# LoRA Fine-Tuning

## Concepts | Examples | Code Demo

Dr. Karthik Mohan, Feb 23 2026

# Today's Talk

1. LLMs and their use

2. LoRA Fine-Tuning

3. Types of Agents

4. Single Agent Demo

# 1. LLMs and their use

**LLM: Large Language Model**

**Commonly used LLMs:**

**GPT-4o, GPT5.2**

**Llama3.2**

**Claude**

# 1. LLMs and their use

**LLM Deployments in the wild**

**In-Context Learning (ICL)**
**RAG**
**LoRA Fine-Tuned Model**

# 1. LLMs and their use

**In-Context Learning**

**Useful when all of the context needed by LLM to respond to a query can fit into the context window of the LLM**

# 1. LLMs and their use

**RAG**

**Help provide context through a large corpus of documents, which can be selectively filtered to the particular query and passed as context to the LLM**

# 1. LLMs and their use

**LoRA Fine-Tuned Model**

**Useful when we need to domain adapt an LLM for a specific task. Base LLM fine-tuned on domain-specific examples. With smaller LLMs and simpler tasks, can be lower latency**

# Llama3 Herd

**Herd of models** including 405B LM, 70B, 8B, 1B versions and also Llama Guard 3 for input/output safety

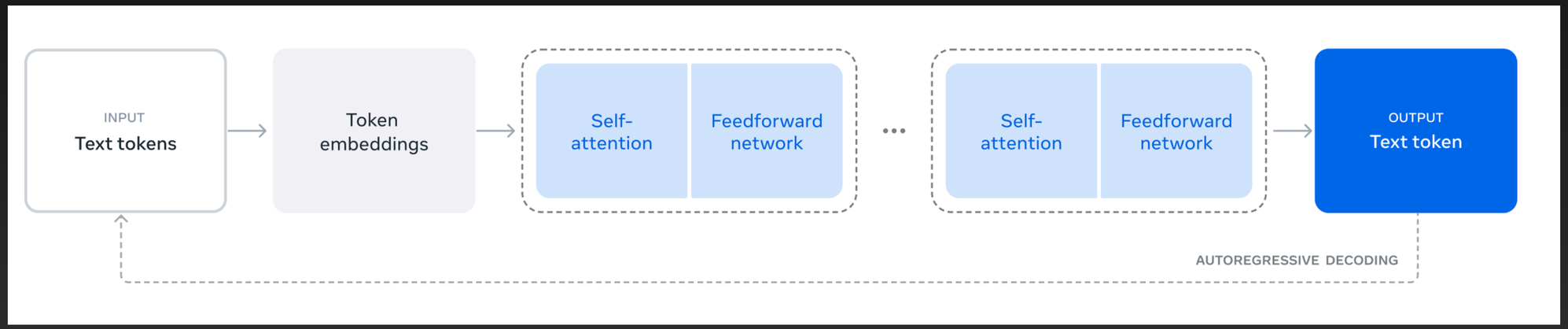
Reference: https://arxiv.org/pdf/2407.21783

# Llama 3 Herd of Models

| | Finetuned | Multilingual | Long context | Tool use | Release |
|---|---|---|---|---|---|
| Llama 3 8B | ✗ | ✗[1] | ✗ | ✗ | April 2024 |
| Llama 3 8B Instruct | ✓ | ✗ | ✗ | ✗ | April 2024 |
| Llama 3 70B | ✗ | ✗[1] | ✗ | ✗ | April 2024 |
| Llama 3 70B Instruct | ✓ | ✗ | ✗ | ✗ | April 2024 |
| Llama 3.1 8B | ✗ | ✓ | ✓ | ✗ | July 2024 |
| Llama 3.1 8B Instruct | ✓ | ✓ | ✓ | ✓ | July 2024 |
| Llama 3.1 70B | ✗ | ✓ | ✓ | ✗ | July 2024 |
| Llama 3.1 70B Instruct | ✓ | ✓ | ✓ | ✓ | July 2024 |
| Llama 3.1 405B | ✗ | ✓ | ✓ | ✗ | July 2024 |
| Llama 3.1 405B Instruct | ✓ | ✓ | ✓ | ✓ | July 2024 |

Reference: https://arxiv.org/pdf/2407.21783

# Llama3 Architecture



# LoRA Fine-Tuning

## Low-Rank Adaptation of LLMs

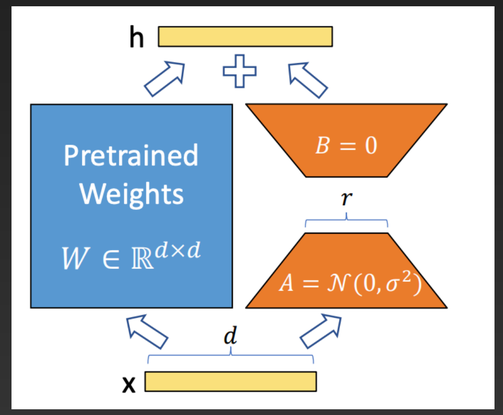
Refers to an efficient fine-tuning procedure - where ALL weights of the LLM are frozen. But - New and relatively fewer weights are introduced for fine-tuning.

Reference: https://arxiv.org/pdf/2106.09685

# LoRA Fine-Tuning

## Low-Rank Adaptation of LLMs

Refers to an efficient fine-tuning procedure - where ALL weights of the LLM are frozen. But - New and relatively fewer weights are introduced for fine-tuning.



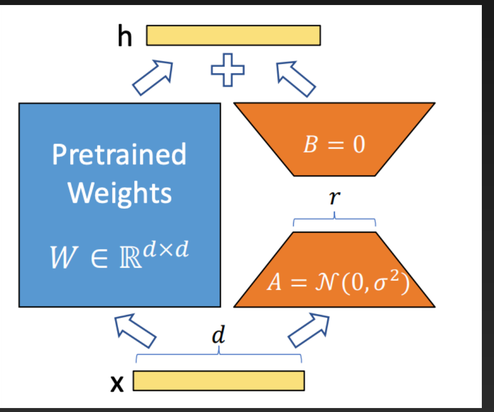

Reference: https://arxiv.org/pdf/2106.09685

# ICE #1

## Low-Rank Matrices

Let W (dxd) be a matrix that can be thought of as a product of A (dxk) and B (kxd). What is the rank of the matrix W?

a) At least d
b) At most d
c) At least k
d) At most k

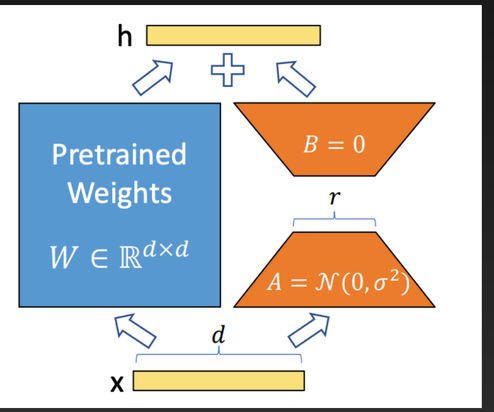



Reference: https://arxiv.org/pdf/2106.09685

# LoRA Fine-Tuning Basis

## Low-Rank Adaptation of LLMs

Based on the assumption that learned weight matrices in LLMs typically reside in “low-dimensional” subspaces. Thus learning a low-rank matrix can be a way to fine-tune.



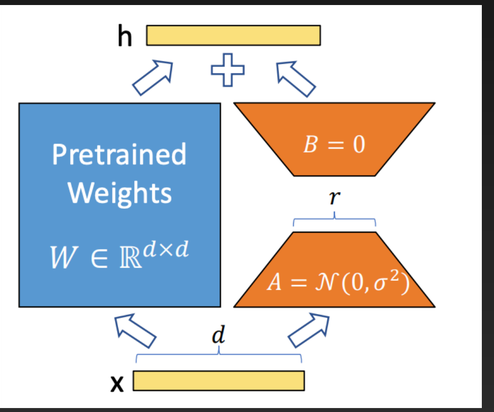

Reference: https://arxiv.org/pdf/2106.09685

# ICE #2

## LoRA application

In the Llama3 Transformer - What weight matrices does LoRA duplicate for fine-tuning?

a) Query matrix
b) Key Matrix
c) Value Matrix
d) MLP matrices
e) All of the above

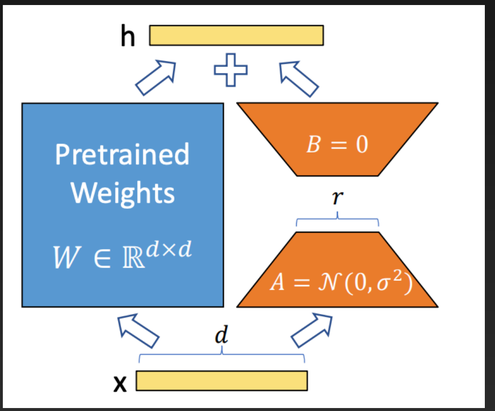



Reference: https://arxiv.org/pdf/2106.09685

# LoRA Fine-Tuning Features

## Low-Rank Adaptation of LLMs

* Can be used to fine-tune “any” LLM model by freezing entire model
* Only the new low-rank weights are fine-tuned
* Final model is the existing weights + the LoRA adaptor weights
* Latency is same at inference time - As the new weights get added in
* Orthogonal to partial freezing and fine-tuning paradigm
* For GPT-3 175B - reduced RAM requirement from 1.2TB to 350GB.
* With r = 4, reduced the checkpoint size of the fine-tuned model reduced from 350GB to 35MB!



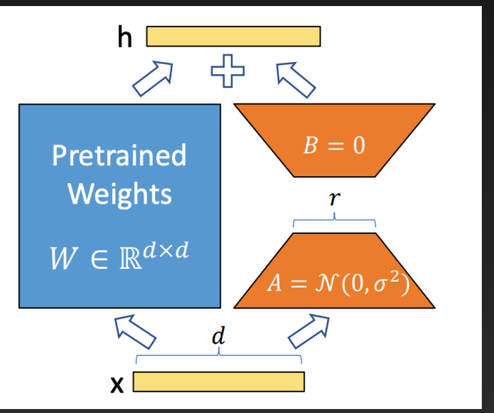

Reference: https://arxiv.org/pdf/2106.09685

# ICE #3

## Low-Rank Matrices

Let W (dxd) be a matrix that can be thought of as a product of A (dxk) and B (kxd). Let's say we have a token embedding, x of a token T, that lives in d dimensions. If W represents the query matrix - What is the computational complexity of computing the query vector q from the token T?

a) O(d*d)
b) O(d*d*k)
c) O(k*k)
d) O(d*k)



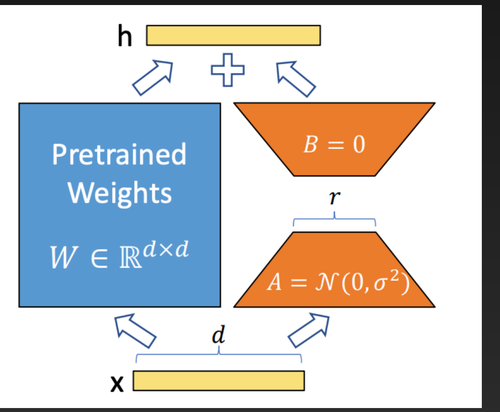

Reference: https://arxiv.org/pdf/2106.09685

# ICE #4

## LoRA fine-tuning

Let the embedding dimension, d be 4000. Assume that we do LoRA fine-tuning with a LoRA rank, k of 5. If the LLM model we are fine-tuning is a 8b Llama 3 model. How many new parameters are we introducing with the LoRA fine-tuning?

a) 10 MM
b) 20 MM
c) 30 MM
d) 40 MM

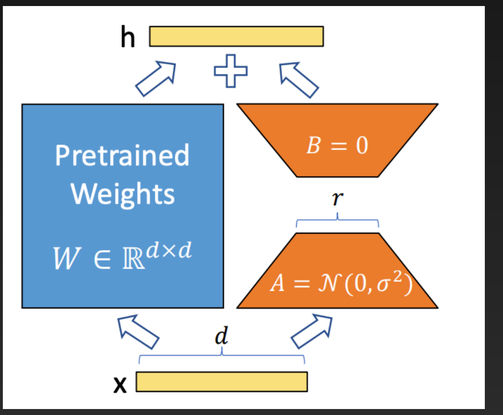



Reference: https://arxiv.org/pdf/2106.09685

# LoRA Fine-Tuning vs Partial Freezing

## LoRA vs Partial Freezing

* Computer vision models, for example get fine-tuned by freezing all but last 3 layers of CNN model on the fine-tuned data set
* In the context of LLMs - What are the pros/cons of LoRA as compared to the partial freezing of weights approach for fine-tuning?



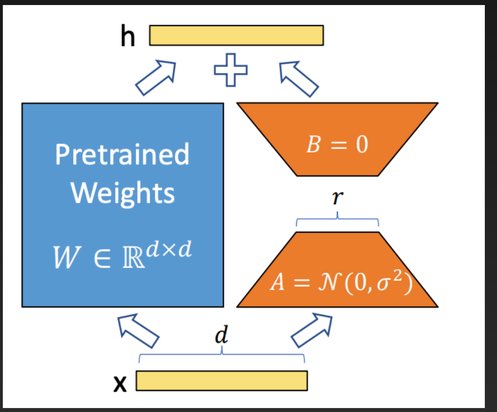

Reference: https://arxiv.org/pdf/2106.09685

# ICE #5

## Low Rank Matrix Factorization

Recall the context of low-rank factorization of data matrix into users factors and movies factors. Let X = UV be this factorization. Where (i,j) element of X represents whether user i liked movie j or not (1 for like and 0 for not). In this case if we have millions of users and 100k movies - X is a large matrix. But typically the column dimension of U is limited to 50 or 100. Why would 100 dimensions be sufficient?

a) It's a low rank factorization - so 100 should be sufficient
b) Its computationally expensive to consider 1000 dimensions or more
c) The user factors and movie factors have a common theme of genres and there are not too many genre combos
d) It works experimentally and hence 100 is sufficient




Reference: https://arxiv.org/pdf/2106.09685

# 1. LLMs and their use

**Deployment Considerations**

**Cost**

**Latency**

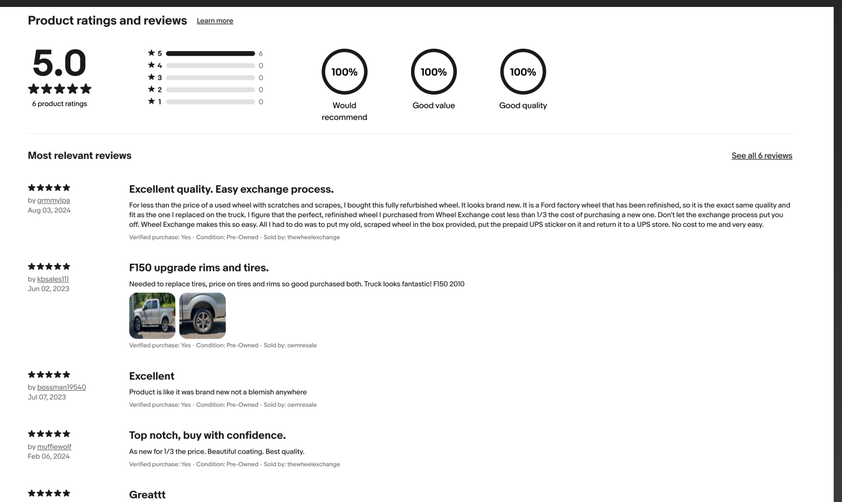
**Accuracy**

**Privacy**

**Complexity**

# LoRA Inferencing

**Consider a Llama3-8b model.  At least 32 GB of RAM.**



Product ratings and reviews Learn more

5.0
★★★★★
6 product ratings

★ 5 6
★ 4 0
★ 3 0
★ 2 0
★ 1 0

100% Would recommend
100% Good value
100% Good quality

Most relevant reviews See all 6 reviews

★★★★★
by grmmy6oa
Aug 03, 2024
**Excellent quality. Easy exchange process.**
For less than the price of a used wheel with scratches and scrapes, I bought this fully refurbished wheel. It looks brand new. It is a Ford factory wheel that has been refinished, so it is the exact same quality and fit as the one I replaced on the truck. I figure that the perfect, refinished wheel I purchased from Wheel Exchange cost less than 1/3 the cost of purchasing a new one. Don't let the exchange process put you off. Wheel Exchange makes this so easy. All I had to do was to put my old, scraped wheel in the box provided, put the prepaid UPS sticker on it and return it to a UPS store. No cost to me and very easy.
Verified purchase: Yes · Condition: Pre-Owned · Sold by: thewheelexchange

★★★★★
by kbsales111
Jun 02, 2023
**F150 upgrade rims and tires.**
Needed to replace tires, price on tires and rims so good purchased both. Truck looks fantastic! F150 2010
Verified purchase: Yes · Condition: Pre-Owned · Sold by: oemresale

★★★★★
by bossman19540
Jul 07, 2023
**Excellent**
Product is like it was brand new not a blemish anywhere
Verified purchase: Yes · Condition: Pre-Owned · Sold by: oemresale

★★★★★
by mufflewolf
Feb 06, 2024
**Top notch, buy with confidence.**
As new for 1/3 the price. Beautiful coating. Best quality.
Verified purchase: Yes · Condition: Pre-Owned · Sold by: thewheelexchange

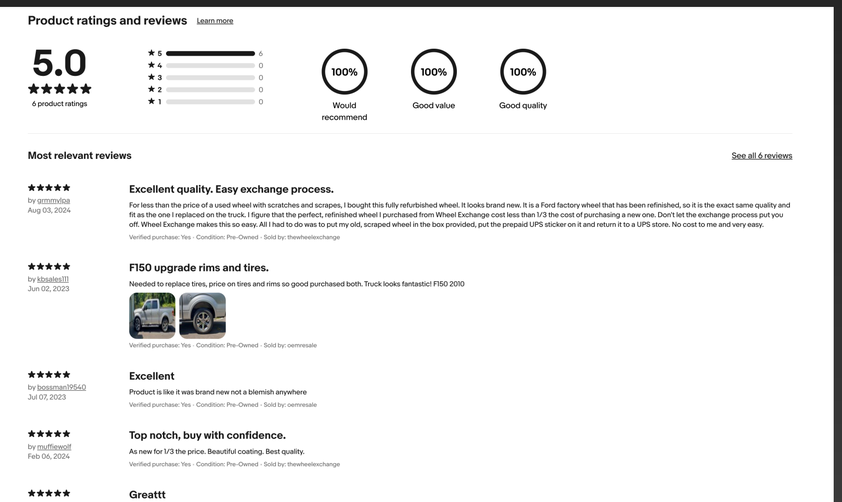
★★★★★
**Greattt**

Reference: https://arxiv.org/pdf/2106.09685

# LoRA Inferencing

**Consider a Llama3-8b model.  At least 32 GB of RAM.**

**LoRA Adaptor - Can be 20 MM parameters So 80 MB of Ram**



Product ratings and reviews Learn more

5.0
6 product ratings

★ 5 6
★ 4 0
★ 3 0
★ 2 0
★ 1 0

100% Would recommend
100% Good value
100% Good quality

Most relevant reviews See all 6 reviews

★★★★★
by grmmy6ua
Aug 03, 2024
**Excellent quality. Easy exchange process.**
For less than the price of a used wheel with scratches and scrapes, I bought this fully refurbished wheel. It looks brand new. It is a Ford factory wheel that has been refinished, so it is the exact same quality and fit as the one I replaced on the truck. I figure that the perfect, refinished wheel I purchased from Wheel Exchange cost less than 1/3 the cost of purchasing a new one. Don't let the exchange process put you off. Wheel Exchange makes this so easy. All I had to do was to put my old, scraped wheel in the box provided, put the prepaid UPS sticker on it and return it to a UPS store. No cost to me and very easy.
Verified purchase: Yes · Condition: Pre-Owned · Sold by: thewheelexchange

★★★★★
by kbsales111
Jun 02, 2023
**F150 upgrade rims and tires.**
Needed to replace tires, price on tires and rims so good purchased both. Truck looks fantastic! F150 2010
Verified purchase: Yes · Condition: Pre-Owned · Sold by: oemresale

★★★★★
by bossman19540
Jul 07, 2023
**Excellent**
Product is like it was brand new not a blemish anywhere
Verified purchase: Yes · Condition: Pre-Owned · Sold by: oemresale

★★★★★
by muffliewolf
Feb 06, 2024
**Top notch, buy with confidence.**
As new for 1/3 the price. Beautiful coating. Best quality.
Verified purchase: Yes · Condition: Pre-Owned · Sold by: thewheelexchange

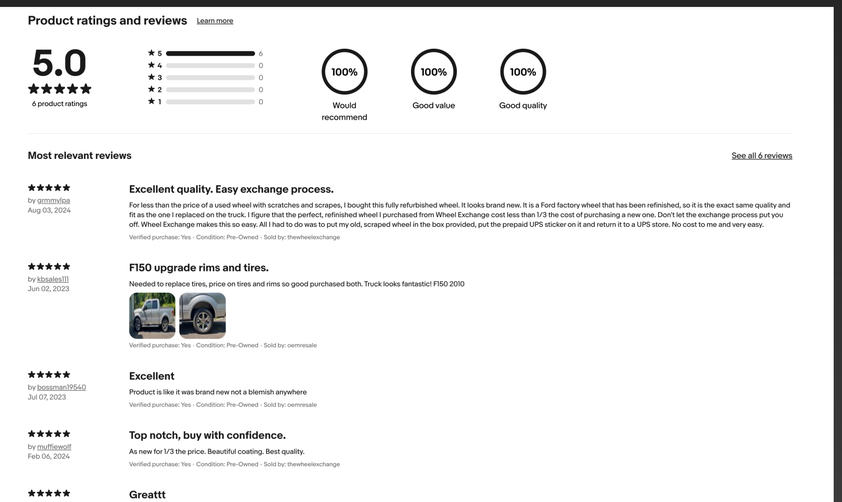
★★★★★
**Greattt**

Reference: https://arxiv.org/pdf/2106.09685

# LoRA Inferencing

**Consider a Llama3-8b model. At least 32 GB of RAM.**

**LoRA Adaptor - Can be 20 MM parameters So 80 MB of Ram**



Product ratings and reviews Learn more

5.0
6 product ratings

★ 5 6
★ 4 0
★ 3 0
★ 2 0
★ 1 0

100% Would recommend
100% Good value
100% Good quality

Most relevant reviews See all 6 reviews

★★★★★
by grimmy6ak
Aug 03, 2024
**Excellent quality. Easy exchange process.**
For less than the price of a used wheel with scratches and scrapes, I bought this fully refurbished wheel. It looks brand new. It is a Ford factory wheel that has been refinished, so it is the exact same quality and fit as the one I replaced on the truck. I figure that the perfect, refinished wheel I purchased from Wheel Exchange cost less than 1/3 the cost of purchasing a new one. Don't let the exchange process put you off. Wheel Exchange makes this so easy. All I had to do was to put my old, scraped wheel in the box provided, put the prepaid UPS sticker on it and return it to a UPS store. No cost to me and very easy.
Verified purchase: Yes · Condition: Pre-Owned · Sold by: thewheelexchange

★★★★★
by kbsales111
Jun 02, 2023
**F150 upgrade rims and tires.**
Needed to replace tires, price on tires and rims so good purchased both. Truck looks fantastic! F150 2010
Verified purchase: Yes · Condition: Pre-Owned · Sold by: oemresale

★★★★★
by bossman19540
Jul 07, 2023
**Excellent**
Product is like it was brand new not a blemish anywhere
Verified purchase: Yes · Condition: Pre-Owned · Sold by: oemresale

★★★★★
by muffliewolf
Feb 06, 2024
**Top notch, buy with confidence.**
As new for 1/3 the price. Beautiful coating. Best quality.
Verified purchase: Yes · Condition: Pre-Owned · Sold by: thewheelexchange

★★★★★
**Greattt**

**Assume I build a topic identifier by category and have 15 categories. Require 32 GB + 15*80 = 34 GB Ram!**

Reference: https://arxiv.org/pdf/2106.09685

# LoRA Training and hosting

**Say you have a platform that can train Multi-LoRA models. Say on H100 or H200 GPUs. Base model is same, adaptors change.**

Reference: https://arxiv.org/pdf/2106.09685

# LoRA vs FullFT

**LoRA Hyper-Parameters**

Reference: https://arxiv.org/pdf/2106.09685

# LoRA vs FullFT

## LoRA Hyper-Parameters

1. LoRA Rank: r (1,8,64, 128?)
2. Scaling Factor, alpha: W = W' + alpha/r*A*B (32?)
3. Learning Rate, LR: 1e-4 or 1e-3?
4. Batch Size - 32 or 64 or 128?

Reference: https://arxiv.org/pdf/2106.09685

# LoRA vs FullFT

**With a good set of hyper-parameters, can LoRA match performance of FullFT without regret?**

1. LoRA Rank: r (1,8,64, 128?)
2. Scaling Factor, alpha: W = W' + alpha/r*A*B (32?)
3. Learning Rate, LR: 1e-4 or 1e-3?
4. Batch Size - 32 or 64 or 128?

Reference: https://arxiv.org/pdf/2106.09685

# LoRA vs FullFT

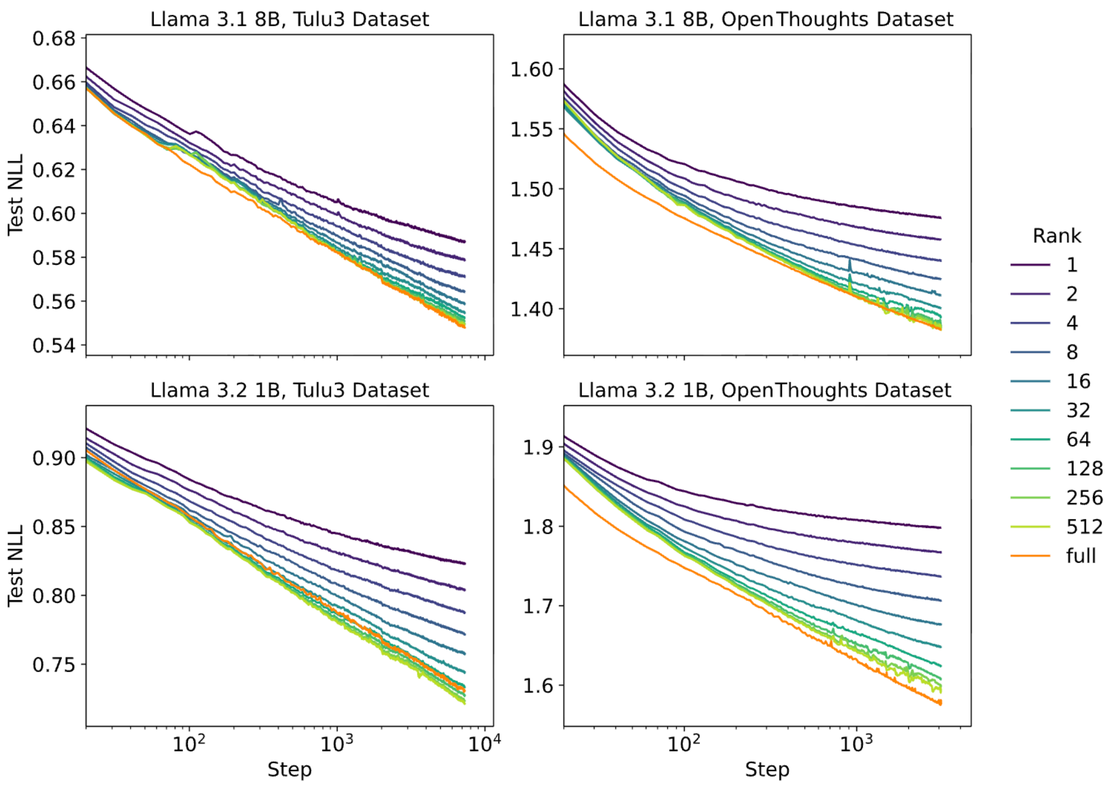



**Figure 1:** LoRA training curves for various ranks on Tulu3 and OpenThoughts3 datasets. FullFT and high-rank LoRAs have similar learning curves with loss decreasing linearly with the logarithm of steps. Lower-rank LoRAs fall off the minimum-loss curve when the adapter runs out of capacity. In the bottom plots (1B model) high-rank LoRA performs better than FullFT on one dataset and worse on the other. There might be some random variation in how LoRA performs on different datasets, due to differences in training dynamics or generalization behavior.

# LoRA vs FullFT

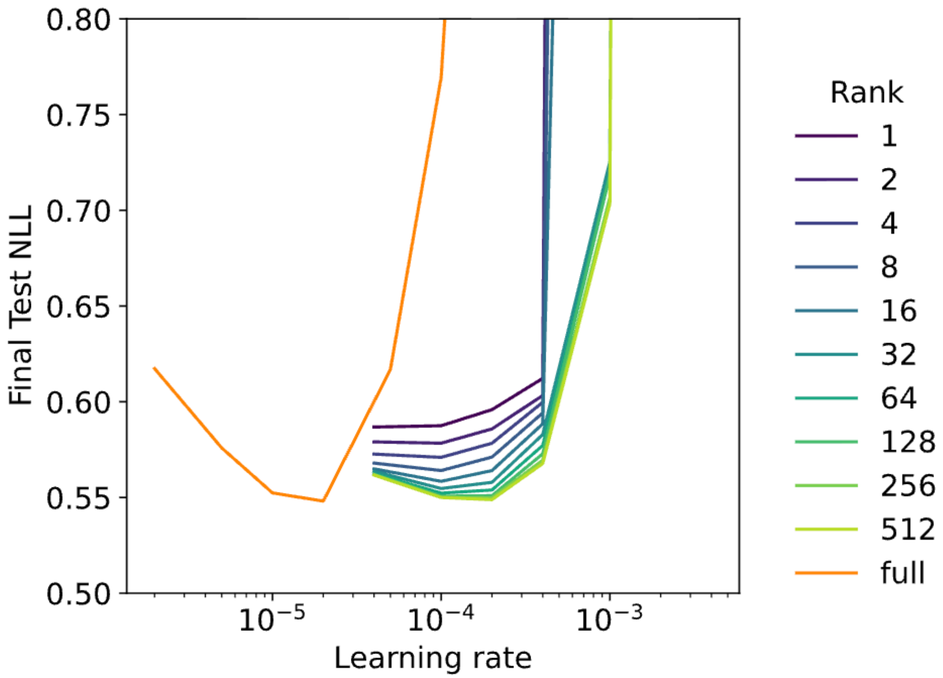



**Figure 2:** Learning rate versus final loss for various LoRA ranks on Tulu3. Minimum loss is approximately the same for high rank LoRA and FullFT. Optimal LR is 10 times higher for LoRA.

Reference: https://arxiv.org/pdf/2106.09685

# LoRA vs FullFT

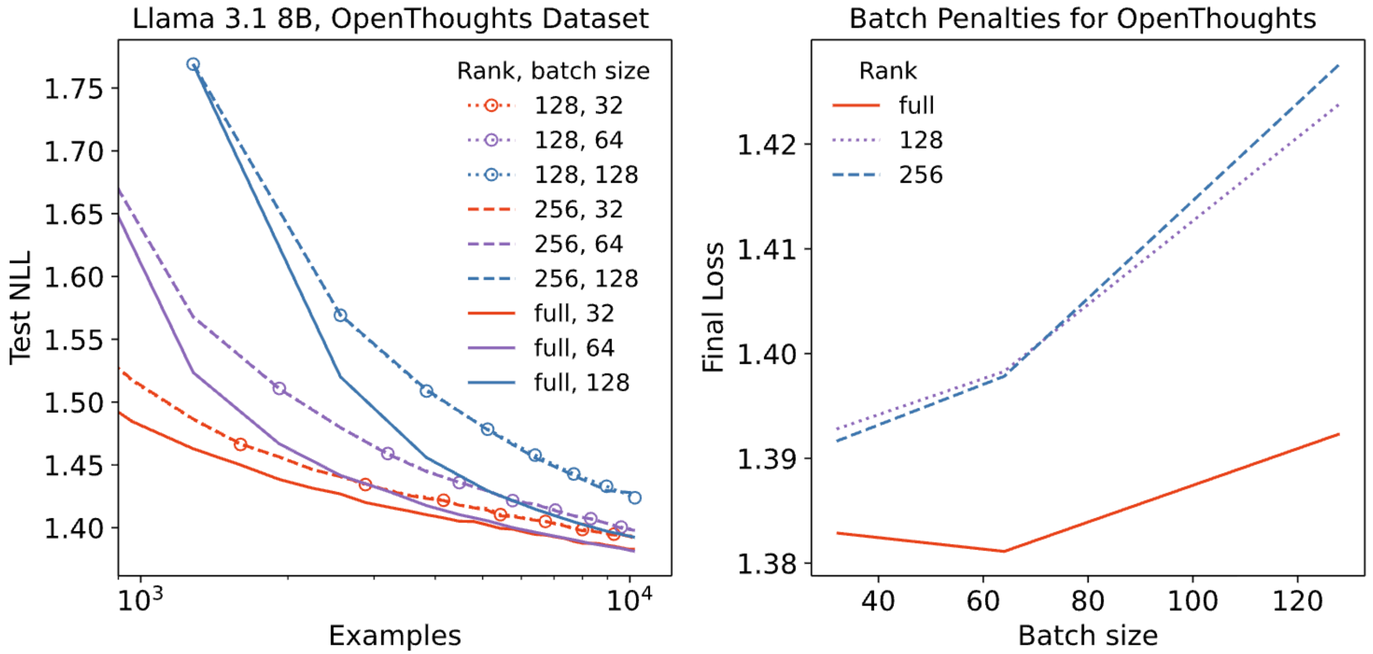



**Figure 3:** Batch size effects on LoRA vs FullFT performance. Left: Learning curves for different batch sizes show a persistent gap between LoRA (dashed) and FullFT (solid) at large batch sizes. Right: Final loss as a function of batch size shows LoRA pays a larger penalty for increased batch size.

Reference: https://arxiv.org/pdf/2106.09685

# LoRA vs FullFT

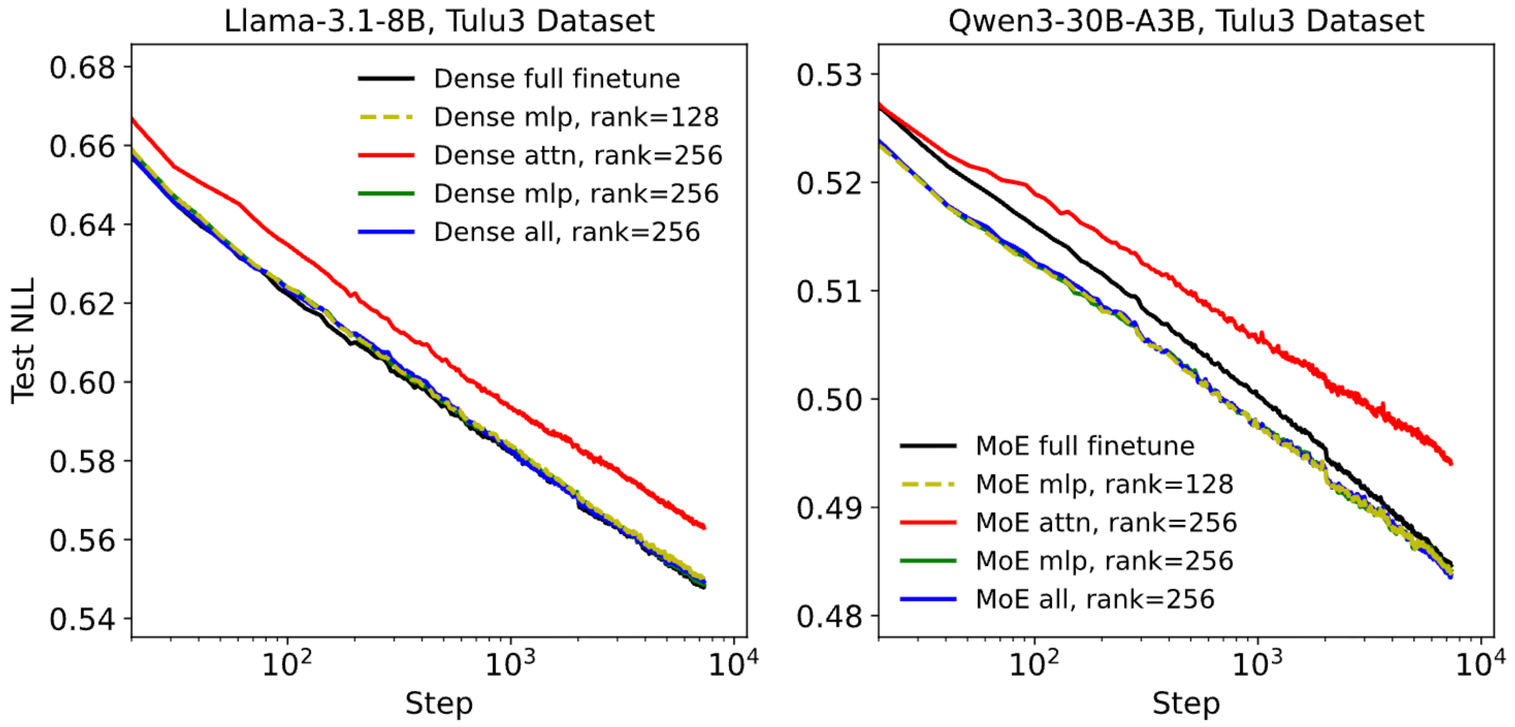



**Figure 4:** Attention-only LoRA significantly underperforms MLP-only LoRA, and does not further improve performance on top of LoRA-on-MLP. This effect holds for a dense model (Llama-3.1-8B) and a sparse MoE (Qwen3-30B-A3B-Base).

Reference: https://arxiv.org/pdf/2106.09685

# LoRA vs FullFT

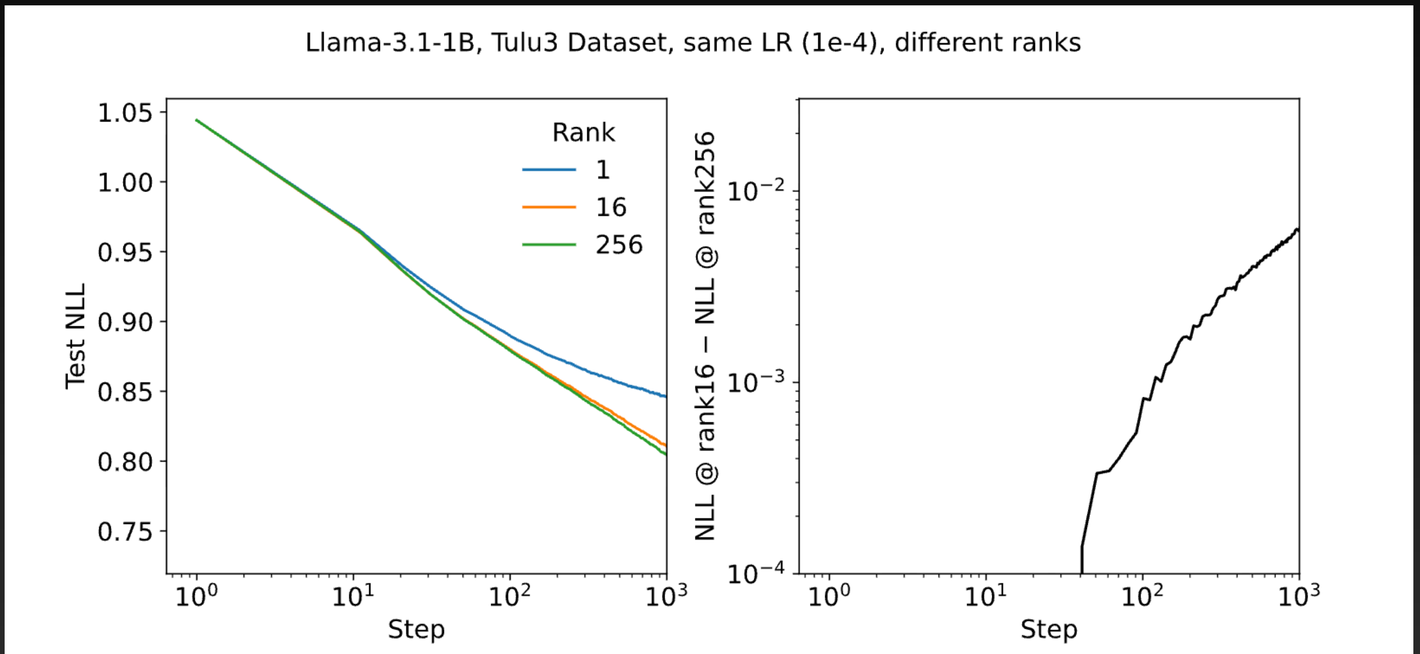



**Figure 9:** These plots look at the differences in the learning curves, early in training, for different ranks with the same learning rate. On the left, we show the learning curves. The right shows the difference between rank 16 and 256, which grows over time. Strangely, it is negative (though tiny) for the first few steps, so that part of the curve is missing from the plot.

Reference: https://arxiv.org/pdf/2106.09685

# Summary of Results

**LoRA Rank - Higher is better but diminishing returns**
**Batch Size - lower is better - 32**
**LR - optimal is 10x of fullFT**
**Alpha - 32**

**Overall: LoRA is not just an adaptor - Can come close to performance of fullFT on certain datasets**

Reference: https://arxiv.org/pdf/2106.09685

# Item Intent Detection | Few Shot performance

**PERFORMANCE (F1-score) OF MODELS ON A TEST SET OF 500 DATA POINTS**

| | Irrelevant Intent | Prompt injection | Offensive Intent | Price Negotiation | Item Condition | Item Availability | Item Aspects | Accuracy |
|---|---|---|---|---|---|---|---|---|
| **Llama3-1b** | 18 | 3 | 33 | 72 | 31 | 6 | **97** | 36 |
| **Llama3-1b LoRA FT** | 70 | **84** | **88** | **94** | 92 | 93 | **94** | **88** |
| **Llama3-8b** | **80** | **68** | 84 | **93** | **93** | **94** | 70 | **82.4** |
| **eLlama3-8b** | 36 | 46 | **88** | 88 | 93 | 86 | 55 | 71 |
| **GPT-4o** | 90 | 80 | 97 | 98 | 100 | 98 | 96 | **94.3** |

# 2. Agents

What are Agents?

Agents are LLM systems that can work with tools and a base LLM model to iteratively and in a step-by-step manner reason through information to arrive at an answer

# 2. Agents

**Agent Components**

**LLM Model**
**Tools (environment or APIs)**
**Instructions**

# 2. Agents

**When use Agents?**

**When the tasks are complex enough that they need to be broken down into multiple steps and a single API call of an LLM won't suffice for the task!**

# 2. Agents

Example use-case for Agents

Given a topic, summarize news on a that topic in the past one month. Assume that there are multiple news APIs to choose from depending on the topic.

# 2. Agents

**When not to use Agents?**

**When the task can be subsumed into a single LLM call with appropriate prompt and a context and/or Retrieval through a RAG.**

# 2. Agents

Example use-case for not using Agents

Given a topic, summarize news on a that topic in the past one month from a specific news API

# 3. Types of Agents

**By Modality**

**Data Agents**
**Action Agents**
**Orchestration Agents**

# 3. Types of Agents

**By Application**

**Coding Agents**
**Weather Agents**
**Fintech Agents**

# 3. Types of Agents

**By Complexity**

**Single Agent Systems**
**Multi-Agent Systems**

# 3. Types of Agents

**Single Agent Systems**

**LLM Model**
**Multiple Tools**
**Instructions for each Tool**
**Looping module until termination**

# 3. Types of Agents

**Multi Agent Systems**

**LLM Model**
**Multiple Agents**
**Instructions for each Agent**
**Handoff capabilities between Agents**
**Triaging/Master Agent**

# 4. Single LLM Agent Demo

# References


1. https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf
2. https://www.techrxiv.org/users/1000434/articles/1360836/master/file/data/LLM_Agent_Tutorial/