LLM Best Practises

ODSC Conference: Module 1

Oct 31, 2023

Sanyam Bhutani, Sr Data Scientist H2O.ai

Agenda

- [Hands-on] Fine-Tuning your first LLM
  Creating your own GPT
- Understanding building blocks of LLMs
- Training from Scratch
- How to evaluate LLMs
- Prompting Vs Fine-Tuning
- Best Practises

Democratizing AI and LLMs with H2O.ai

50% OF THE FORTUNE 500

❤ H2O

8 OF THE TOP 10 BANKS

7 OF THE TOP 10 INSURANCE COMPANIES

6 OF THE TOP 10 MANUFACTURING COMPANIES

Customer Obsession
Maker Culture

30+
Kaggle Grandmasters

World's #1, #3, #5, and #9

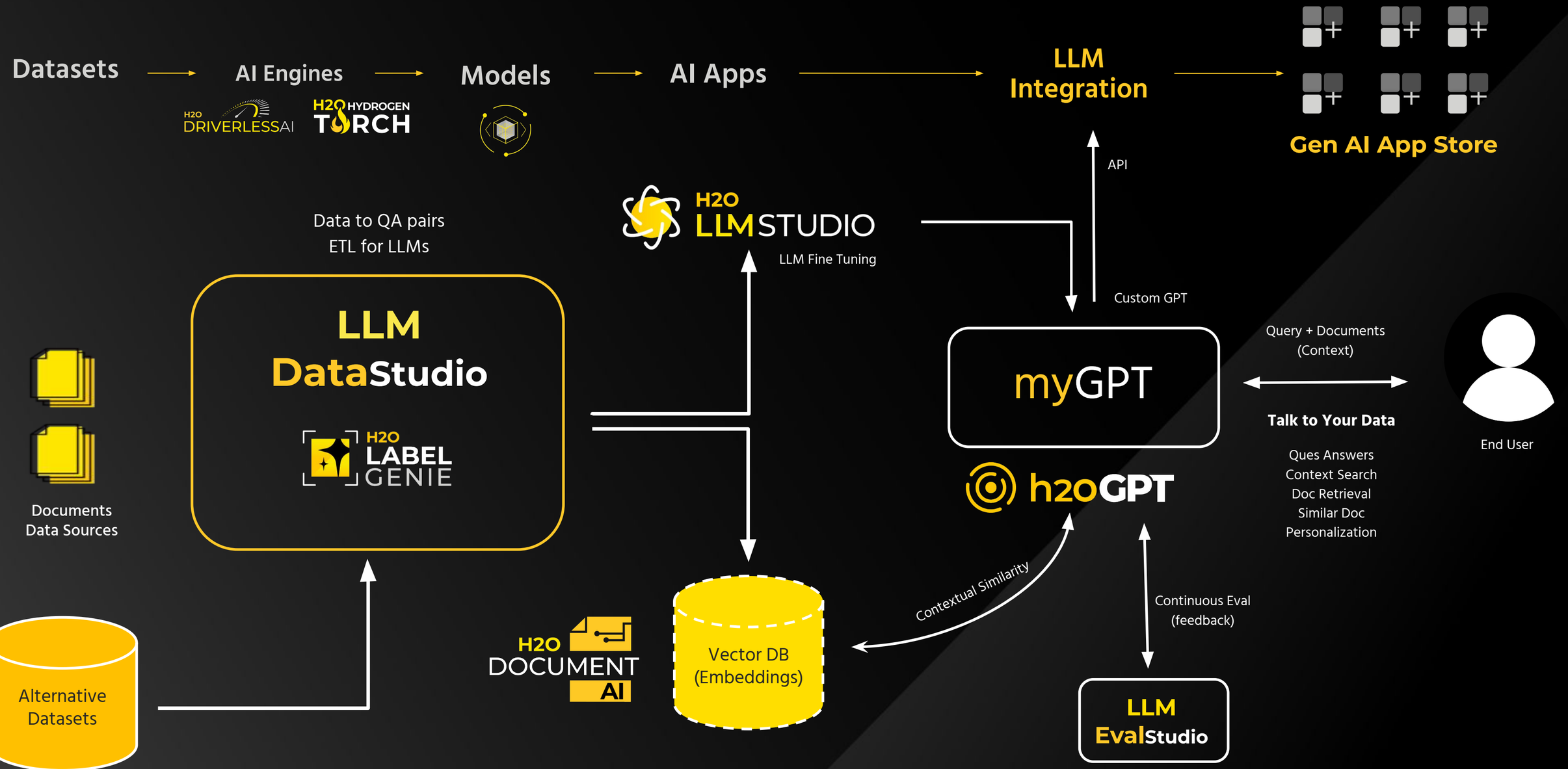
2.5M+
Community

100K+
h2ogpt requests per month

Innovation inspired and powered by World's Top 10% Data Scientist

Your projects are backed by 10% of the World's Data Science Grandmasters and a Team of Experts who are relentless in solving your critical problems.

H2O AI & GenAI Suite



Build Vs Buy

- Is this the right question to ask?
- Another alternate question: Host your model or call an API?
- Another question: Create a foundation model for your domain or fine-tune to your domain?
- Another question: How many M$ do you have for the problem?
- Another question: How much is this going to cost?

Building Blocks of LLMs

Memory + Agents

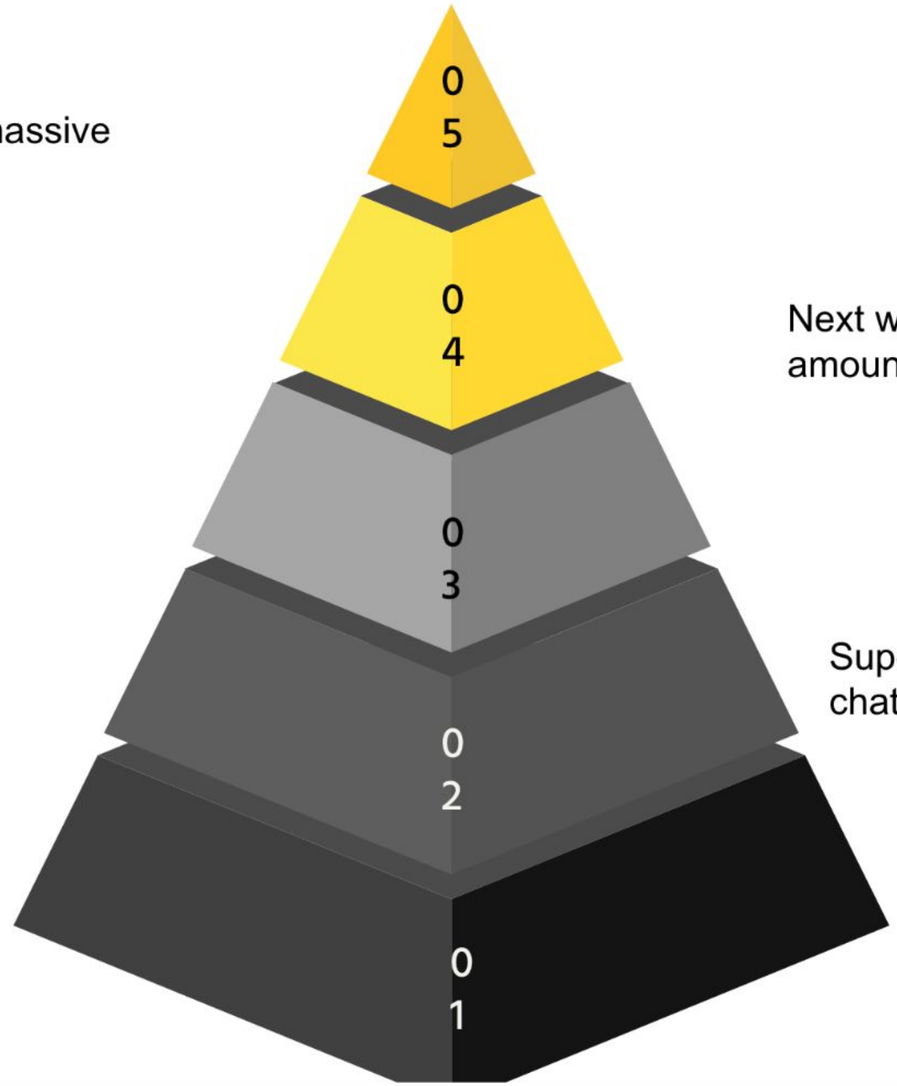
Next word prediction trained on massive amounts of data

RLHF

To “align” model to human preference

Foundation

Next word prediction trained on massive amounts of data



Database

Next word prediction trained on massive amounts of data

Fine Tuning

Supervised Fine-Tuning to create chatbot behaviour

Foundation Models

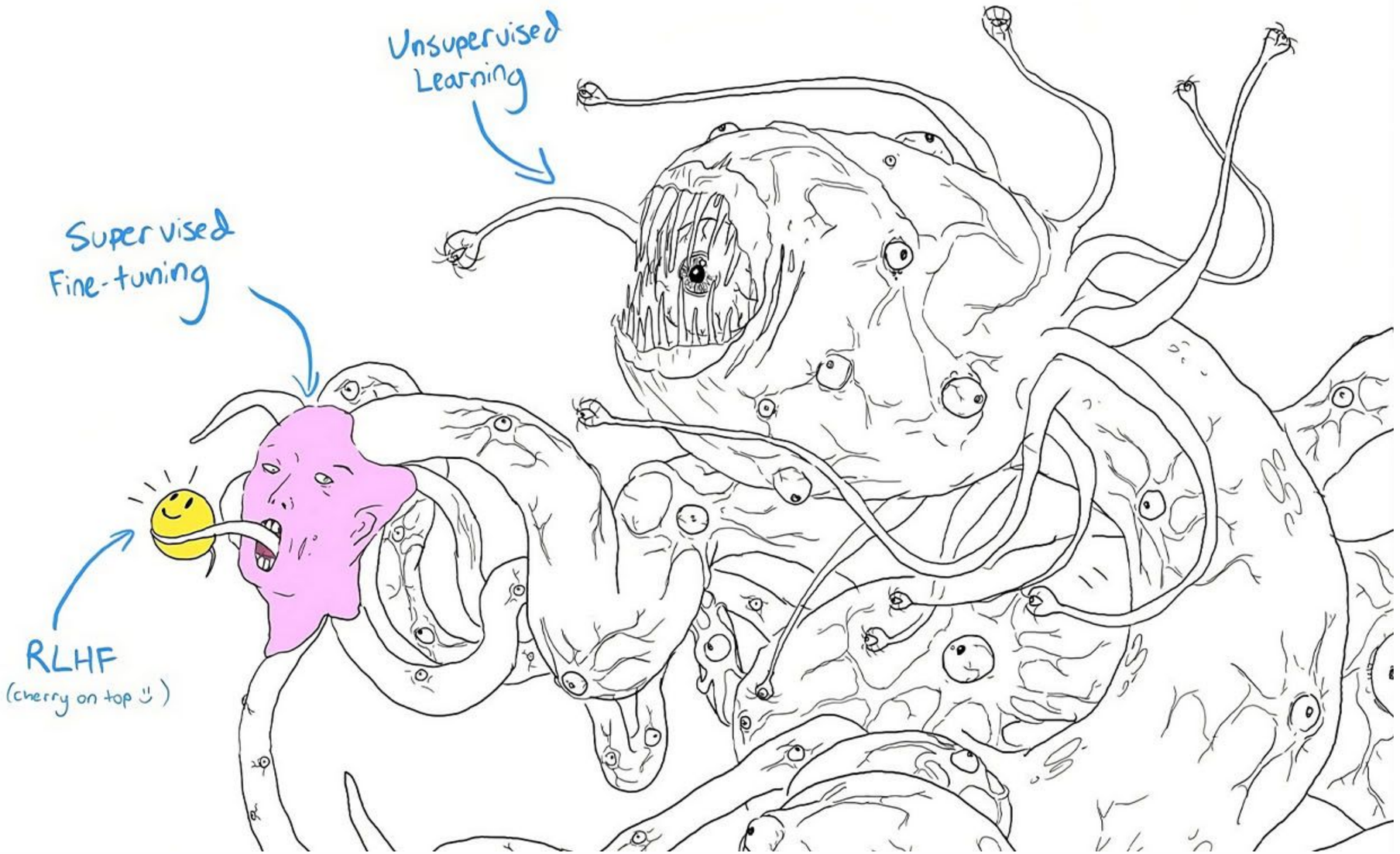
- T-5: Encoder-Decoder family
- Llama-2 (Base)
- GPT-3.5
- Would you say BERT is a foundation Model?

What does Fine-Tuning enable?

Fine-Tuning

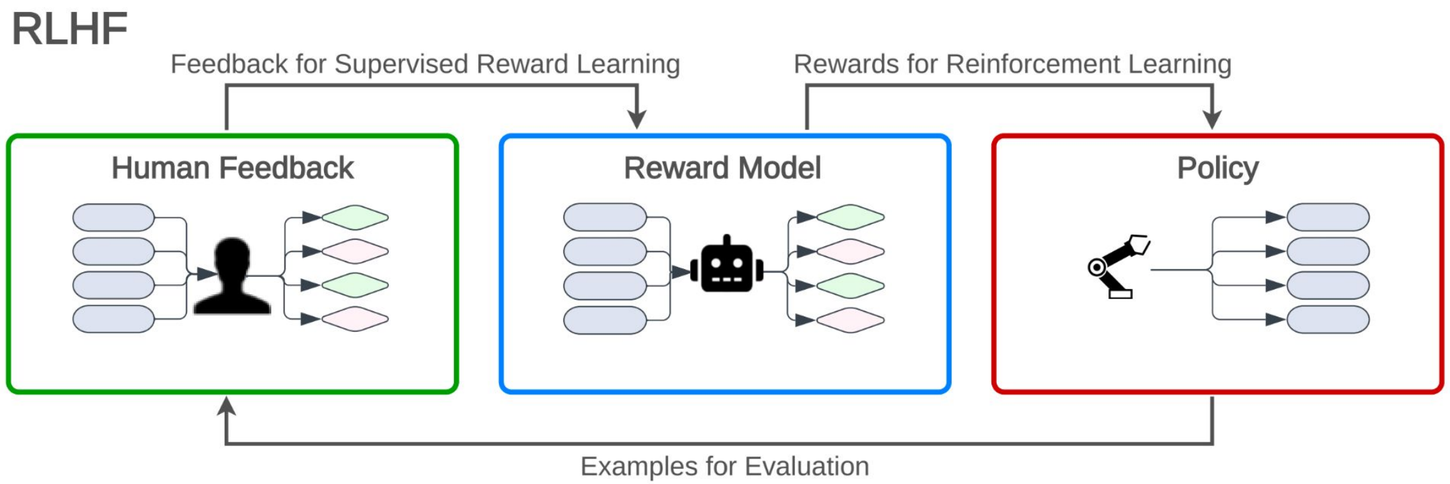
- Add existing knowledge/capabilities
- Open Question on how does it affect older capabilities?
- Helps increase chat following capabilities
- Sets “focus” on domain/create specialised models

RLHF



RLHF

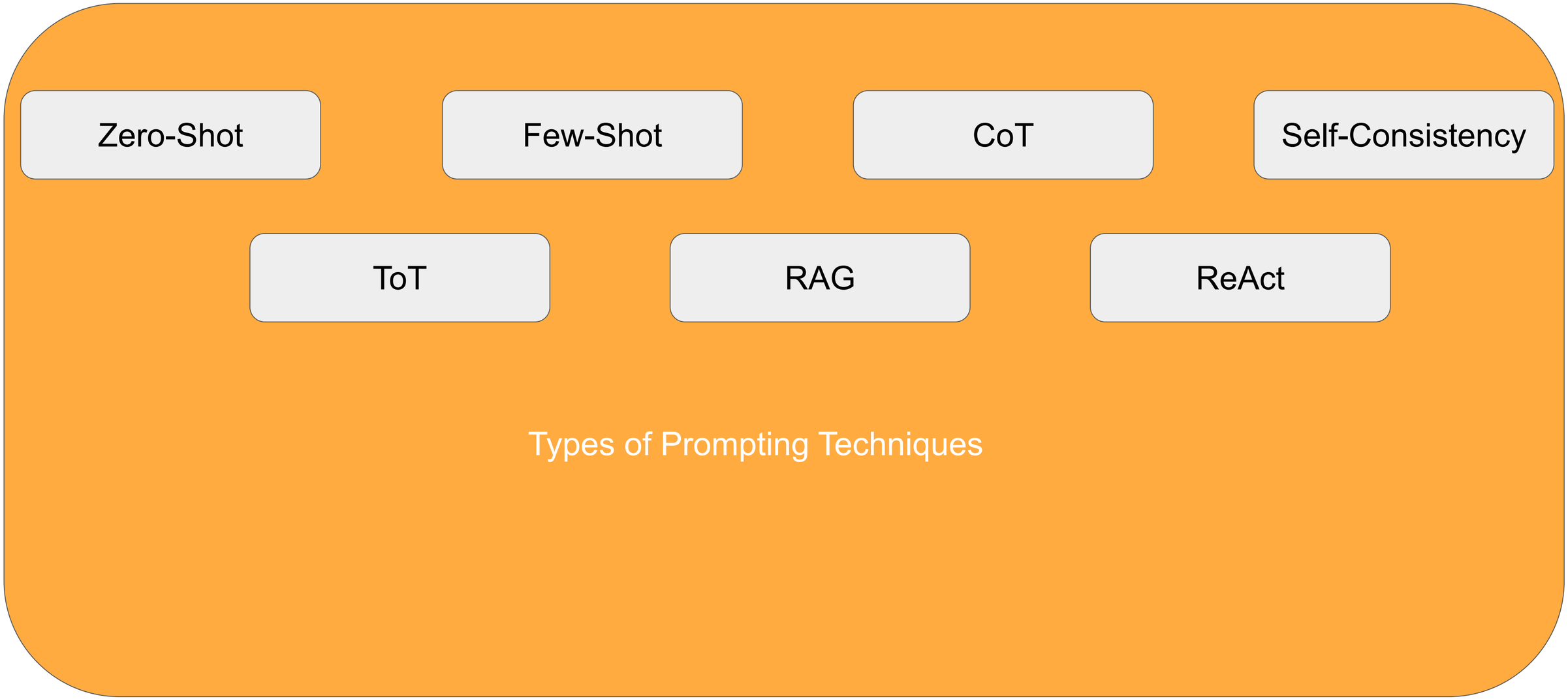
- “Alignment” stage of creating Large Language Models
- Human Annotation provides “preferred” answer from list of possible answers
- A Reward Model learns how to predict this
- A “Policy” Model is then trained to be included on the “Instruct” model



What does Prompting Enable?

- Bring out existing knowledge
- LLMs predict next tokens based on probability distributions
- Prompting enables finding the *right distribution*
- Set Tone/Style of responses

Prompting Techniques



Zero, n-shot

- Give 0 examples and ask LLM to solve a problem
- Give n examples of expected input and output
- Pro-Tip: 2-4 examples are enough
- If you give 10+ examples, it tends to restrict “reasoning”

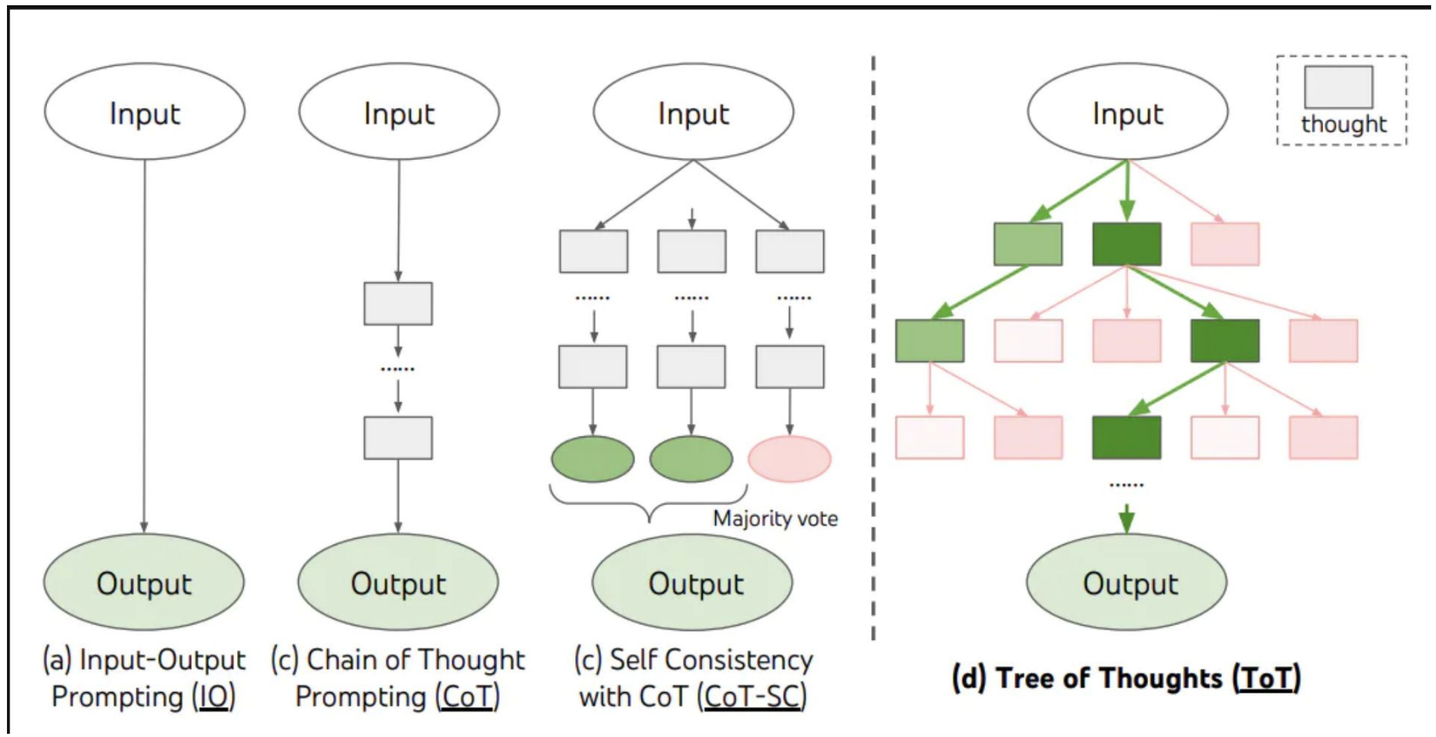
CoT

- “Think Step by Step”
- Allows the model to reason better by modelling more tokens
- In other words, “as the model thinks”, it gets more computation to reason about the question
- Great improves Math and Reasoning abilities

Self-Consistency

- Provide examples of CoT instead of getting LLM to do processing
- Realistic answer: You need a LLM to do CoT and use in-context learning
- The above is **expensive** and **high latency** requirement
- CoT is a better and more useful system

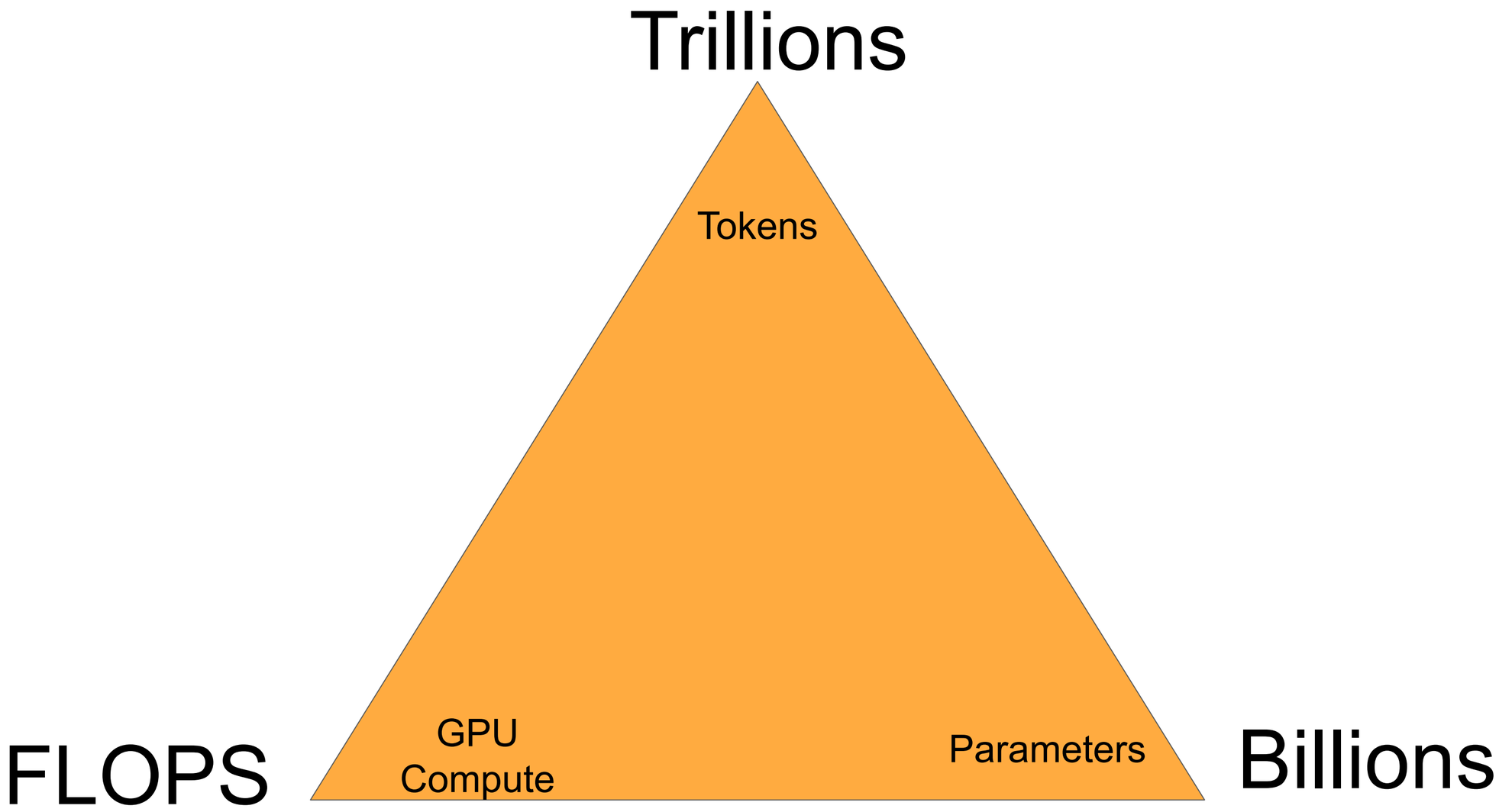
Tree of Thoughts



- Takes inspiration from Data Structures
- Allows better planning abilities
- Extremely helpful for Math abilities
- Hardest to implement in production

Case Study: FinGPT ChatLaw

The LLM Triangle



Measures of LLMs

- Compute: Measured in Training FLOPs, usually a replacement for number of epochs
- Why?
- Usually models are trained for 1-4 epochs and measuring orders of FLOPS is a better measure
- Think: 10,000 GPUs running at maximum utilisation for 1-4 months

Measures of LLMs

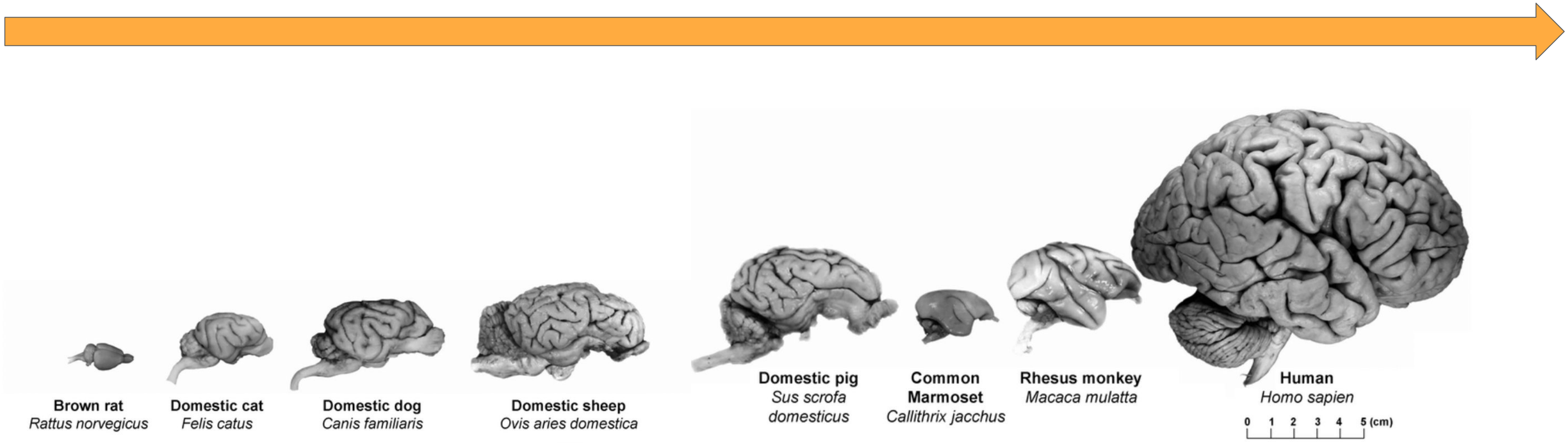
- Tokens: Measure of number of characters in input
- 100,000 Tokens ~=75,000 words
- Trillion-Scale is used to denote the number of words LLMs are exposed to during training
- Llama-2: 2T Tokens for pre-training
- Denotes pre-training strength

Measures of LLMs

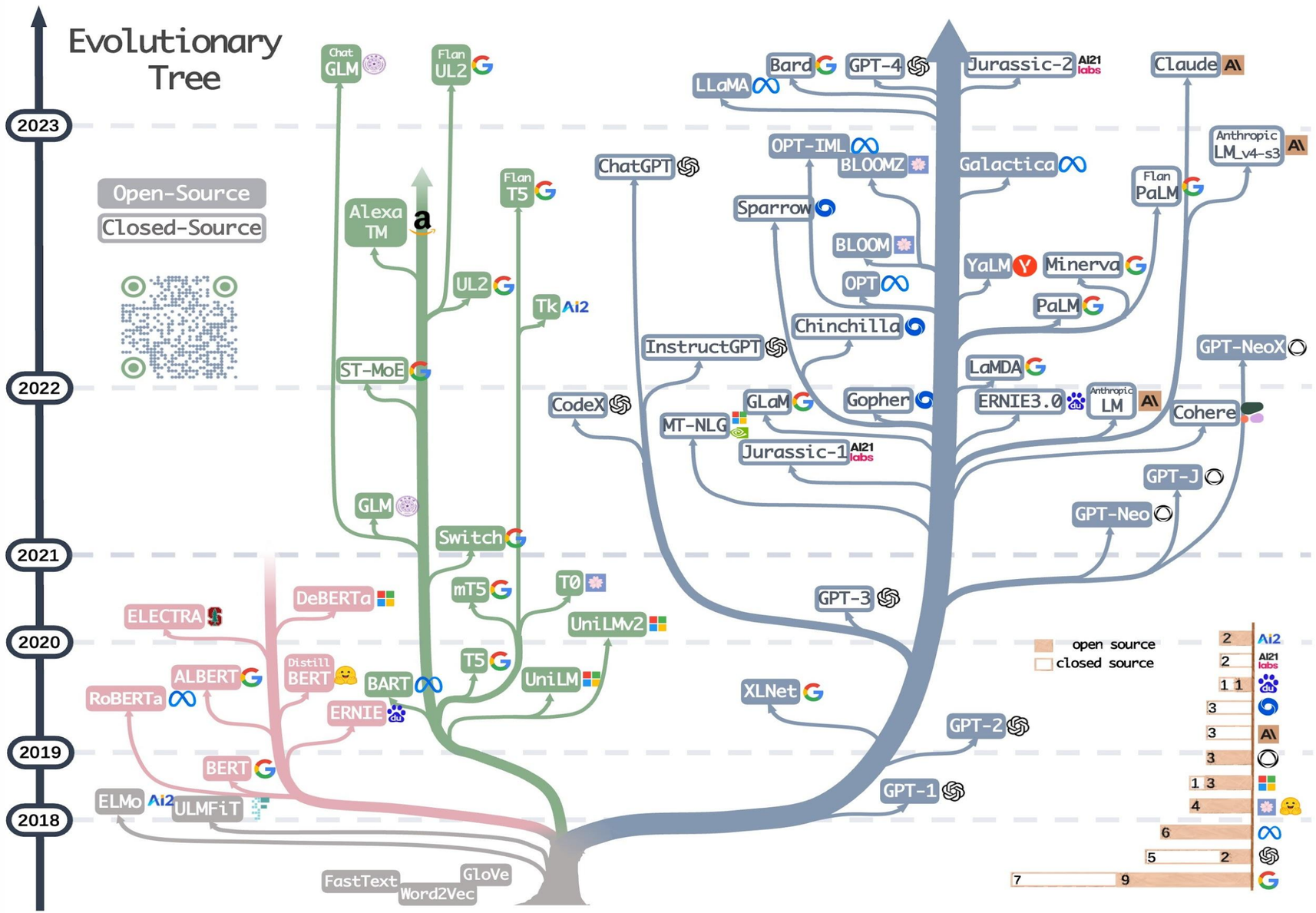
- Parameters: Billion-Scale measure
- Used to measure number of activations inside a Neural Network
- Larger Parameters not always denote stronger models
- Most common models: 7B family
- Why?
- This is what runs best in Colab and T4 GPUs
- 70B fits into 40GB GPU using int4 quantisation

of Parameters

Tiny Stories	Pythia	Red Pajama	Falcon	llama-2	GPT-3.5	GPT-4
0.3B	**1.3B**	**7B**	**40B**	**70B**	**175B**	**220B * 8**



Credit: https://www.researchgate.net/figure/Gross-comparative-neuroanatomy-of-various-large-animal-species-used-to-model-cerebral_fig2_323764775


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
Alexa TM
Flan T5
UL2
Tk
ST-MoE
GLM
Switch
mT5
T0
DeBERTa
ELECTRA
UniLMv2
Distill BERT
ALBERT
T5
BART
UniLM
RoBERTa
ERNIE
BERT
ELMo
ULMFiT
FastText
GloVe
Word2Vec
ChatGPT
Bard
GPT-4
LLaMA
Jurassic-2
Claude
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
LaMDA
GPT-NeoX
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source

Enterprise LLMs

- **Context Length:**
  Maximum length of Tokens that can be consumed
  #100,000 tokens ~= 75,000 words

- **Parameters:**
  Usually remains undisclosed. GPT-3.5 is 175B
  GPT-4 is rumoured to be 220B*8

- **Cost**:
  Console access remains free
  API is the key to access but remains expensive

  Claude-2 is ~2.5x more expensive than GPT-4 but permits 3x more tokens

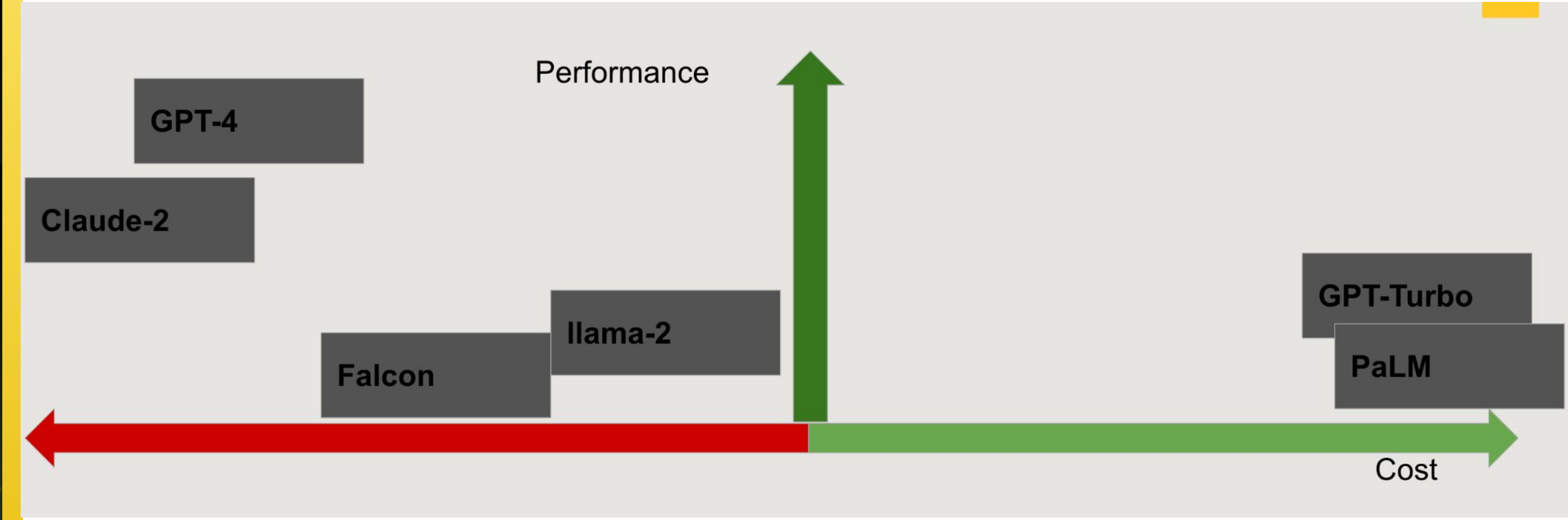
Enterprise LLMs

- **Code Capabilities:**
  GPT-4 Code interpreter is widely regarded as “GPT-4.5”

- **API Following Capabilities:**
  Anthropic and GPT-4 allow API following

- **Multimodal:**
  BARD and GPT-4 are the only multi-modal systems as of now

Enterprise LLMs



LLMs: Cost Vs Performance

- GPT-4 has the most resources
- Claude has the longest context length
- Llama-2 is the best open source model
- Llama-2 is horrible compared to GPT-3.5
- GPT-3.5 16k has the best price/performance tradeoff

LLMs: Cost Vs Performance

- GPT-4 32k is the best in code, reasoning and performance
- Claude-2 100k is the second best in enterprise grade LLMs
- Remember, you probably don’t need an Enterprise LLM
- Small, special models are all you need!

myGPT vs Closed models

	Closed Models LLMs	myGPT
Data Privacy and Security	Data is shared on cloud, Leakage Risk	No data (query, prompts, response) is shared outside
AI Governance	Limited Transparency	Full visibility and transparency
Customization	Limited Customization	Fully customizable
Cost of Scalability	High enterprise adoption cost	Fixed Cost
Access and Availability	Downtime Risk	Control over Downtime
Ownership	Owned by third party	Owned by Customer
Replicability	Need to start from scratch	Replicable for other groups

PEFT

- What is PEFT?
- Why do we require PEFT in LLMs?
- Parameters in LLM
- Types of PEFT
- Decide best technique

How did it all start?

Freeze or fine-tune?

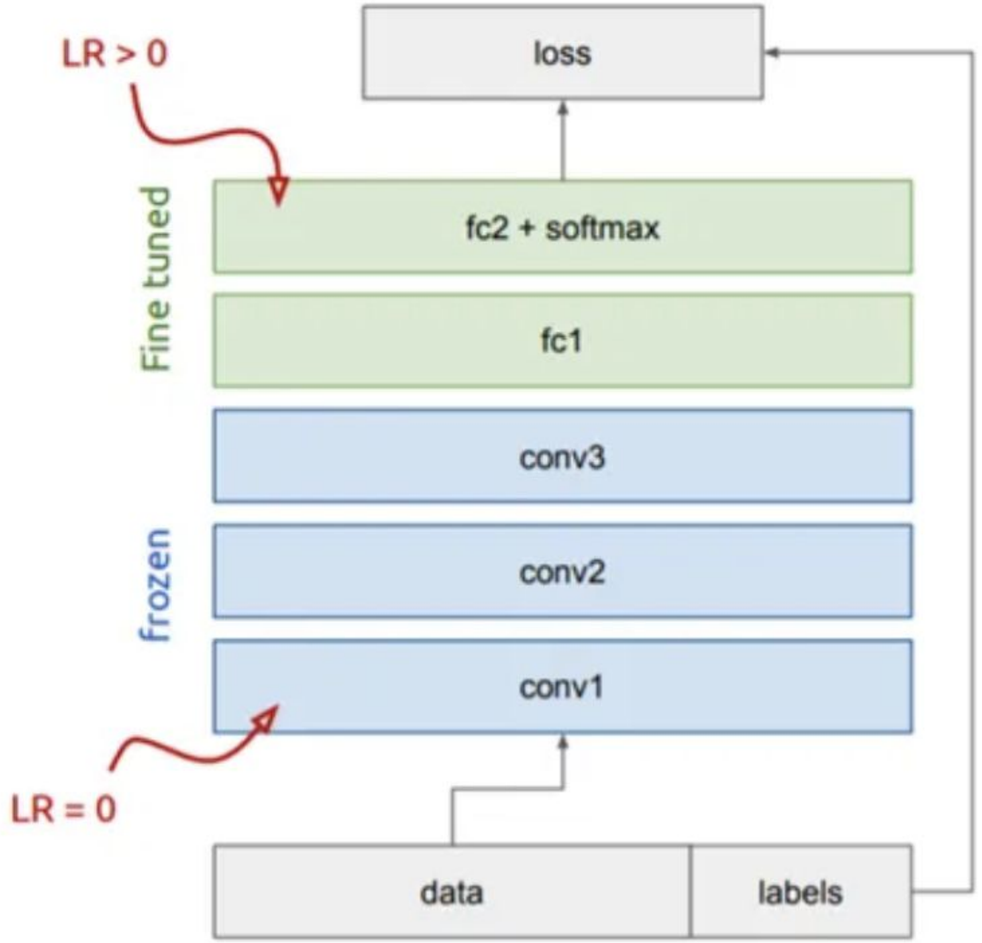
Bottom *n* layers can be frozen or fine tuned.

- **Frozen**: not updated during backprop
- **Fine-tuned**: updated during backprop

Which to do depends on target task:

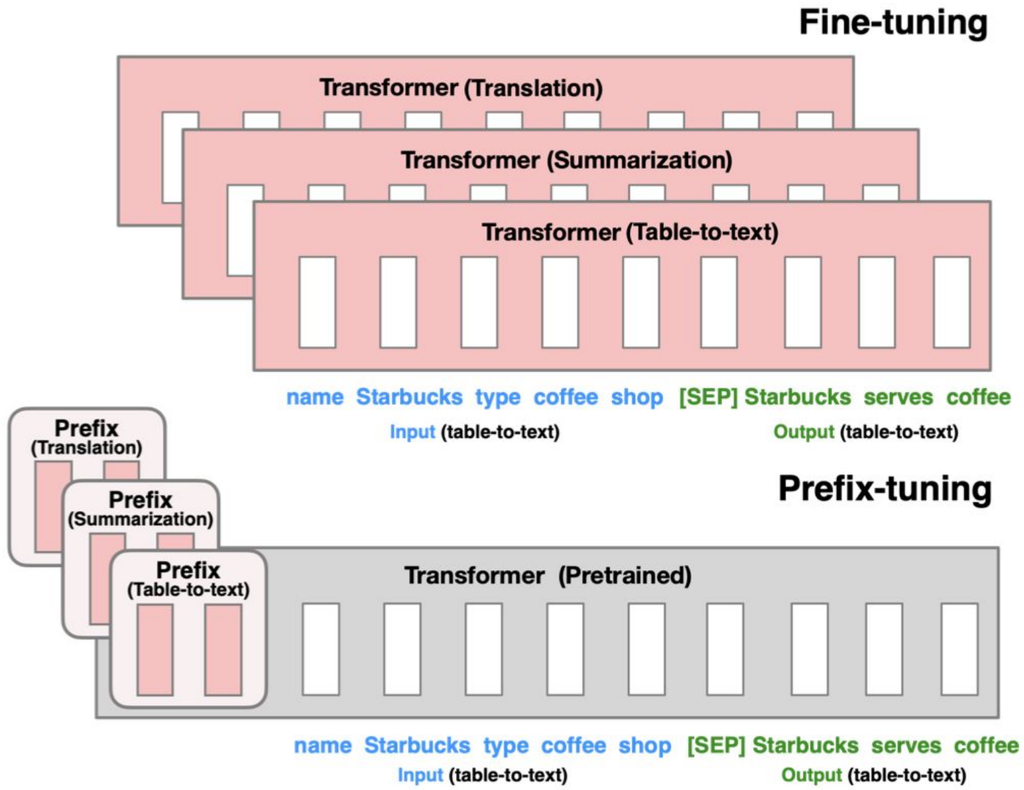
- **Freeze**: target task labels are scarce, and we want to avoid overfitting
- **Fine-tune**: target task labels are more plentiful

In general, we can set learning rates to be different for each layer to find a tradeoff between freezing and fine tuning

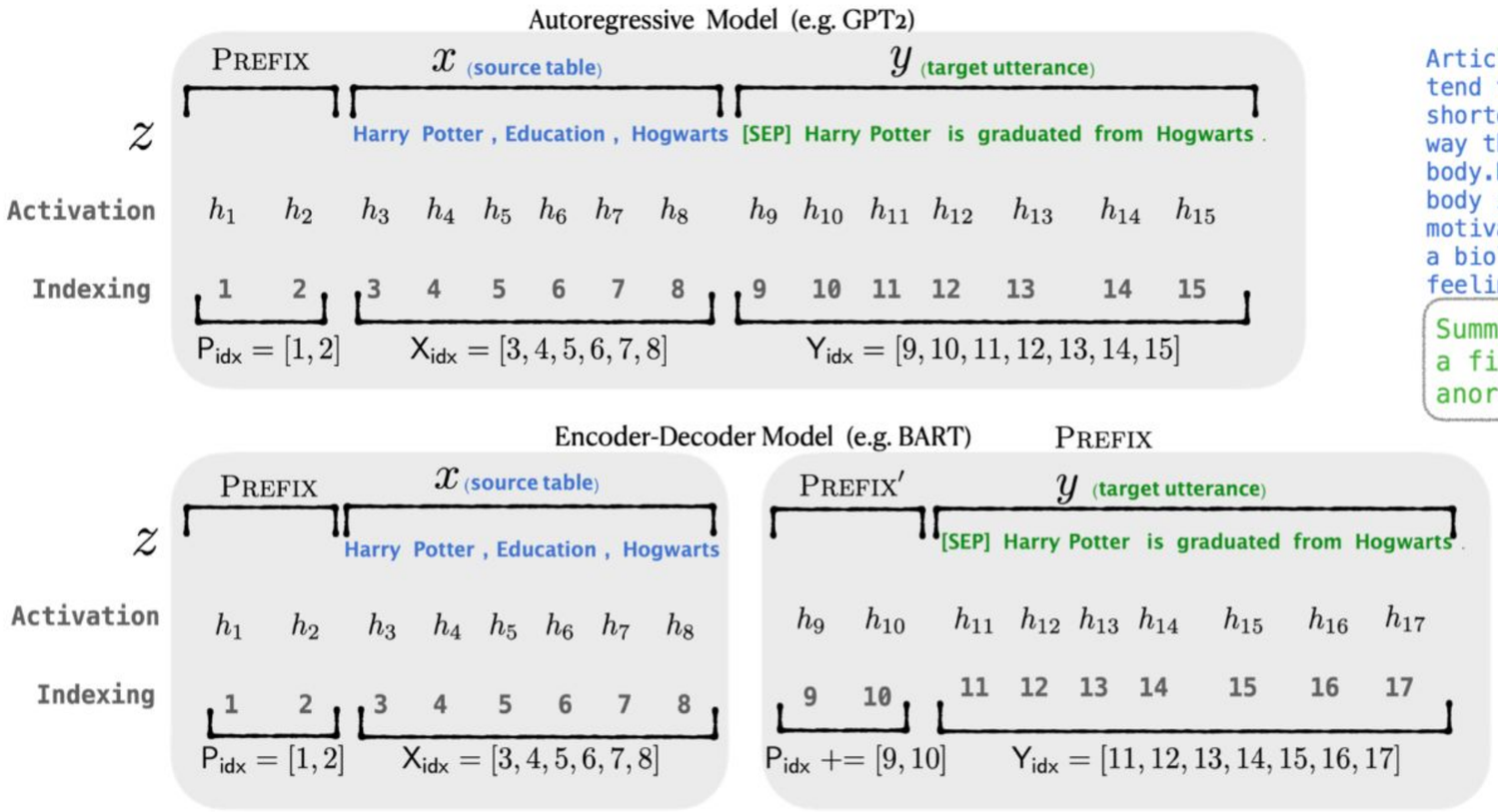


Prefix Tuning

- **Prefix**: Task specific Vectors
- **0.x %** parameters learn
- Lightweight Finetuning
- Prefix Length



Prefix Tuning



Summarization Example

```
Article: Scientists at University College London discovered people
tend to think that their hands are wider and their fingers are
shorter than they truly are.They say the confusion may lie in the
way the brain receives information from different parts of the
body.Distorted perception may dominate in some people, leading to
body image problems ... [ignoring 308 words] could be very
motivating for people with eating disorders to know that there was
a biological explanation for their experiences, rather than
feeling it was their fault."

Summary: The brain naturally distorts body image -
a finding which could explain eating disorders like
anorexia, say experts.
```

Table-to-text Example

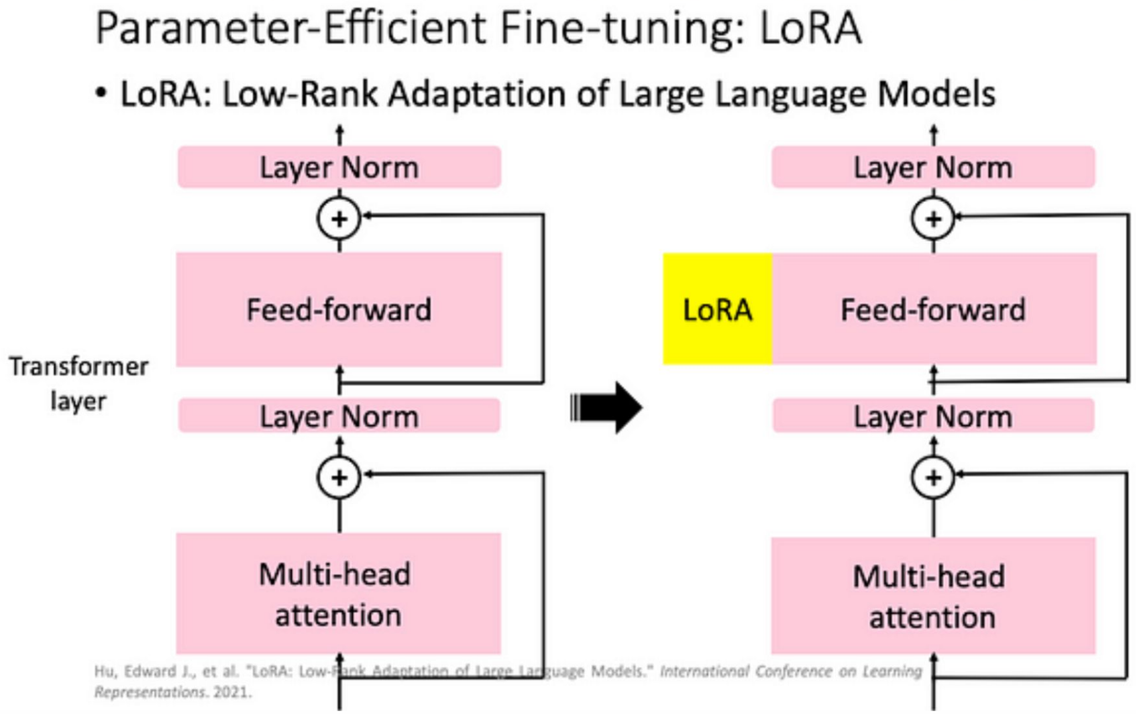
```
Table:  name[Clowns] customer-
rating[1 out of 5]  eatType[coffee
shop] food[Chinese] area[riverside]
near[Clare Hall]

Textual Description: Clowns is a
coffee shop in the riverside area
near Clare Hall that has a rating
1 out of 5 . They serve Chinese
food .
```

LoRA

- Task Specific Attention parameters
- Low rank matrix factorization
- Preserve Pretrained Weights
- 16 Bit FT

Slides credit: Cheng-Han Chiang, Yung-Sung Chuang, Hung-yi Lee, "AACL_2022_tutorial_PLMs," 2022



QLoRA

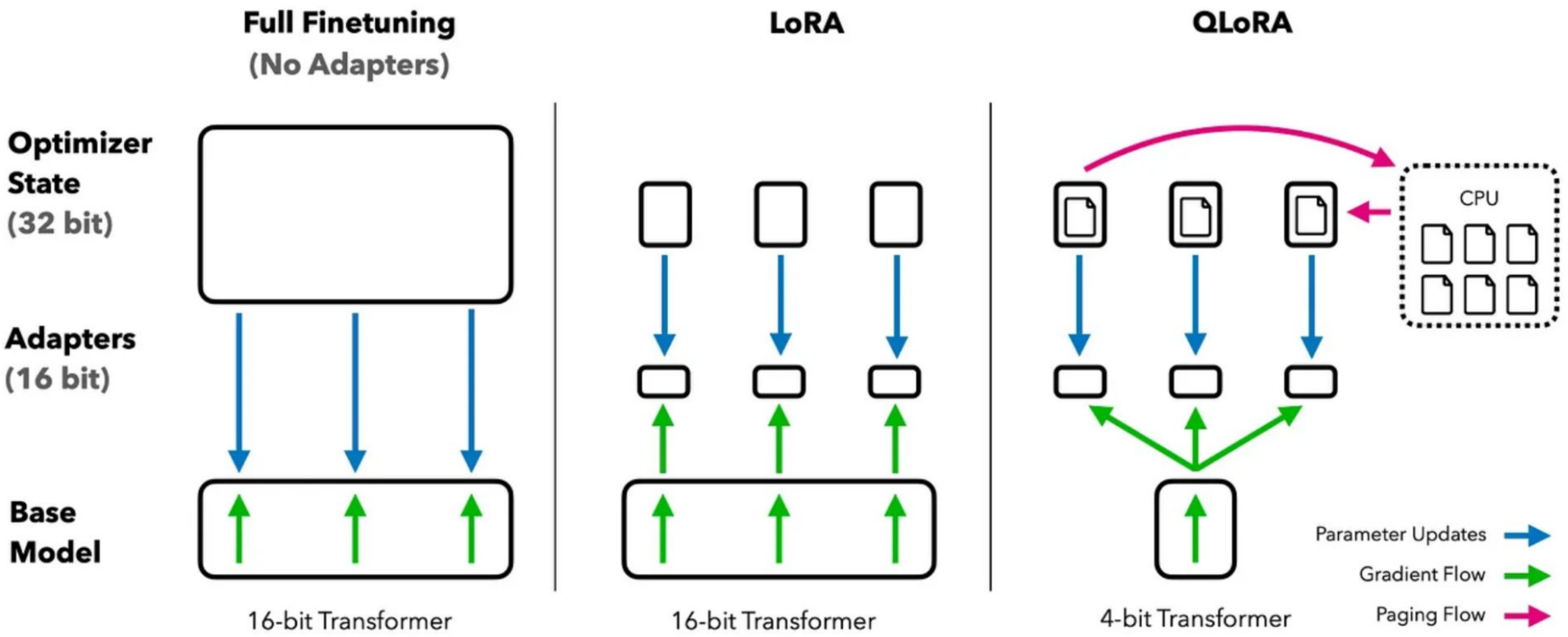


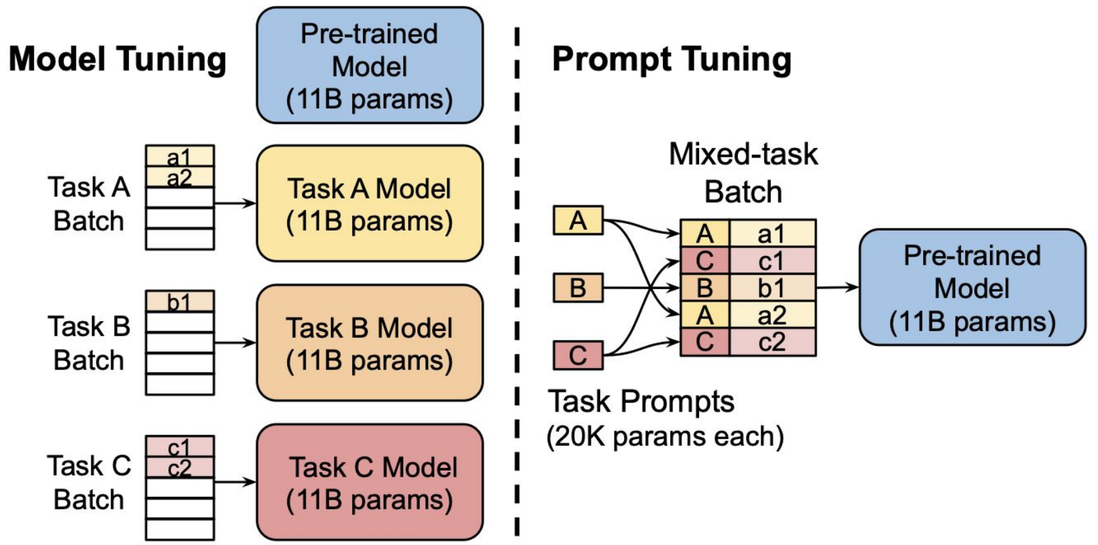
Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

QLoRA >>> LORA ???

- 4-bit quantized pretrained language model
- Introduces 4-bit NormalFloat (NF4)
- Applies Double Quantization, saving about 0.37 bits per parameter
- Less Memory Required

Prompt Tuning

- Frozen model
- Soft Prompts



Quiz!

https://tinyurl.com/ODSCLLMQ1

Demo in Practise

Thank You