|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Scope** | **Select** | **Adapt & Align Model** | | **Application Integration** | |
| Define the Problem | Choose the model | Prompt Engineering | Evaluate | Optimize & Deploy model for inference | Augment model & build LLM powered application |
| Fine Tuning |
| Align with human feedback |

**Fine Tuning**

**Fine tune and LLM with Instruction Prompts:**

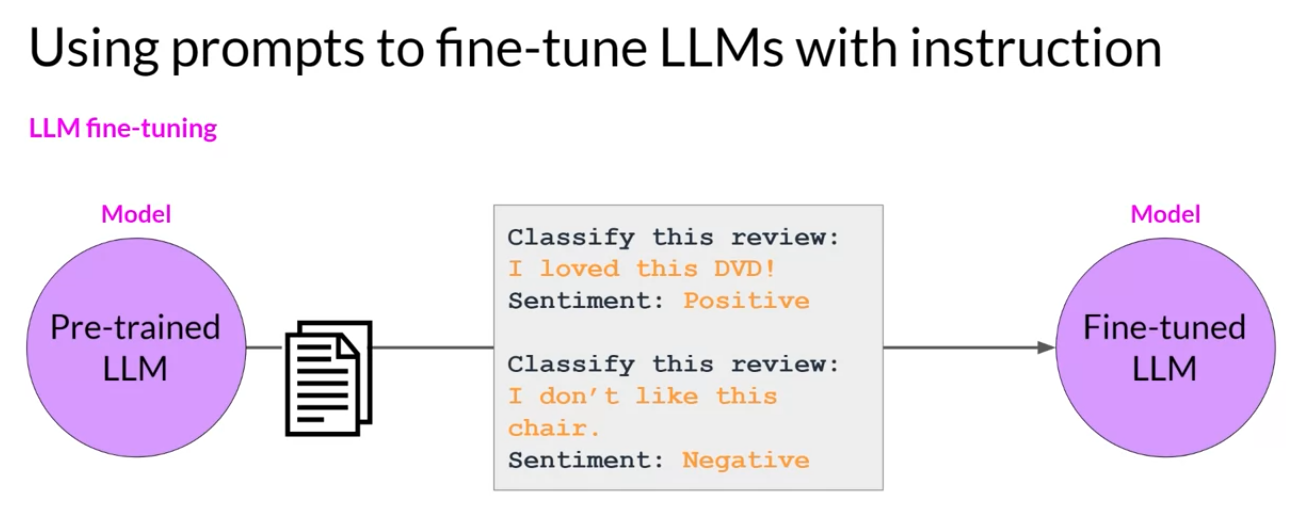
There are limitations of in context learning (one shot, few shot) because:

* For smaller models it doesn’t work even with five or six examples are included.
* Examples take up space in context window (where prompt is written).
* Instead try fine tuning the model.

In pre training where we train the model using unstructured textual data via self-supervised learning.

Fine tuning is supervised learning process where you use a dataset of labeled examples to update the weights of the LLM. The labeled examples are prompt, completion pair.

**Instruction fine tuning:** Trains the model using example that demonstrates how it should respond to a specific instruction.

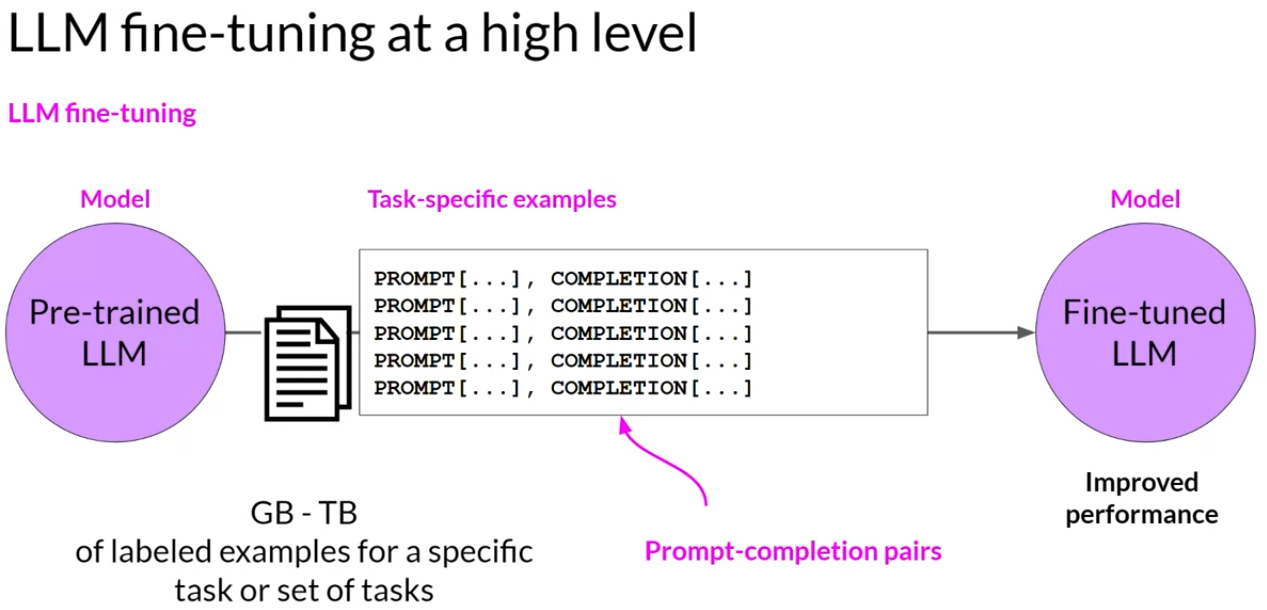


Where all the model weights are updated is known as full fine tuning.

How do you go about Instruction fine tuning an LLM:

* Prepare your training data.
* Divide the dataset into Training [prompt, completion], validation [prompt, completion], and test set [prompt, completion].

The fine tuning results in new version of base model often called instruct model that is better at task you are interested in. Fine tuning with instruct prompt is the most common way to fine tune LLMs these days. Fine tuning is also called instruct fine tuning.



**Fine Tuning on Single task:**

Down side is that model can be affected by Catastrophic forgetting, when fine tuning the LLM on single task it can perform better but it can degrade performance on other tasks.

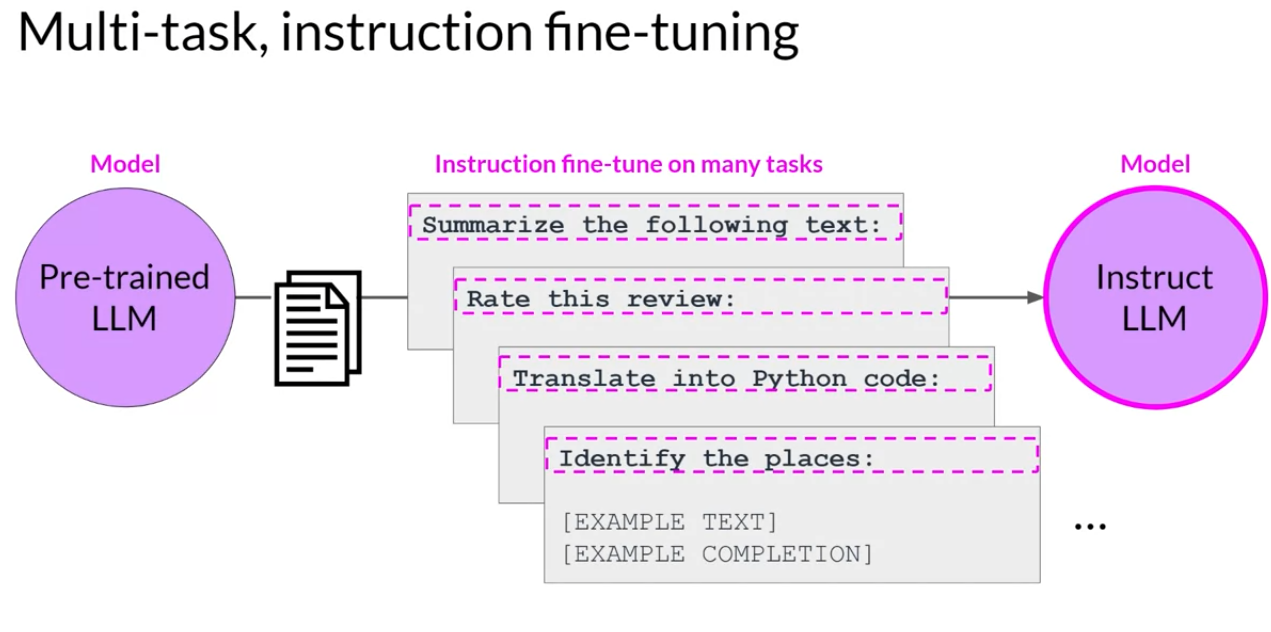
**To avoid catastrophic forgetting:**

Fine tune the model on multiple tasks.

Consider parameter efficient Fine tuning (PEFT) instead of full fine tuning. PEFT preserves the weights of original LLM and trains only a small number of task specific layers.

**Multi-task fine tuning:**

It is extension of single task fine tuning where dataset is comprised of example inputs and outputs for multiple tasks.



**Model Evaluation:**

LLM Evaluation Metrics:

1. ROUGE(Recall Oriented Understudy for Gisting Evaluation):

* Used for text summerization.
* Compares summaries to one or more reference summeries.

1. BLEU(Biligual Evaluation Understudy):

* Used for text translation.
* Compares to human generated translation.

Uni-gram = one word

Bi-gram = two words

N-gram= n no of words

**LLM Evaluation-Metrics-ROUGE:**

**ROUGE-1**

Reference(human):

It is cold outside.

Generated output:

It is very cold outside.

ROUGE-1 Recall = unigram matches / unigram in reference = 4/4 = 1.0

ROUGE-1 Precision = unigram matches / unigram in output = 4/5 = 0.8

ROUGE-1 F1 = 2 X precision X recall / precision + recall = 2\* 0.8/1.8 = 0.89

**ROUGE-2**

Reference(human):

It is cold outside.

|  |  |  |
| --- | --- | --- |
| It is | Is cold | Cold outside |

Generated output:

It is very cold outside.

|  |  |  |  |
| --- | --- | --- | --- |
| It is | Is very | Very cold | Cold outside |

ROUGE-2 Recall = unigram matches / unigram in reference = 2/3 = 0.67

ROUGE-2 Precision = unigram matches / unigram in output = 2/4 = 0.5

ROUGE-2 F1 = 2 X precision X recall / precision + recall = 2\* 0.335/1.17 = 0.57

**ROUGE-L**

Reference(human):

It is cold outside.

Generated output:

It is very cold outside.

Longest common subsequent (LCS):

|  |
| --- |
| It is |

|  |
| --- |
| Cold outside |

ROUGE-L Recall = LCS(Gen,Ref) / unigram in reference = 2/4 = 0.5

ROUGE-L Precision = LCS(Gen,Ref) / unigram in output = 2/5 = 0.4

ROUGE-L F1 = 2 X precision X recall / precision + recall = 2\* 0.2/0.9 = 0.44

**LLM Evaluation-Metrics-BLEU:**

BLEU metrics = Avg(pricision across range of n-gram sizes)

How many n-grams in machine generated translation match those in the reference translation. To calculate the score you average precision score across range of different n-gram sizes.