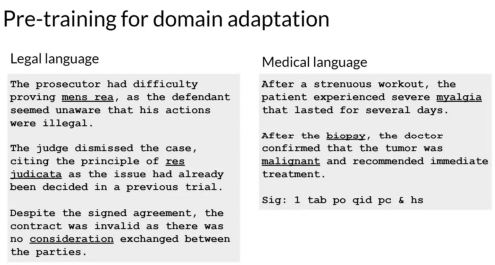
Module 2

Monday, June 17, 2024 10:03 PM

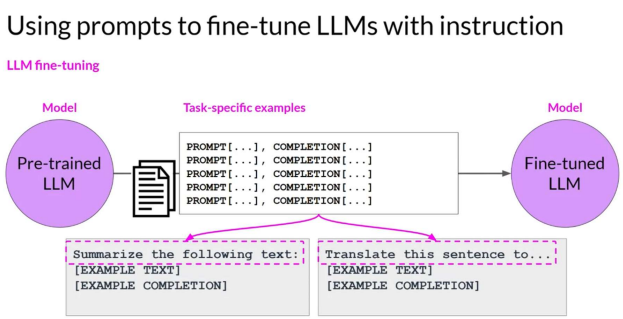
Pretraining for Domain adaptation:

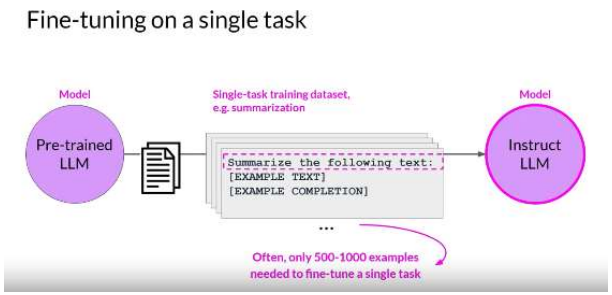


Why domain specific model?
→ Because language in that domain is different. Example

When scaling data, we need to increase both data and number of parameters of model.

Fine-Tuning is supervised method to change LLM behavior





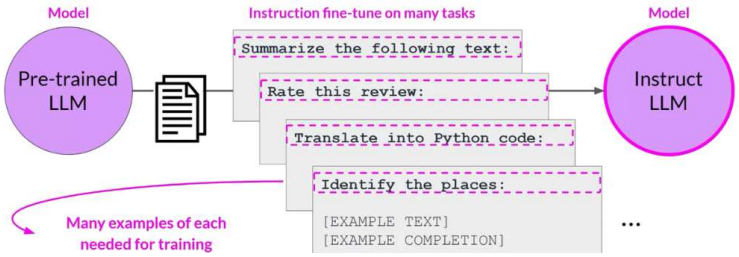
- Single-task training dataset
- Often only 500-1000 examples are needed to fine-tune a single task.

Problem: Catastrophic forgetting

Solution: Fine-tune on multiple tasks (or) Parameter Efficient Fine-tuning (PEFT)

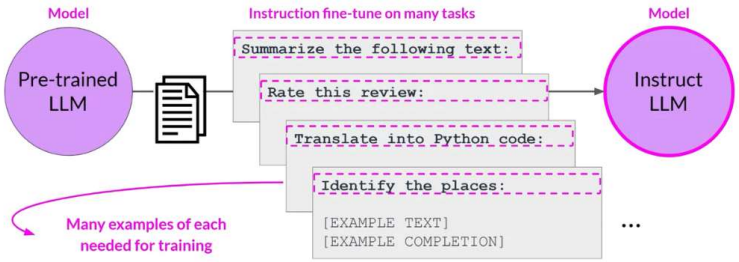
Multi-task, instruction fine-tuning:

→ Contains example of different instruction fine-tuning on many tasks.



- Major Problem is that it requires lot of data, but it is worth it!

Multi-task, instruction fine-tuning



• Major problem ... data, but it is worth it!

Instruction fine-tuning with FLAN

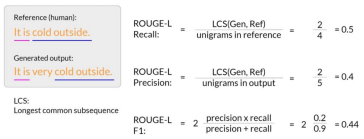
→ Fine Tuned Language Net: It is a specific set of instruction used to perform instruction fine-tuning.

Model Evaluation:

ROUGE

Recall-Oriented Understudy for Gisting Evaluation

- Used for text summarization
- Compare a summary to one or more reference summaries



BLEU Score (Bilingual evaluation understudy)

- Used for text translation
- Compare to human-generated translation

Benchmarks:

GLUE: General language understanding tasks
→ Contains sentiment analysis and question answering

SuperGLUE: Multi-sentence reasoning and reading comprehension

Big-Bench: Contains different dataset sizes to test again-

HELM: Includes metrics like Fairness, Bias, Toxicity

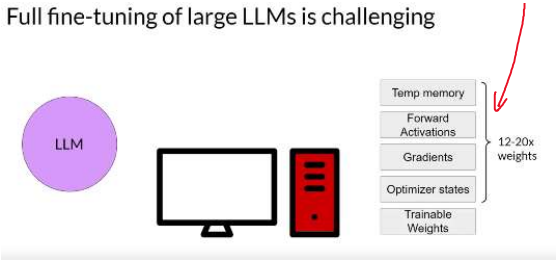
MMLU: Massive Multitask Language Understanding
→ Designed specifically for modern LLMs.

Parametric Efficient Fine-Tuning (PEFT)

- While fine tuning models, it involves storing model variables that are not usually loaded for inferencing. For fine tuning, we need atleast

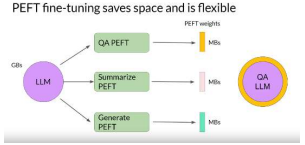
Full fine-tuning of large LLMs is challenging

Full fine-tuning of large LLMs is challenging

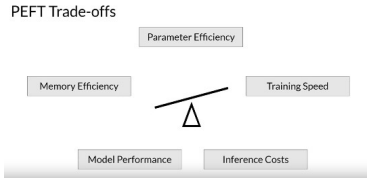


PEFT → Fine tuning a small subset of parameters.

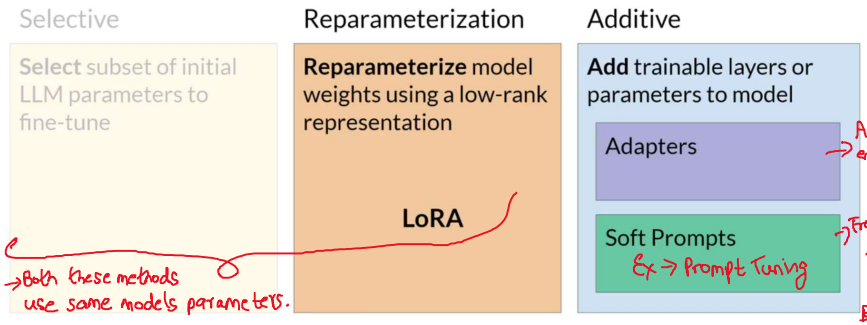
- Either freeze the model and train subset of layers.
- or freeze the whole model and add new layers.



- We can then add new layers on top of original layer for our usecase.



PEFT methods



→ Both these methods use same models parameters.

→ Add new trainable layers to encoder/decoder architecture

→ Freezes architecture, and changes the input type to improve performance

By adding trainable parameters to the prompt embeddings (or) keeping the input fixed and retraining the embedding weights.

Source: Lialin et al. 2023, "Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning",

Low-Rank Adaptation of large language Models

- Parametric Efficient Fine-Tuning on single GPU
- Choosing rank of lower matrices is Still Active Research (4-32 is good enough)

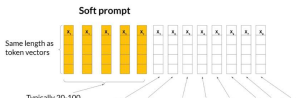
→ Widely used

Prompt Tuning

Prompt Tuning is not prompt engineering

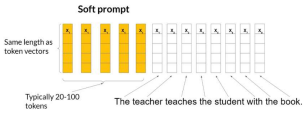
→ Limitations of Prompt Engineering:
- → Requires lot of effort to write and try different prompts.
- → Also limited by the length of context window.

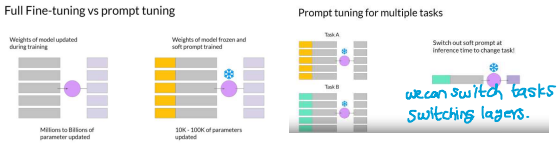
Prompt tuning adds trainable "soft prompt" to inputs



↳ Limitations of Prompt Engineering:
→ Requires lot of effort to write and try different prompts.
→ Also limited by the length of context window.

With prompt Tuning → we add additional trainable tokens to the prompt.





we can switch tasks by switching layers.