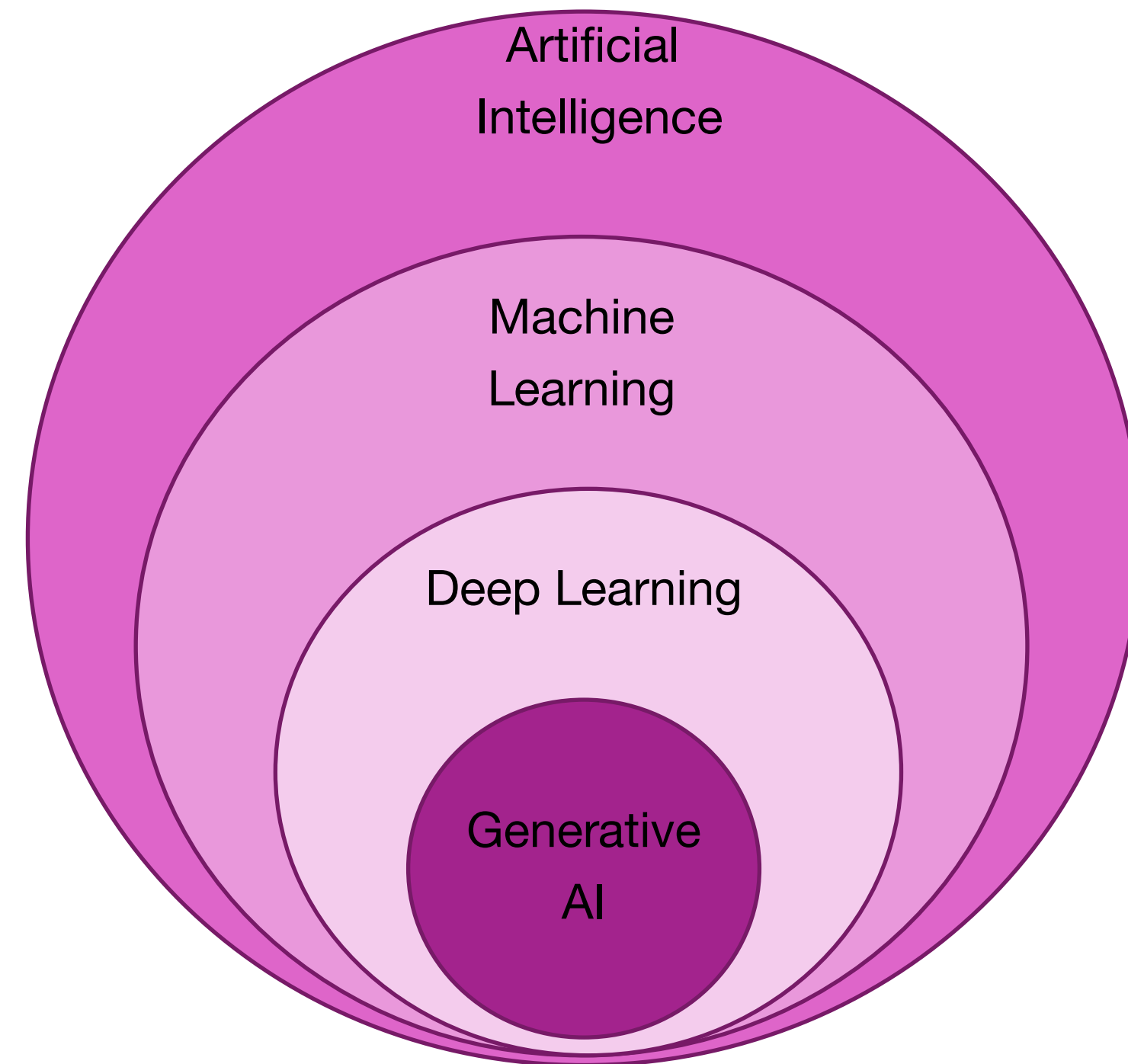
Foundation Models

NLP Timeline

Core Concepts

Limitations of Generative AI

Research Directions

Generative AI for Tamil

CFP: Abstract

What is Generative AI?

Artificial Intelligence

Machine Learning

Deep Learning

Generative AI

- **AI:** Build intelligent agents that can act like humans autonomously.

- **ML:** A machine learns the patterns in the data by training a model.

  - **Supervised learning** – Use labeled data, train models, predict on unseen data. Classification/Regression.

  - **Unsupervised learning** – Use unlabelled data to identify groups or clusters.

  - **Semi supervised learning** – uses little labeled data and more unlabelled data to train models.

  - **Reinforcement Learning** – An agent performs actions based on the environment and learns through trial and error (either rewarded or punished).

What is Generative AI?
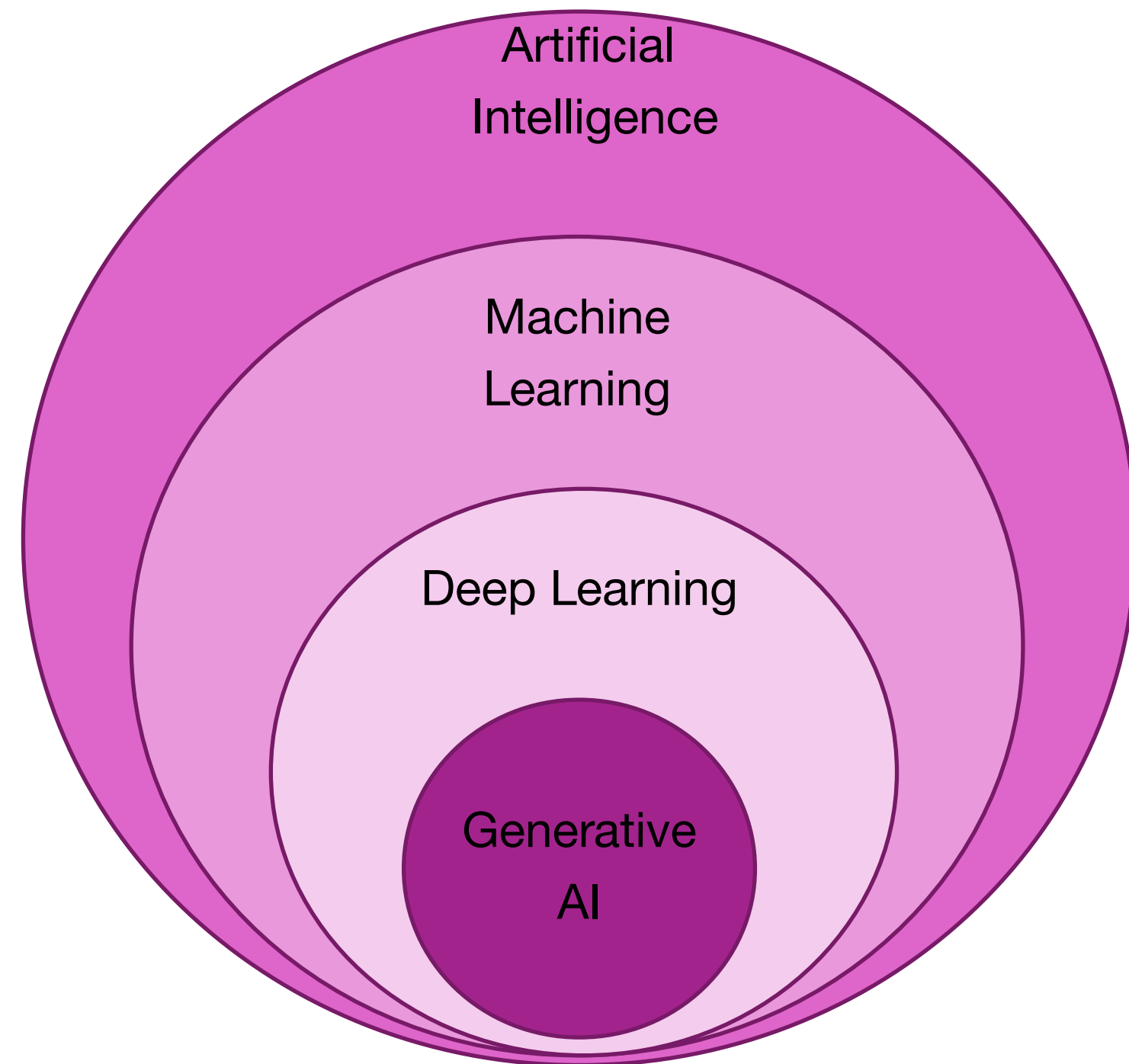
- **DL:** A neural network with interconnected nodes and layers is trained to learn complex patterns in the data. Uses the following methods of learning.

  1. Supervised,
  2. Unsupervised, and
  3. Semi supervised

- **Generative AI:** It is a type of AI that can create new content, such as text, images, audio, and video.

  - Learns from existing data
  - Uses existing knowledge to generate new and unique outputs
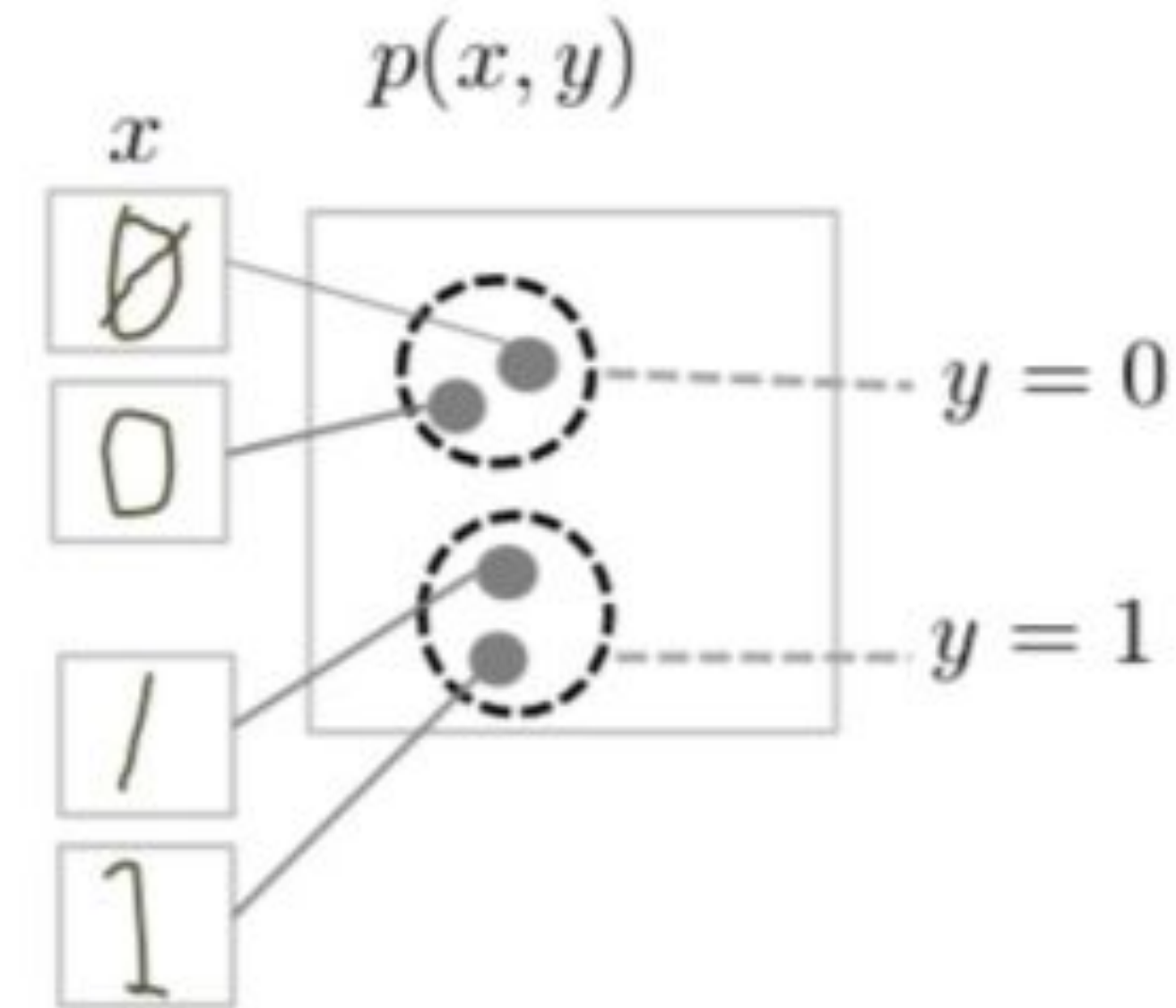
Types of Models

Discriminative Models:

- Discriminates between different classes.

- Approximates the decision boundary or the distribution function, given the data points it predicts the labels.
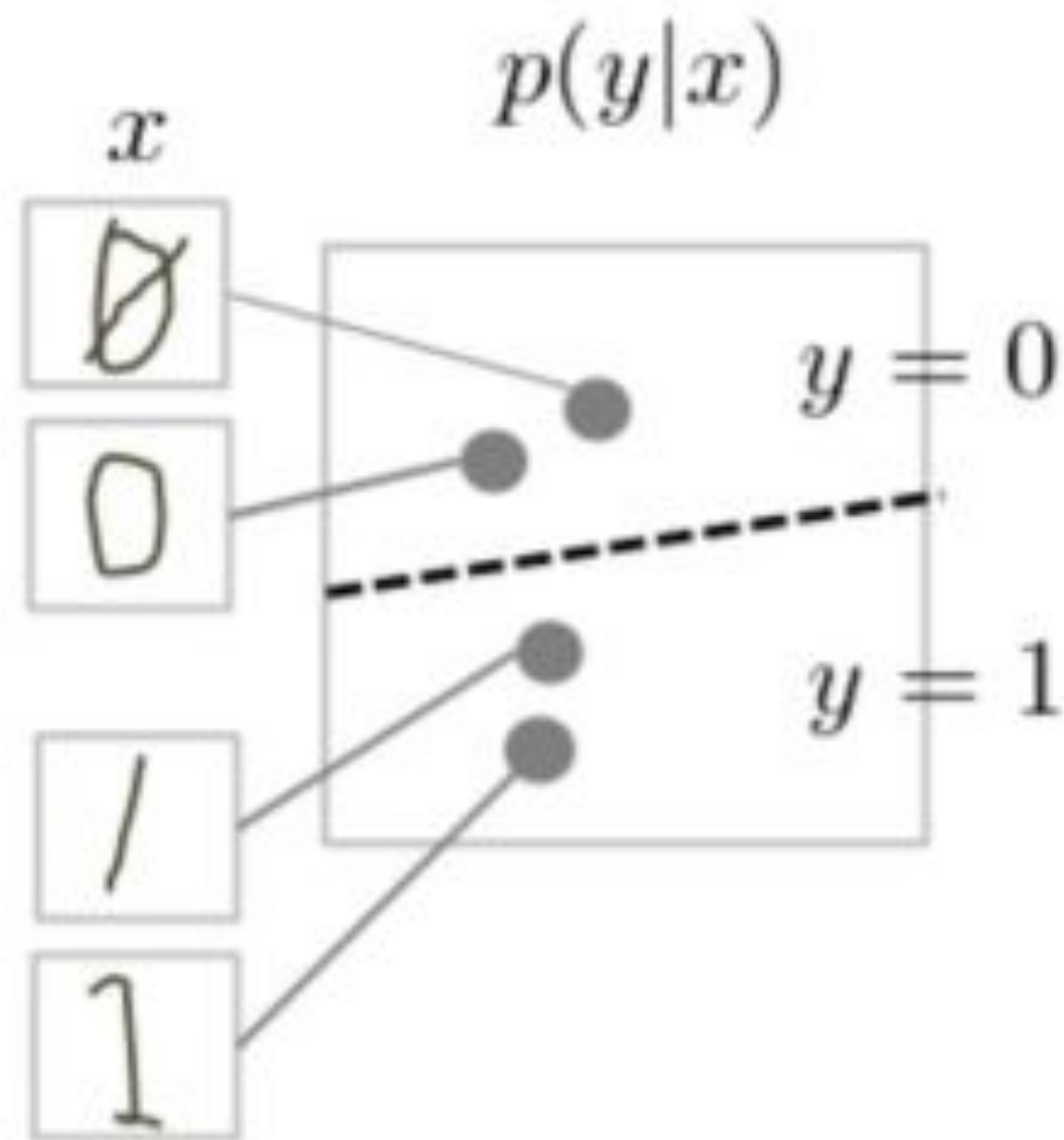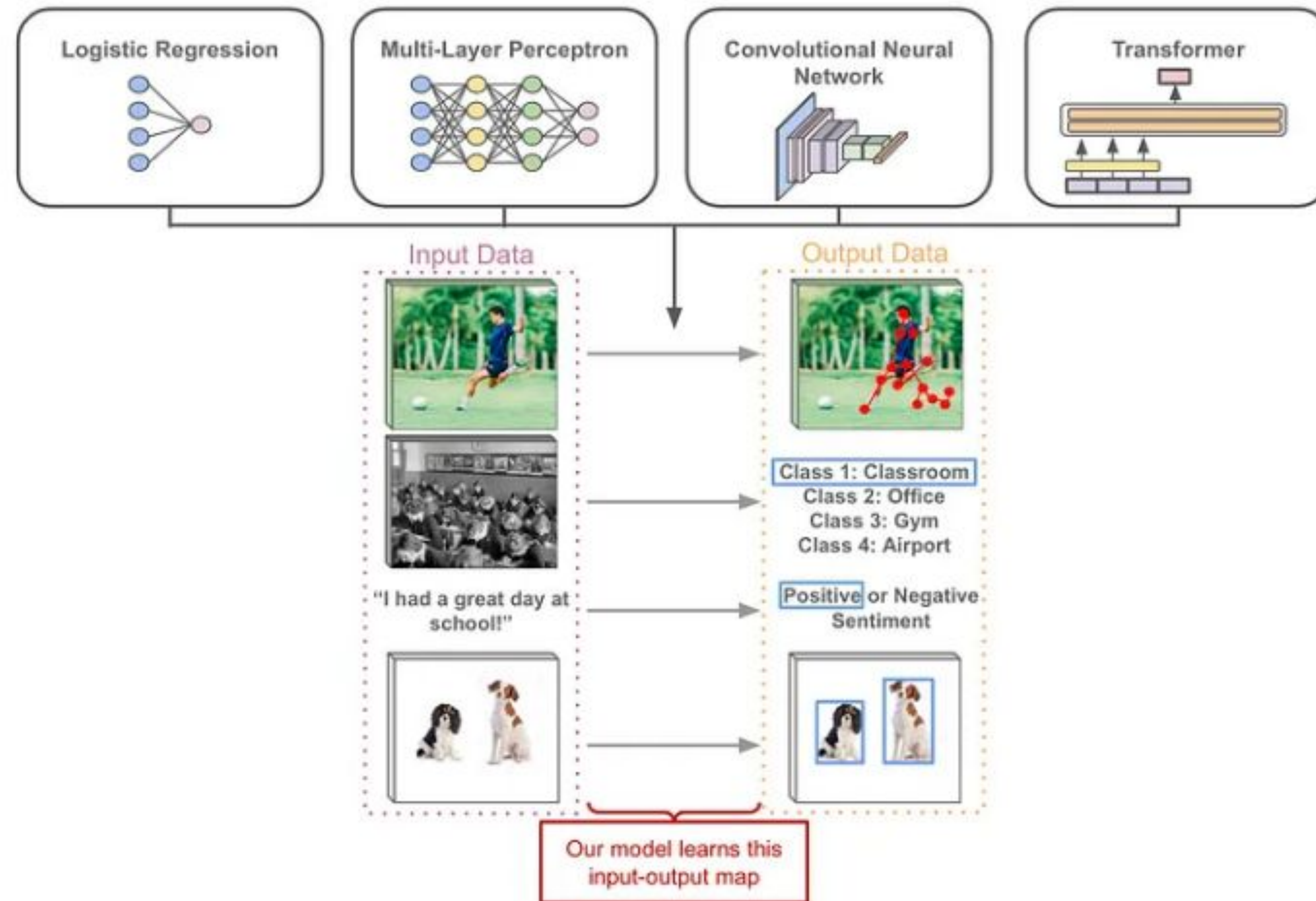
Generative Models:

- Generates new data points.

- Assumes the data distribution and produces convincing data points the are close to its real counterparts in the space.

Traditional Approach

Transfer Learning

Step 1: Pre-training

- Use large amounts of generic data and train on a specific objective function.

- Unlabelled data is used to train on the language modelling objective like MLM.

Step 2: Fine-tuning

- Fine-tuning is done using task-specific objective function.

- Labelled data is used to fine-tune model on the downstream tasks.

https://www.ruder.io/recent-advances-lm-fine-tuning/
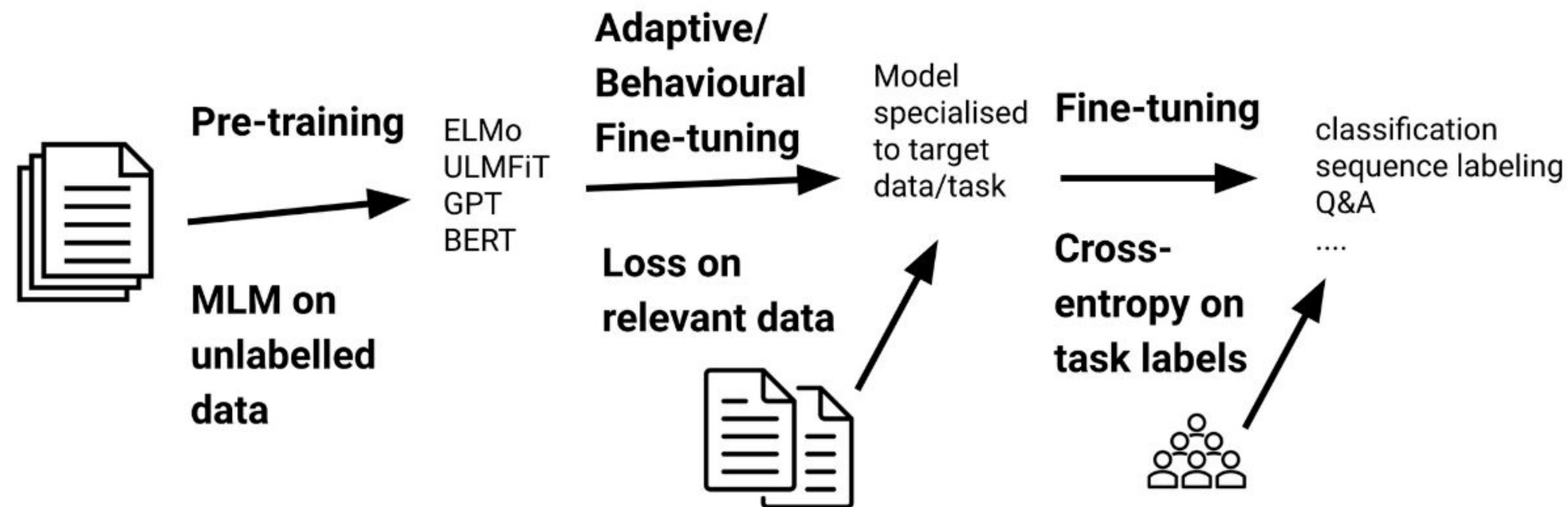
Types of Fine Tuning

Category	Methods	Motivation
Adaptive Fine-tuning	Domain/task/ language adaptive fine-tuning	Specialise to target domain
Behavioural Fine-tuning	Intermediate-task training, self-supervised, frame as MLM	Specialise to target task
Parameter-efficient Fine-tuning	Adapters, sparse parameter permutations, pruning	Reduce space of fine-tuned models
Text-to-text Fine-tuning	Frame as text-to-text, prompt engineering, controllable NLG	Effectively use large autoregressive pre-trained LMs
Mitigating Fine-tuning Instabilities	Stop runs early, use a small lr, regularisation, avoid random init	Reduce variance of fine-tuning runs

https://www.ruder.io/recent-advances-lm-fine-tuning/
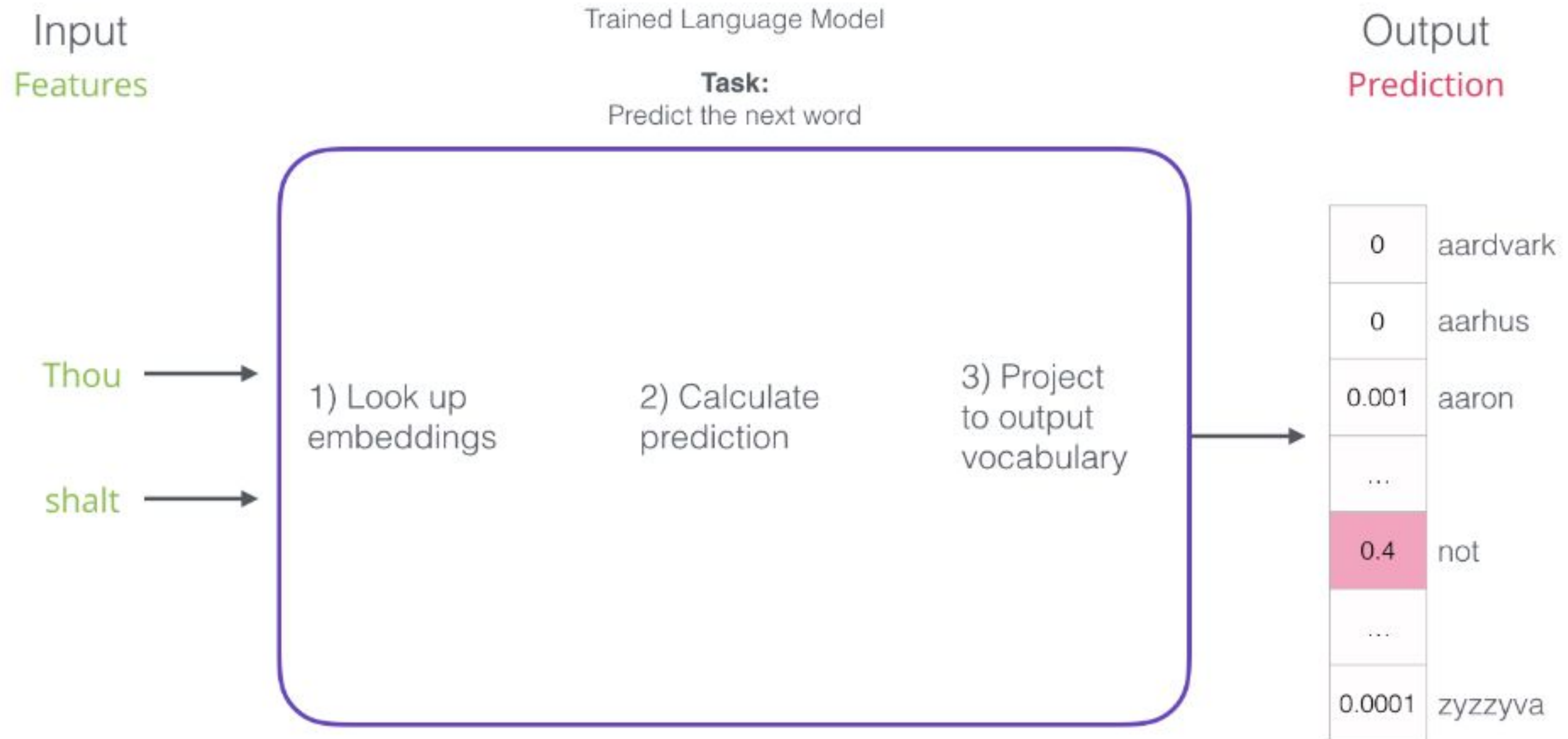
Foundation Models

A foundation model is a large AI model pre-trained on a vast quantity of unlabelled data that was "designed to be adapted" (or fine-tuned) to a wide range of downstream tasks, such as sentiment analysis, image captioning, and object recognition.
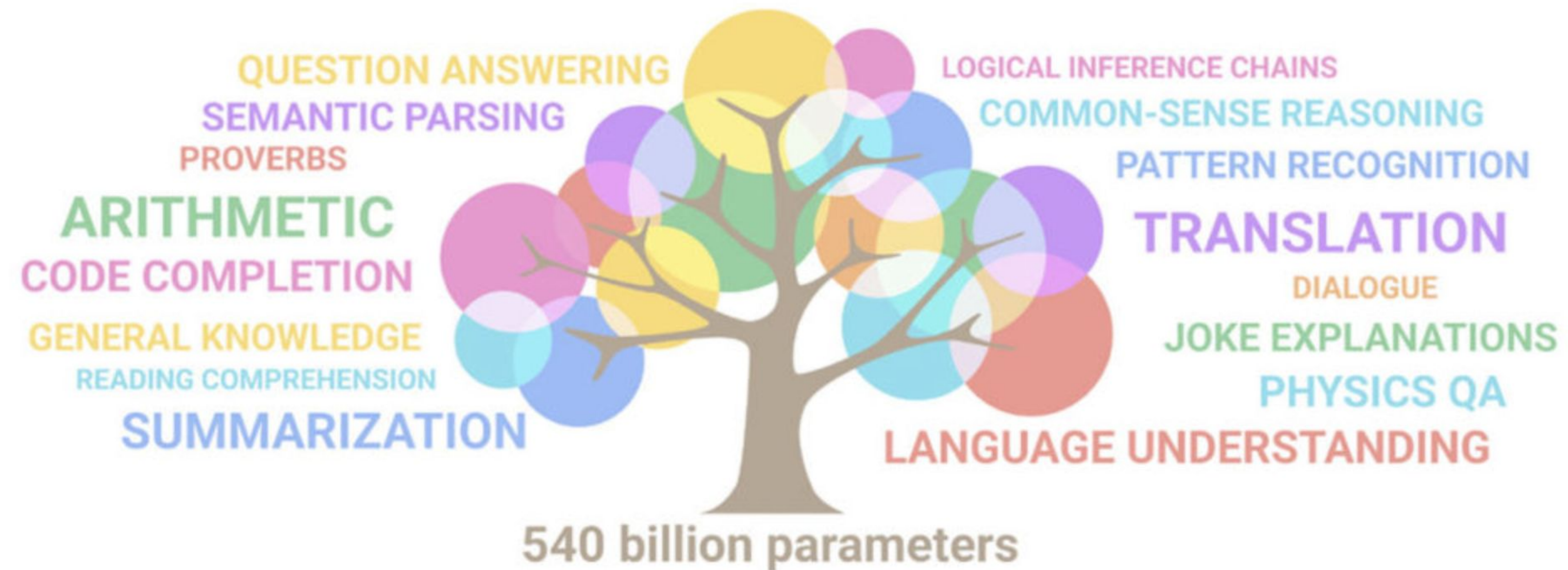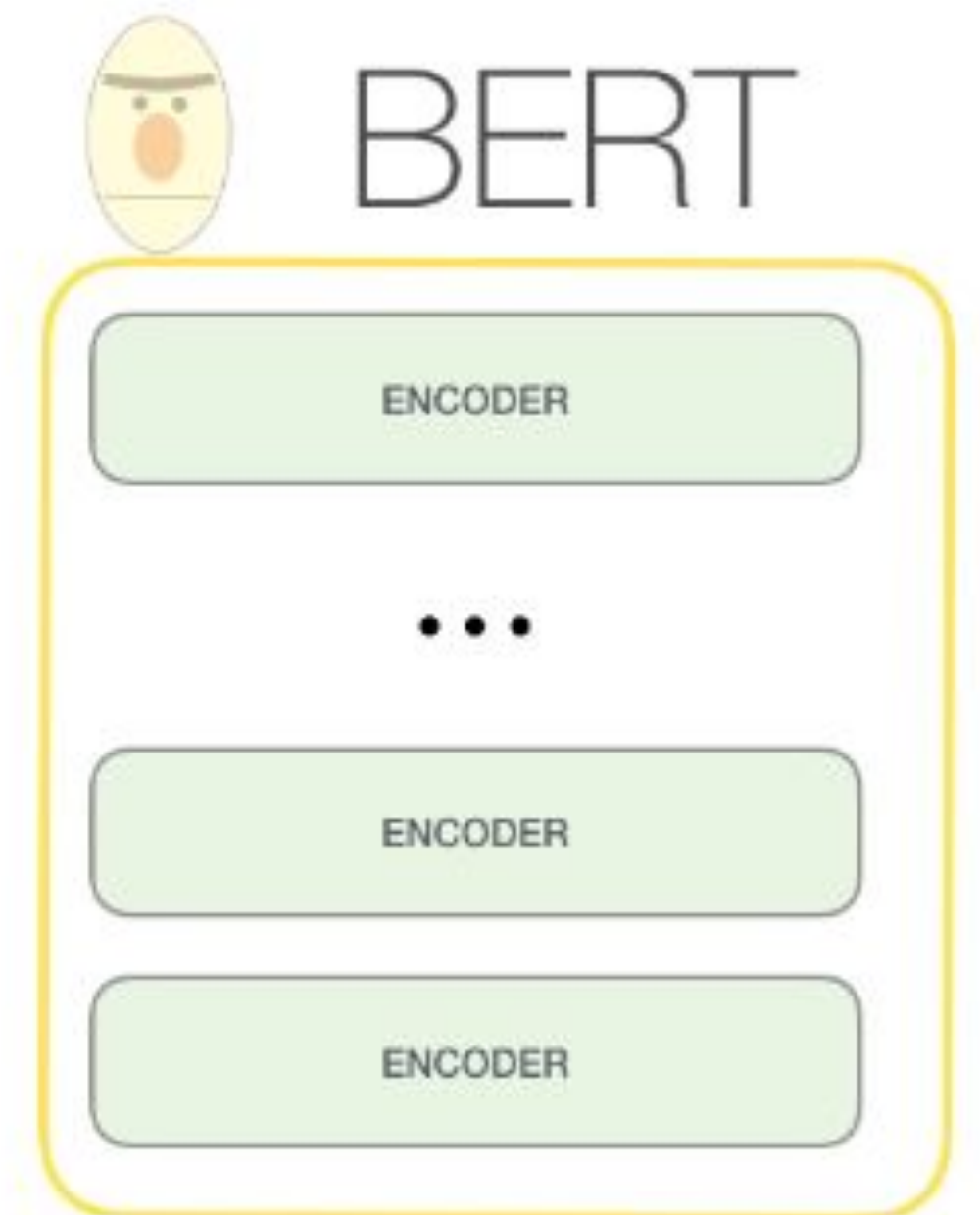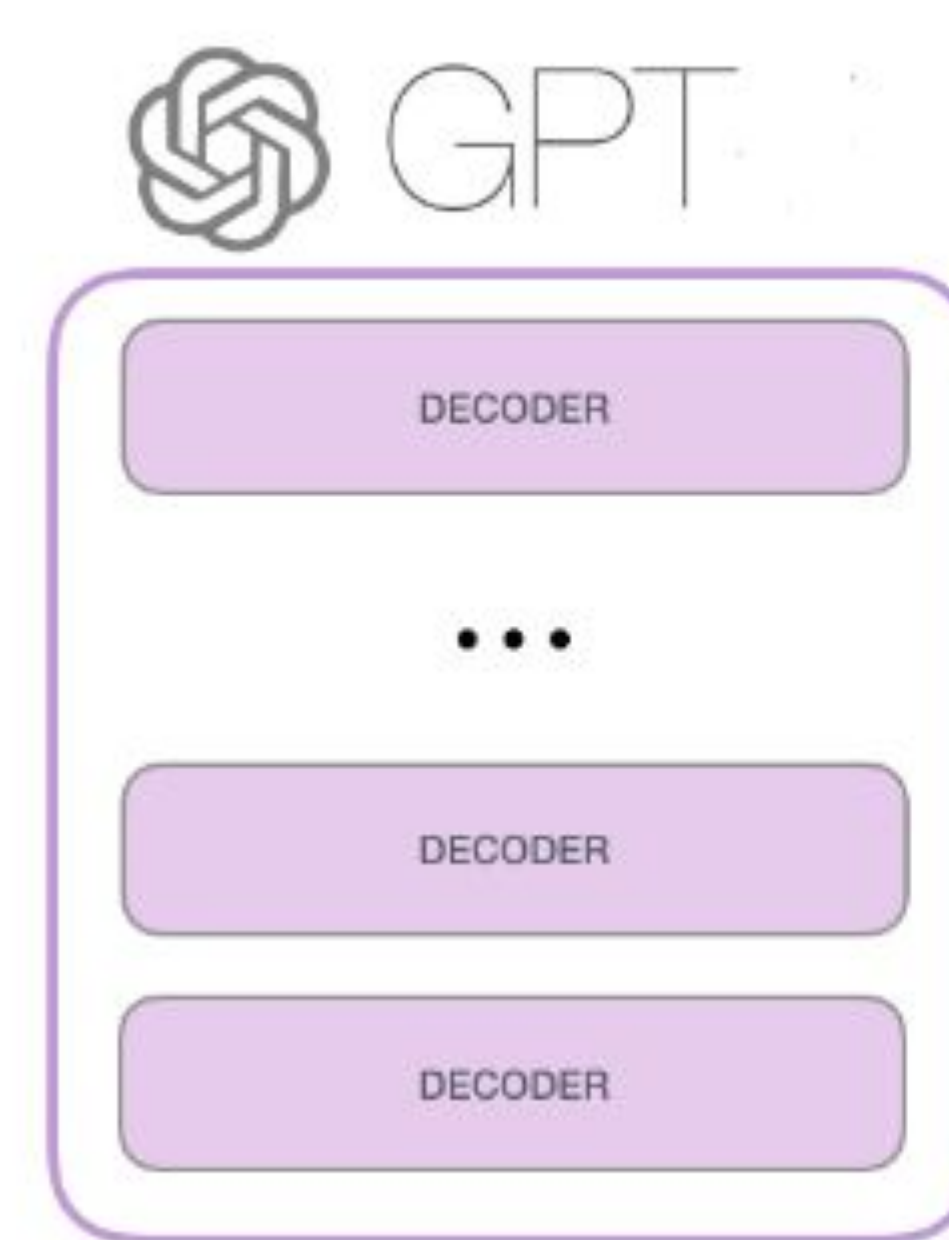
NLP Timeline
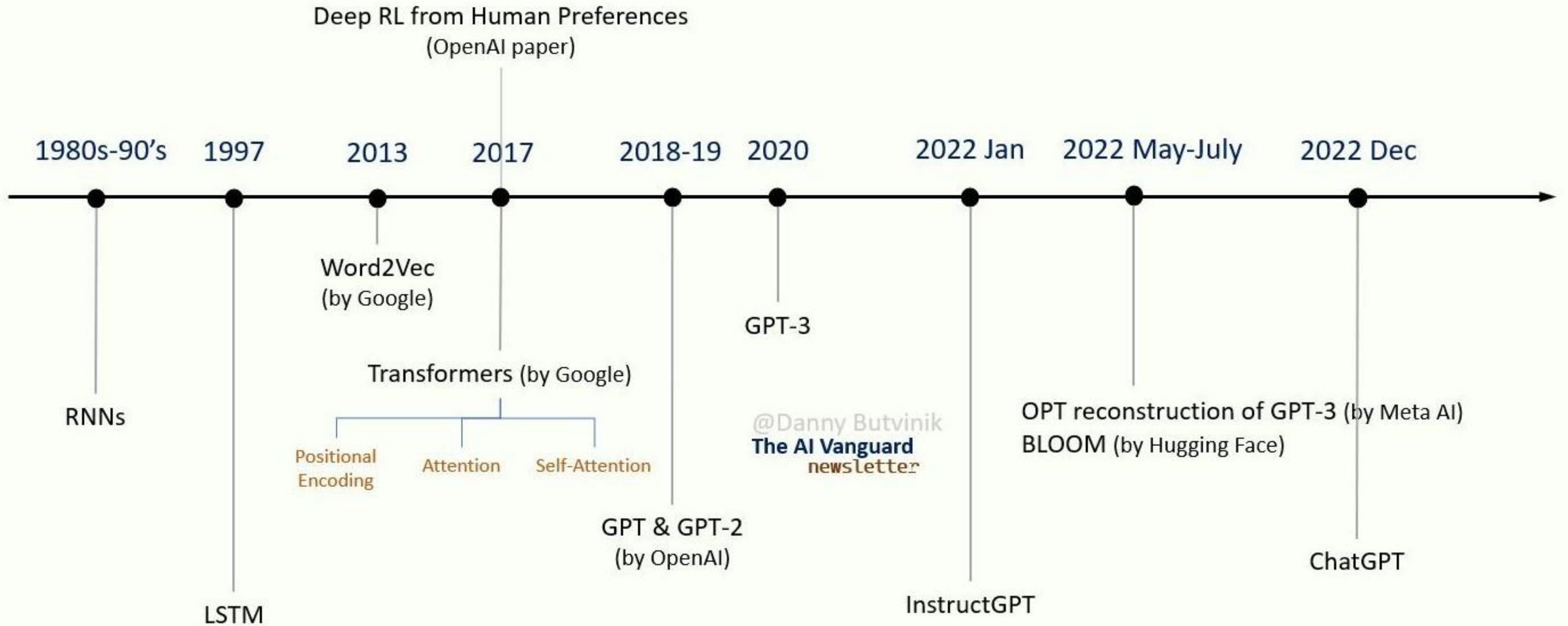
Concepts: Language Modelling

Concepts: Encoder / Decoder

Concepts: RLHF
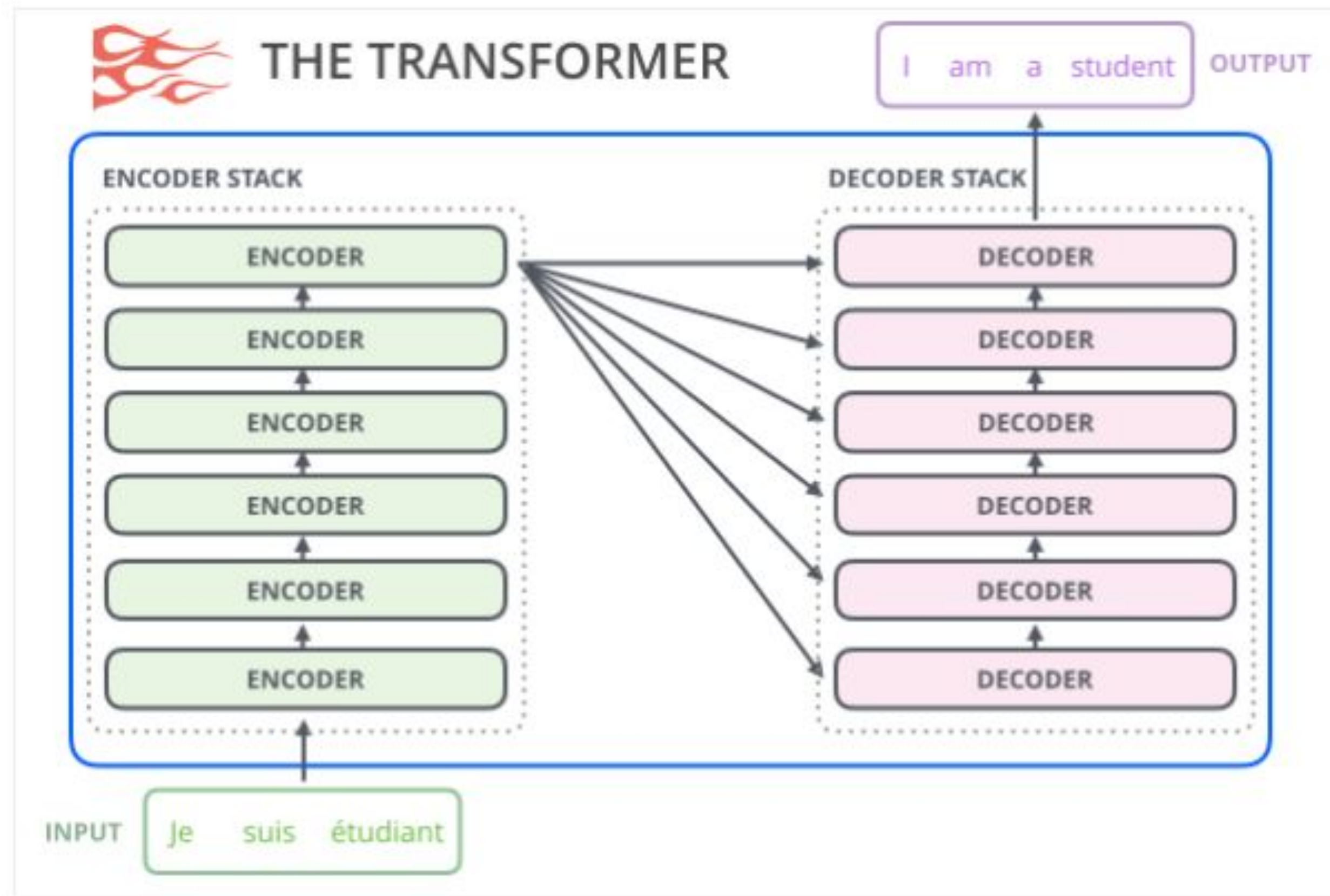
https://huggingface.co/blog/rlhf
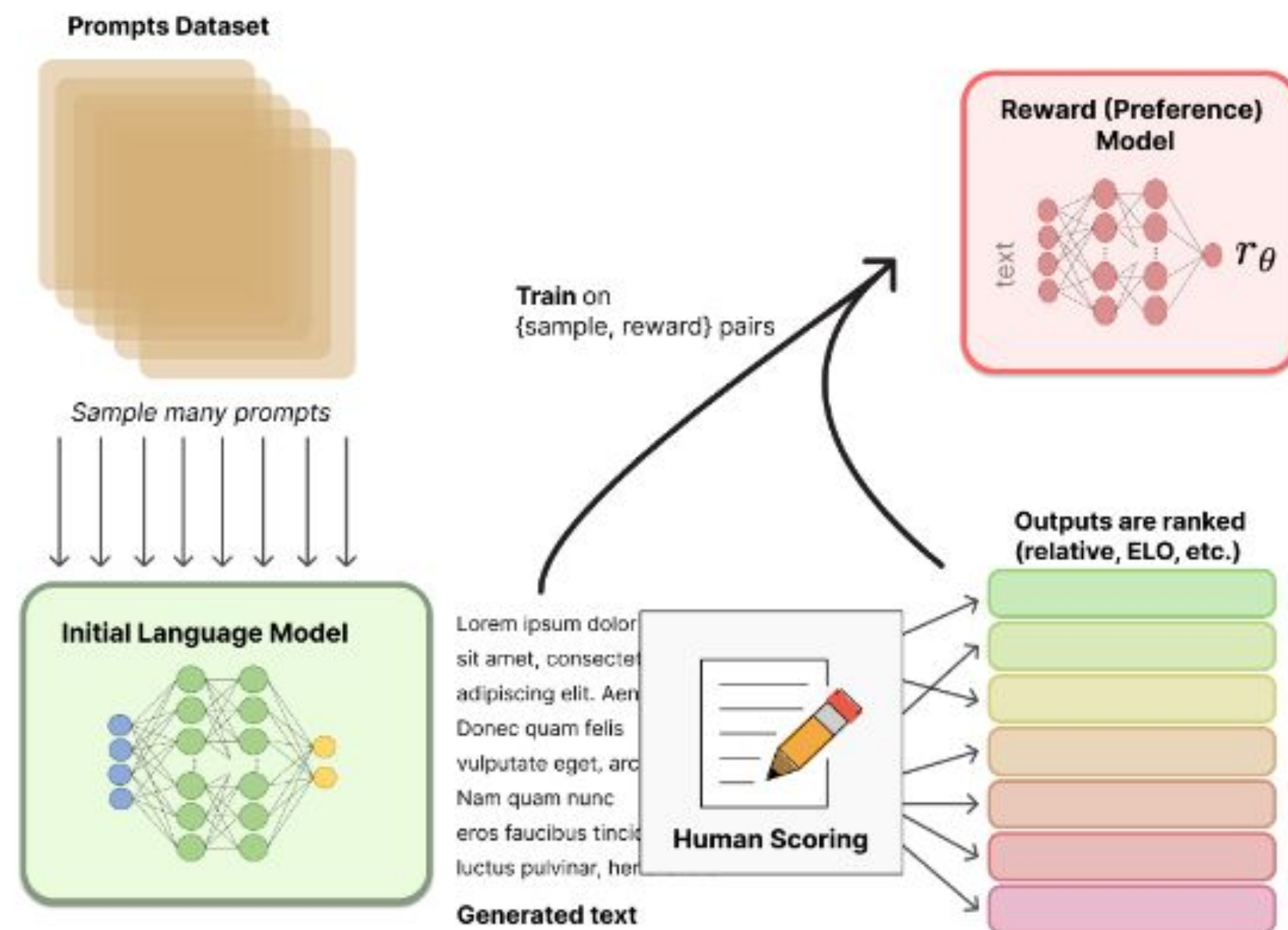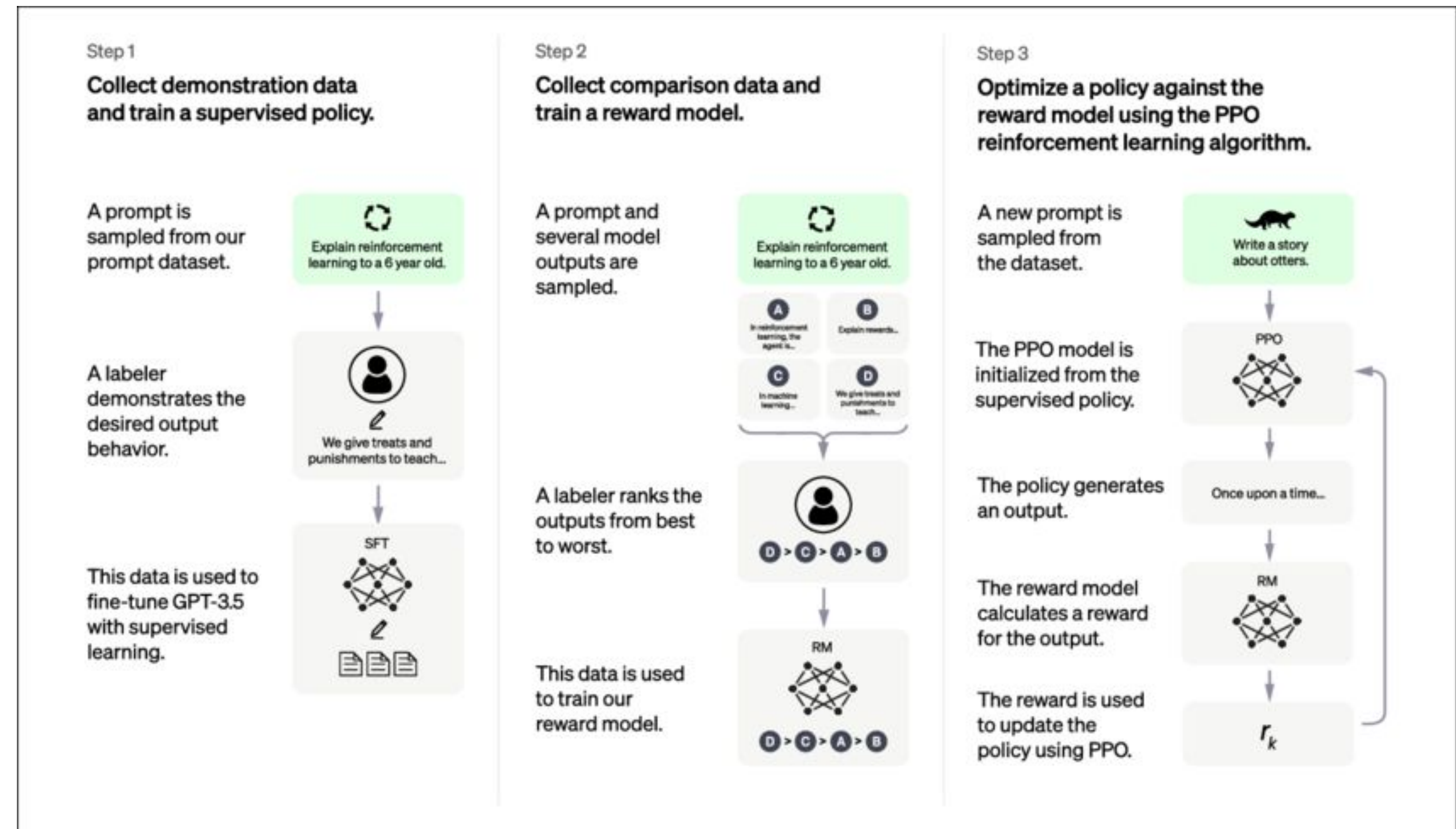
The ChatGPT training process. The figure is from OpenAI (2022a).

Concepts: Prompt

A prompt is a short piece of text that is given to the large language model as input, and it can be used to control the output of the model in many ways.

A prompt contains any of the following elements:

Instruction - a specific task or instruction you want the model to perform

Context - external information or additional context that can steer the model to better responses

Input Data - the input or question that we are interested to find a response for

Output Indicator - the type or format of the output.

Concepts: Prompt Engineering

Zero-shot Prompting

Prompt:

> Classify the text into neutral, negative or positive.
> Text: I think the vacation is okay.
> Sentiment:

Output:

> Neutral

Few-shot Prompting

Prompt:

> A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
> We were traveling in Africa and we saw these very cute whatpus.
> To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

Output:

> When we won the game, we all started to farduddle in celebration.

https://www.promptingguide.ai/techniques

Concepts: Prompt Engineering

Chain-of-Thought Prompting

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✔

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:
_____
(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:
_____
(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✔

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is
_____
(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: **Let's think step by step.**
_____
(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✔

https://www.promptingguide.ai/techniques

Concepts: RAG

https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models-customize-rag.html
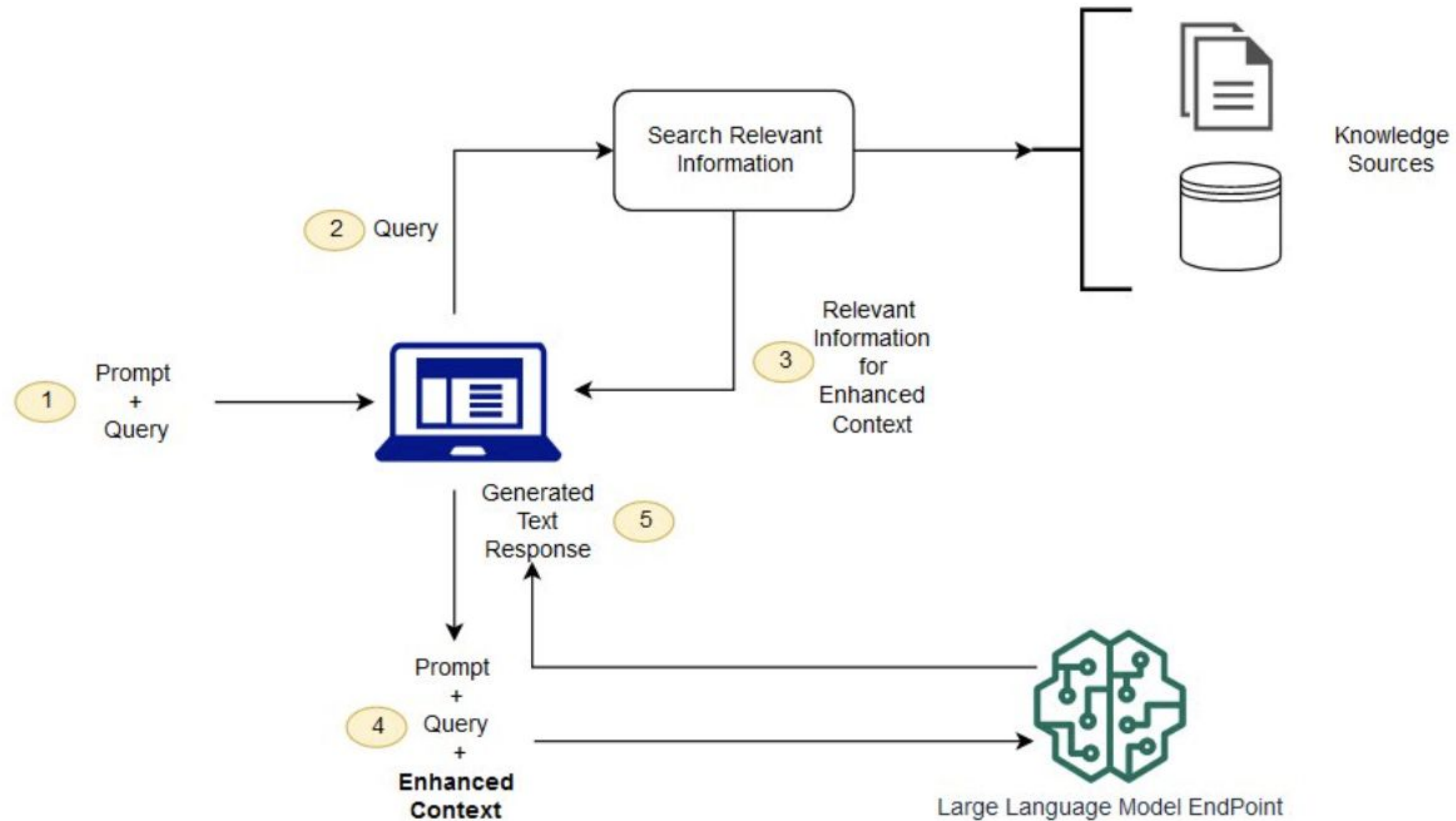
Limitations of Generative AI

- Hallucinations are words or phrases that are generated by the model that are often nonsensical or grammatically and factually incorrect.
  - The model is not trained on enough data. Misleading information.
  - The model is trained on noisy or dirty data. Garbage in => Garbage out!
  - The model is not given enough context. Incomplete information.
  - The model is not given enough constraints. Anyone can use it.

- Ethical concerns – what if the models are biased and are used for unintended purpose.

- Productionizing the LLMs is difficult.
  - Cost — Infrastructure
  - Time — Takes longer to build your own LLMs. Pre-training vs Fine-tuning vs Prompting.

- Explainability is difficult.

- Responsible AI - Security & Privacy.

Research Directions

Open challenges in LLM research

1. Reduce and measure hallucinations
2. Optimize context length and context construction
3. Incorporate other data modalities
4. Make LLMs faster and cheaper
5. Design a new model architecture
6. Develop GPU alternatives
7. Make agents usable
8. Improve learning from human preference
9. Improve the efficiency of the chat interface
10. Build LLMs for non-English languages

Hardest is to build LLMs for non-English languages!

Generative AI for Tamil

**Tamil is a culturally rich language with a great wealth of
information spanning across literature, music and the fine arts!**

What does it take to build a Tamil Generative AI?

- Building a community and educate about AI - [AI Tamil Nadu](#) (some chapters have been doing it for many years).

- Foster open source projects that digitises Tamil literature like [Project Madurai](#), [Kaniyam](#), etc.

- Managing infrastructure cost through crowd-sourcing or CSR activities.

- Motivating and teaching people to contribute high quality data (initiatives ike [Aya by Cohere for AI](#)) to solve for specific problems faced by the Tamil community.

- Imparting technical knowledge (NLP) and do fundamental research for Tamil computing.

- Put up regulations to handle the ethical and societal issues (involve government).

Global South in AI - CFP

Title: Generative AI: A boon or bane to the Tamil community

Abstract: Tamil is one of the oldest classical languages in the world, with a rich history in literature, music, and the fine arts. Approximately 1.06% of the world's population speaks Tamil, with around 84.12 million being native speakers. There are many grassroots-level creators who are writers, poets, thought leaders, educators, and innovators contributing to the preservation and perpetuation of Tamil's heritage and culture. The language is at risk of losing its roots due to technological advancement and English being the widely used mode of business communication across the globe. Adding to that fear is the introduction of generative AI (GenAI), which is a fundamental paradigm shift from the traditional AI which is largely about organizing information created by humans for easy retrieval and ranking of relevant information and providing actionable insights and recommendations. But GenAI models are projected as a replacement for human creative capabilities owing to their ability to emulate skills that only humans may possess, and they pose a threat to the day-to-day lives of creators. There is no stopping GenAI other than by educating communities to embrace them and use them for the better while addressing and mitigating the potential threats they pose. A community's culture, values, and way of life are captured, communicated, preserved, and passed on via its literature, music, and various fine arts, including visual media. GenAI's impact on these aspects will have profound significance for how a community and its citizens evolve.

Thank You :)

Questions?