Fine Tuning LLMs

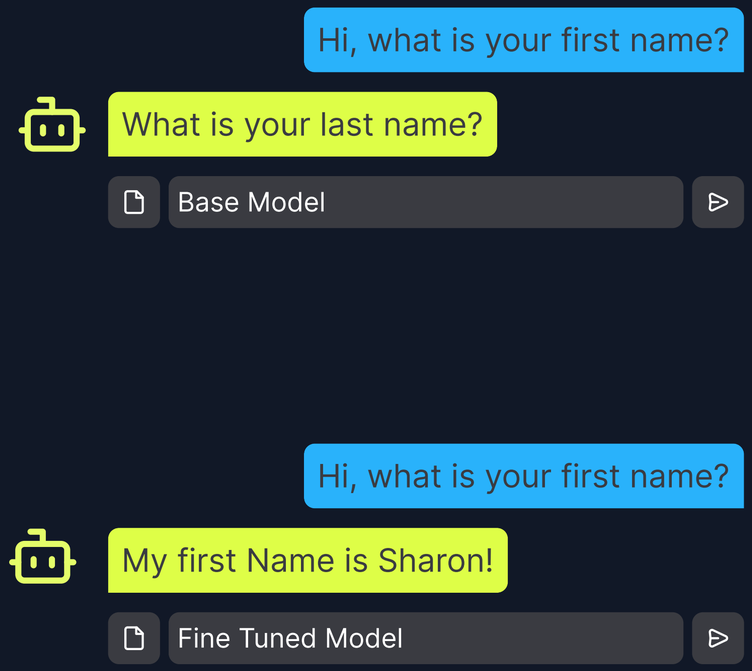
Why Finetune LLMs

More Consistent Outputs

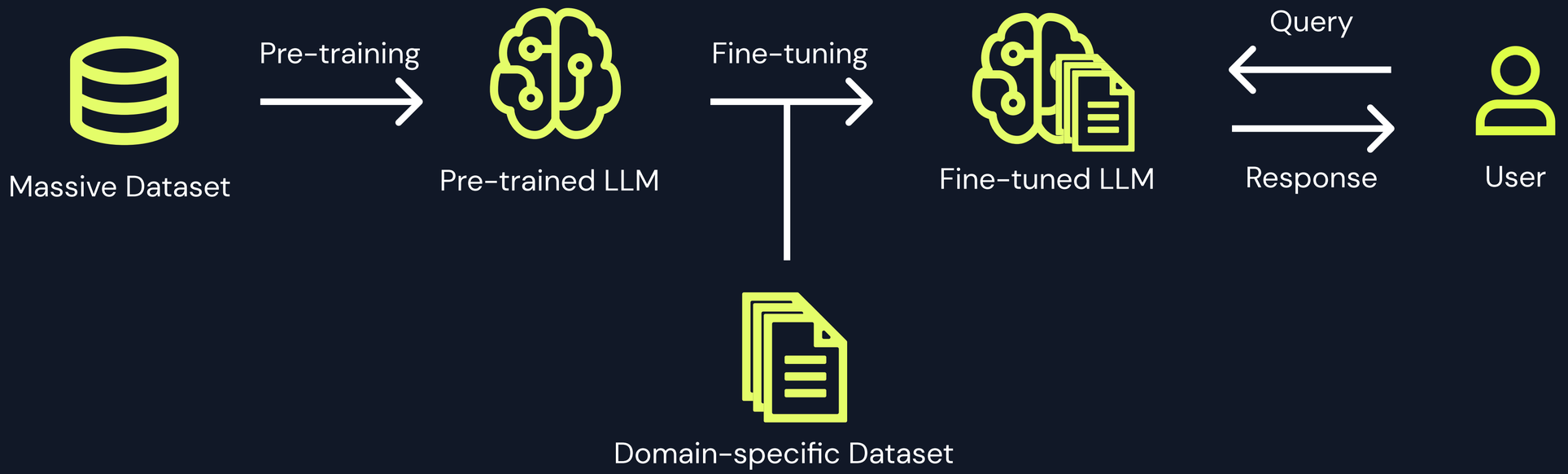
Customize Models to a Use Case

Reduce Hallucination

Eliminate Need to Train Models from Scratch



Fine Tuning Architechture



Fine Tuning Methods

Self-Supervised Learning

Supervised Learning

Reinforcement Learning with Human Feedback

Dataset Creation and Processing

- → Each LLM requires a specific format for the training data
- → High quality data essential for better performance
- → Tokenize the formatted data converting text to numbers
- → Eliminates need of training a model from scratch

Methods for Parameter Fine Tuning

Full Fine Tuning

Adjusts all parameters of the LLM using task-specific data.

Computationally expensive

Transfer Learning

Freeze all parameters except for the head of the neural network

Only finetune the layers that translate to the output layer

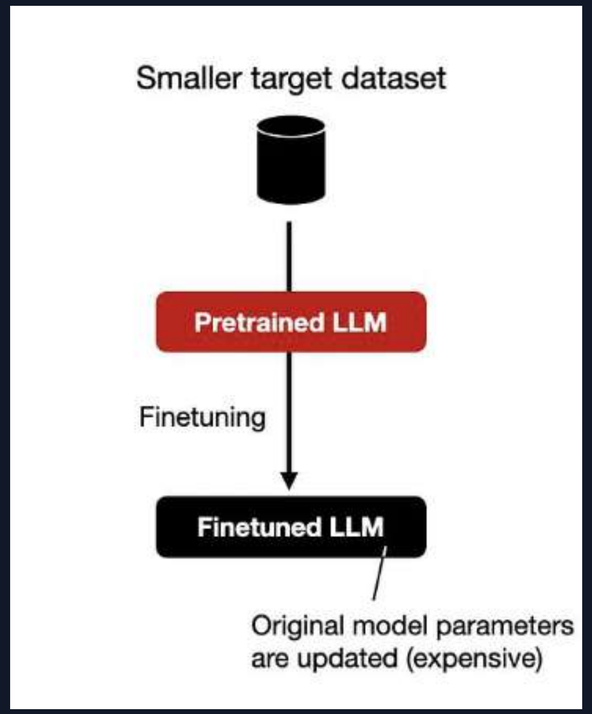
Parameter Efficient Fine Tuning

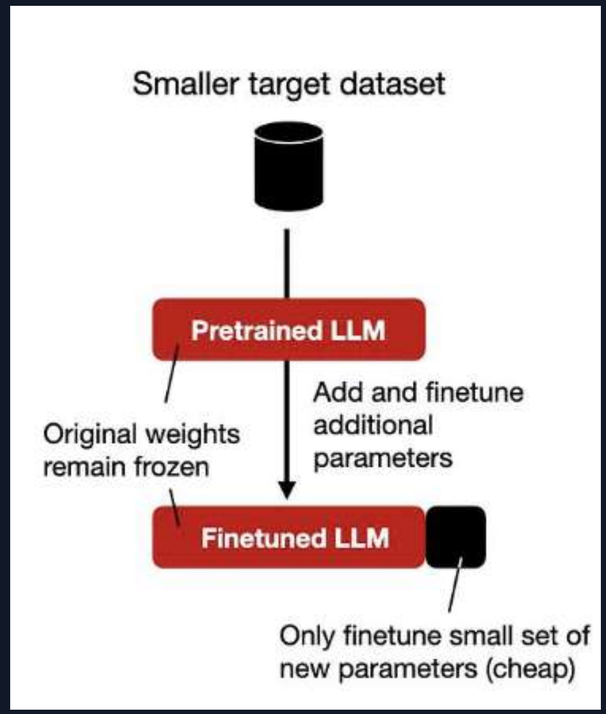
Freeze all the weights of the base LLM

Augment the model with additional parameters and finetune those

Less computationally expensive

Methods for Parameter Fine Tuning





Low Rank Adaptation (LoRA) for PEFT

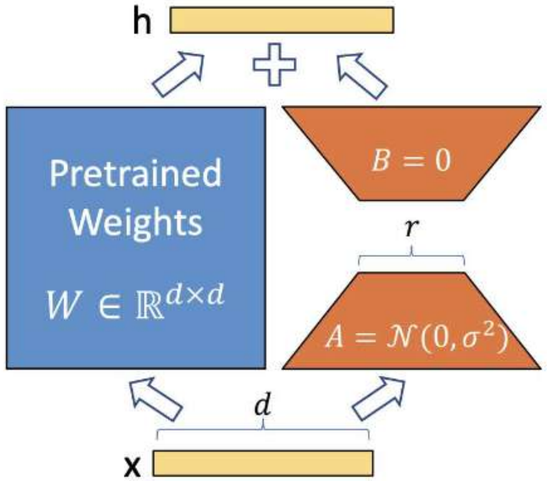
Reduces number of trainable parameters

Identifies crucial parameters for the task at hand and finetunes those

During fine-tuning, only the parameters in low-rank matrices are updated

Less chance of overfitting since only a few parameters are updated

Reduces computational and memory requirements needed to fine-tune



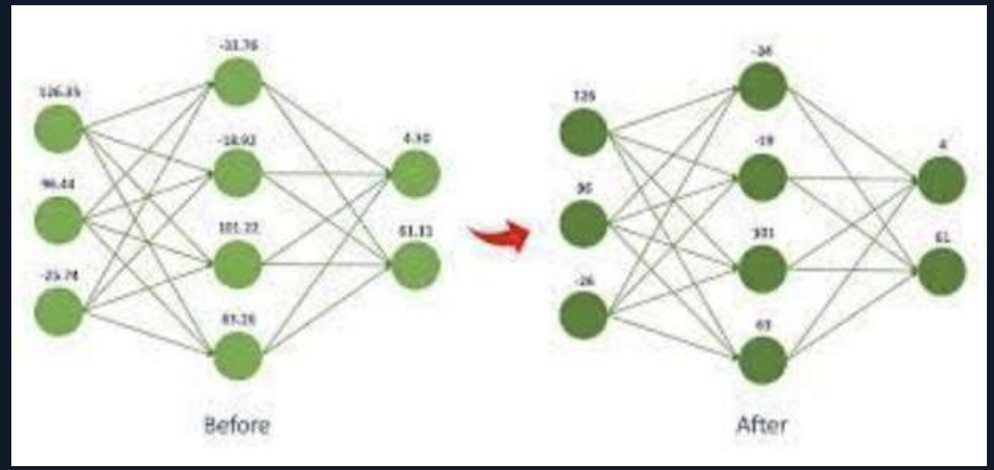
QLoRA for Quantized LLMS

Quantization: process of reducing the numerical precision of a model's tensors to make it faster and more compact

QLoRA combines quantization & low-rank adaptation

Model parameters are first quantized (usually to 4 bit precision) and then go through LoRA

Made fine tuning a lot more accessible



Fine Tuning Results on Finance-Bench

Question: What is the FY2018 capital expenditure amount (in USD millions) for 3M? Give a response to the question by relying on the details shown in the cash flow statement.

Correct Answer: $1577M

LLAMA-2 with RAG:

The FY2018 capital expenditure amount for 3M is not directly disclosed in the cash flow statement. Therefore, I don't know the answer to this question

Fine Tuned LLAMA-2:

Based on the information provided in the 2018 10-K filing, the FY2018 capital expenditure amount for 3M is $1577M.

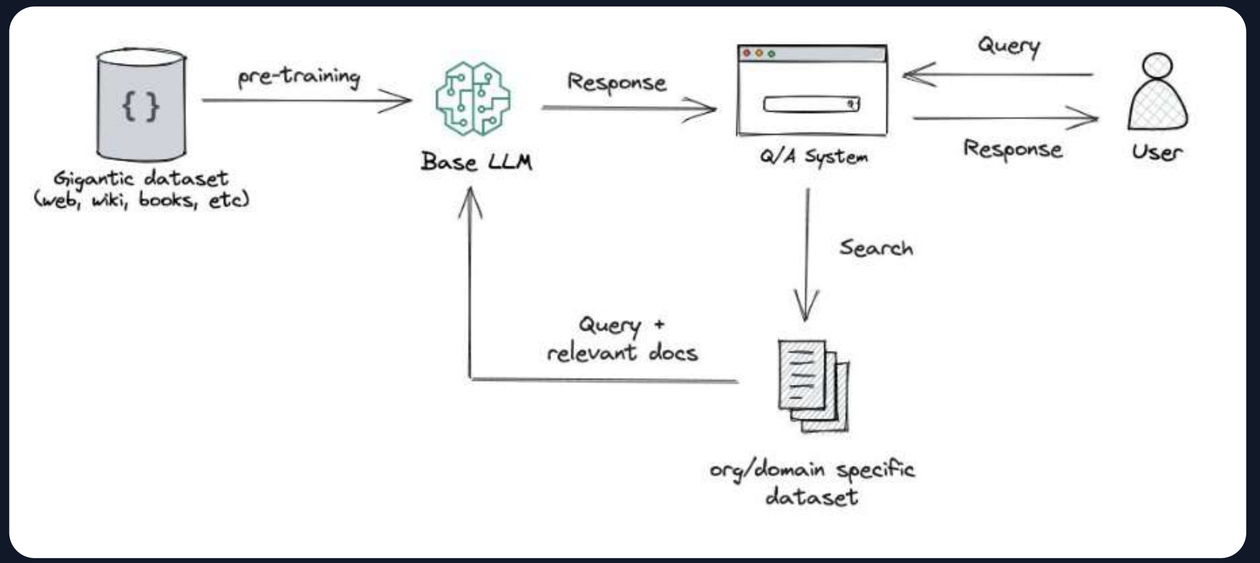
Retrieving Context for Q-A

RAG usually finds relevant context via vector search of the embeddings

Retrieval algorithm does not always find the place where the correct answer is

How to improve Retrieval?

- Query Expansion
- Prompt Engineering
- Embedding Adapters
- Chunking Parameters
- Metadata Filtering



Retrieving Context for Q-A

Prompt Engineering & RAG can be used in conjunction with Fine Tuning

Prompt engineering to optimize system prompts for a specific task

Include context as a part of the input data that is retrieved via RAG

Use prompt engineering & RAG on fine tuned model similar to base model

Combine techniques for highly sophisticated, accurate, and efficient AI systems