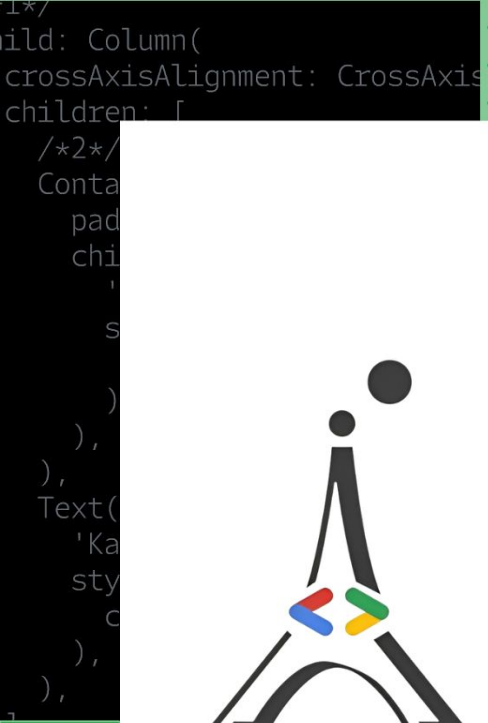


MASTERING AI WORKSHOP

Google Developer Student Clubs Paris-Saclay

Lucas J. Velôso

GDSC Lead

14/12/2023

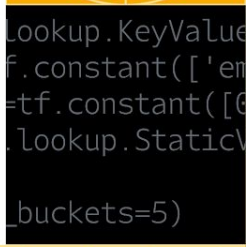
Google Developer Student Clubs

Google Developer Student Clubs

Agenda

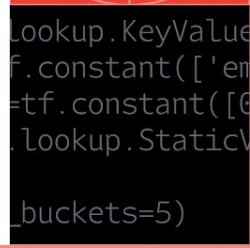
Google Developer Student Clubs

1. What are LLMs?
2. How LLMs work?
3. What is Fine-tuning?
4. What is Distillation?
5. What is **PEFT**?
6. What is **LoRA**
7. What is **QLoRA**?
8. Coding time...



Google Developer Student Clubs

What are LLMs?



1. **Large Language Models (LLMs)** are advanced AI tools designed for processing and generating human-like text.

2. They **leverage vast amounts of data** to understand context, semantics, and language nuances.

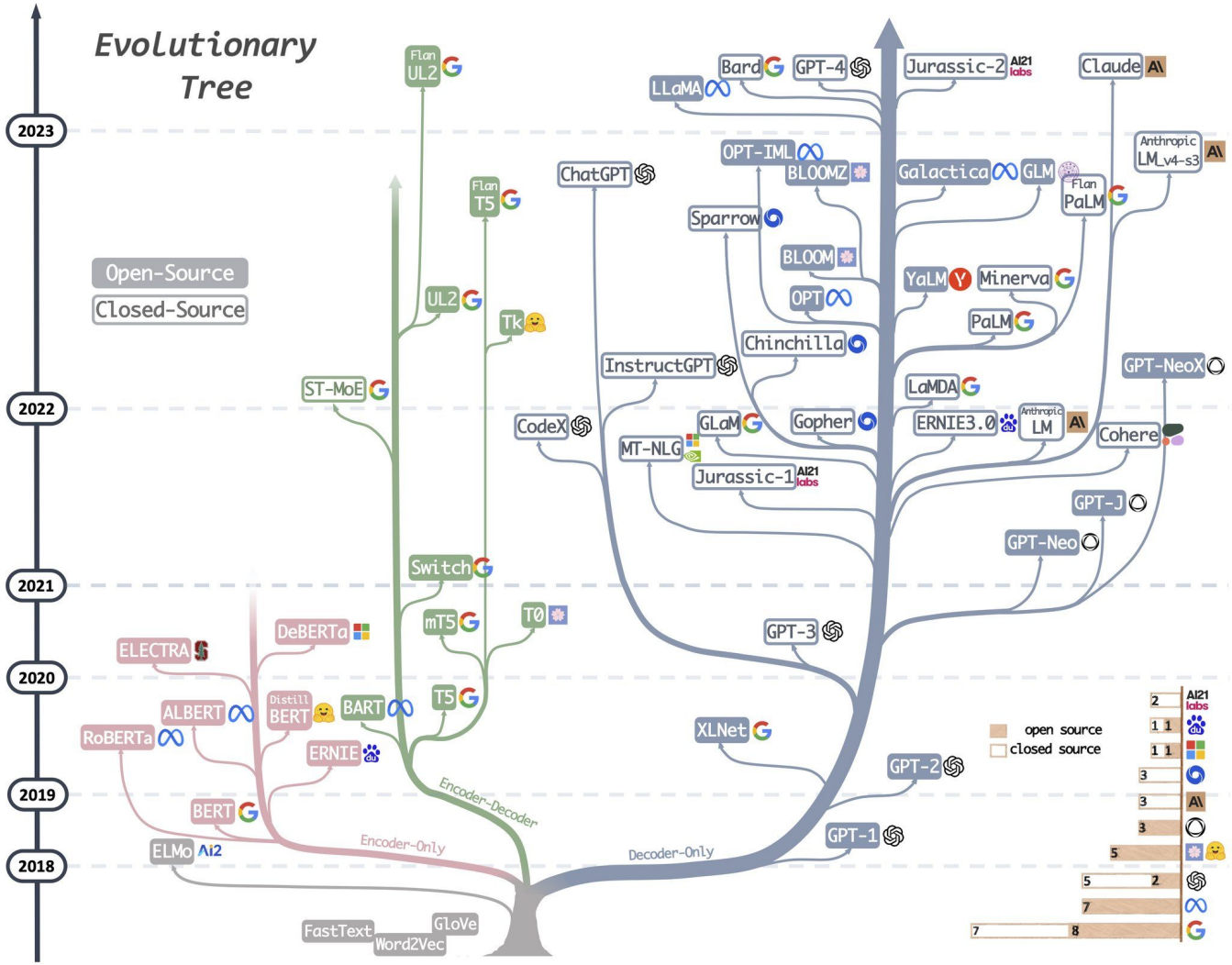
3. LLMs are used in various applications like **chatbots, content creation, and language translation**.

Gemini

LLaMA by Meta

Google Bard

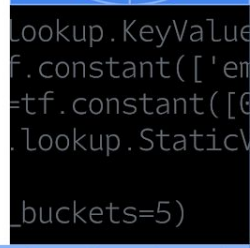
Although various Large Language Models (LLMs) have existed since 2018, they only gained widespread popularity among the general public starting in 2022, with the advent of **ChatGPT**.

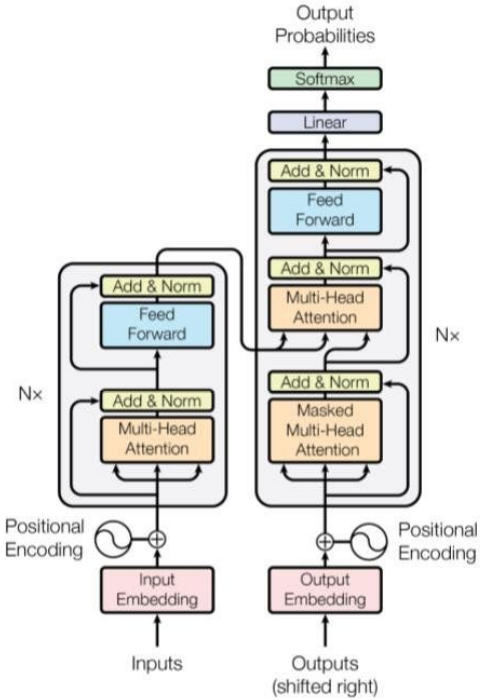


Google Developer Student Clubs

Google Developer Student Clubs

How LLMs work?

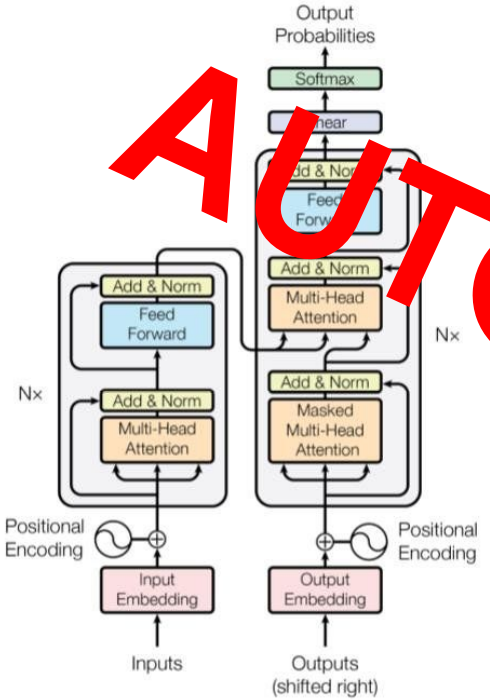




1. **Deep learning models**, primarily Transformers (Google 2017).

2. Multiple layers of processing (attention mechanisms, etc.)

3. Solves the task: **Given the phrase with is the next word?**



AUTOCOMPLETE

1. **Deep learning models**, primarily Transformers (Google 2017).
2. Multiple layers of processing (attention mechanisms, etc.)
3. Solves the task: **Given the phrase with is the next word?**

Stochastic Parrots

Google Developer Student Clubs

What is Fine-Tuning?

Fine-Tuning

Gigantic web-scale dataset

Fine-Tuning

Gigantic web-scale dataset

Base LLM

Fine-Tuning

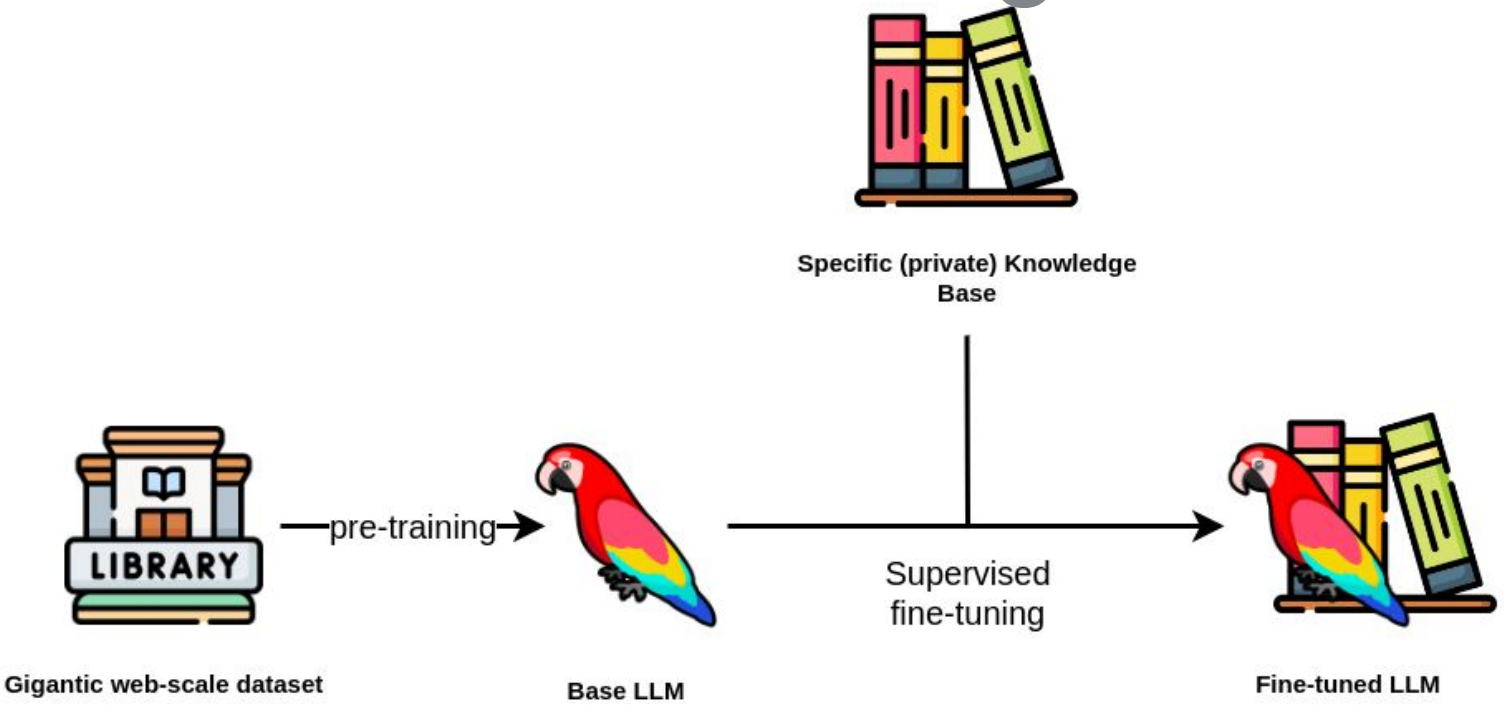
Specific (private) Knowledge Base

—pre-training→

Gigantic web-scale dataset

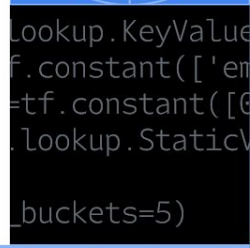
Base LLM

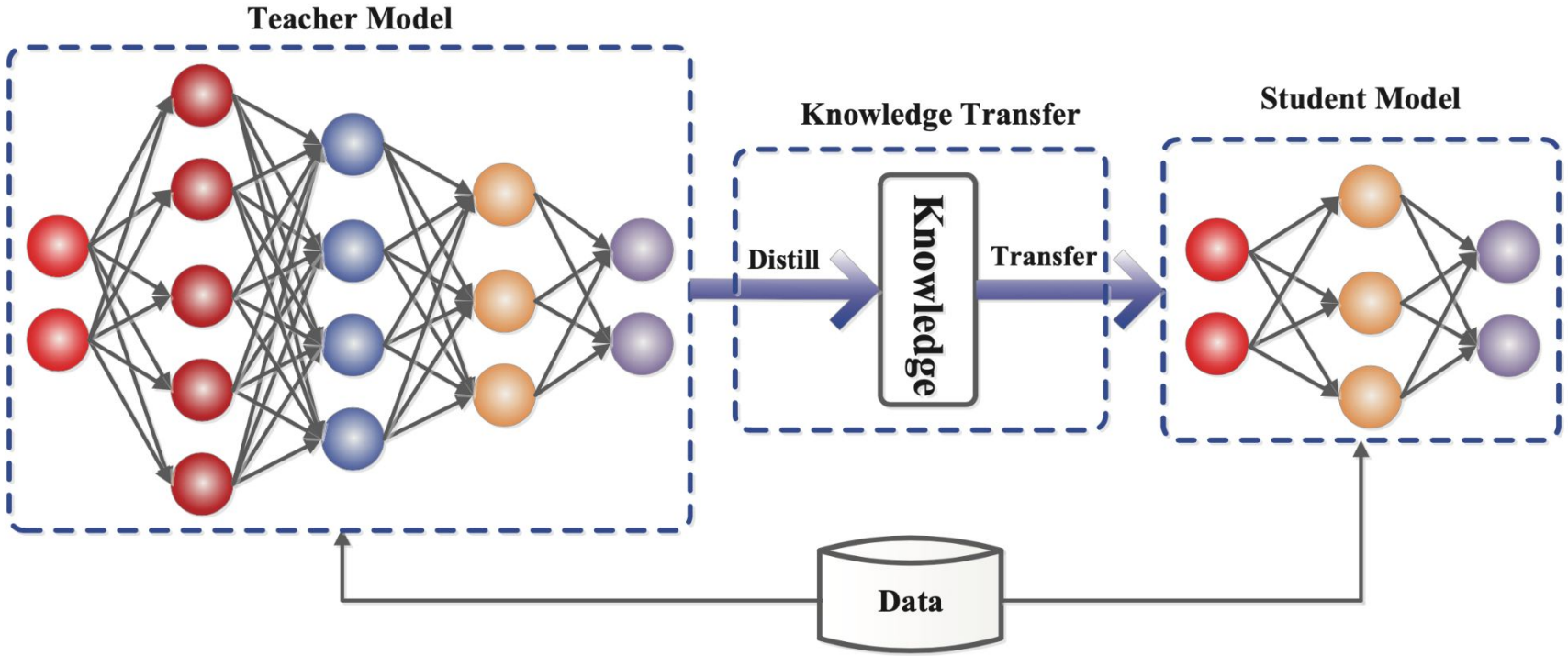
Fine-Tuning

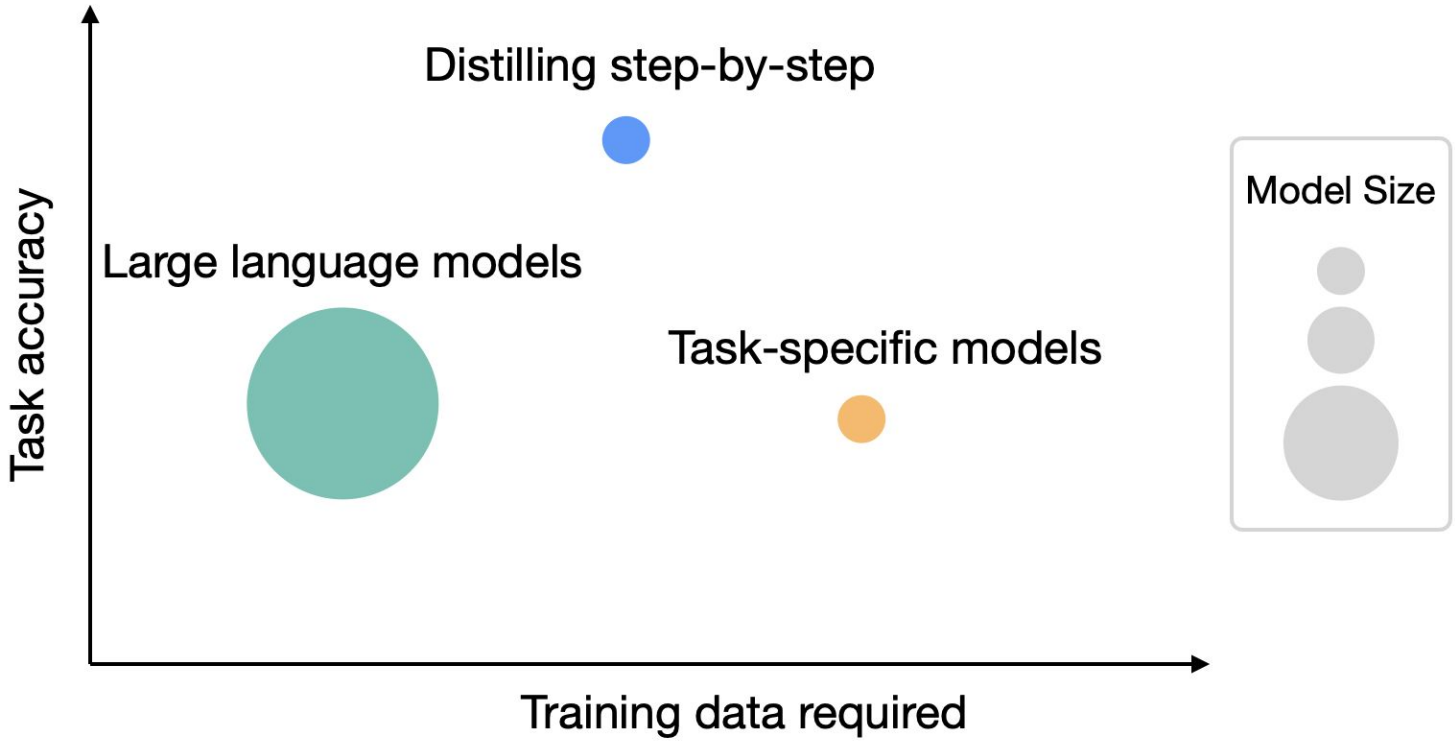

Specific (private) Knowledge Base
LIBRARY
pre-training
Supervised fine-tuning
Gigantic web-scale dataset
Base LLM
Fine-tuned LLM

Google Developer Student Clubs

What is Distillation?

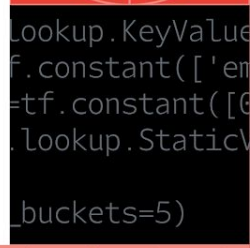



Teacher Model
Knowledge Transfer
Student Model
Distill
Knowledge
Transfer
Data


Distilling step-by-step
Large language models
Task-specific models
Task accuracy
Training data required
Model Size

Google Developer Student Clubs

What is PEFT?



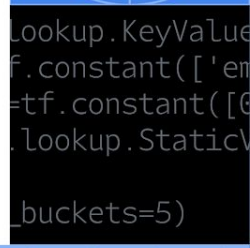
Parameter-Efficient Fine-Tuning (PEFT)

Parameter-Efficient Fine-Tuning (PEFT) by **Hugging Face** is a technique to adapt large models for specific tasks by **fine-tuning only a few parameters**, reducing computational and storage demands

Hugging Face

Google Developer Student Clubs

What is LoRA?



LoRA

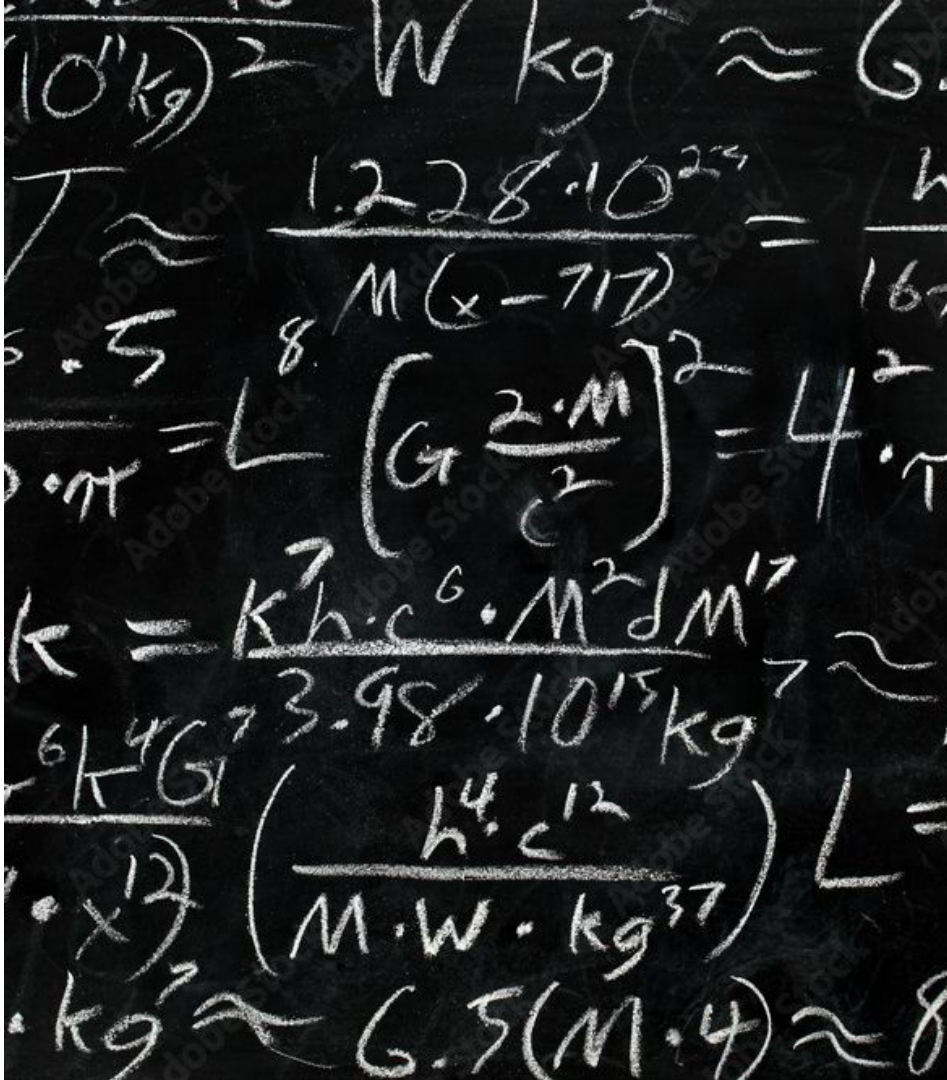
Finetuned Weights

Weight Update

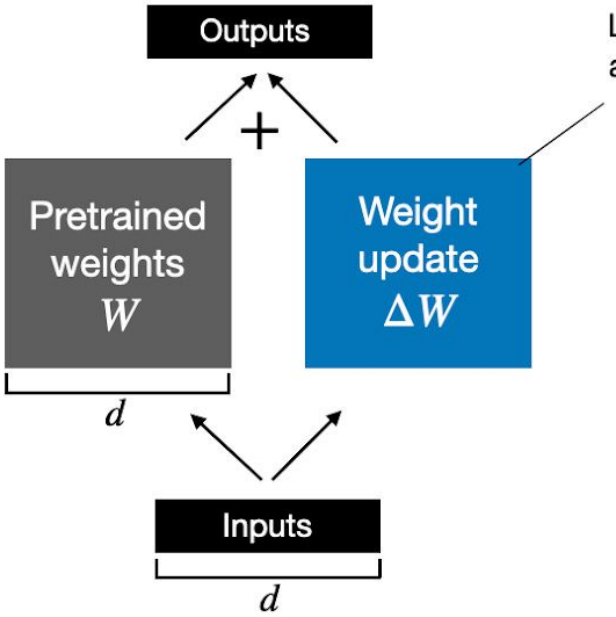
$$W_{\mathbf{ft}} = W_{\mathbf{pt}} + \Delta W$$

Pretrained Weights

Google Developer Student Clubs

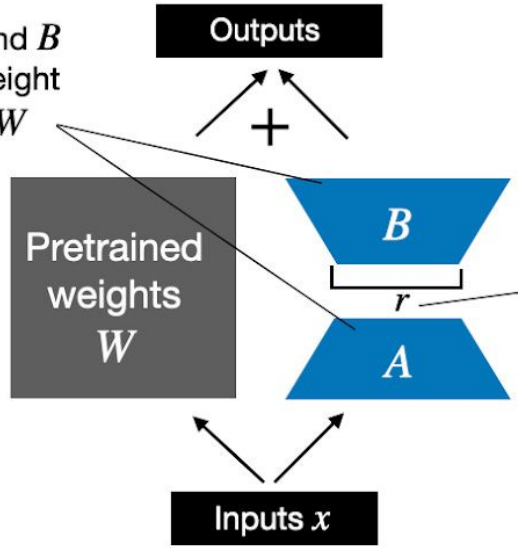


Weight update in regular finetuning



LoRA matrices A and B approximate the weight update matrix ΔW

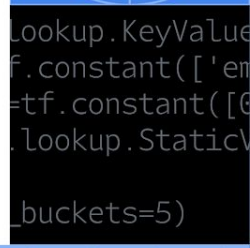
Weight update in LoRA



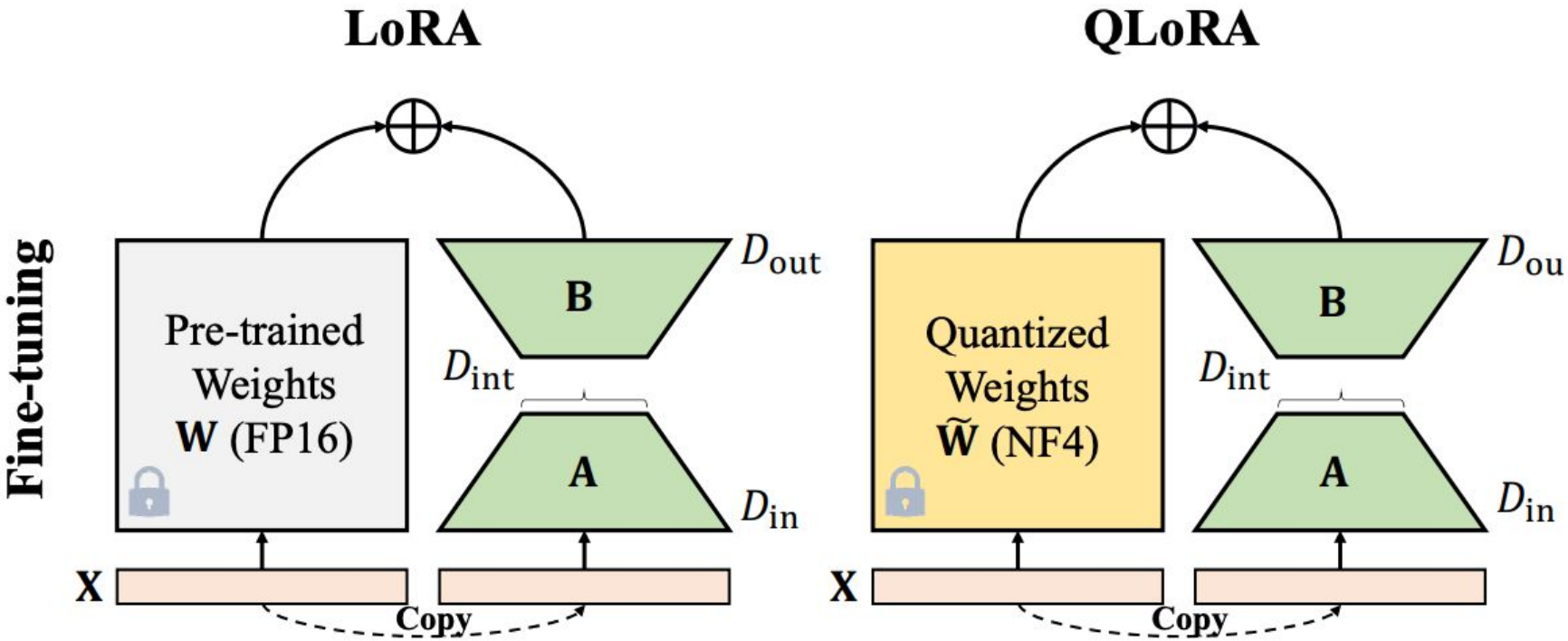
The inner dimension r is a hyperparameter

Google Developer Student Clubs

What is QLoRA?

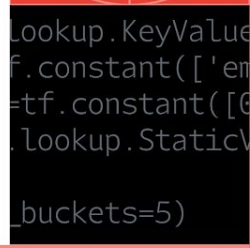


QLoRA


LoRA
QLoRA
Fine-tuning
Pre-trained
Weights
W (FP16)
B
A
D_{out}
D_{int}
D_{in}
X
Copy
Quantized
Weights
$\widetilde{\mathbf{W}}$ (NF4)
B
A
D_{ou}
D_{int}
D_{in}
X
Copy

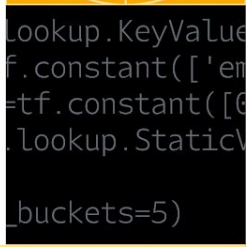
Google Developer Student Clubs

Coding Time...



Google Developer Student Clubs

STARTUP CASE



Google Developer Student Clubs

Google Developer Student Clubs

Google Developer Student Clubs

1. It **must** run in a cell phone!!

Google Developer Student Clubs

1. It **must** run in a cell phone!!
2. It **must** run offline!!

Google Developer Student Clubs

1. It **must** run in a cell phone!!
2. It **must** run offline!!
3. The cost must be at **maximum** $1

Thank you for joining!

Google Developer Student Clubs