Building Custom LLMs

a step by step guide and pros and cons of custom LLMs

Large Language Model

A large language model (LLM) is a computerized language model consisting of an artificial neural network with many parameters (tens of millions to billions), trained on large quantities of unlabeled text using self-supervised learning or semi-supervised learning.

Introduction

Accuracy: 0.9124235

Accuracy: 0.0000271

Accuracy: 0.9872232

Min Maung

linkedin.com/in/minmaung

MVP
Microsoft®
Most Valuable
Professional

Lwin Maung

linkedin.com/in/lmaung

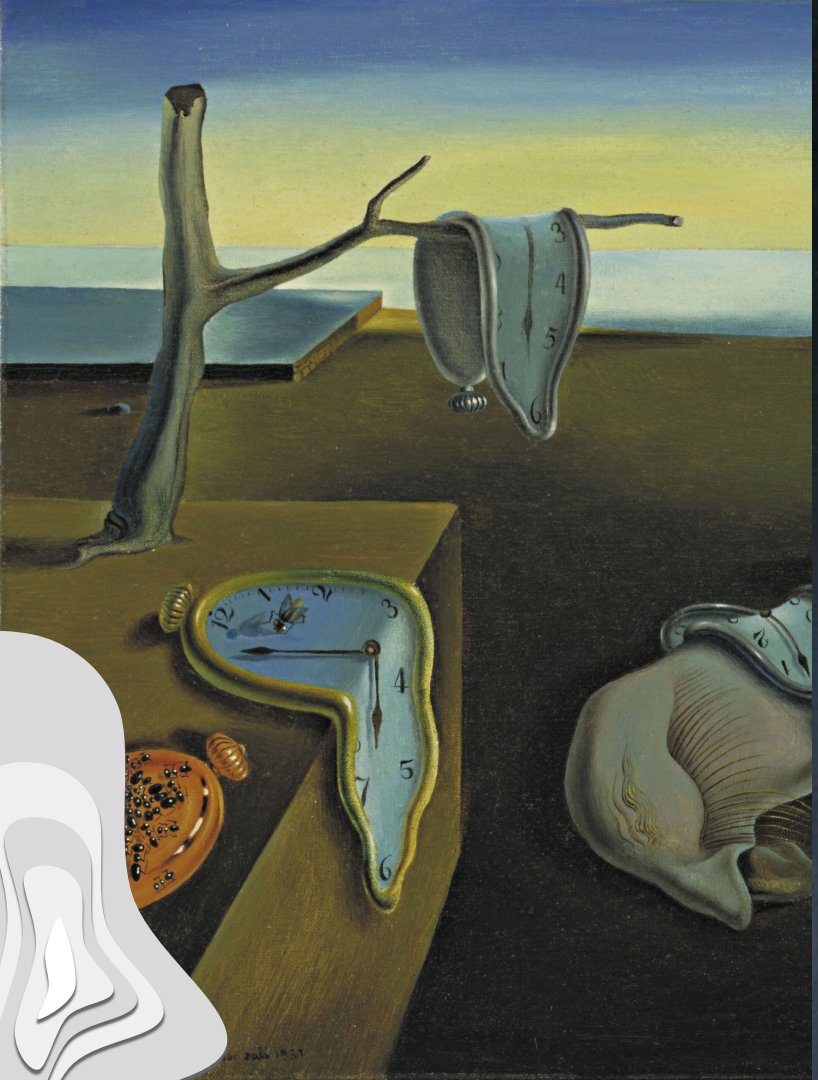
MVP
Microsoft®
Most Valuable
Professional

Here -> There

Reality is...

Reality is...

Reality is...



$$D = \frac{1}{c}\frac{dl}{dt} = \frac{1}{c}$$

$$D^2 = \frac{1}{P^2}\frac{P_0 - P}{P} \sim P^2$$

$$D^2 \frac{K\rho}{3}\frac{P_0 - P}{P_0} \sim 1 K\rho$$

$$D^2 \sim 10^{-53}$$

$$\rho \sim 10^{-26}$$

$$P \sim 10^{8}\ \text{Д.Э}$$

$$t \sim 10^{10}(10^{11})\ \text{Э}$$

LIVE DEMOS

LLAMA 3 8B – MLOPS

- Fine Tune
- Inference

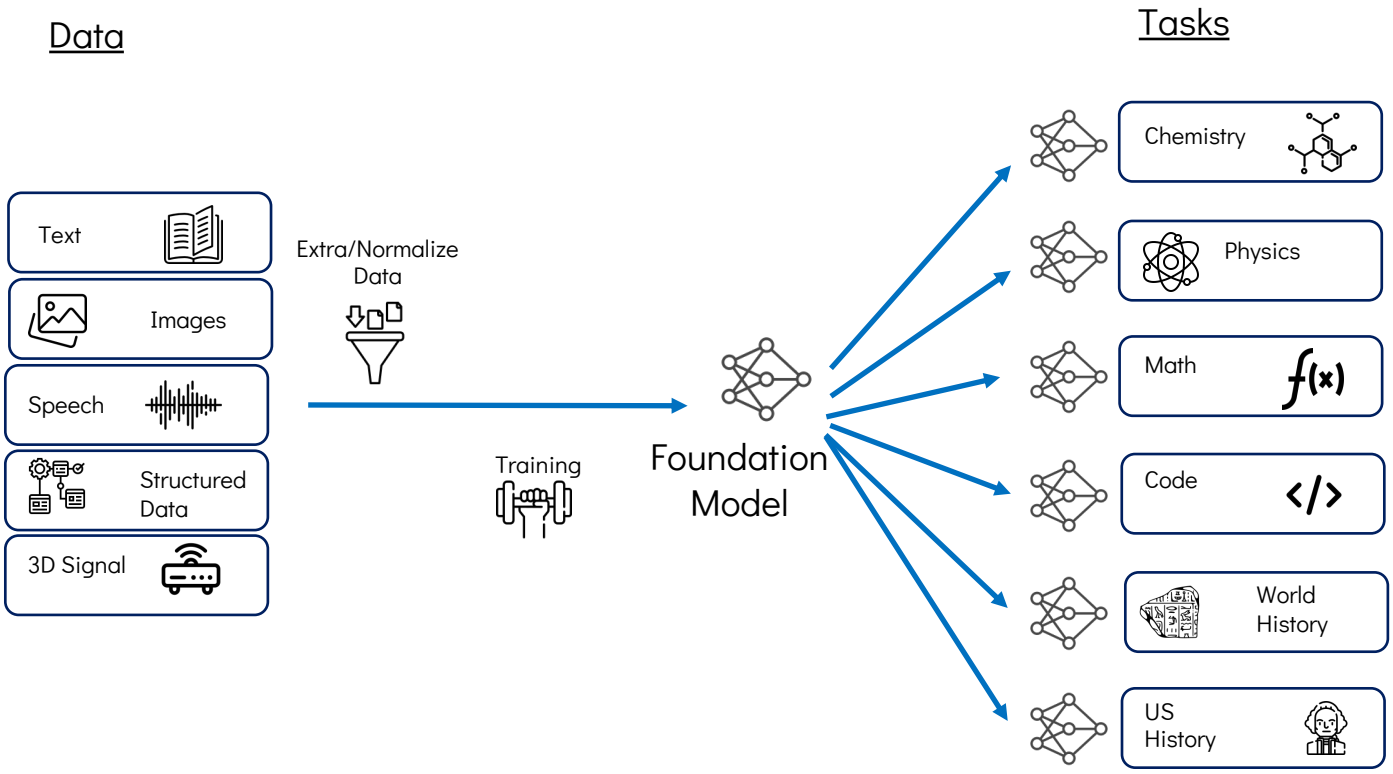
LLAMA2 70B – LOCAL

- Inference

PHI 3 Medium 128k – LOCAL

- Convert
- Fine Tune
- Inference

Models


Data
Tasks
Text
Images
Speech
Structured Data
3D Signal
Extra/Normalize Data
Training
Foundation Model
Chemistry
Physics
Math
Code
World History
US History

Models

LLM / SLM

φ	2 – 3	
φ	3 – 3	26
🦙	2 – 7	
🦙	2 – 13	
🦙	2 – 70	138
🦙	3 – 8	16
φ	3 – 14	56
🦙	3 – 70	141

Benchmark

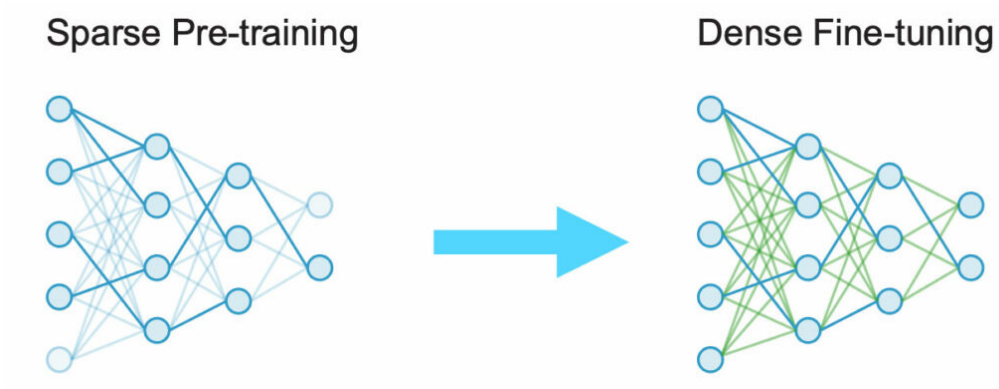
Benchmark	Llama 3 8B	Llama 2 70B
GPQA (0-shot)	**34.2**	**21.0**
HumanEval (0-shot)	**62.2**	**25.6**
GSM-8K (8-shot, CoT)	**79.6**	**57.5**
MATH (4-shot, CoT)	**30.0**	**11.6**

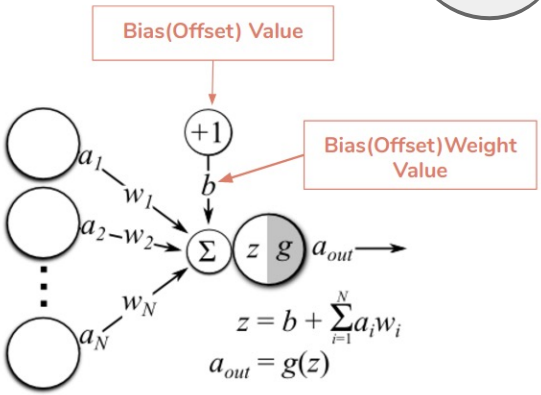
Category	Benchmark	Phi-3-Medium		Mistral-8x22B	Llama-3-70B-Instruct	GPT3.5-Turbo-1106	Claude-3 Sonnet	Gemini 1.0 Pro
		Phi-3-Medium-4K-In	Phi-3-Medium-128K-In					
Popular Aggregate Benchmarks	MMLU (5-shot)	78.0	76.6	76.2	80.2	71.4	73.9	66.7
Language Understanding	HellaSwag (5-shot)	82.4	81.6	79.0	82.6	78.8	79.2	76.2
Reasoning	WinoGrande (5-shot)	81.5	78.9	75.3	83.3	68.8	81.4	72.2
	Social IQA (5-shot)	80.2	79.0	78.2	81.1	68.3	80.2	75.4
	TruthfulQA (MC2) (10-shot)	75.1	74.3	67.4	81.9	67.7	77.8	72.6
	MedQA (2-shot)	69.9	67.6	67.9	78.5	63.4	67.9	58.2
Factual Knowledge	TriviaQA (5-shot)	73.9	73.9	84.5	78.5	85.8	65.7	80.2
Math	GSM8K CoT (8-shot)	91.0	87.5	83.8	93.5	78.1	79.1	80.4
Code generation	HumanEval (0-shot)	62.2	58.5	39.6	78.7	62.2	65.9	64.4
	MBPP (3-shot)	75.2	73.8	70.7	81.3	77.8	79.4	73.2

Tokens

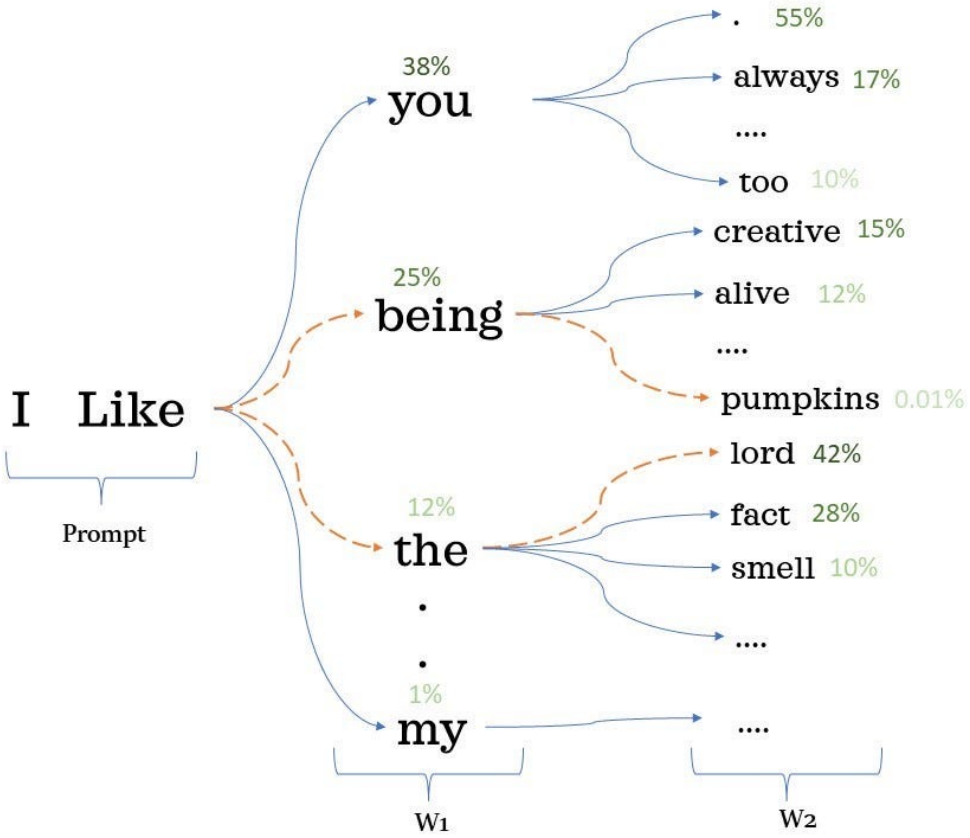
A Large Language Model (LLM), like OpenAI's GPT-3 or GPT-4, operate based on a process called tokenization. Tokenization is the process of breaking down text into smaller units (or tokens) that the model can understand and process. Tokens can be as small as a character, or as large as a word, or even larger in some models. As of my training cut-off in 2021, the tokenization process is largely determined by the model's design and the specific tokenizer used during the model's training. In the case of GPT-3 and GPT-4, they use a Byte Pair Encoding (BPE) tokenizer. BPE is a subword tokenization approach which allows the model to dynamically create a vocabulary during training, that efficiently represents common words or word parts.

Weights

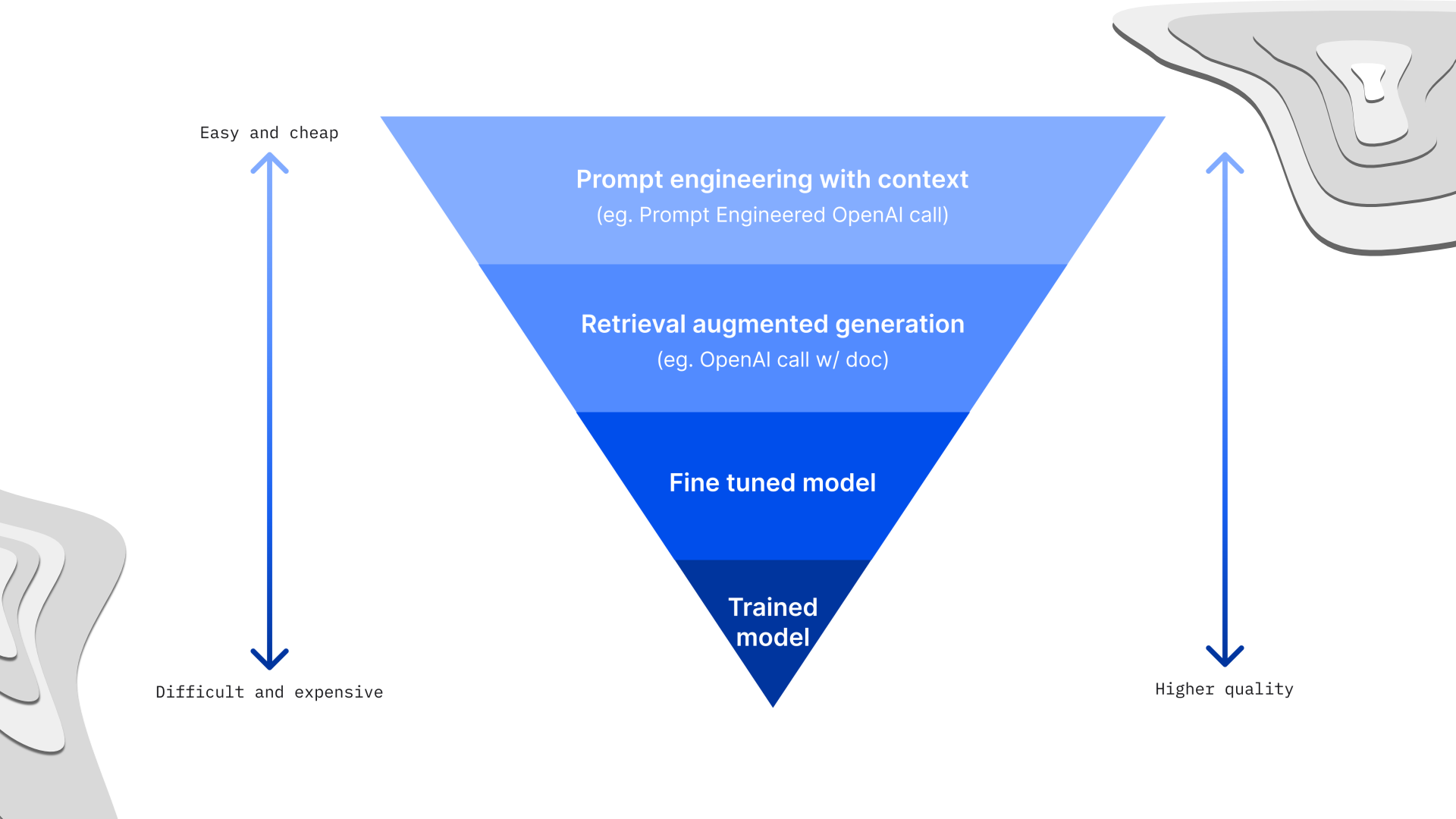

Sparse Pre-training
Dense Fine-tuning


Bias(Offset) Value
Bias(Offset)Weight Value
+1
b
a_1
a_2
a_N
w_1
w_2
w_N
Σ
z
g
a_{out}
$z = b + \sum_{i=1}^{N} a_i w_i$
$a_{out} = g(z)$

Temperature



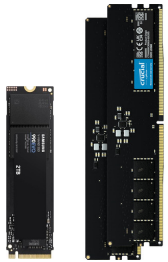

STOP

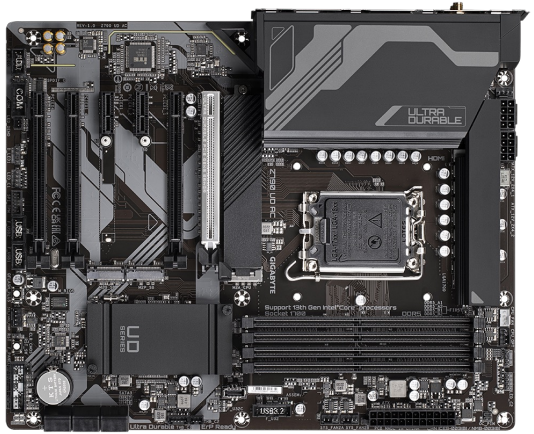

Easy and cheap
Prompt engineering with context
(eg. Prompt Engineered OpenAI call)
Retrieval augmented generation
(eg. OpenAI call w/ doc)
Fine tuned model
Trained model
Difficult and expensive
Higher quality

Complex Path

Is this for you?

Is this for you?






ZOTAC
ZOTAC
msi
msi
ASUS
ASUS

Proud Father

Cost

GPU

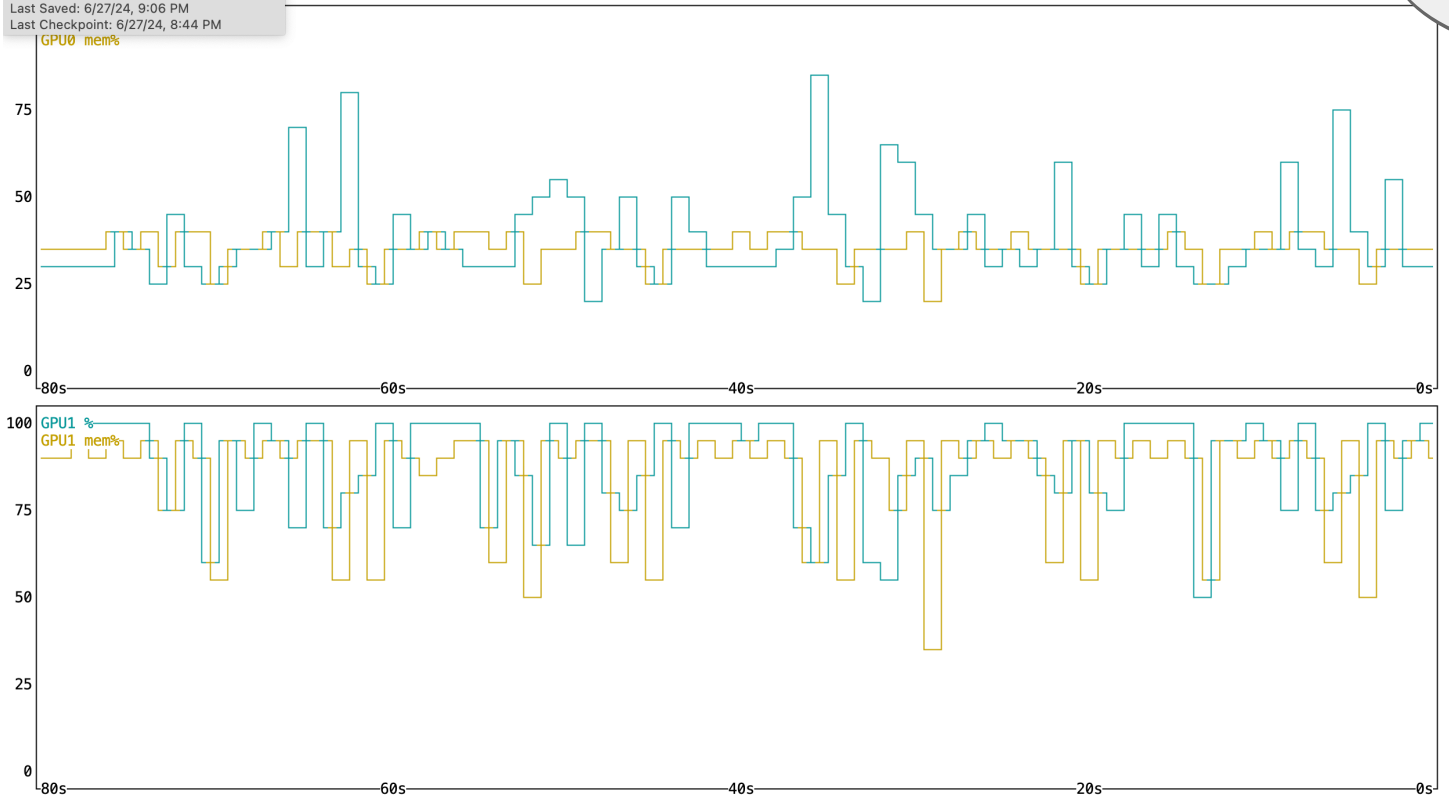
RTX 4090	Nvidia A/H100	Clusters	Cloud Services
24GB	40GB-80GB	8 x Nvidia A100 / H100	Azure ML
$1800	$25,000-$45,000	$19,835/month	$27.197/hour

Device 0 [NVIDIA GeForce RTX 4090] PCIe GEN 3@16x RX: 1.344 GiB/s TX: 4.883 MiB/s Device 1 [NVIDIA GeForce RTX 4090] PCIe GEN 3@ 8x RX: 0.000 KiB/s TX: 0.000

4°C FAN 30% POW 113 / 450 W GPU 2835MHz MEM 10251MH TEMP 45°C FAN 30% POW 256 / 450 W

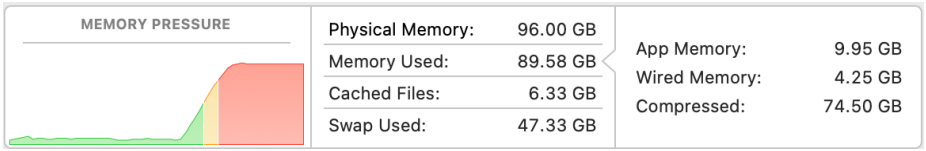
28%] MEM[|||||||||||||| 8.730Gi/23.988Gi] GPU[||||||||||||||||||||||||||||||100%] MEM[||||||||||||||||||21.823Gi/23.98

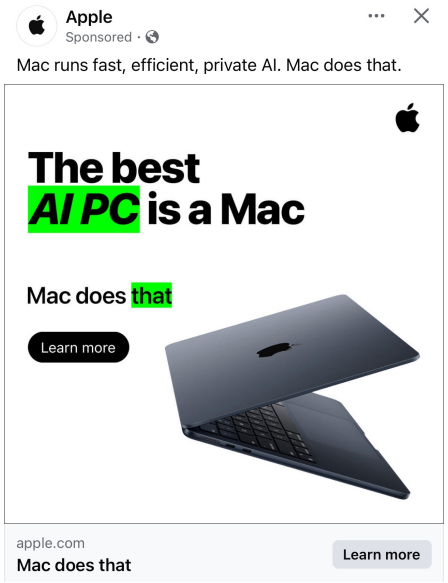
Name: FineTune8B.ipynb
Path: Desktop/Notebooks/FineTune8B.ipynb
Last Saved: 6/27/24, 9:06 PM
Last Checkpoint: 6/27/24, 8:44 PM



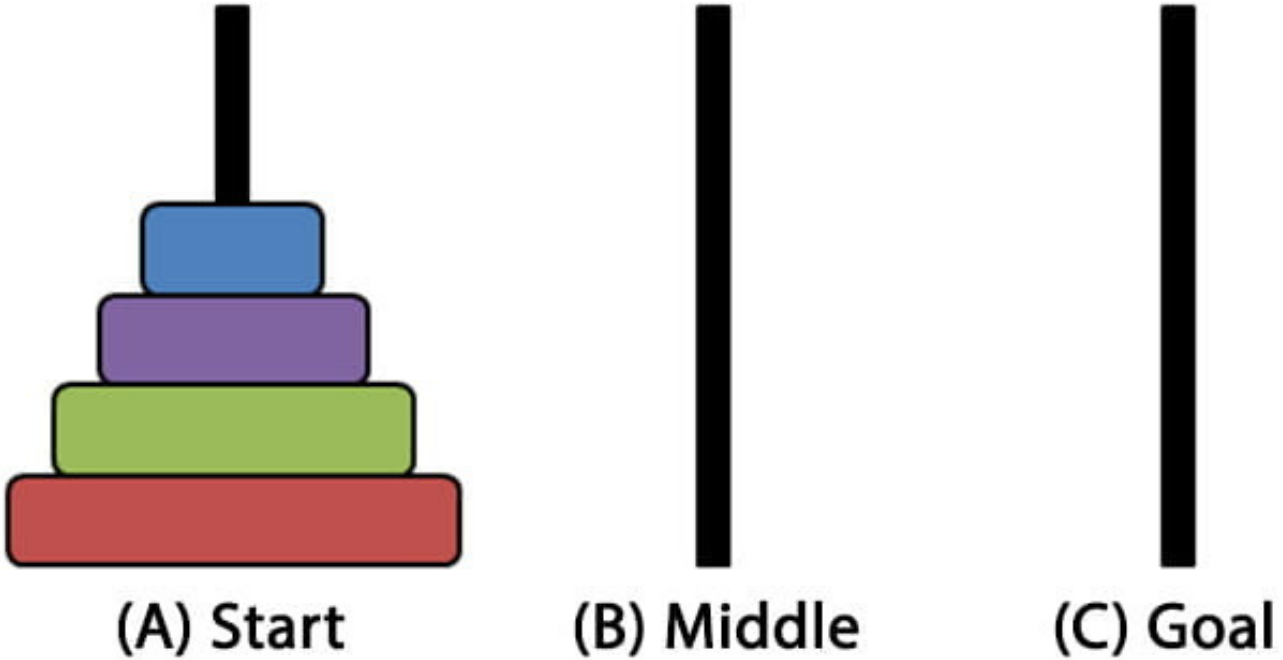
PID	USER	DEV	TYPE	GPU	GPU MEM	CPU	HOST MEM	Command
115519	mlops	1	Compute	63%	21694MiB	88%	0%	1503MiB	/opt/jupyterhub/bin/python3 -m ipykernel_launcher -f /home/mlops/.local/share/jupyter/runtime/kernel-
115519	mlops	0	Compute	28%	7838MiB	32%	97%	1503MiB	/opt/jupyterhub/bin/python3 -m ipykernel_launcher -f /home/mlops/.local/share/jupyter/runtime/kernel-
1687	mlops	0	Graphic	0%	494MiB	2%	0%	112MiB	/usr/lib/xorg/Xorg vt2 -displayfd 3 -auth /run/user/1000/gdm/Xauthority -nolisten tcp -background non
1687	mlops	1	Graphic	18%	133MiB	1%	0%	112MiB	/usr/lib/xorg/Xorg vt2 -displayfd 3 -auth /run/user/1000/gdm/Xauthority -nolisten tcp -background non
1991	mlops	0	Graphic	0%	109MiB	0%	0%	105MiB	/usr/bin/gnome-shell

Apple Silicon





Layers


(A) Start
(B) Middle
(C) Goal

LIVE DEMOS

LLAMA 3 8B – MLOPS

- Fine Tune
- Inference

LLAMA2 70B – LOCAL

- Inference

PHI 3 Medium 128k – LOCAL

- Convert
- Fine Tune
- Inference

Q & A

Large Language Model

A large language model (LLM) is a computerized language model consisting of an artificial neural network with many parameters (tens of millions to billions), trained on large quantities of unlabeled text using self-supervised learning or semi-supervised learning.