LLM Instruction Fine-Tuning & Evaluation

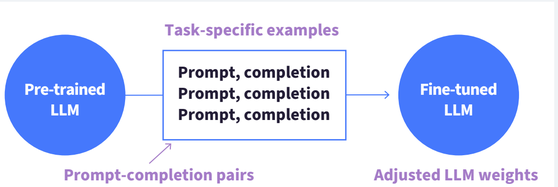
INSTRUCTION FINE-TUNING

In-Context Learning Limitations:

- May be insufficient for very specific tasks.
- Examples take up space in the context window.

Instruction Fine-Tuning

The LLM is trained to estimate the next token probability on a cautiously curated dataset of high-quality examples for specific tasks.



- The LLM generates better completions for a specific task
- Has potentially high computing requirements

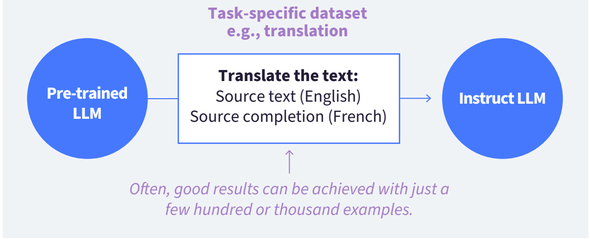
Steps:

1. Prepare the training data.
2. Pass examples of training data to the LLM (prompt and ground-truth answer).

3. Compute the cross-entropy loss for each completion token and backpropagate.

TASK-SPECIFIC FINE-TUNING

Task-specific fine-tuning involves training a pre-trained model on a particular task or domain using a dataset tailored for that purpose.



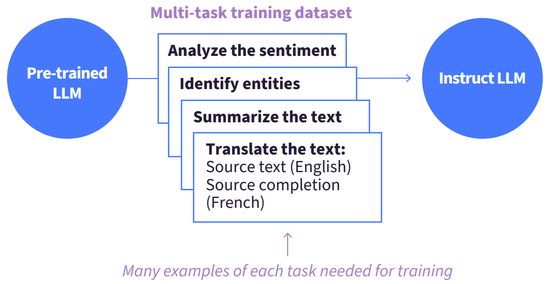
Fine-tuning can significantly increase the performance of a model on a specific task, but can reduce the performance on other tasks ("**catastrophic forgetting**").

Solutions:

- **It might not be an issue** if only a single task matters.
- **Fine-tune for multiple tasks concurrently** (~50K to 100K examples needed).
- **Opt for** Parameter Efficient Fine-Tuning (PEFT) instead of full fine-tuning, which involves training only a small number of task-specific adapter layers and parameters.

MULTI-TASK FINE-TUNING

Multi-task fine-tuning diversifies training with examples for multiple tasks, guiding the model to perform various tasks.



Drawback: It requires a lot of data (around 50K to 100K examples).

Model variants differ based on the datasets and tasks used during fine-tuning.

Example of the FLAN family of models
FLAN, or Fine-tuned LAnguage Net, provides tailored instructions for refining various models, akin to dessert after pre-training.

FLAN-T5 is an instruct fine-tuned version of the T5 foundation model, serving as a versatile model for various tasks.

FLAN-T5 has been **fine-tuned on a total of 473 datasets** across **146 task categories.** *For instance, the SAMSum dataset was used for summarization.*

A specialized variant of this model for chat summarization or for custom company usage could be developed through additional fine-tuning on specialized datasets (e.g., DialogSum or custom internal data).

MODEL EVALUATION

Evaluating LLMs Is Challenging

(e.g., various tasks, non-deterministic outputs, equally valid answers with different wordings).

→ Need for automated and organized performance assessments

Various approaches exist, but there are a few examples:

ROUGE & BLEU SCORE

- **Purpose:** To evaluate LLMs on narrow tasks (summarization, translation) when a reference is available
- Based on n-grams and rely on precision and recall scores (multiple variants)

BERT SCORE

- **Purpose:** To evaluate LLMs in a task-agnostic manner when a reference is available.
- Based on token-wise comparison, a similarity score is computed between candidate and reference sentences.

LLM-as-a-Judge

- **Purpose:** To evaluate LLMs in a task-agnostic manner when a reference is available.
- Based on prompting an LLM to assess the equivalence of a generated answer with a ground-truth answer.

To measure and compare LLMs more holistically, use **evaluation benchmark datasets** specific to model skills.

E.g., *GLUE*, *SuperGLUE*, *MMLU*, *Big Bench*, *Helm*