

THE 23RD INTERNATIONAL SEMANTIC WEB CONFERENCE

November 11, 2024 – November 15, 2024

Live! Casino & Hotel Maryland

Slot 4: Fine-tuning

ISWC 2024

Part 1: Basics

Why and when do we fine-tune?

Basics

- Pre-trained LLMs
 - “*Stochastic parrots*”
 - Repeat knowledge
- Why?
 - Trained on million documents of data
 - Fixed weights
- Offer huge potential
 - Accelerate your work
 - KG creation, augmentation

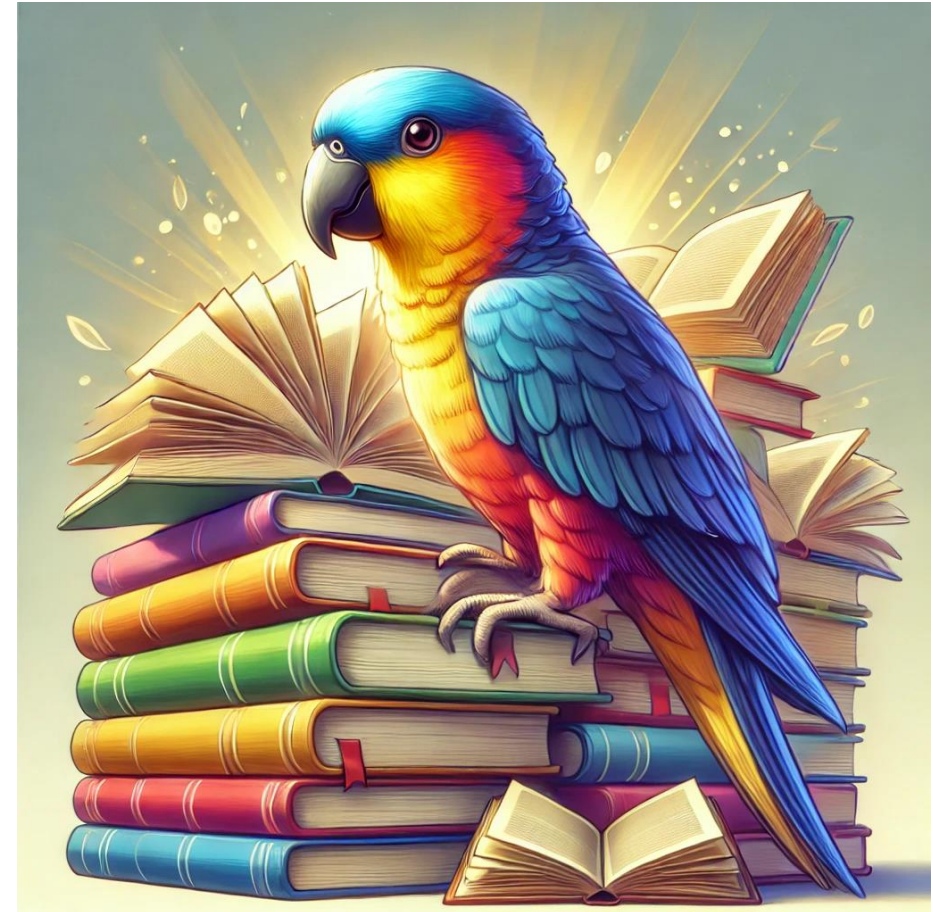


Image created with DALL-E using ChatGPT 4o

Basics

- Let the LLM
 - Understand complex queries / schemas
 - Write queries
 - Do tedious tasks
- KG application
 - Creation e.g. entity extraction / disambiguation
 - Augmentation
- Recall session 1

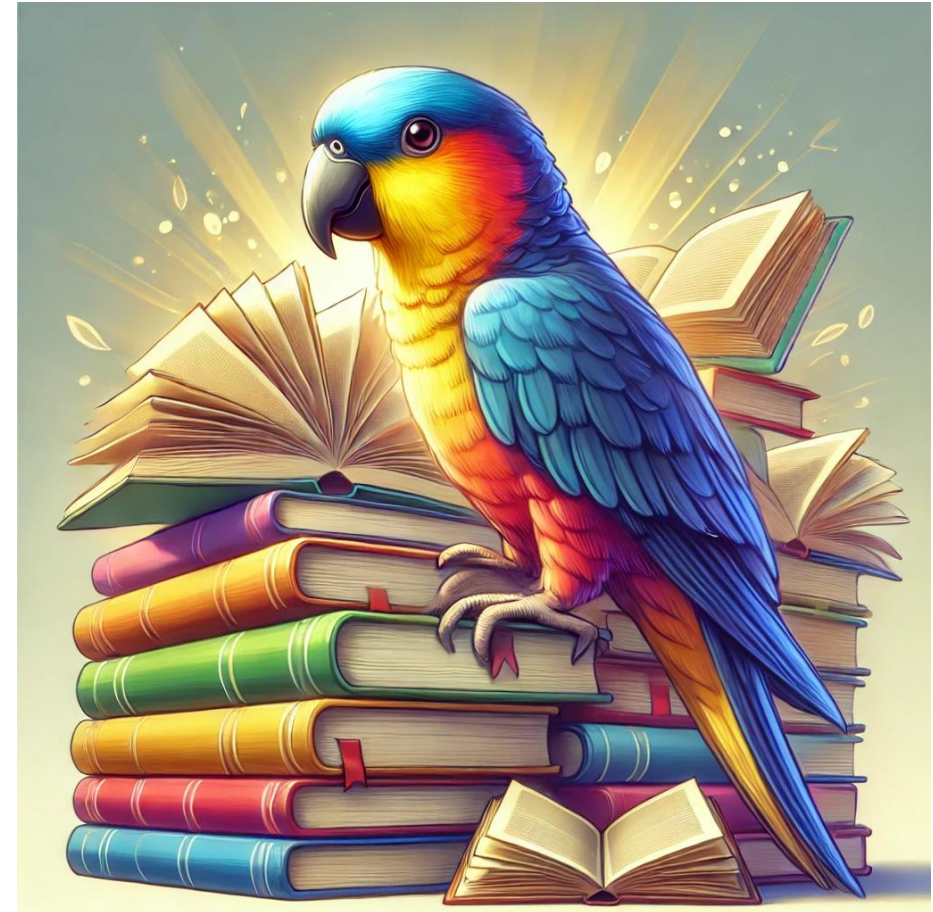


Image created with DALL-E using ChatGPT 4o

Showcase

Showcasing the result for the base model

The solution: fine-tuning

- Fit the LLM to **your needs**
 - Adapt to specific use case
- Let it learn **new knowledge**
- All you need to do: adapt weights



Image created with DALL-E using ChatGPT 4o

The solution: fine-tuning?

- Fine-tuning is **expensive**
- Alternatives (non-exhaustive)
 - RAG
 - Prompt-tuning
 - Knowledge injection (via prompt)
- Fine-tuning as a last resort



Image created with DALL-E using ChatGPT 4o

Part 2: How to fine-tune

Common steps

- Set goal
- Model selection
- Dataset selection
 - Data augmentation
- Fine-tuning
 - Iterative enhancements

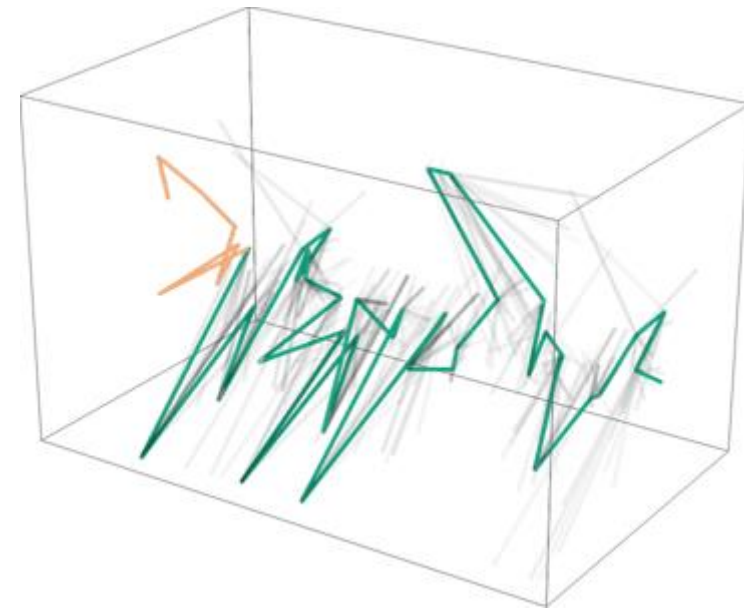


Image taken from
<https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>

LORA

- Fine-tuning costs TIME
- Fine-tuning costs RESOURCES
- Add extra weights
- Train these instead
- Reduction in parameters



Image based on

https://huggingface.co/docs/peft/developer_guides/lora

LORA

- Fine-tuning costs TIME
- Fine-tuning costs RESOURCES
- Add extra weights
- Train these instead
- Reduction in parameters

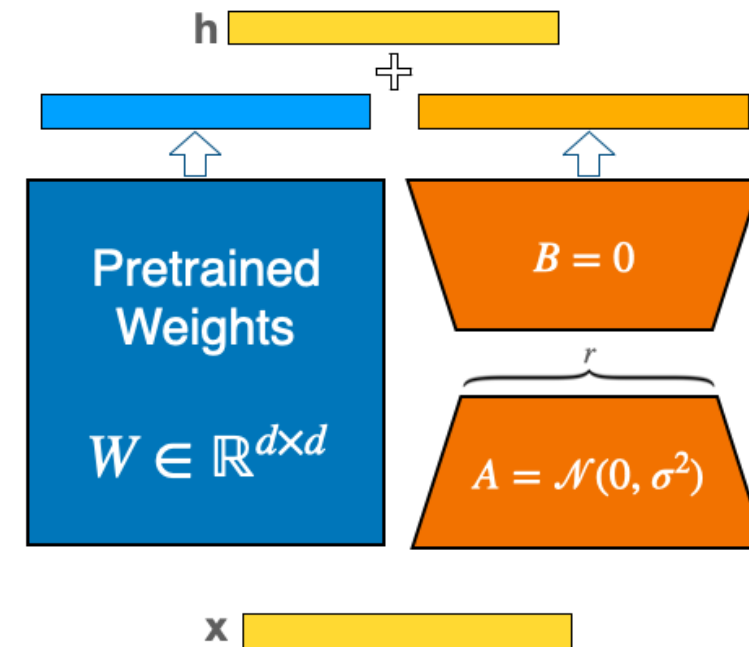


Image based on

https://huggingface.co/docs/peft/developer_guides/lora

LORA

- Fine-tuning costs TIME
- Fine-tuning costs RESOURCES
- Add extra weights
- Train these instead
- Reduction in parameters

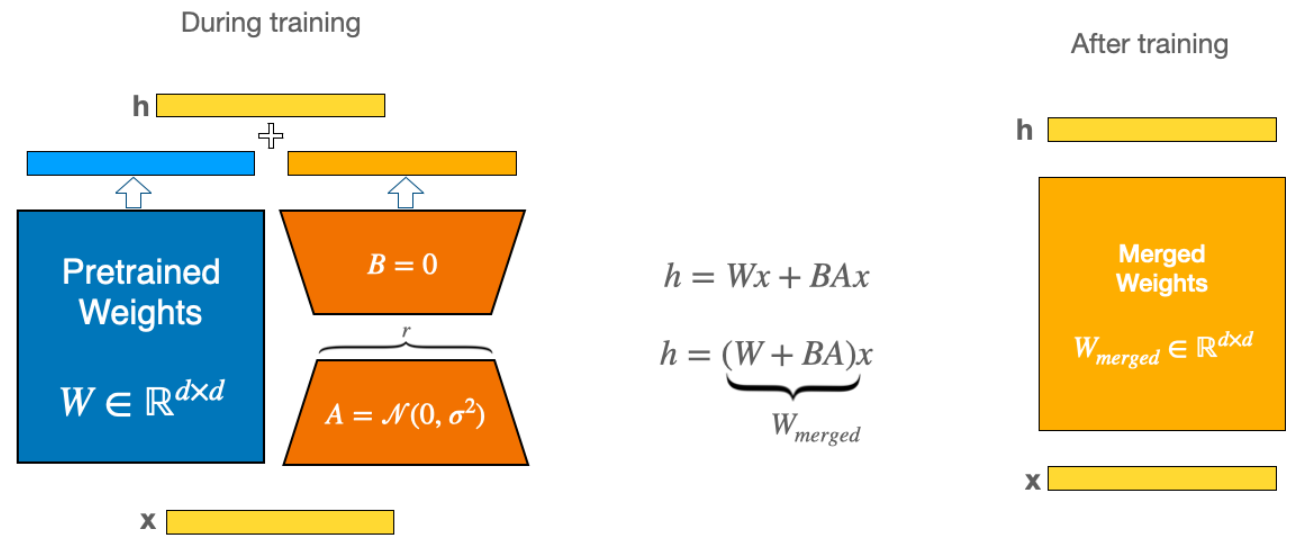


Image taken from

https://huggingface.co/docs/peft/developer_guides/lora

Showcase

Fine-tuning Llama3 on DBLP-QUAD using LORA

Take-aways

- Small steps
- This guide is non exhaustive
- Further enhancements
 - QLORA

Be sure fine-tuning is what you need