

LLM Hands on

Super AI Engineer SS5

Norapat Buppodom

Outline


1. Overview
2. LLM Fine-tuning SFT on Lanta
3. LLM Fine-tuning GRPO on Lanta
4. LLM Inference (Fast!)

How to Fine-tune LLM



Example Dataset

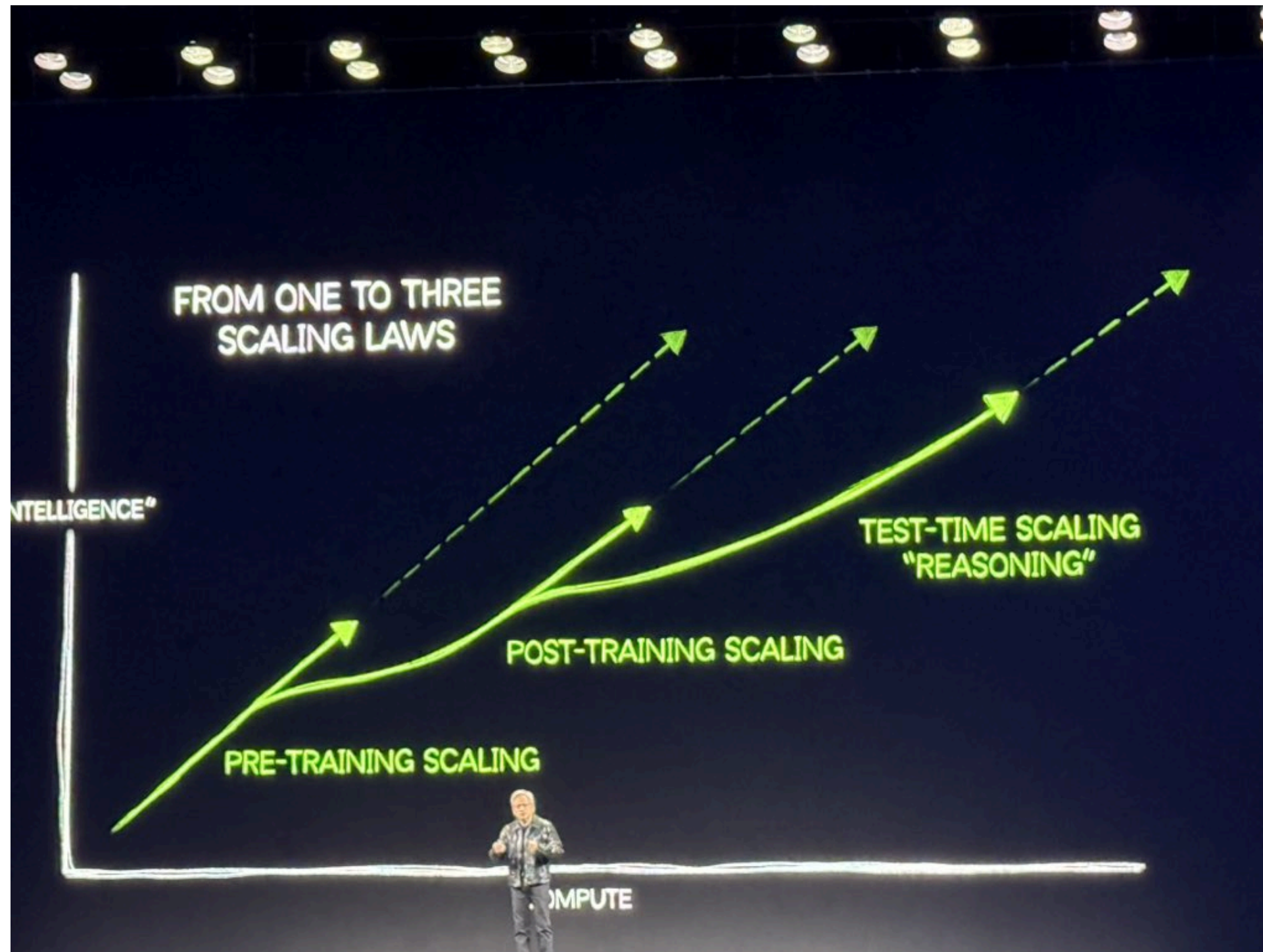
AG News

# Class Index	△ Title	△ Description
Consists of class ids 1-4 where 1-World, 2-Sports, 3-Business, 4-Sci/Tech	Contains title of the news articles	Contains description of the news articles
 14	7568 unique values	7594 unique values
3	Fears for T N pension after talks	Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken...
4	The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com)	SPACE.com - TORONTO, Canada -- A second\team of rocketeers competing for the #36;10 million Ansari ...

4 class, 7k examples

<https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset>

3 types of LLM



1. Base Model
2. Instruct Model
3. Reasoning Model

Base vs Instruct Model

Base Model

- Text Completion
- **For fine-tuning large dataset**
- Less Bias

Instruct/Chat Model

- Chat / Instruction Following
- **For fine-tuning small dataset**
- More Bias

Quick Model Selection

Base vs Instruct

300 - 1,000 Rows: Instruct or Base

1,000+ Rows: Base

Ref: <https://docs.unsloth.ai/get-started/fine-tuning-guide/what-model-should-i-use>

Base Model

The process of photosynthesis in plants begins when light energy from the sun is absorbed by chlorophyll in the chloroplasts. This energy is used to convert carbon dioxide and water into glucose and oxygen, providing essential nutrients and releasing oxygen as a byproduct.

 Input  Output

Instruct/Chat Model

<|user|>

Explain the process of photosynthesis in plants.

<|assistant|>

Photosynthesis is the process by which plants convert light energy from the sun into chemical energy in the form of glucose. It occurs in the chloroplasts, where chlorophyll absorbs sunlight. Plants take in carbon dioxide from the air and water from the soil, and use the energy from sunlight to convert these into glucose and oxygen. The glucose provides energy for the plant, while oxygen is released as a byproduct.

 Input  Output

Instruct/Chat Model





<|user|>

Title: Fears for T N pension after talks

Description: Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken ...

<|assistant|>

3

# Class Index 	△ Title 	△ Description 
Consists of class ids 1-4 where 1-World, 2-Sports, 3-Business, 4-Sci/Tech	Contains title of the news articles	Contains description of the news articles
 14	7568 unique values	7594 unique values
3	Fears for T N pension after talks	Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken...
4	The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com)	SPACE.com - TORONTO, Canada -- A second\team of rocketeers competing for the #36;10 million Ansari ...

Apply Chat Template

<https://huggingface.co/Qwen/Qwen3-0.6B>

```
# prepare the model input
prompt = "Give me a short introduction to large language model."
messages = [
    {"role": "user", "content": prompt}
]
text = tokenizer.apply_chat_template(
    messages,
    tokenize=False,
    add_generation_prompt=True,
    enable_thinking=True # Switches between thinking and non-thinking
)
```

<|im_start|>user

Give me a short introduction to large language model.<|im_end|>

<|im_start|>assistant

Model Size

- Speed
 - **token/sec**
 - time to first token
- Memory Used
 - **Can it fit on my GPU?**

<https://apxml.com/tools/vram-calculator>

Fine-tuning optimization

- **quantization**
- **Lora**
- **QLora**
- **gradient checkpointing**
- **fused kernel**
- **Flash Attention**

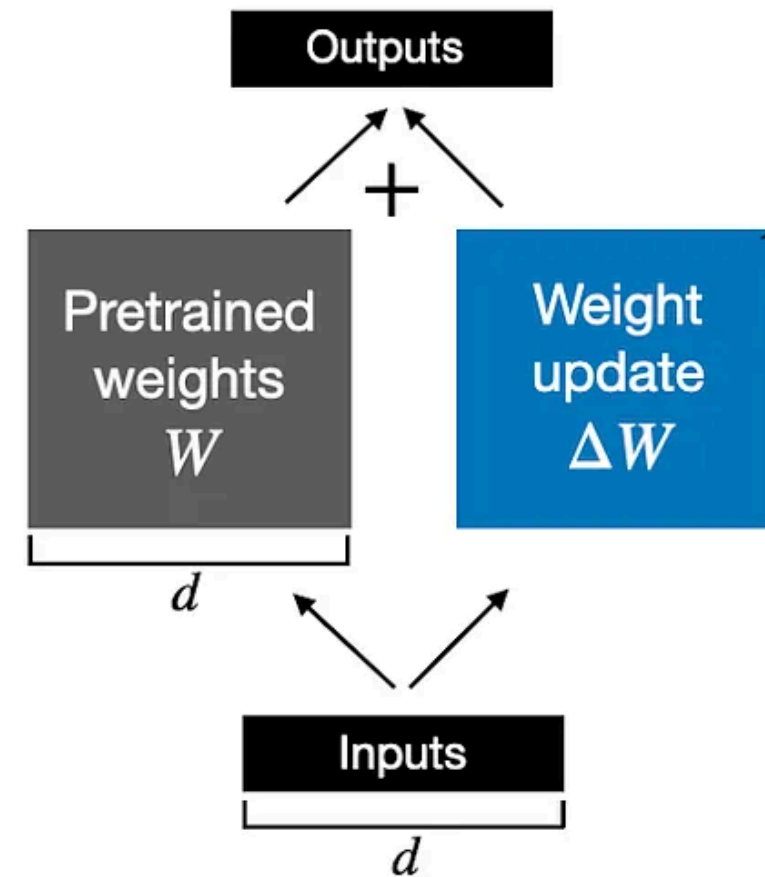
Quantization

16FP → 8 bit → 4 bit

- Impact training Speed ▼
- Impact Performance ▼
- Improve Memory Usage ✓

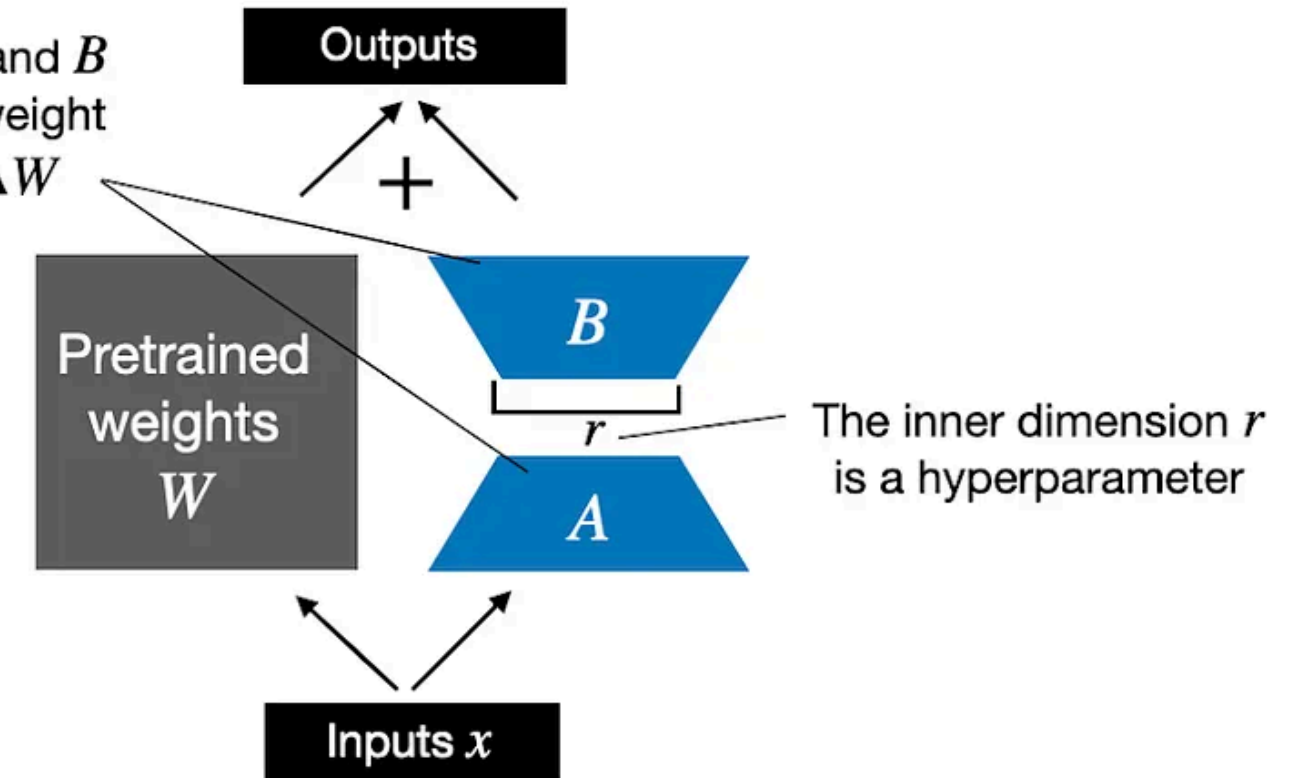
Lora

Weight update in **regular finetuning**



LoRA matrices A and B approximate the weight update matrix ΔW

Weight update in **LoRA**



- Improve Training Speed ✓
- Improve Memory used ✓
- Data Efficient ✓
- Reduce Generalization ▼
- Reduce World Knowledge ▼

QLoRa = Lora + quantization

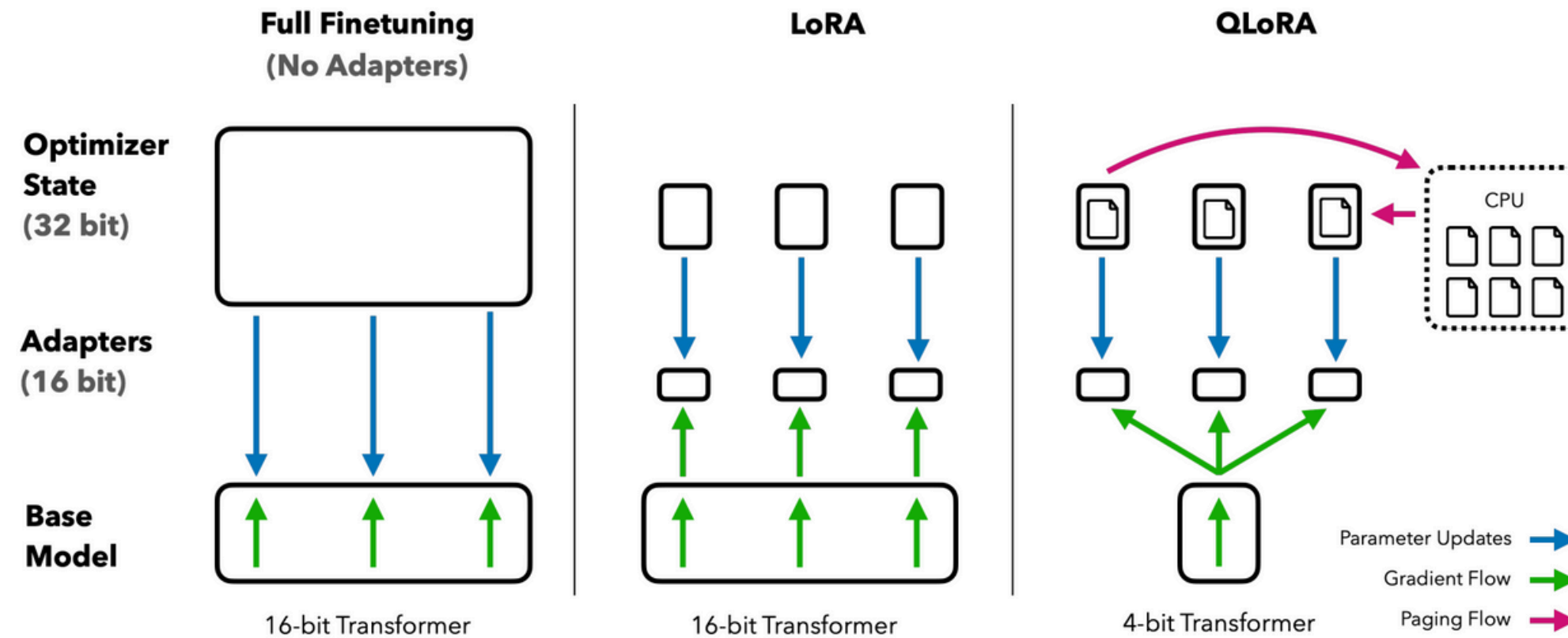
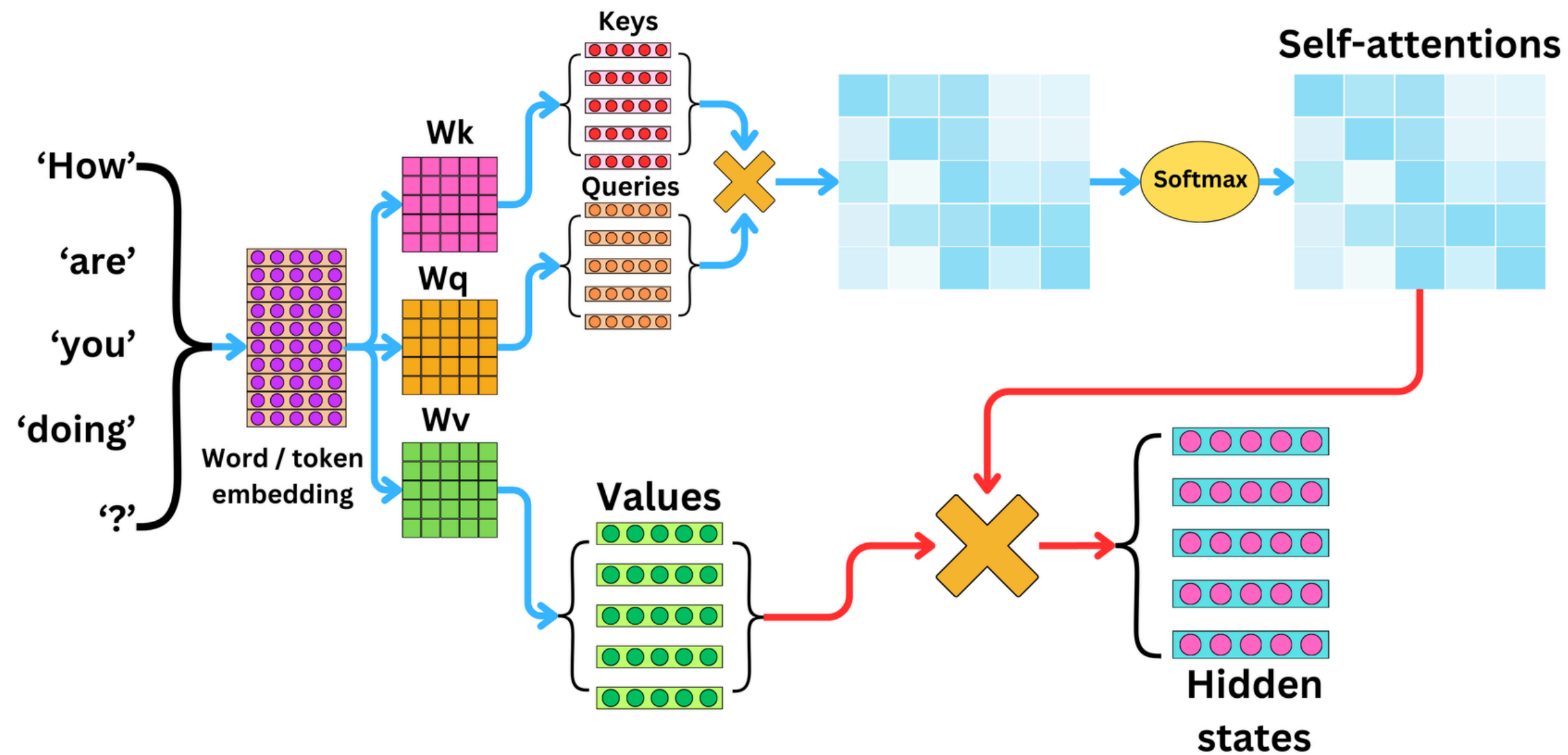


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

- Impact training Speed ▼
- **NOT** Impact Accuracy (Same as Lora) ✓
- Improve Memory Usage ✓

Flash Attention



- **Reduce M**

Multi GPU Training

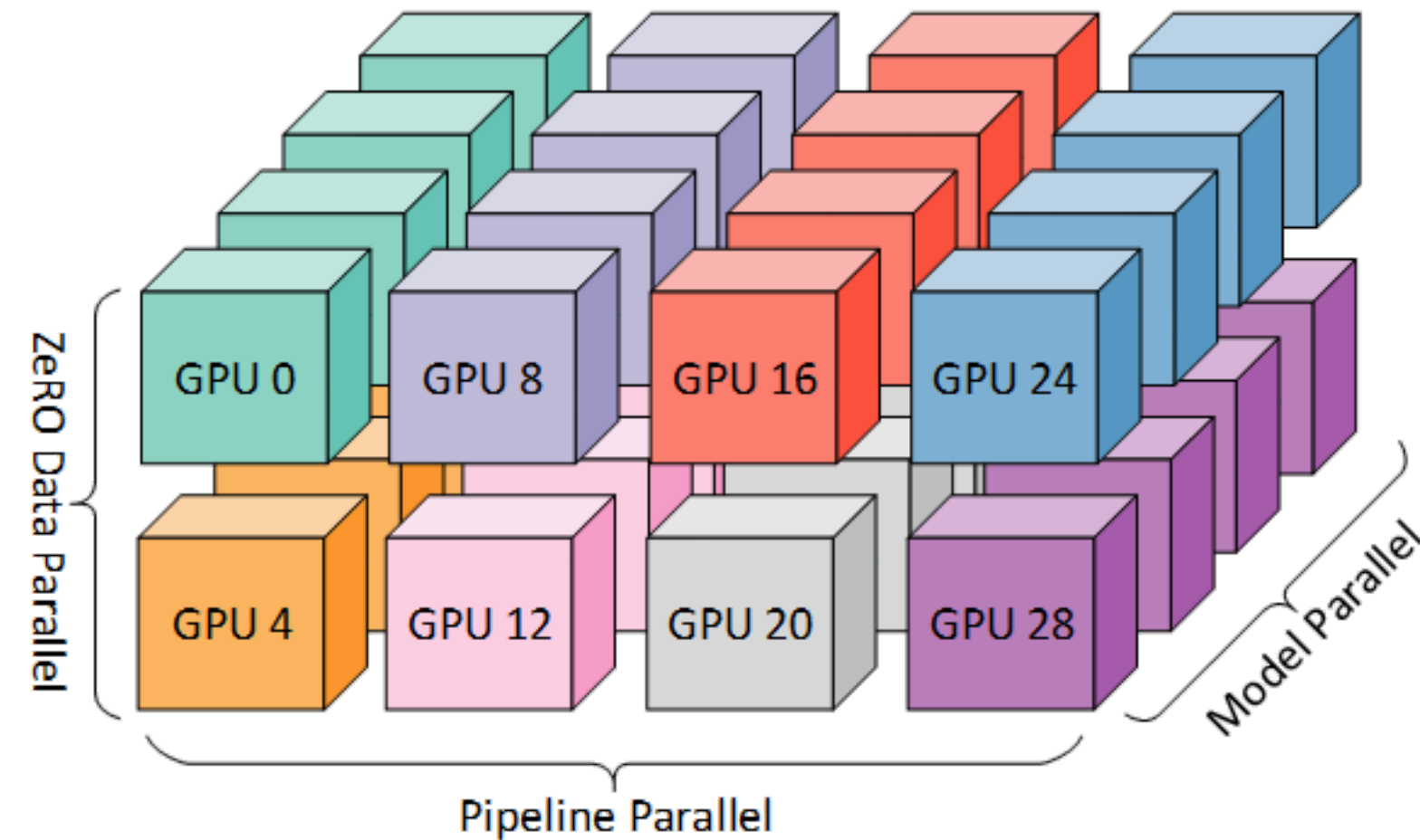
1. Data Parallel - Normal Multi GPU

2. Sharding

a. **Zero Sharding**

b. Pipeline Parallel

c. Model Parallel



Zero Shading



Zero Shading

Memory Usage in LLM (Transformers)

- Static
 - Model Weight **1X**
 - Model Gradient **1X**
 - Optimizer (AdamW) **3X**
- Dynamic
 - Batch size
 - Sequence Length

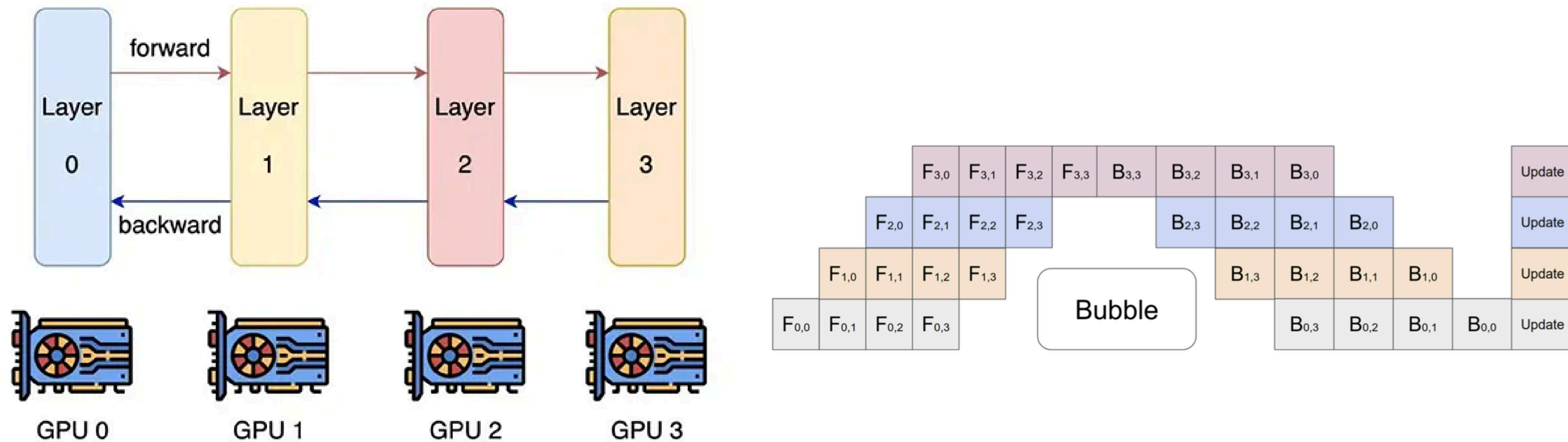
Zero Shading

DeepSpeed Zero Optimizer

- Stage 1: Shard Optimizer (AdamW) 3X
- Stage 2: Shard Optimizer + Gradient
- Stage 3: Shard Optimizer + Gradient + Model Weight

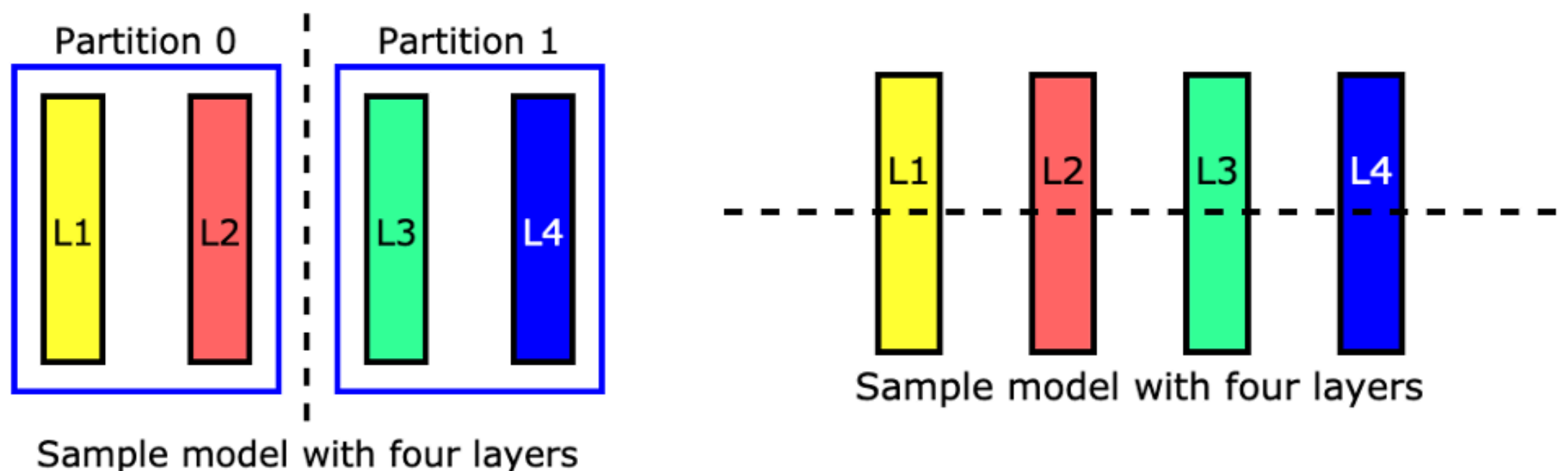
General Recommendation: Go to stage 2 if
model can fit in memory

Pipeline Parallelism



<https://arxiv.org/abs/1811.06965>

Tensor Parallelism



Pipeline Parallelism vs Tensor Parallelism

<https://github.com/NVIDIA/Megatron-LM>

What to adjust when training LLM

1. Batch Size

- **Batch Size = Per Device Batch Size * Gradient Accumulation * Number of GPU**
 - Per Device Batch Size
 - Gradient Accumulation
 - Number of GPU

2. Learning Rate

Learning Rate ↑ Batch Size ↑

What to adjust when training LLM

3. Sequence Length

Seq Length ↑ Memory Usage ↑

4. Lora Parameter

5. Quantization

6. Multi GPU Config/Experiment

Hands-on Fine-tuning