Large Language Models (LLMs)

Dr. Partha Pakray

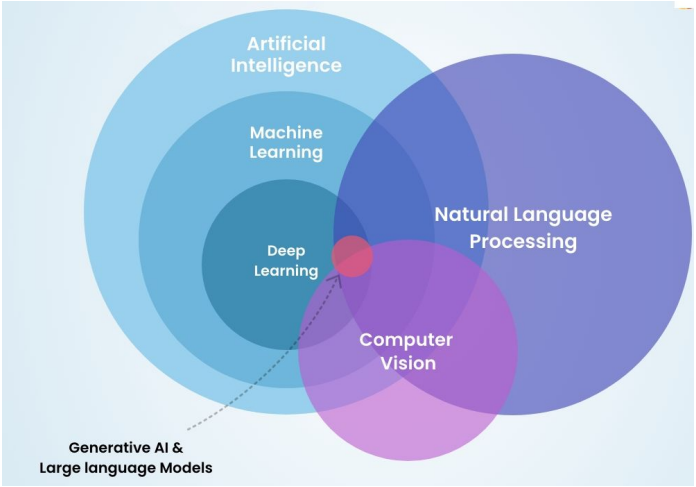
Outline of the talk

- About Large Language Models (LLMs)
- Fine-tuning
- Prompt Engineering
- Retrieval Augmented Generation (RAG)

About LLMs

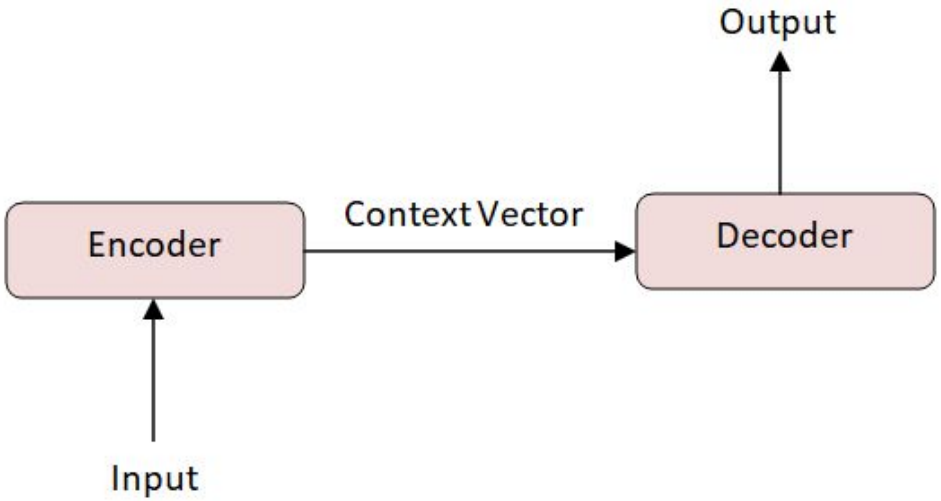
A large language model (LLM) is a language model notable for its ability to achieve general-purpose language generation and other natural language processing tasks such as classification.

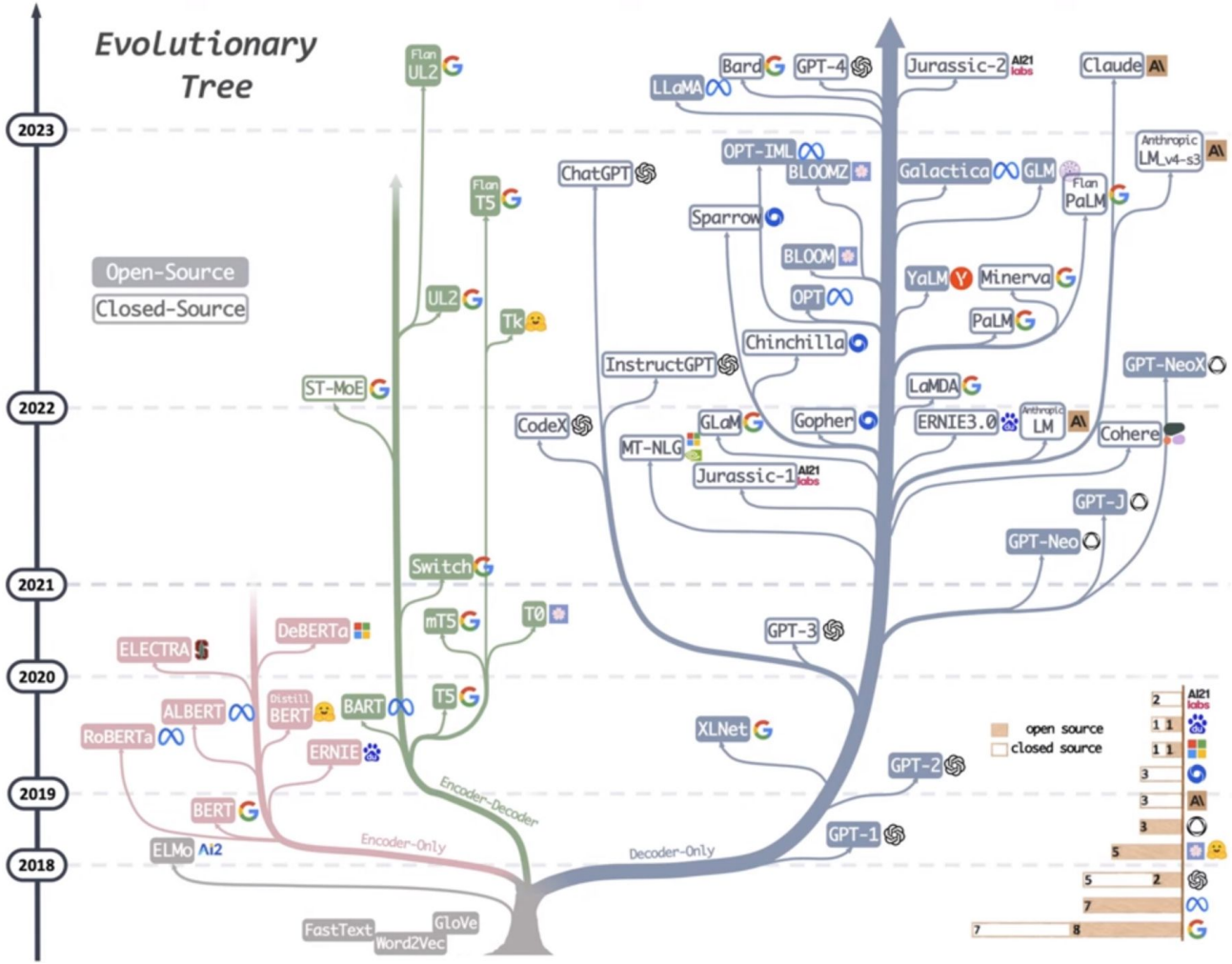
A subset of Deep Learning.



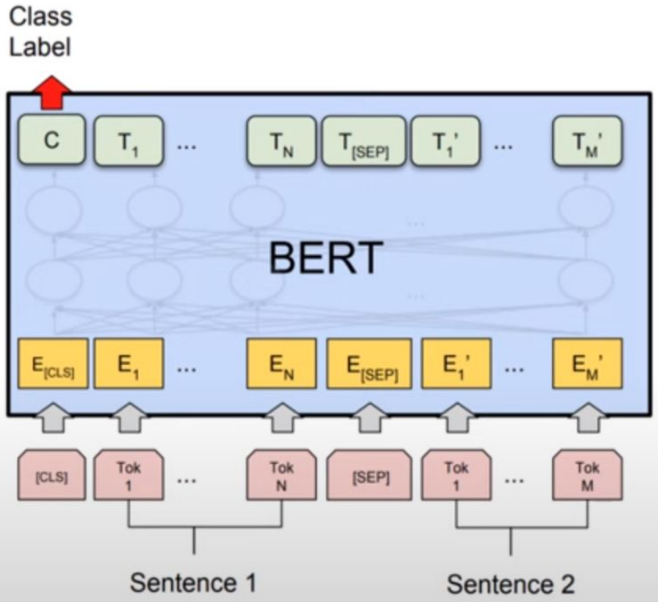
Various Transformer Architecture

- **Encoder-only**
- **Encoder-decoder**
- **Decoder-only**



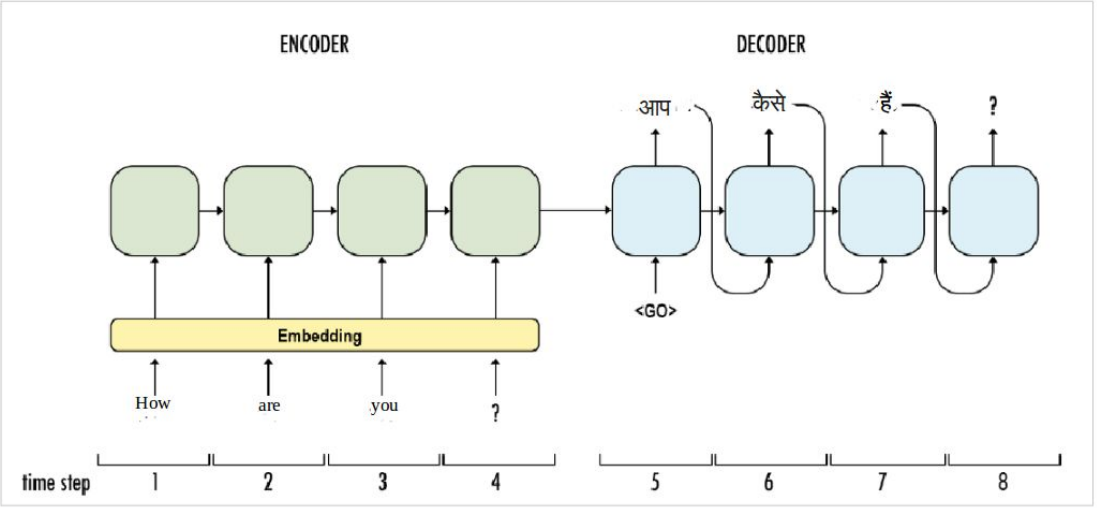

Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Flan UL2
Bard
GPT-4
Jurassic-2
Claude
LLaMA
OPT-IML
BLOOMZ
Galactica
GLM
Anthropic LM_v4-s3
ChatGPT
Flan T5
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
UL2
OPT
Tk
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
ST-MoE
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
Switch
mT5
T0
DeBERTa
GPT-3
ELECTRA
Distill BERT
BART
T5
ALBERT
XLNet
RoBERTa
ERNIE
GPT-2
BERT
Encoder-Decoder
Encoder-Only
Decoder-Only
GPT-1
ELMo
FastText
GloVe
Word2Vec
open source
closed source

Encoder Only

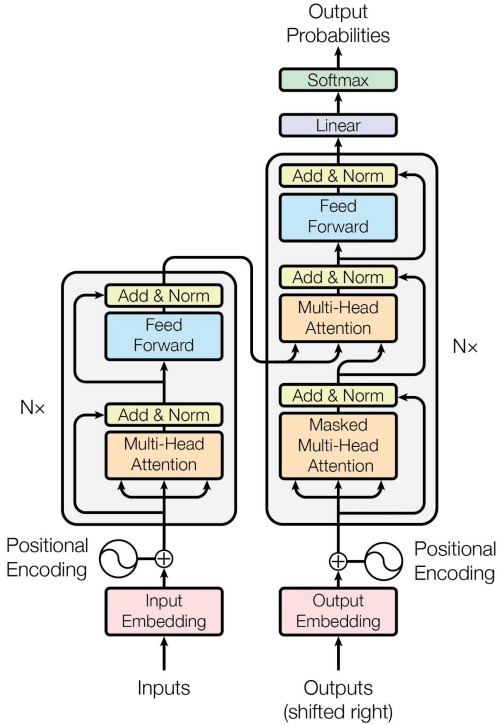


- Good for: classification, sequence tagging (POS tagging, NER), sentiment analysis
- Examples: BERT, RoBERTa, ALBERT, DeBERTa, etc.
- Typically requires fine-tuning for specific tasks
- Cannot generate text (only understand text)

Encoder-Decoder



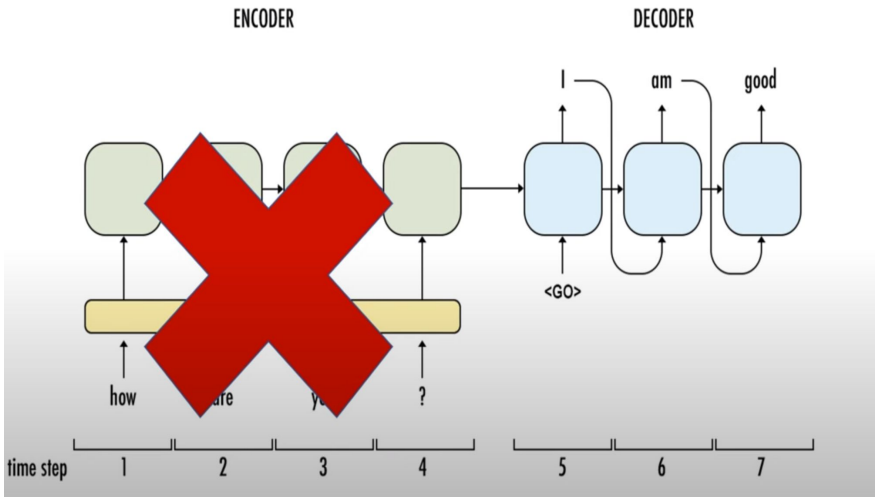
Encoder-Decoder (Transformer Model)



- A transformer is a deep learning architecture developed by Google and based on the multi-head attention mechanism, proposed in a 2017 paper "Attention Is All You Need".
- Text is converted to numerical representations called tokens, and each token is converted into a vector via looking up from a word embedding table.
- At each layer, each token is then contextualized within the scope of the context window with other (unmasked) tokens via a parallel multi-head attention mechanism allowing the signal for key tokens to be amplified and less important tokens to be diminished.
- The transformer paper, published in 2017, is based on the softmax-based attention mechanism proposed by Bahdanau et. al. in 2014 for machine translation, and the Fast Weight Controller, similar to a transformer, proposed in 1992.

– wiki

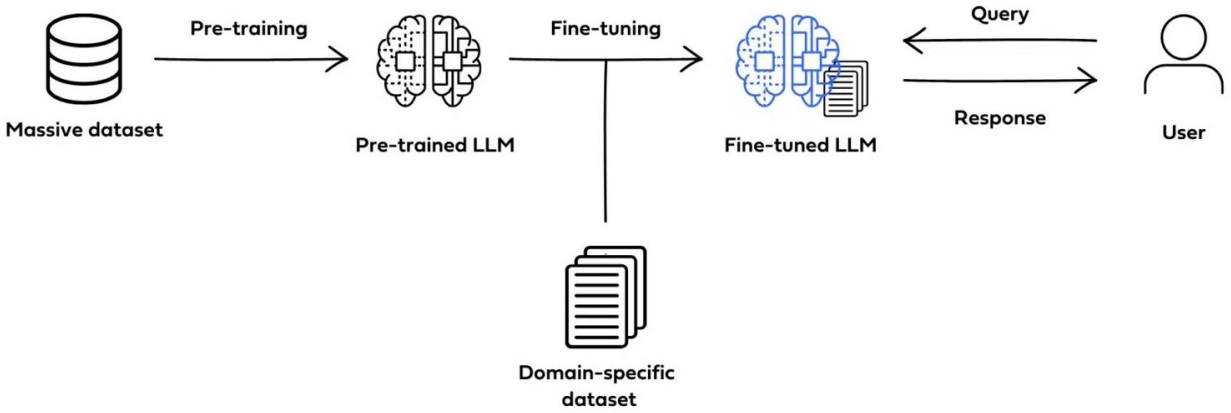
Decoder Only



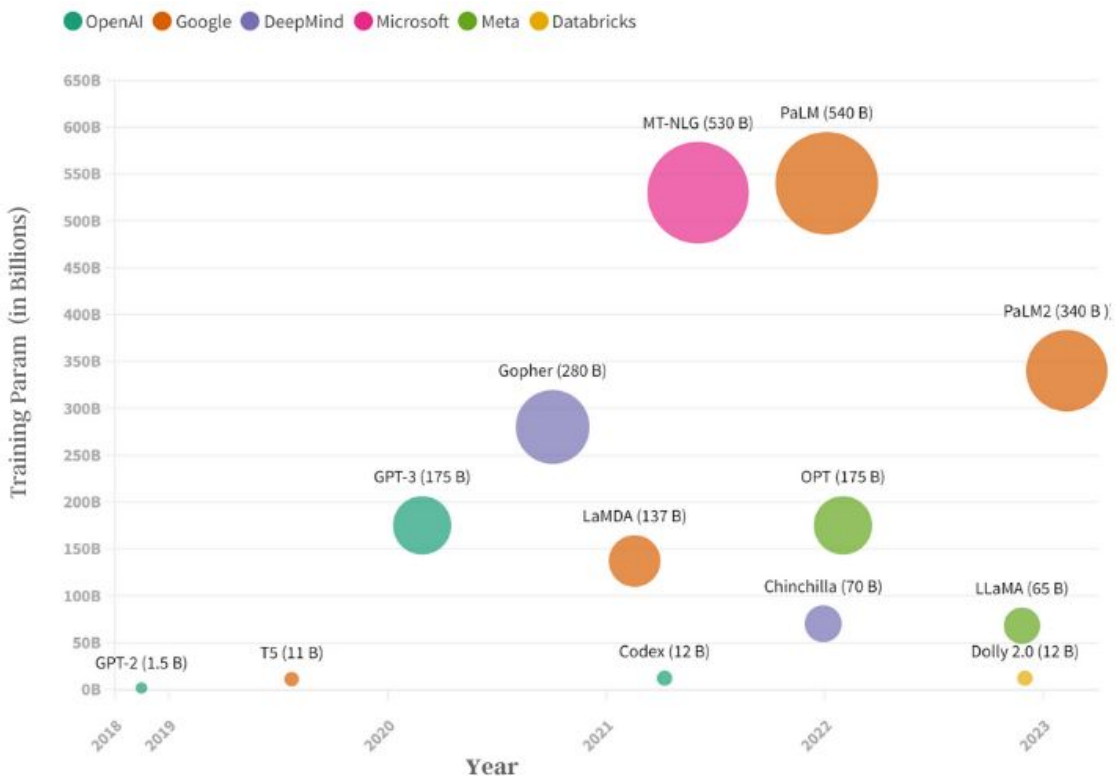
- **OpenAI**: GPT, GPT-2, GPT-3, GPT-4, ChatGPT
- **Google**: PaLM
- **Meta**: LLaMA
- **Deepmind**: Chinchilla

LLMs

- **Pre-trained**: General Purpose Language Model
- **Fine-tuned**: for specific task

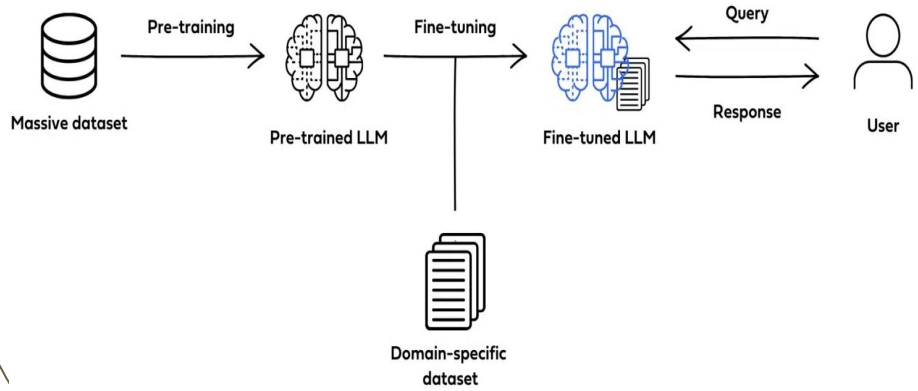


Parameters: What is that?



Transformer Architecture

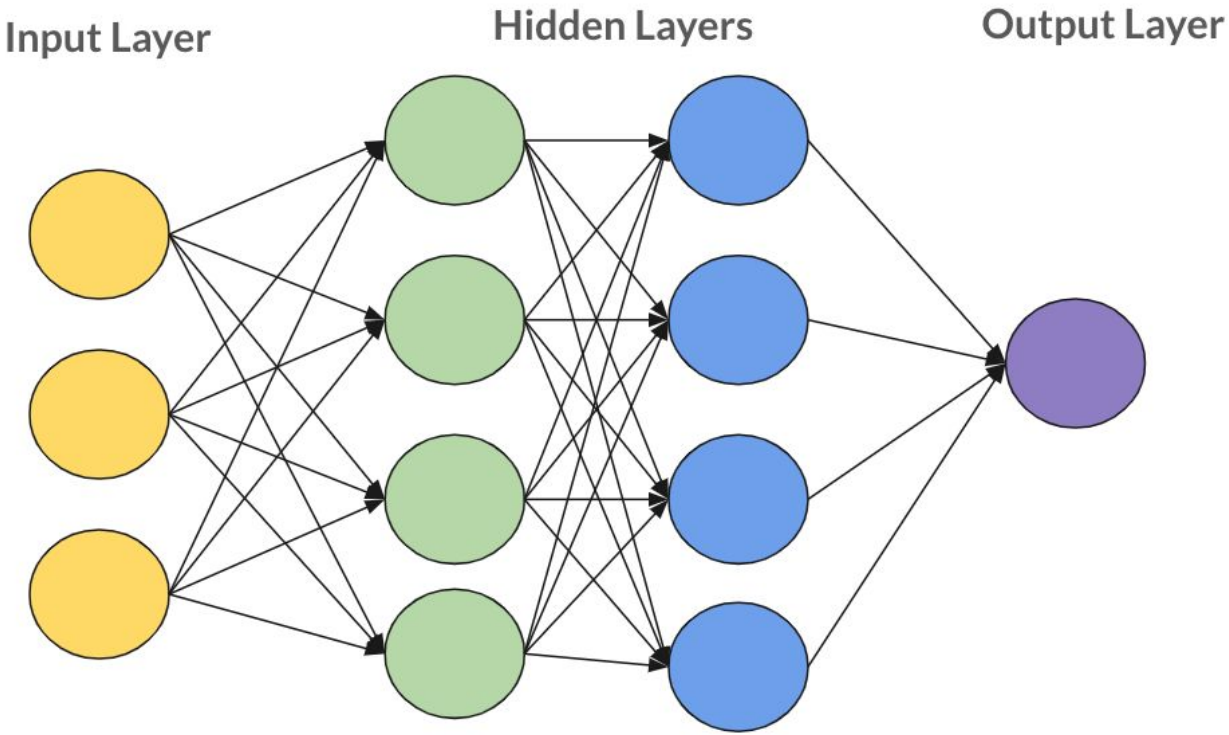
Parameters



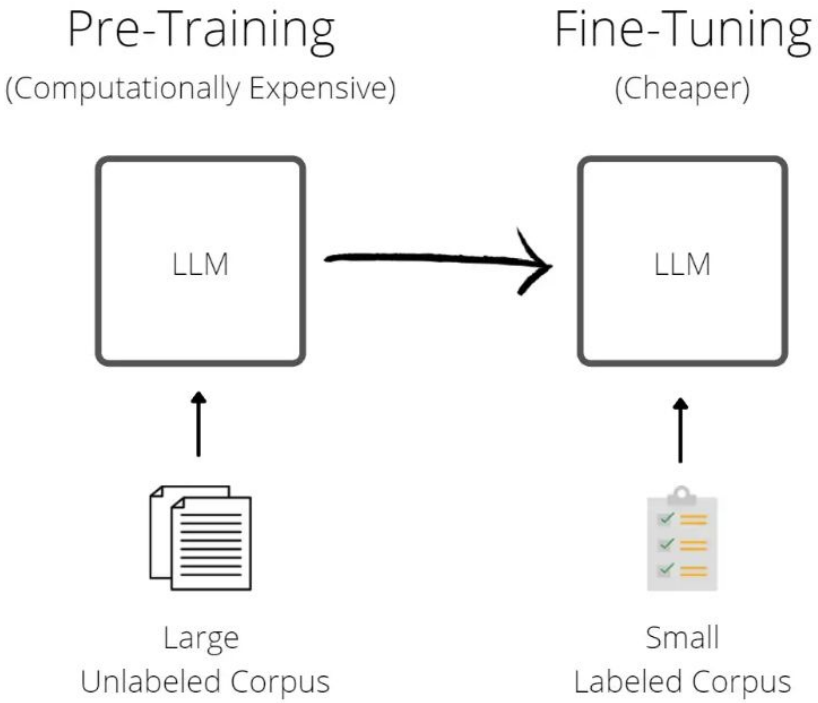
Context Window/ Length

Tokens

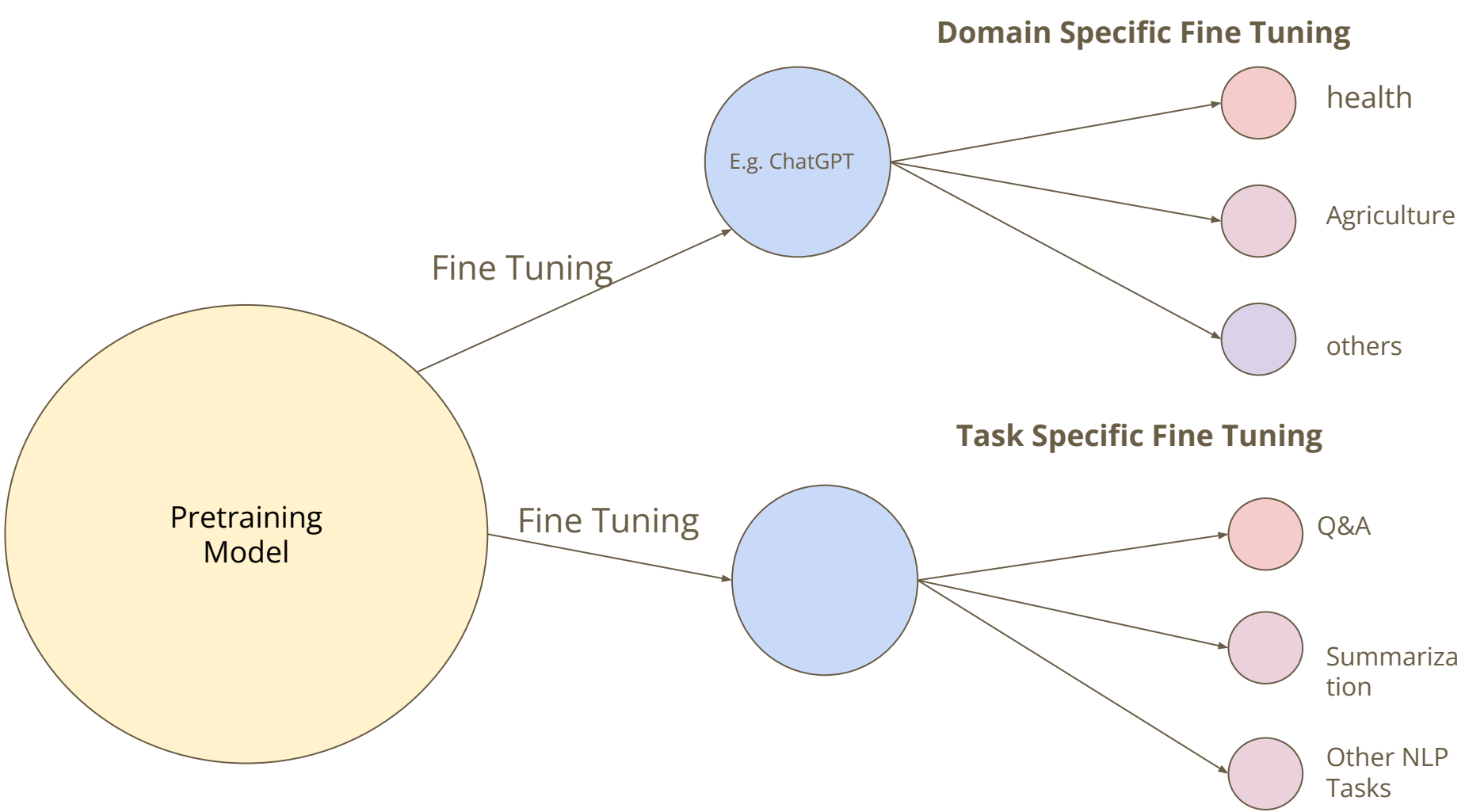
Parameters vs Hyperparameters



Fine Tuning


Pre-Training
(Computationally Expensive)
Fine-Tuning
(Cheaper)
LLM
LLM
Large
Unlabeled Corpus
Small
Labeled Corpus

Fine Tuning


Domain Specific Fine Tuning
E.g. ChatGPT
health
Agriculture
others
Fine Tuning
Pretraining Model
Fine Tuning
Task Specific Fine Tuning
Q&A
Summariza tion
Other NLP Tasks

Full Parameter Fine Tuning

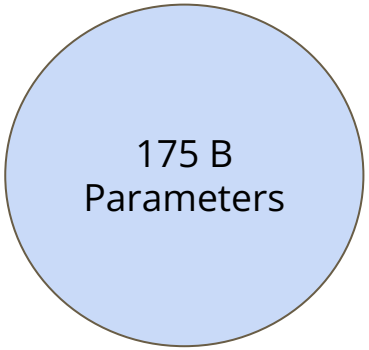
Challenges

- Update all the model weights
- Hardware resource constraints

Full Parameter Fine Tuning

Challenges

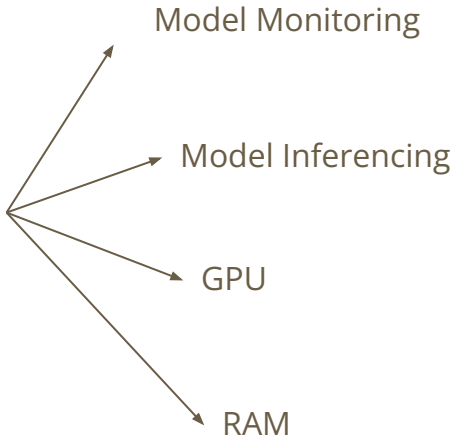
- Update all the model weights
- Hardware resource constraints



weights

Fine-tune: update all the weights

Downstream Task

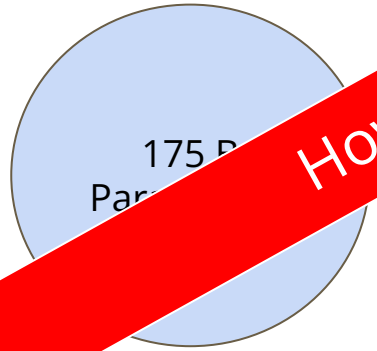


Full Parameter Fine Tuning

Challenges

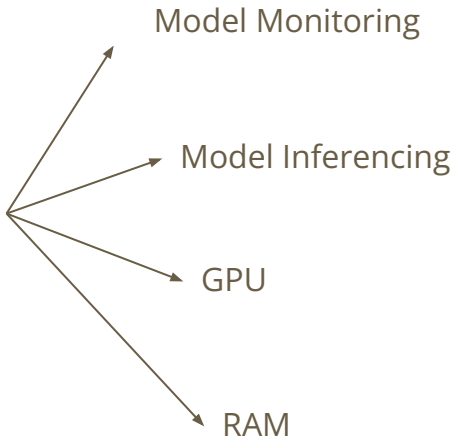
- Update all the model weights
- Hardware resource constraints

How to overcome this challenge...



ne:
date all
the
weights

Downstream Task



Problems in LLM

Factual Inaccuracy and Hallucinations

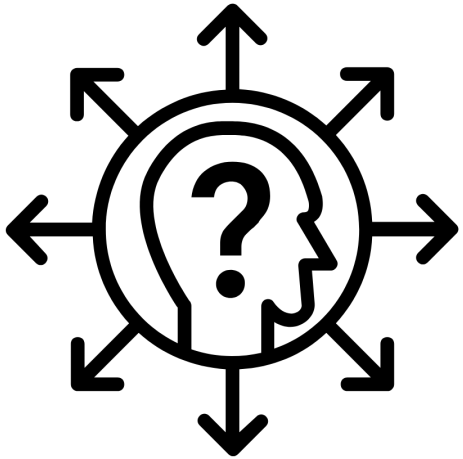
Inherent Biases

Limited Common Sense and Reasoning

Lack of Explainability

Computational and Environmental Costs

Security and Privacy Risks



Factual Inaccuracy and Hallucinations

The Issue: LLMs are trained on massive amounts of text, and this data can include misinformation, errors, and outdated facts. Consequently, LLMs may confidently generate text that is incorrect or misleading, even if it seems plausible. They can "hallucinate" when they don't know the answer but attempt to produce something anyway.

Why It Matters: This poses significant problems for applications where accuracy is vital: news summarization, education, medical or legal advice. It also contributes to the spread of misinformation.