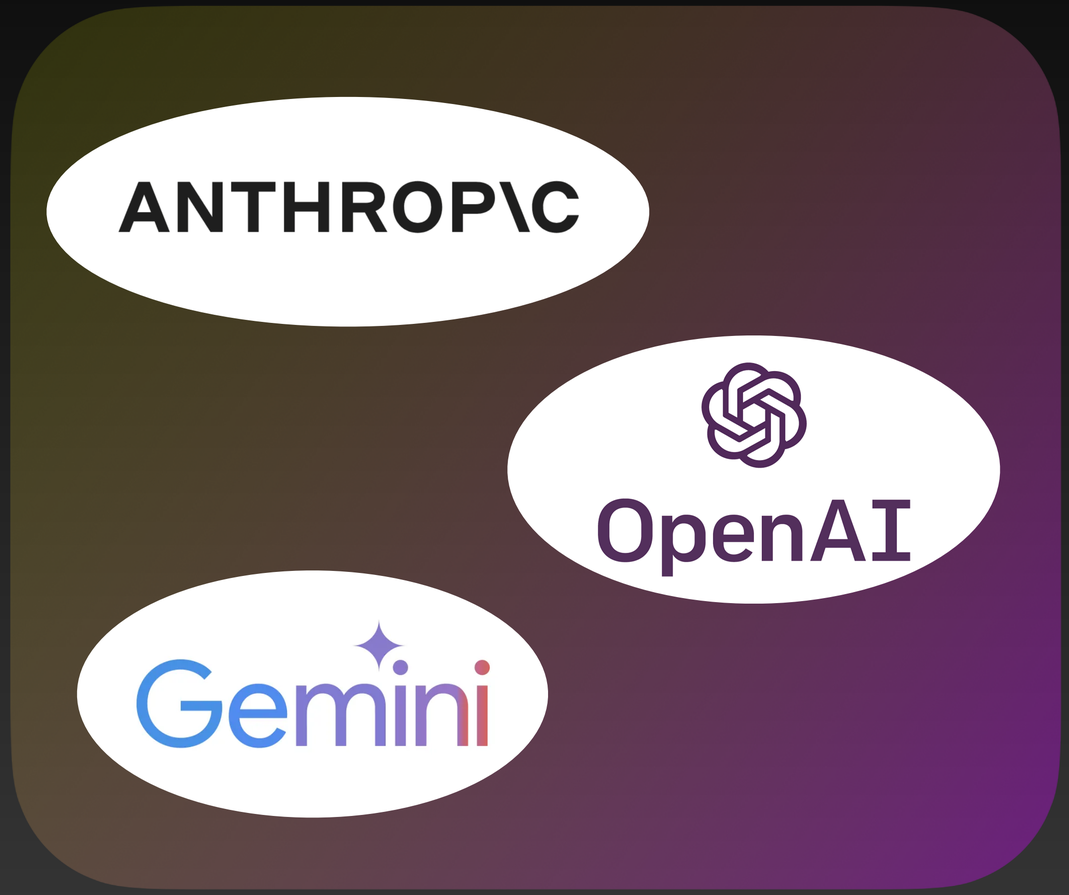
Fine-tuning LLMs on your private data!

Credits- Sri Ranganathan Palaniappan, Mansi Phute, Seongmin Lee, Polo Chau

Foundation Models & LLMS



Closed-Source LLMs

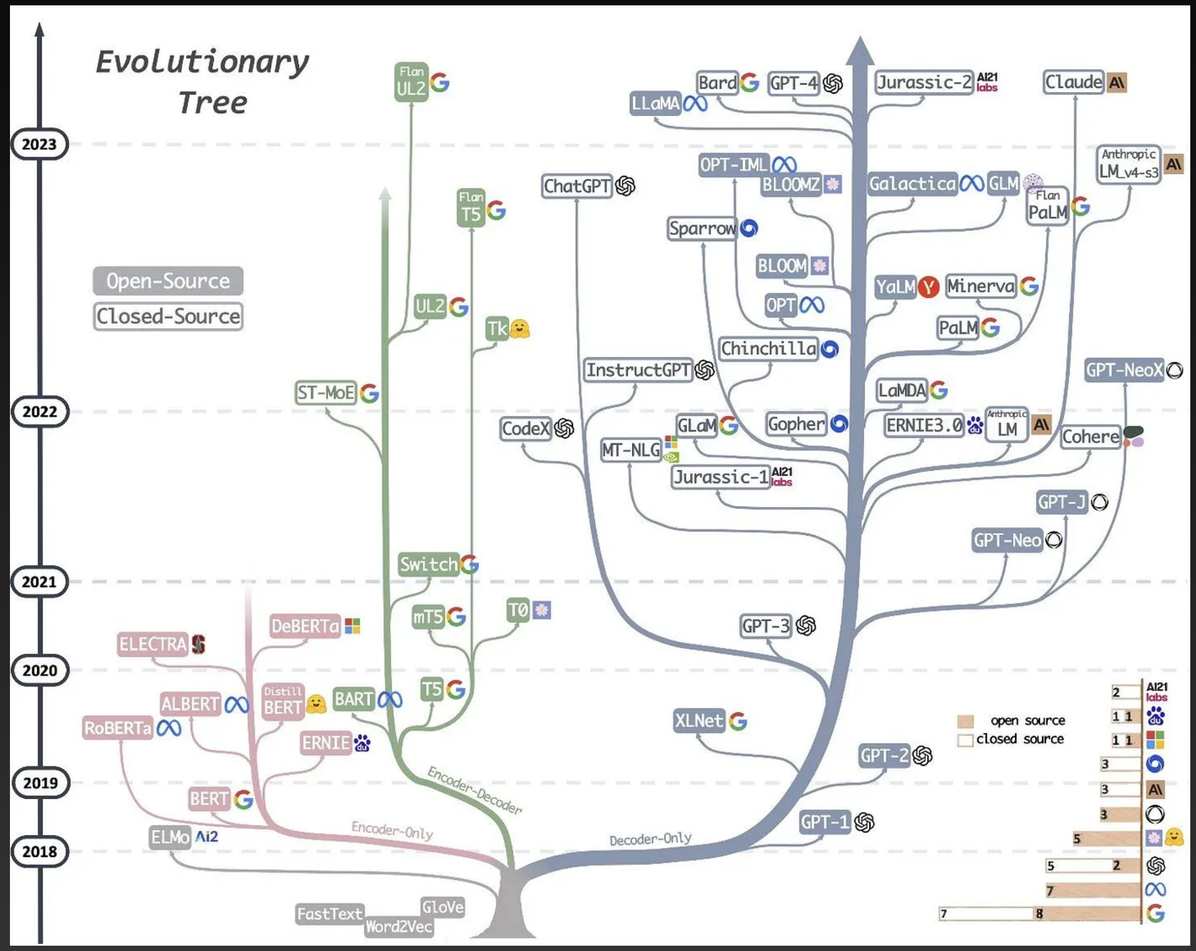
Open-Source LLMs

Rise of Open-sourced LLMs

Why Llama-2?

- Widely used Open-source LLM
- Trained on over 2 trillion tokens of publicly available data
- Llama-chat: optimized for dialogue use-cases, like chatbots
- Accessible yet powerful

Meta



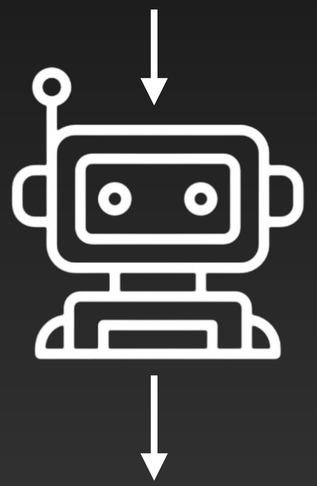
Evolution of Open-source LLMS over the years..

Problems with pre-trained models

Limited domain-specific knowledge

Lack of specific information

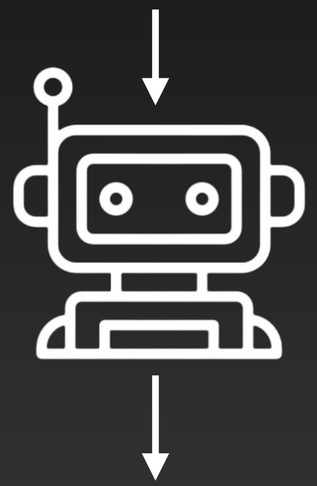
"When did recent Hawaii wildfires take place?"



"I'm sorry, my knowledge is only until June 2022..."

Not trained on private data

"Does my company's HR travel policy reimburse food?"

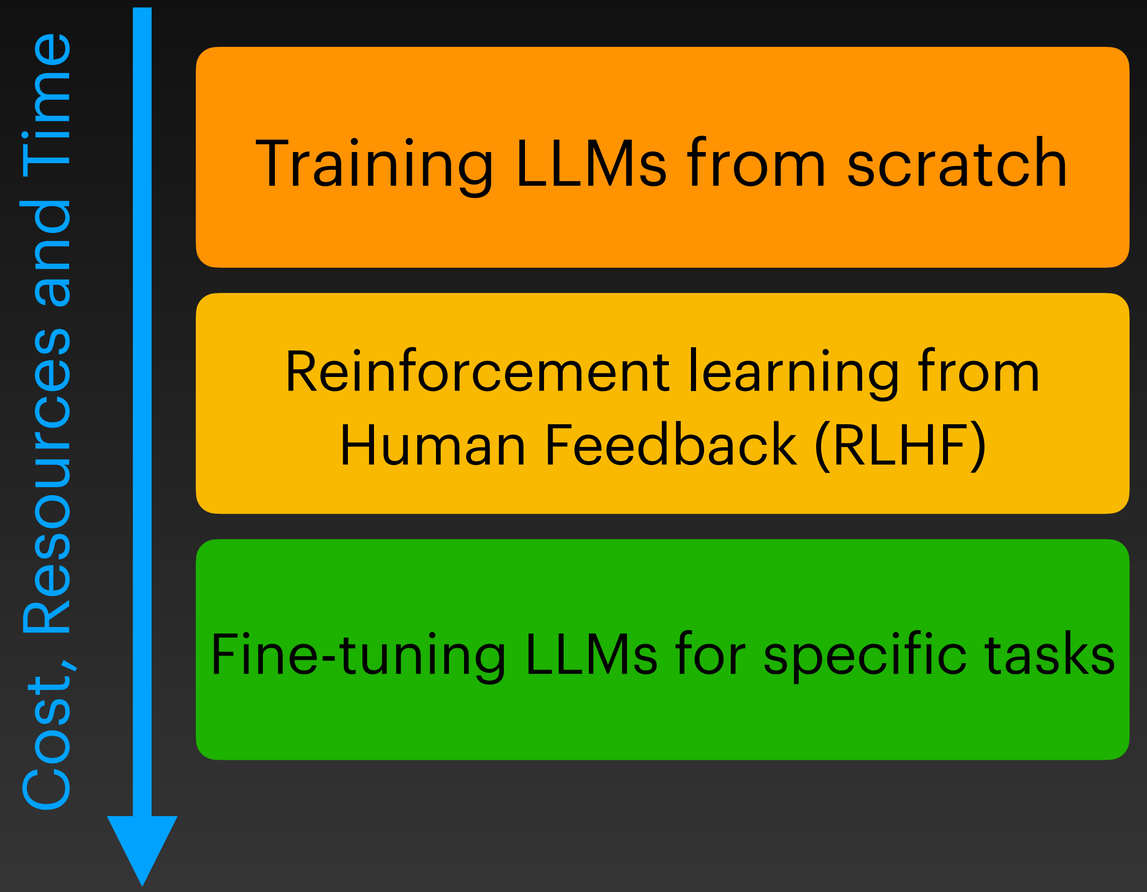


"I'm sorry, I don't have access to your company's internal policies..."

Solution: Fine-tune LLMs

Why Fine-tune?

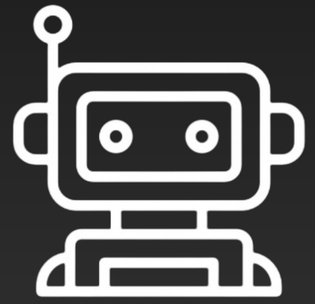
- Requires lower memory & compute power; possible with smaller training data
- Leverages pre-existing knowledge: adopt existing knowledge to specific task
- Enable customization: tailor pre-trained LLM on private data, specific task, etc.



Onboarding Strategies for LLMs

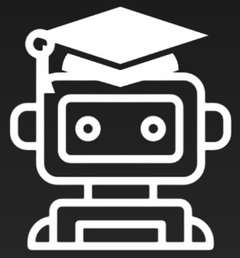
New employee with general knowledge

Train employees on internal data, documents & information



Pre-trained LLMs

Fine-tune LLMs on private, internal data



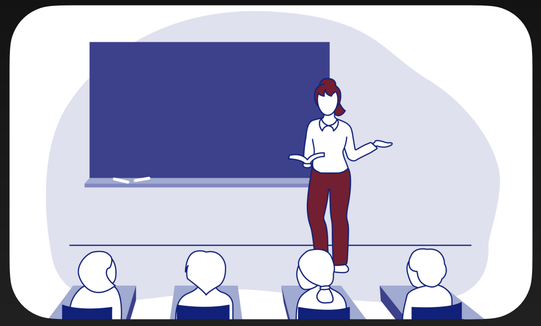
Productive contributions to organization

How’s this tutorial helpful?

Fine-tune on freeform text articles & documents

- No need to manually format data (CSV, JSON, etc.)
- Train on any text extracted from documents, PDFs, and more without any additional work!

Step-by-step walkthrough + demo on real world use case



- No prior technical knowledge required: just run all the cells sequentially to get a fine-tuned model!
- Demo based on a realistic use case: fine-tuning on 2023 Hawaii wildfire report sourced from PDFs

Memory efficient with in-depth explanations

Google Colaboratory

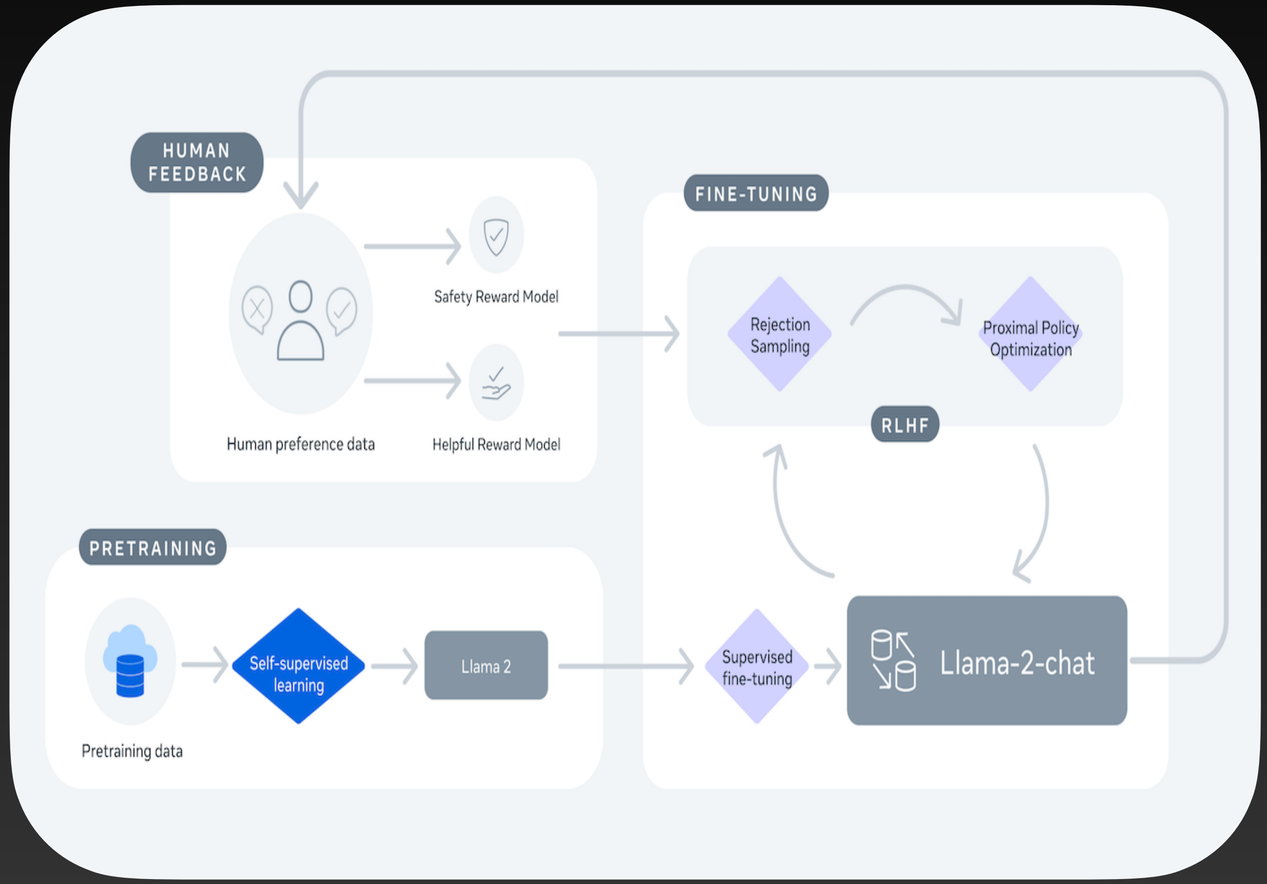
NVIDIA

- Entire fine-tuning process requires less than 14GB!
- Can run on Google Colab (for FREE!) or locally on compatible GPUs
- In-depth explanations to optimize performance for your specific use case

Diving deep into Fine-tuning

Step 1: Choosing the Right Model

- Pre-trained model for fine-tuning: Llama-2-7B-chat
- Refined using RLHF optimized for conversations
- Not expensive or resource-intensive to fine-tune
- Suitable for on-device deployment
- Open source & accessible



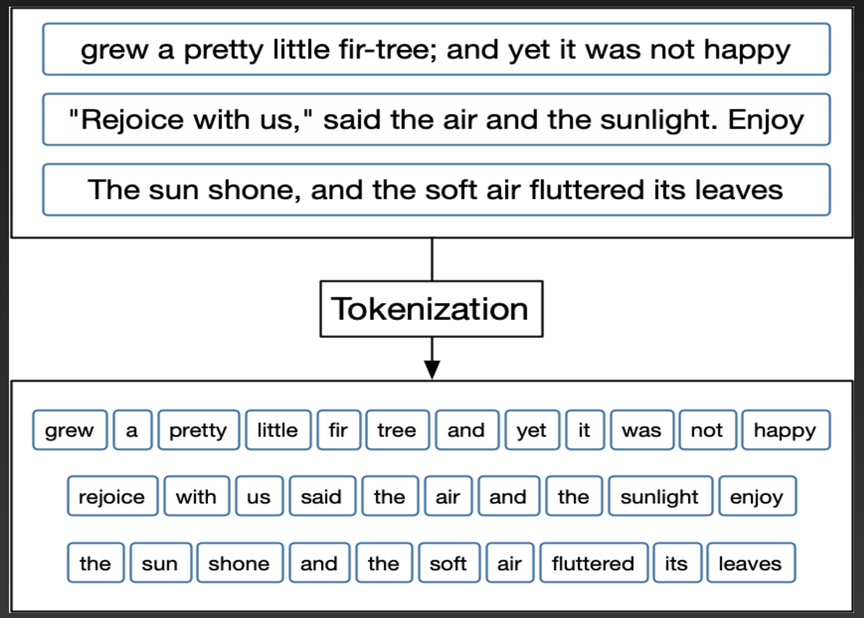
Process of training Llama-2 Chat

Step 2: Prepare Data for Training

- Extract text data from documents, PDFs and more, into raw text (.txt) files
- Use LlamaTokenizer to tokenize text data

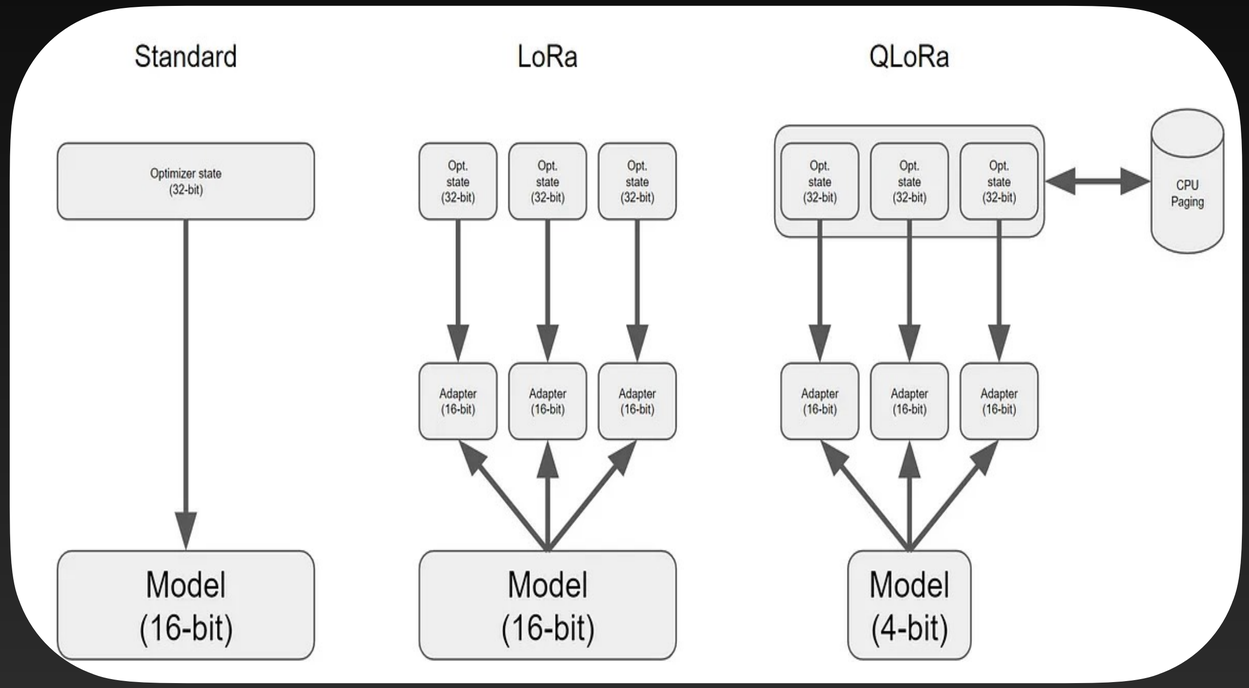
What's tokenization?

- Tokenization: breaking down text into smaller units (tokens) - words, subwords or characters. This enables the model to understand and process text data.



Step 3: 4-bit Quantization using QLoRA

- Significantly reduce memory footprint using 4-bit quantization.
- Non-trainable parameters (<99% of model parameters): loaded in 4-bit
- Trainable parameters (>1% of model parameters): loaded in 16-bit
- Enables fine-tuning on Google Colab (free version!) and other resource-constrained devices
- Maintains model performance while using less memory



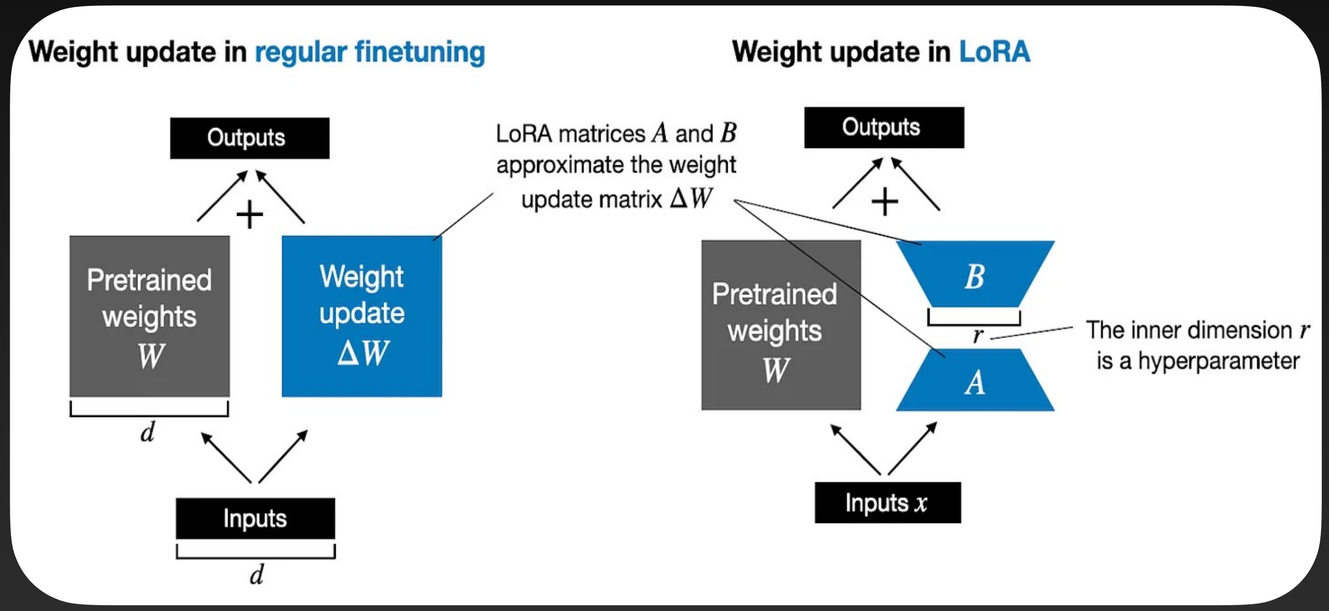
Memory requirements: ~~112 GB~~ **14GB**!

Google Colaboratory

NVIDIA

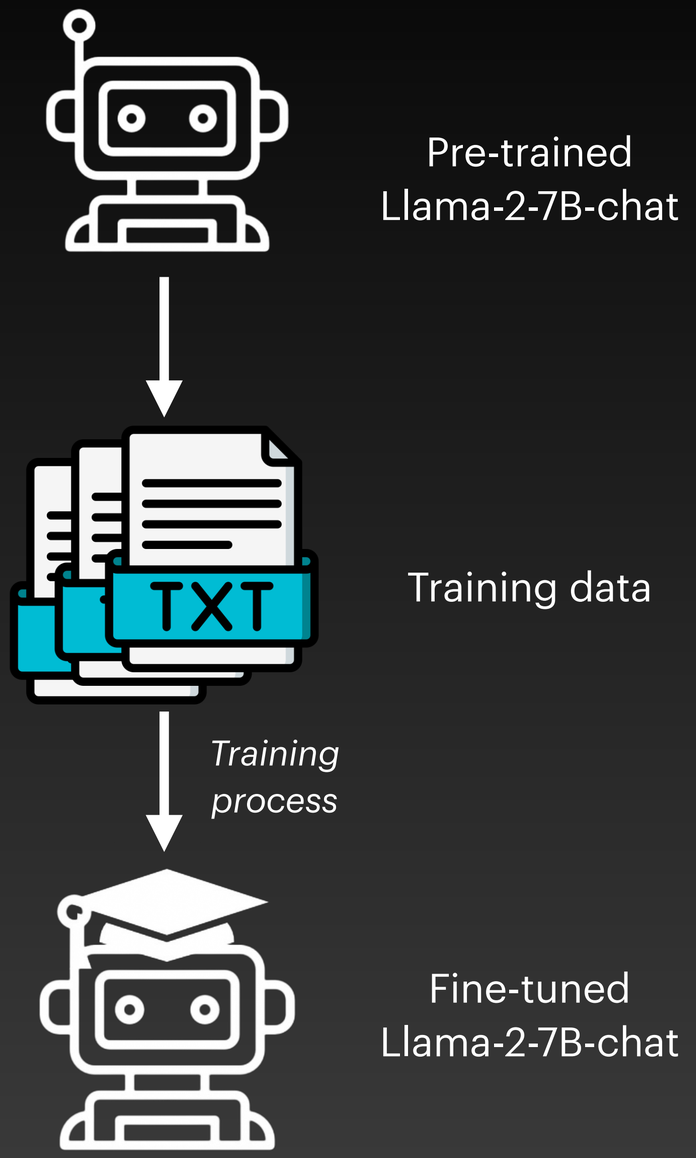
Step 4: Reducing Trainable Parameters with LoRA

- Make fine-tuning more memory and time efficient
- LoRA (Low Rank Adaptation): reduce number of parameters to be trained
- Preserves pre-trained weights and applies changes to separate set of weights
- Speeds up training process & lowers cost



Step 5: Train the Model

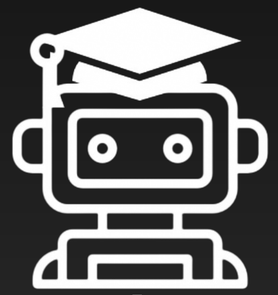
- Exploring some important training arguments:
  - Learning rate: controls the step size during model updates, i.e., degree of weights modified
  - Batch size: number of samples processed before updating the model; larger batch sizes lead to more accurate weight updates at the cost of increased memory usage
  - Epochs: number of complete passes through the training data



Step 6: Your Personal ChatBot!

- You now have your fine-tuned Llama-2 model trained on your private data!
- Chatbot can answer any questions and provide information based on your data personalized and tailored to your needs and use case

“When did the Hawaii wildfires occur in 2023?”



Your custom fine-tuned model

“The Hawaii wildfires incident occurred from August 8th to August 11th, 2023.”