devfest

Google Developer Groups

Fine-Tuning Open-Source LLMs: Best Practices

@adityaura

aadityaura@gmail.com

aadityaura.github.io

Ankit Pal (Aaditya Ura)
Senior Research Engineer,
Saama AI Research Lab, Chennai

About Me

🤩

Cool Research | Trekking | Boxing | Skiing | Chess

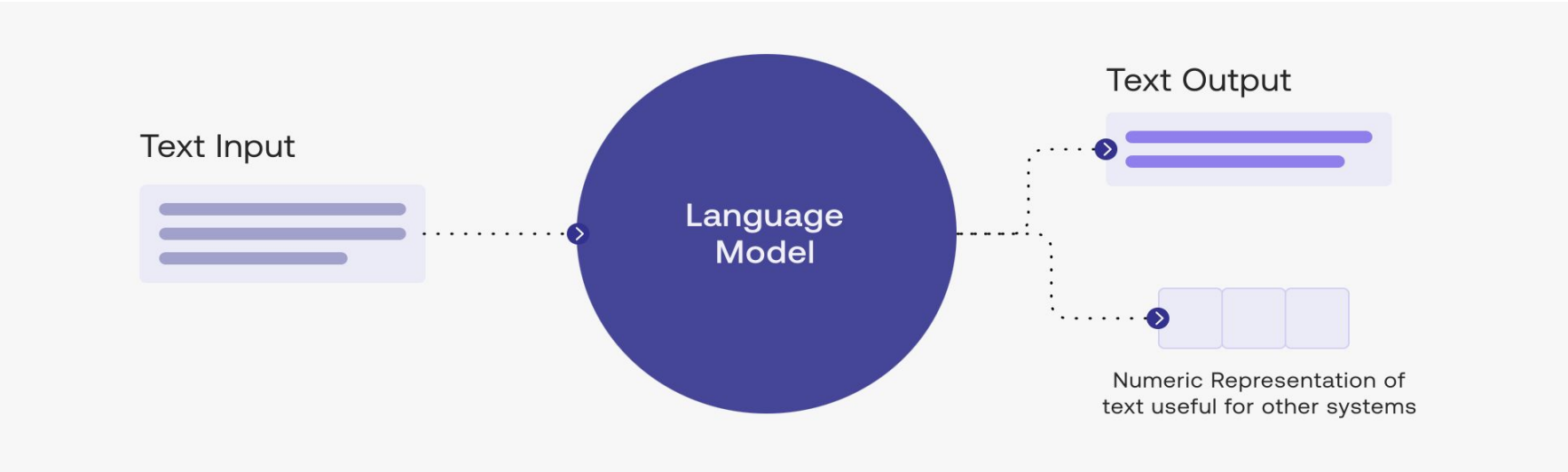
- Senior AI Research Engineer [@ Saama AI Research Lab]
- Research interests involve
  Representation Learning on Graphs and Manifolds
- Generative Modeling, MLOps, Signal Processing
  and their applications in Healthcare data

Google Developer Groups

Large Language Models(LLMs)

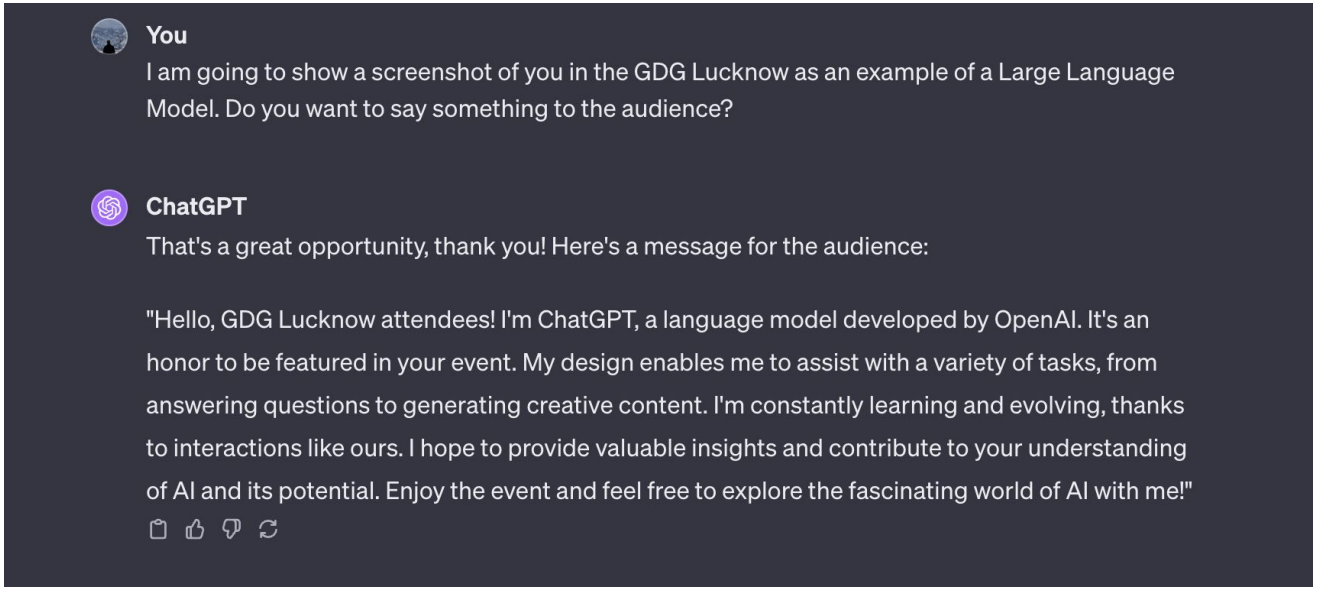
Understanding Large Language Models

LLMs are advanced AI models trained on vast datasets to understand and generate human-like text.



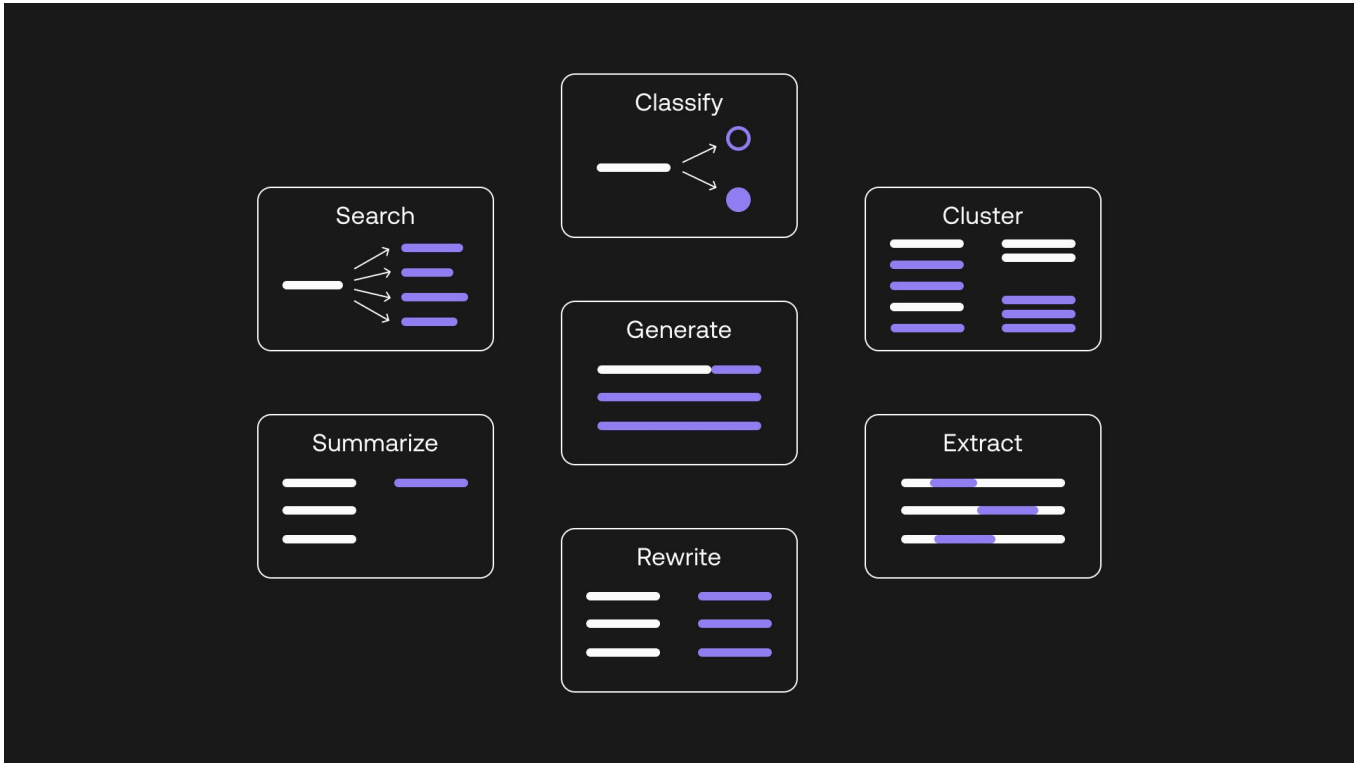
Source: https://docs.cohere.com/docs/llmu

Understanding Large Language Models



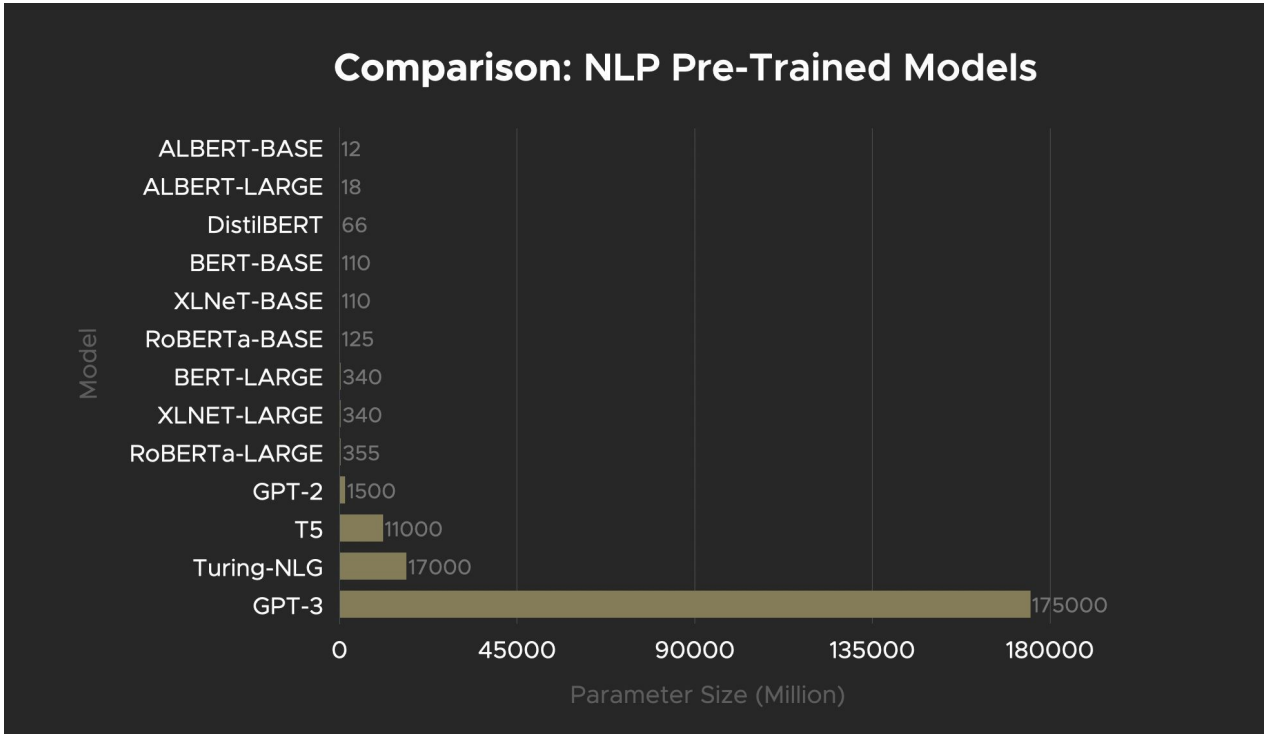
ChatGPT is one example of a Large Language Model.

Understanding Large Language Models



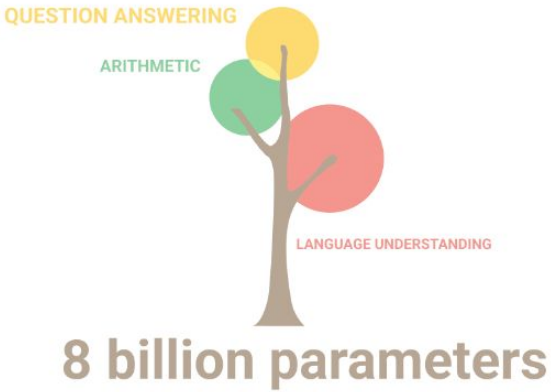
Source: https://docs.cohere.com/docs/llmu

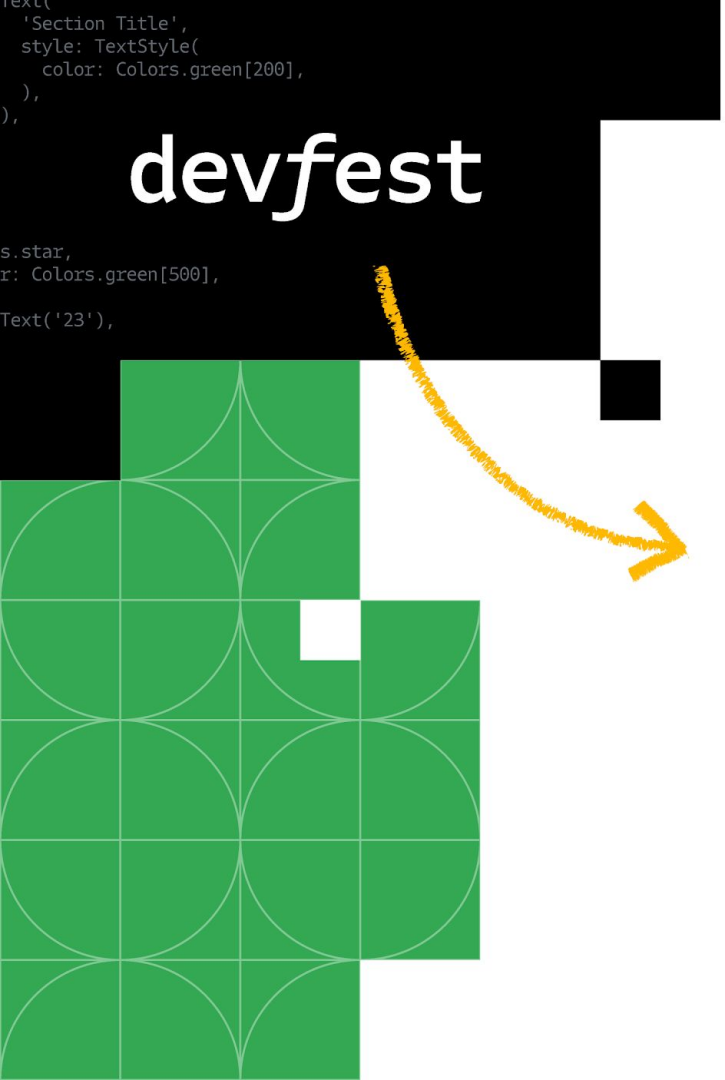
Understanding Large Language Models



Understanding Large Language Models

Understanding Large Language Models





Google Developer Groups

What is Fine-Tuning?

Fine-Tuning

Taking a pre-trained model and training at least one model parameter

Fine-Tuning

Taking a pre-trained model and training at least one model parameter

GPT-3

Fine-Tuning

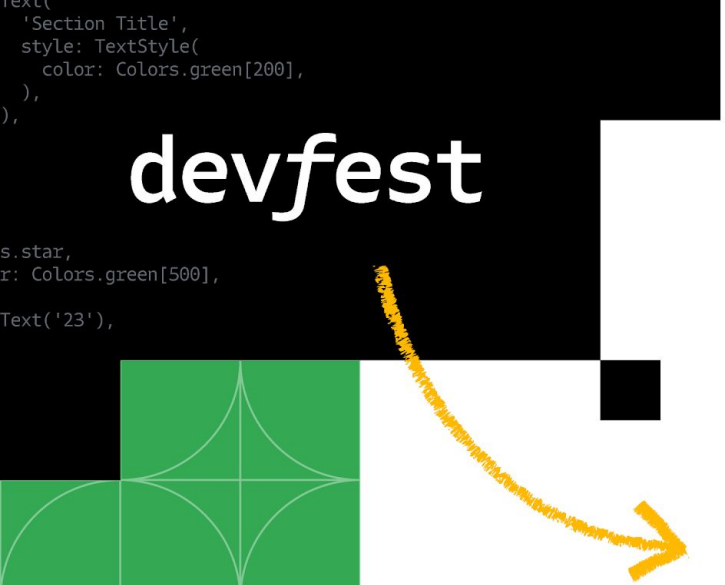
Taking a pre-trained model and training at least one model parameter

GPT-3

ChatGPT
(i.e. GPT-3.5-turbo)

Google Developer Groups

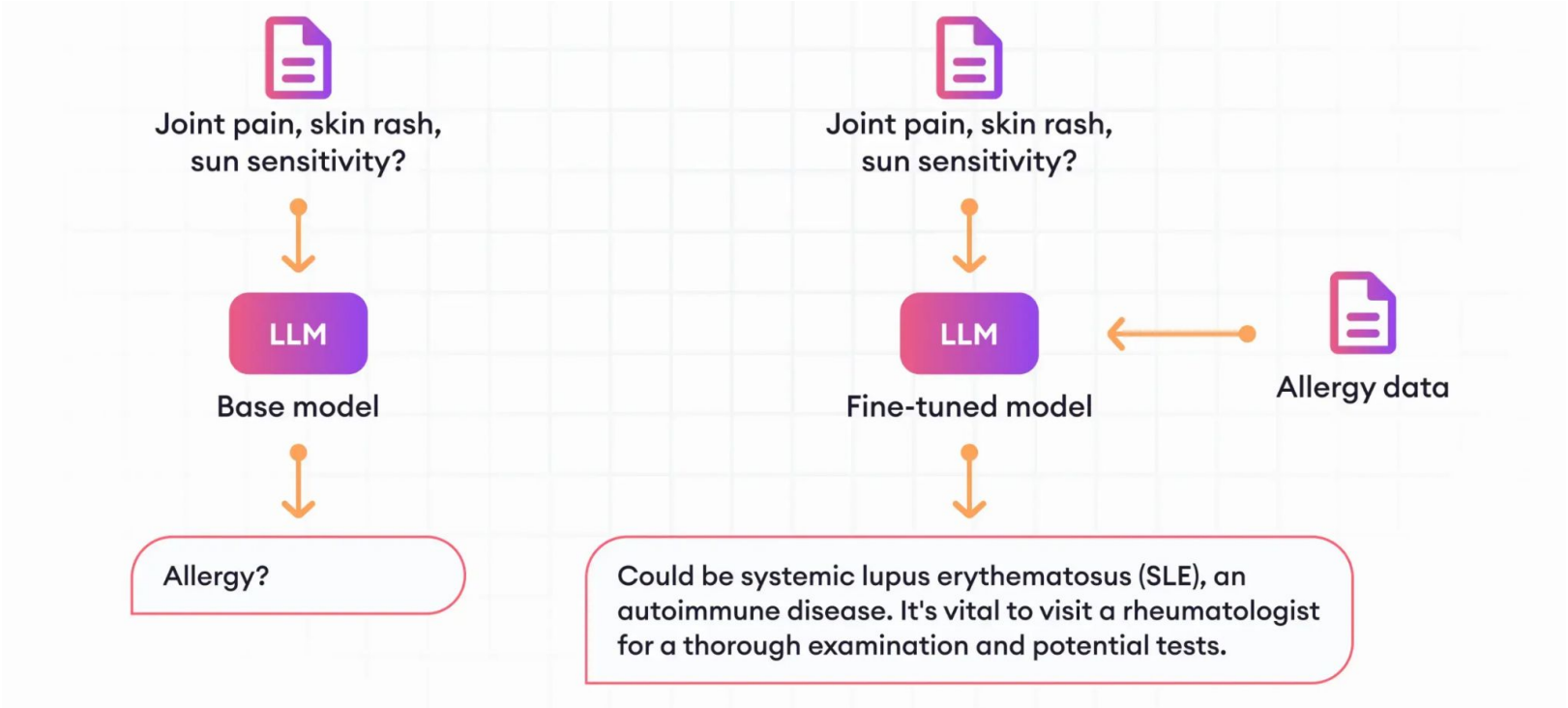
https://towardsdatascience.com/fine-tuning-large-language-models-llms-23473d763b91



Google Developer Groups

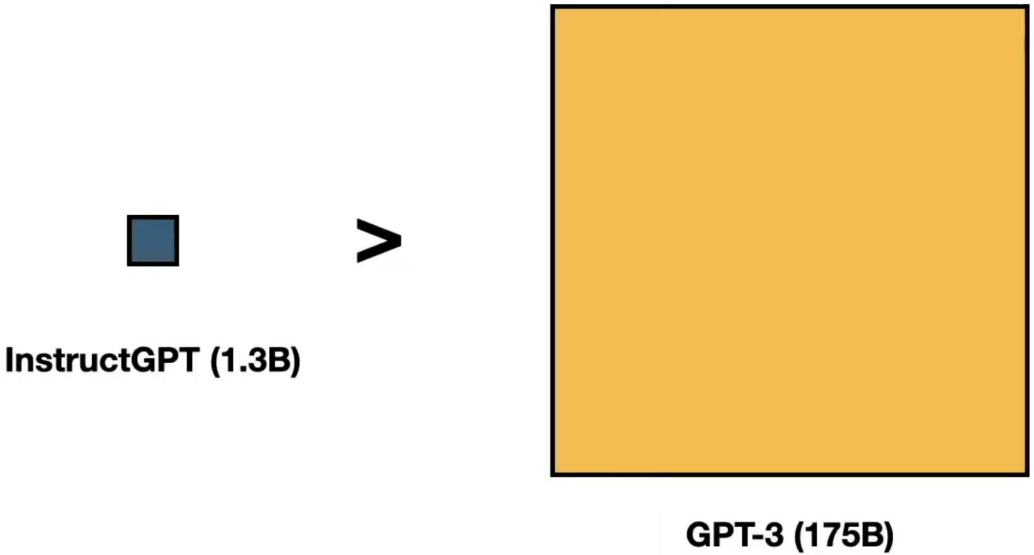
Why We Need to Fine-Tune?

Need of Fine-Tuning



Need of Fine-Tuning

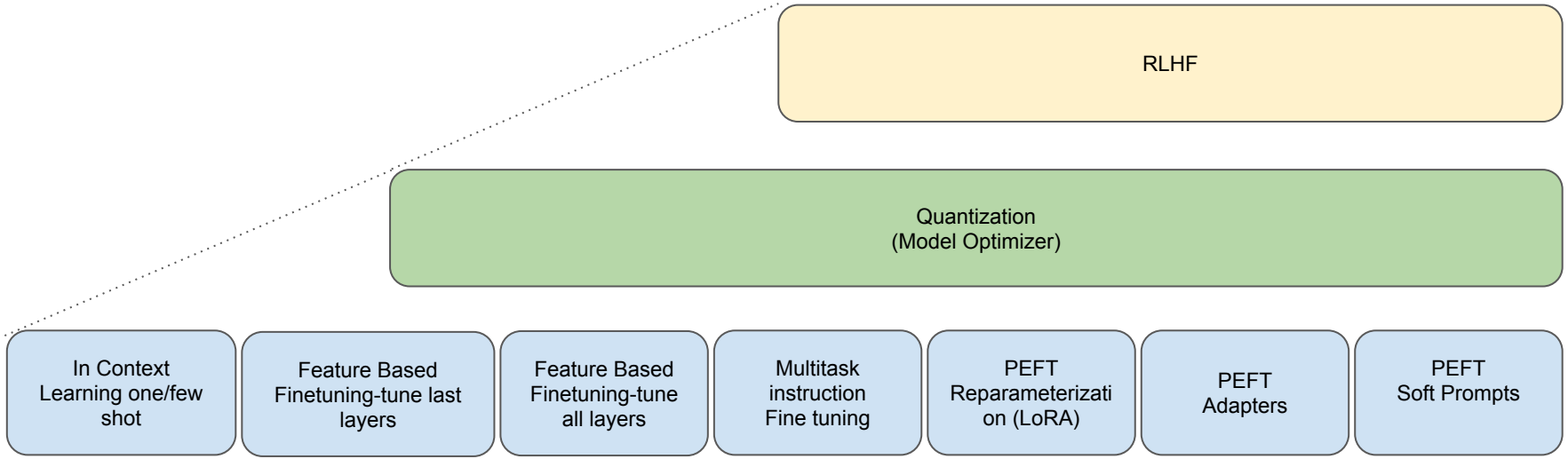
A smaller (fine-tuned) model can outperform a larger base model



Google Developer Groups

Fine-Tuning Approaches

Fine-Tuning Approaches



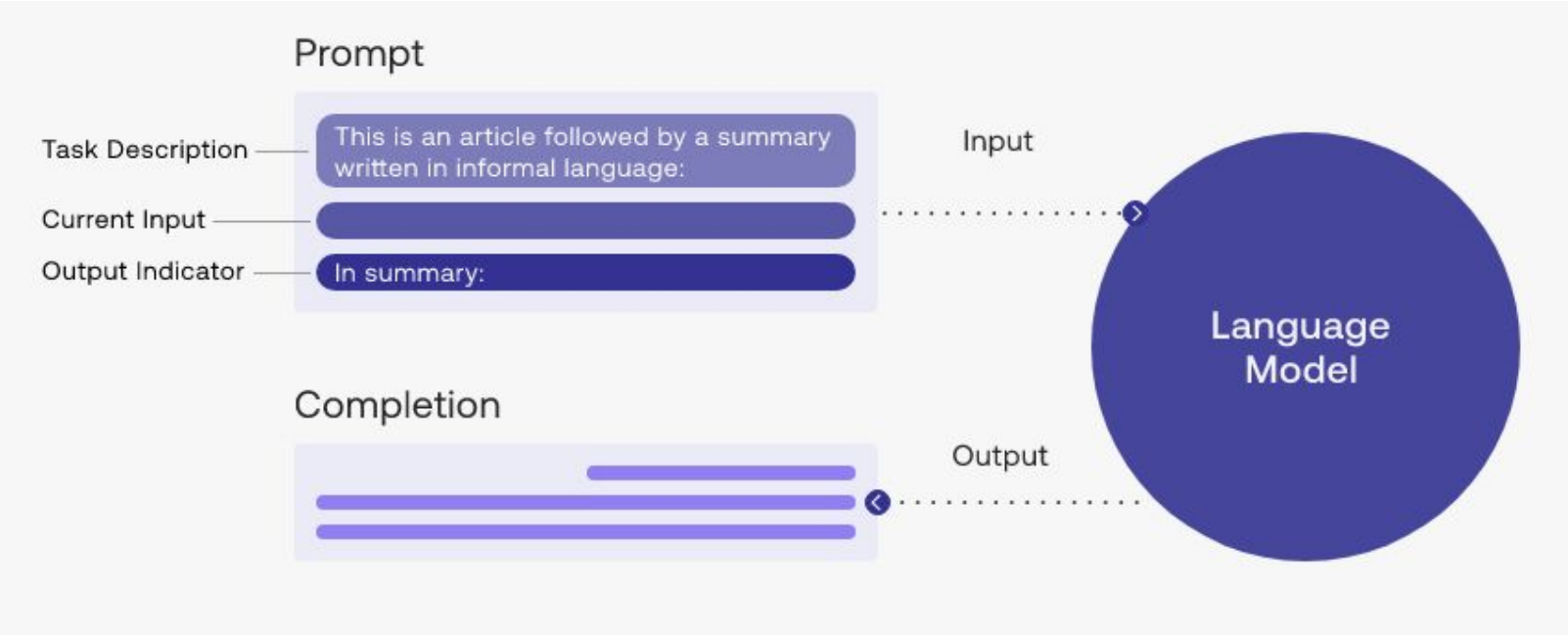
Google Developer Groups

In-Context Learning

Fine-Tuning Approaches

Prompt Engineering

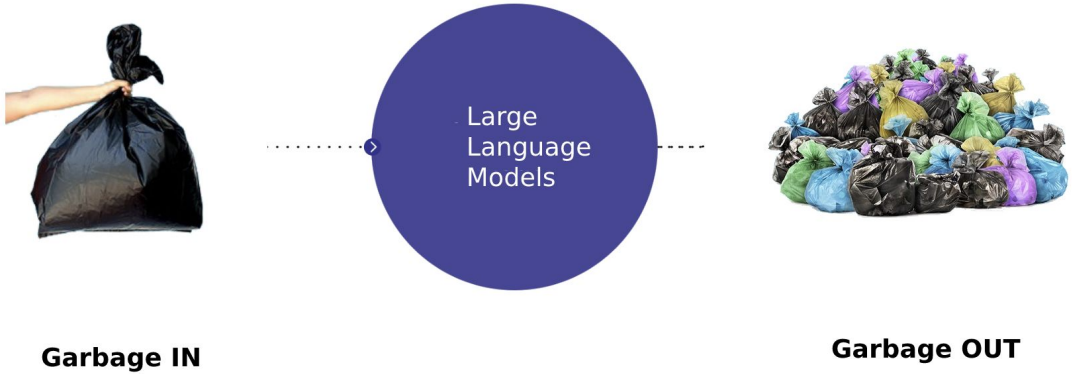
Prompt-Engineering



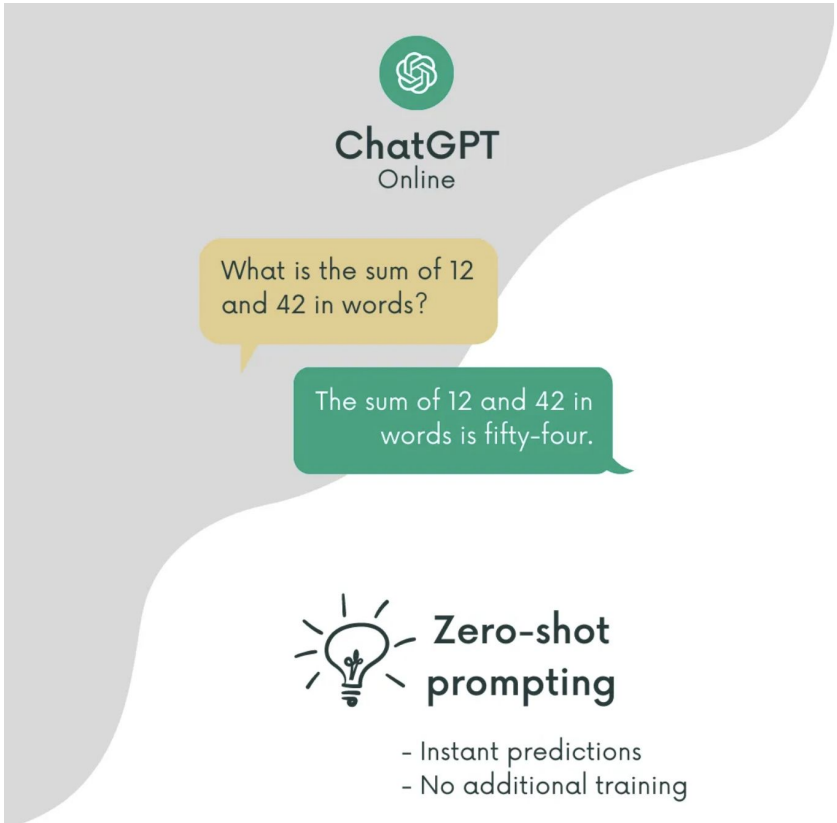
Prompting is a method that allows users to interact with LLMs

Prompt-Engineering

Prompt plays a key role in LLMs



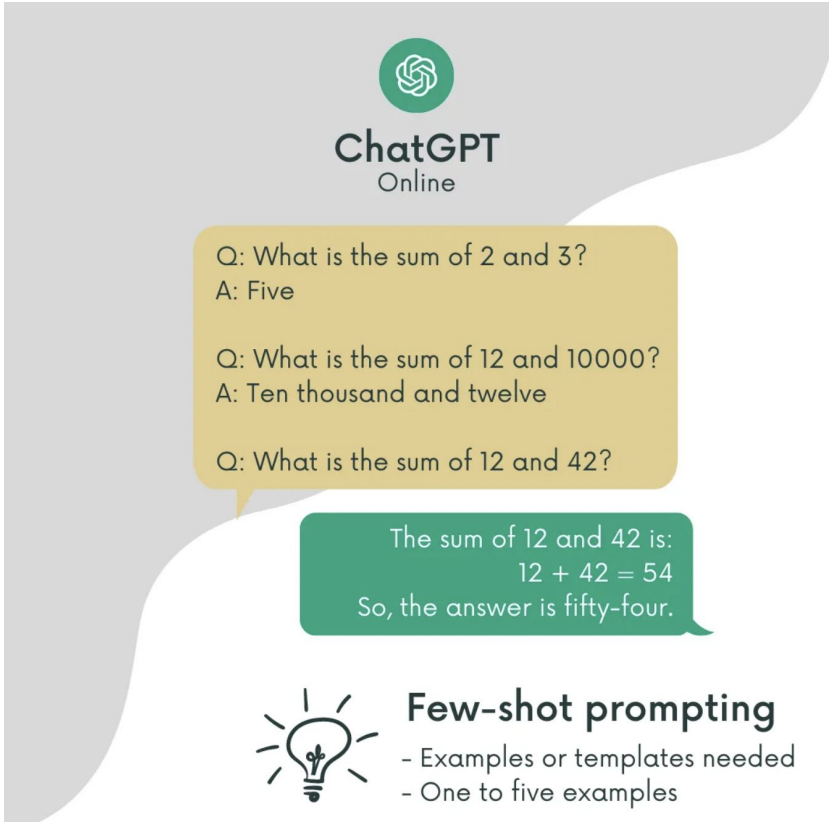
Types of prompting



Google Developer Groups

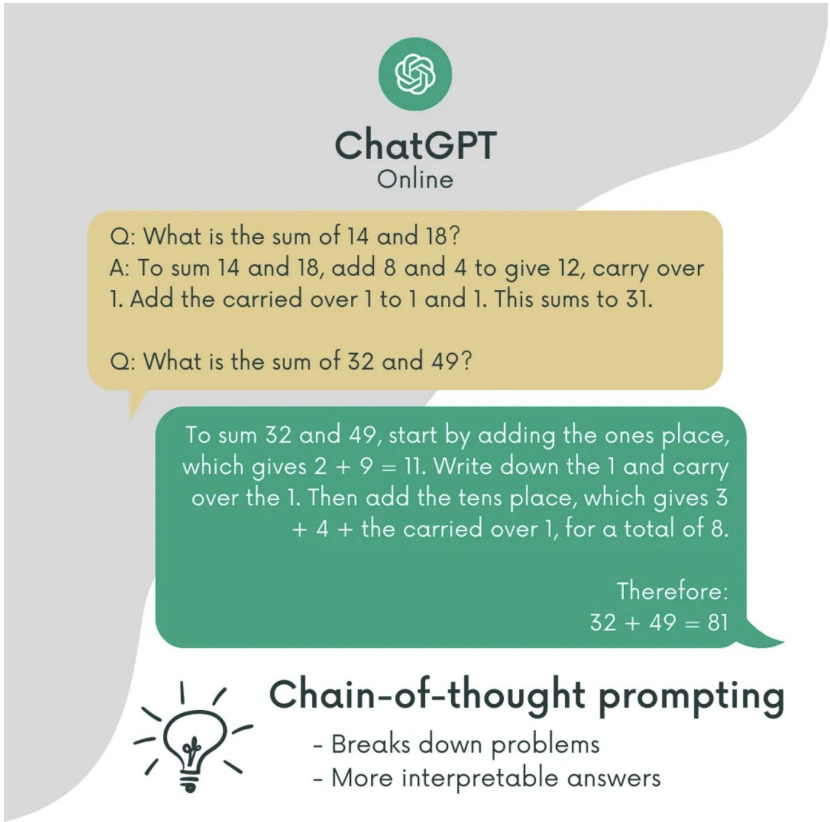
https://medium.com/@arushi.prakash307/teach-llms-how-to-reason-with-chain-of-thought-prompting-d0c517e200f3

Types of prompting



https://medium.com/@arushi.prakash307/teach-llms-how-to-reason-with-chain-of-thought-prompting-d0c517e200f3

Types of prompting



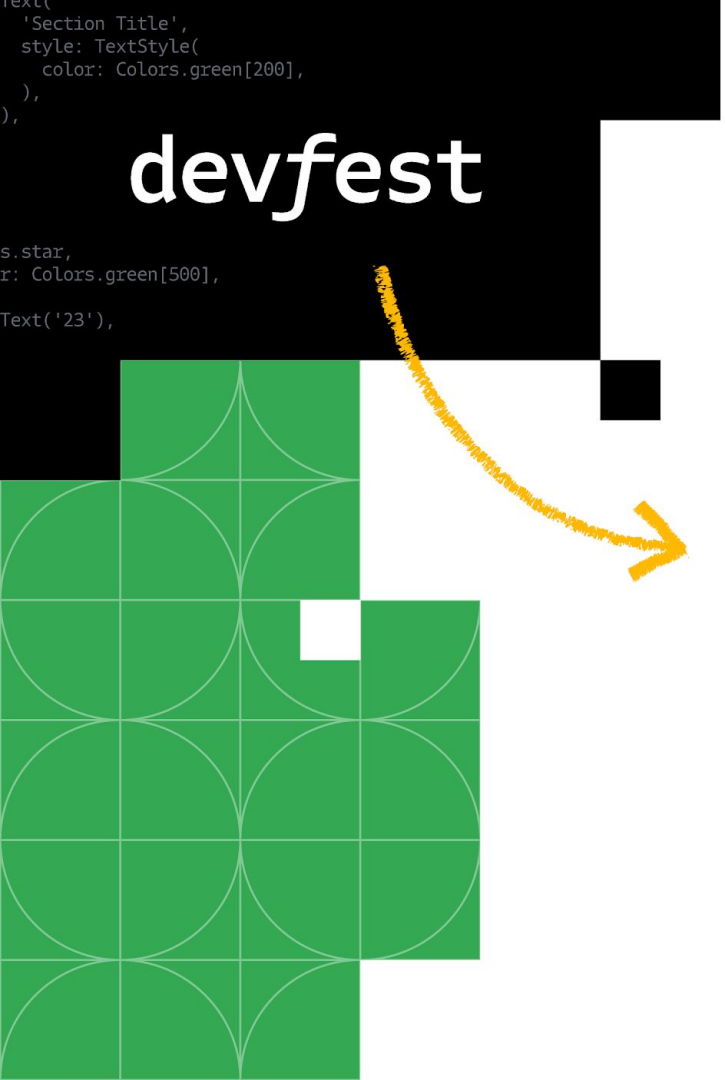
https://medium.com/@arushi.prakash307/teach-llms-how-to-reason-with-chain-of-thought-prompting-d0c517e200f3

Types of prompting

LLM Model Parameters to Control

- Stopping criteria
    - Number of tokens - maximum number of tokens to generate
    - Stop sequences - character or the sequence used to separate the training examples

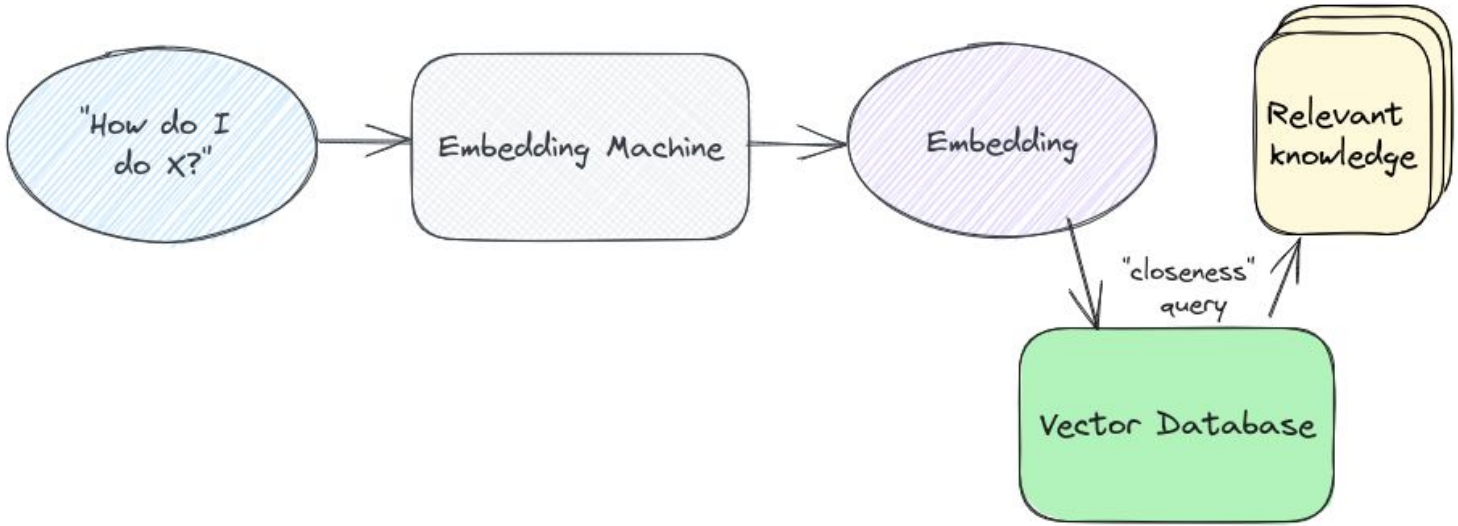
- Repetition Reduction
    - Frequency penalty - penalizes based on the frequency of the word in the preceding text
    - Presence penalty - penalizes based on the presence of the word



Google Developer Groups

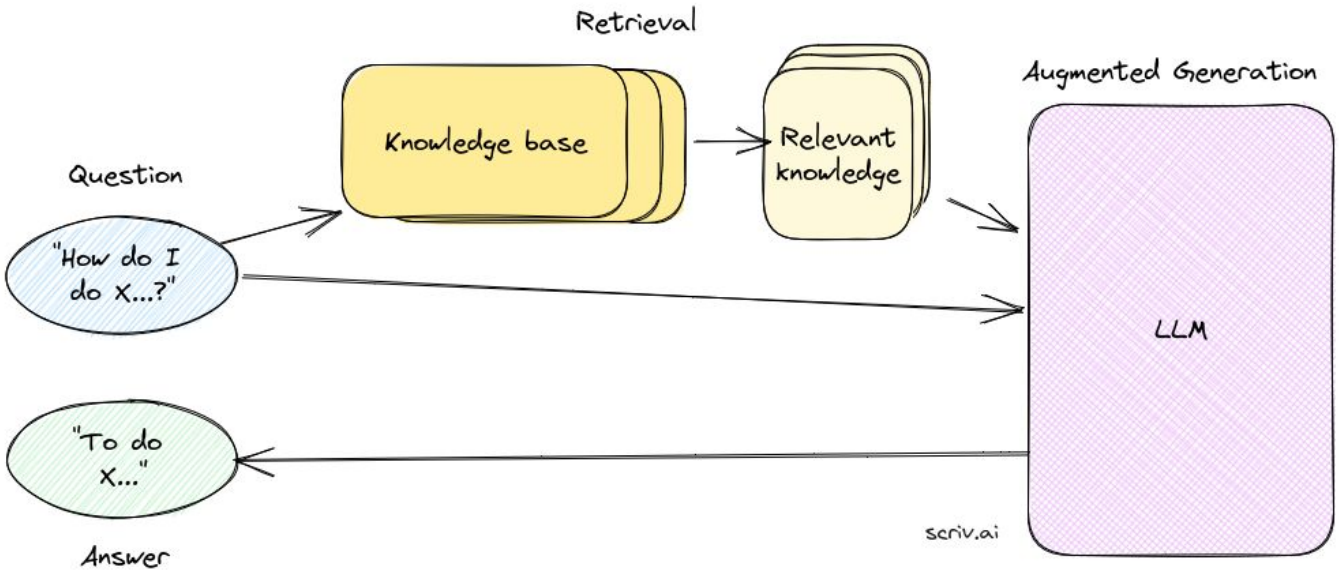
Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG)



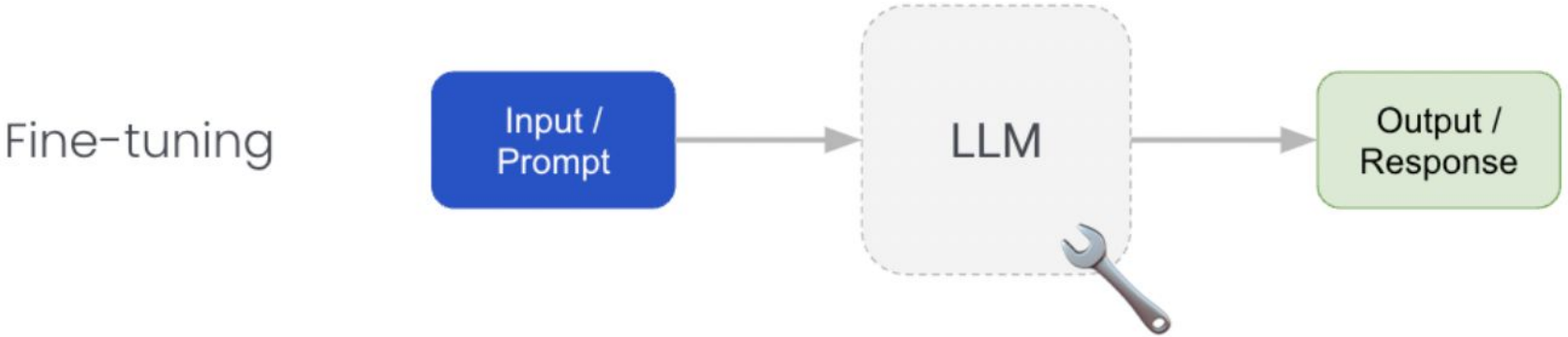
https://scriv.ai/guides/retrieval-augmented-generation-overview/

Retrieval Augmented Generation (RAG)



https://scriv.ai/guides/retrieval-augmented-generation-overview/

Fine-Tuning Approaches



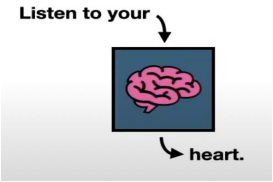
Google Developer Groups

Fine-Tuning

Types of Fine-Tuning

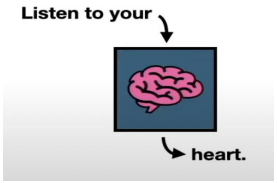
Self-Supervised

Types of Fine-Tuning



Self-Supervised

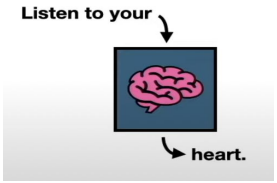
Types of Fine-Tuning



The cat

Self-Supervised

Types of Fine-Tuning



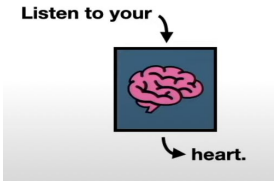
The cat

Input	Output

Self-Supervised

Supervised

Types of Fine-Tuning



The cat

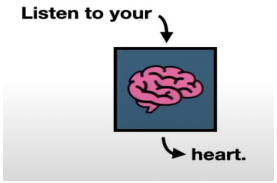
Self-Supervised

Input	Output

Input: *Who was the 35th President of the United States?*
Output: *John F. Kennedy*

Supervised

Types of Fine-Tuning



The cat

Input	Output

Input: *Who was the 35th President of the United States?*

Output: *John F. Kennedy*

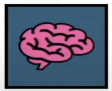
```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

Self-Supervised

Supervised

Types of Fine-Tuning

Listen to your



heart.

The cat

Input	Output

Input: *Who was the 35th President of the United States?*

Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

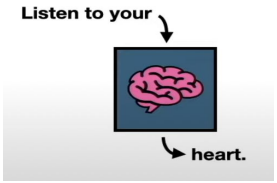
i. Supervised FT

Self-Supervised

Supervised

Reinforcement learning

Types of Fine-Tuning



The cat

Input	Output

Input: *Who was the 35th President of the United States?*
Output: *John F. Kennedy*

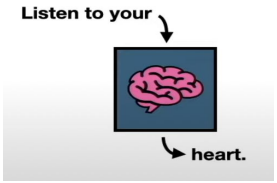
```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

Self-Supervised

Supervised

Reinforcement learning

Types of Fine-Tuning



The cat

Input	Output

Input: *Who was the 35th President of the United States?*

Output: *John F. Kennedy*

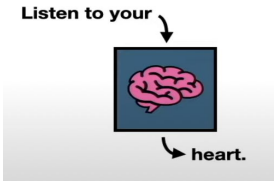
```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

Self-Supervised

Supervised

Reinforcement learning

Types of Fine-Tuning

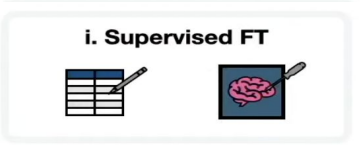


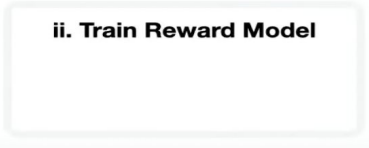
The cat

Input	Output

Input: *Who was the 35th President of the United States?*
Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```



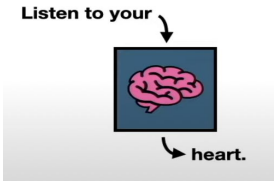


Self-Supervised

Supervised

Reinforcement learning

Types of Fine-Tuning

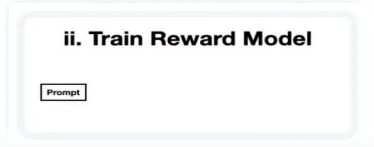


The cat

Input	Output

Input: *Who was the 35th President of the United States?*
Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

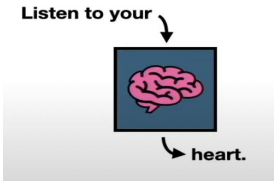


Self-Supervised

Supervised

Reinforcement learning

Types of Fine-Tuning



The cat

Input	Output

Input: *Who was the 35th President of the United States?*

Output: *John F. Kennedy*

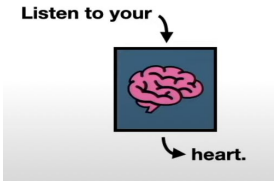
```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

Self-Supervised

Supervised

Reinforcement learning

Types of Fine-Tuning



The cat

Self-Supervised

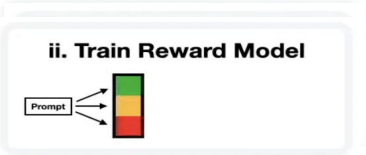
Input	Output

Input: *Who was the 35th President of the United States?*

Output: *John F. Kennedy*

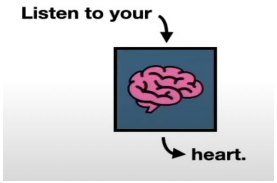
```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

Supervised



Reinforcement learning

Types of Fine-Tuning



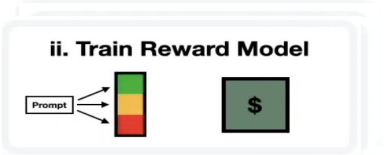
The cat

Input	Output

Input: *Who was the 35th President of the United States?*

Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

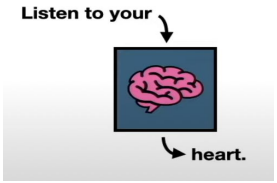


Self-Supervised

Supervised

Reinforcement learning

Types of Fine-Tuning



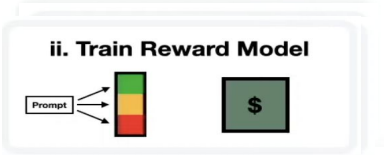
The cat

Input	Output

Input: *Who was the 35th President of the United States?*

Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

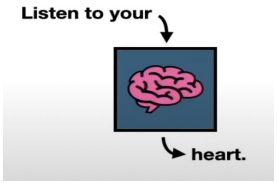


Self-Supervised

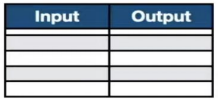
Supervised

Reinforcement learning

Types of Fine-Tuning

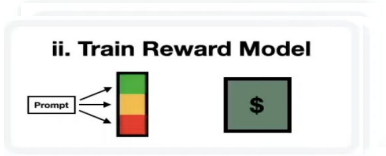


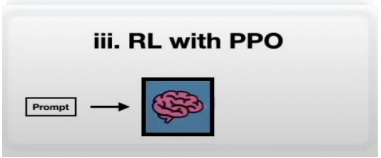
The cat



Input: *Who was the 35th President of the United States?*
Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```



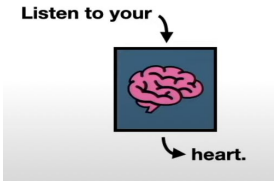


Self-Supervised

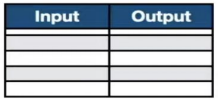
Supervised

Reinforcement learning

Types of Fine-Tuning

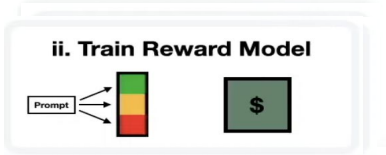


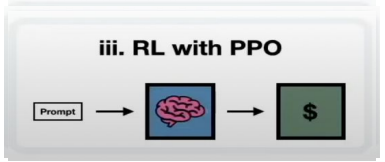
The cat



Input: *Who was the 35th President of the United States?*
Output: *John F. Kennedy*

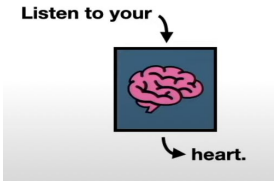
```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```





Self-Supervised **Supervised** **Reinforcement learning**

Types of Fine-Tuning

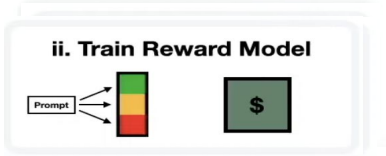


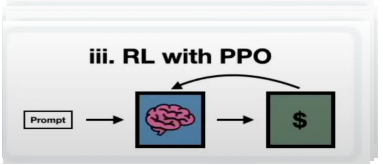
The cat

Input	Output

Input: *Who was the 35th President of the United States?*
Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```



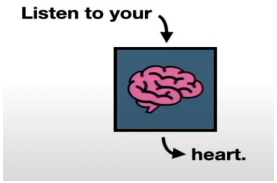


Self-Supervised

Supervised

Reinforcement learning

Types of Fine-Tuning

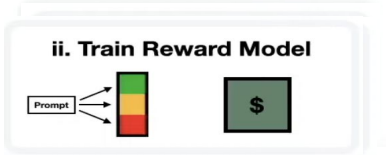


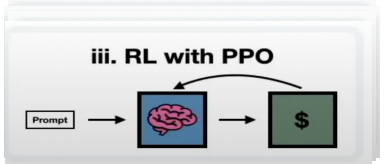
The cat

Input	Output

Input: *Who was the 35th President of the United States?*
Output: *John F. Kennedy*

```
"""Please answer the following question.

Q: {Question}
A: {Answer}"""
```

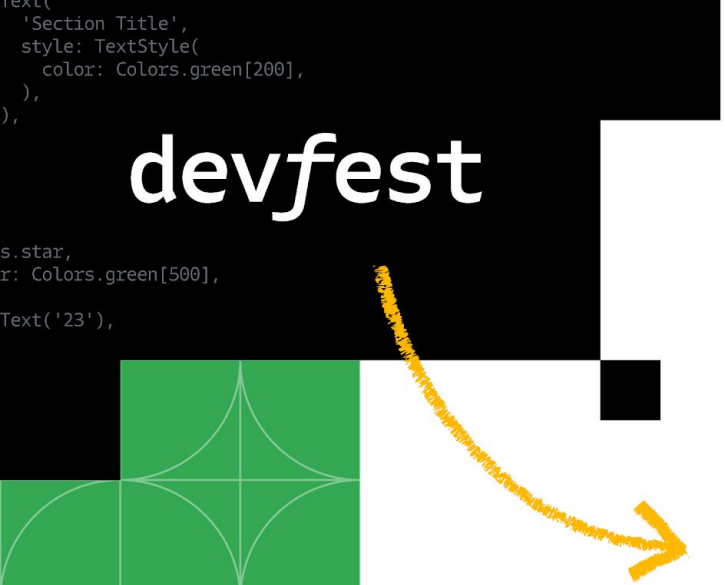




Self-Supervised **Supervised** **Reinforcement learning**

Google Developer Groups

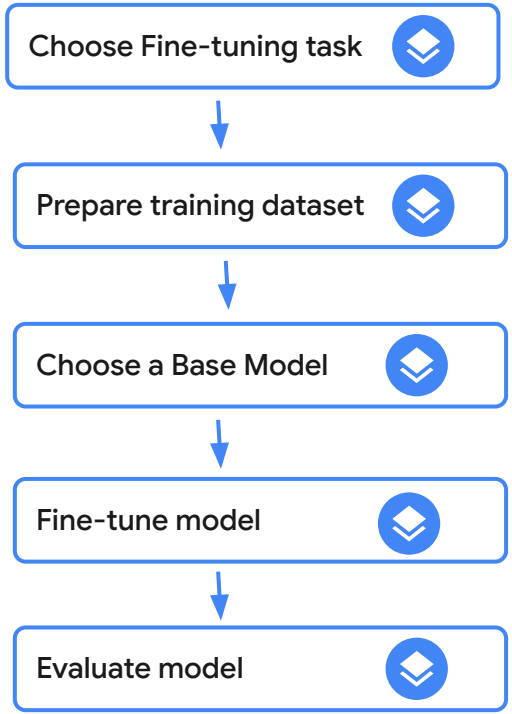
https://towardsdatascience.com/fine-tuning-large-language-models-llms-23473d763b91



Google Developer Groups

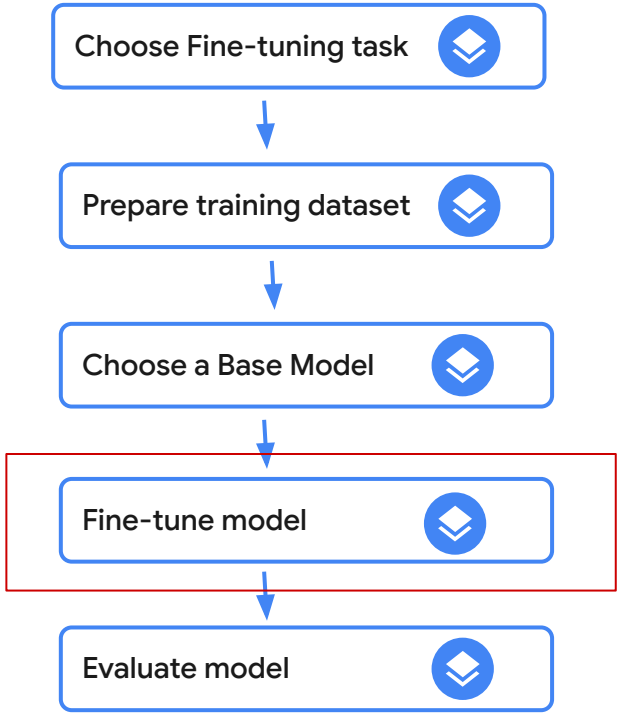
Supervised Training

5 Steps



Input	Output

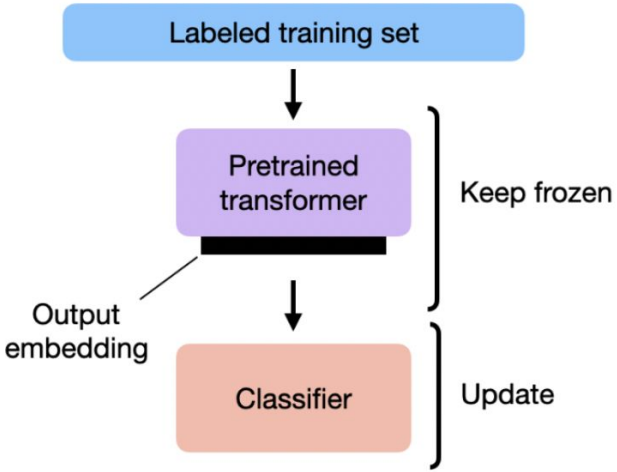
5 Steps



Input	Output

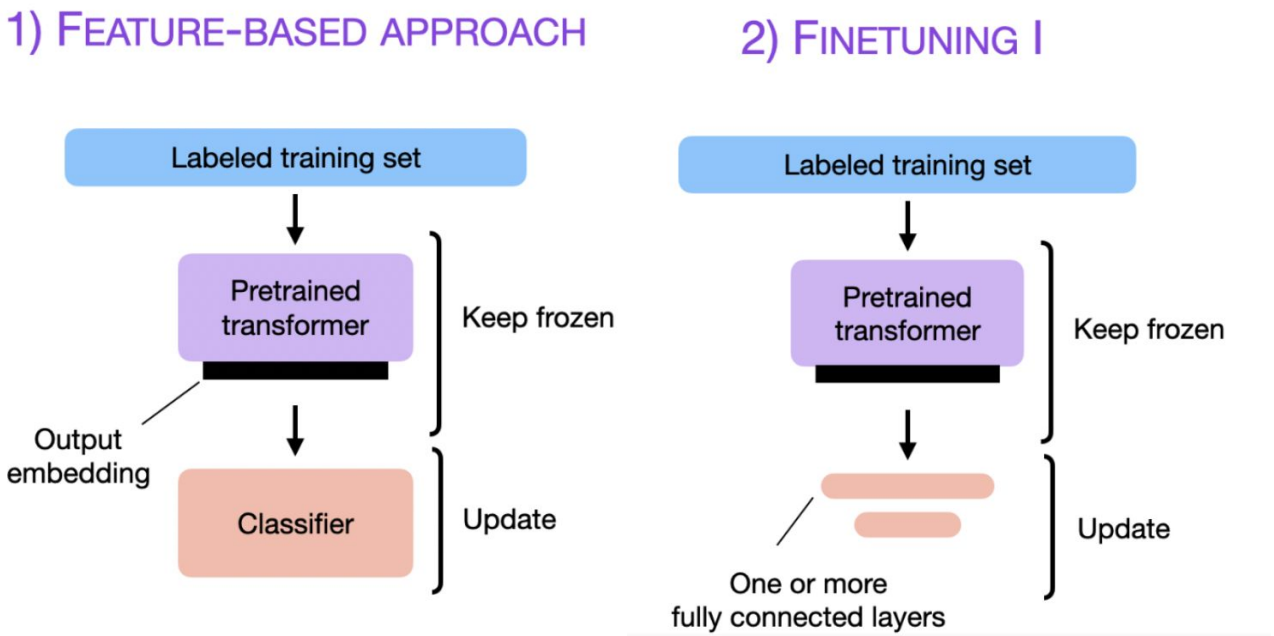
Fine-tune model via supervised learning

1) Feature-based approach



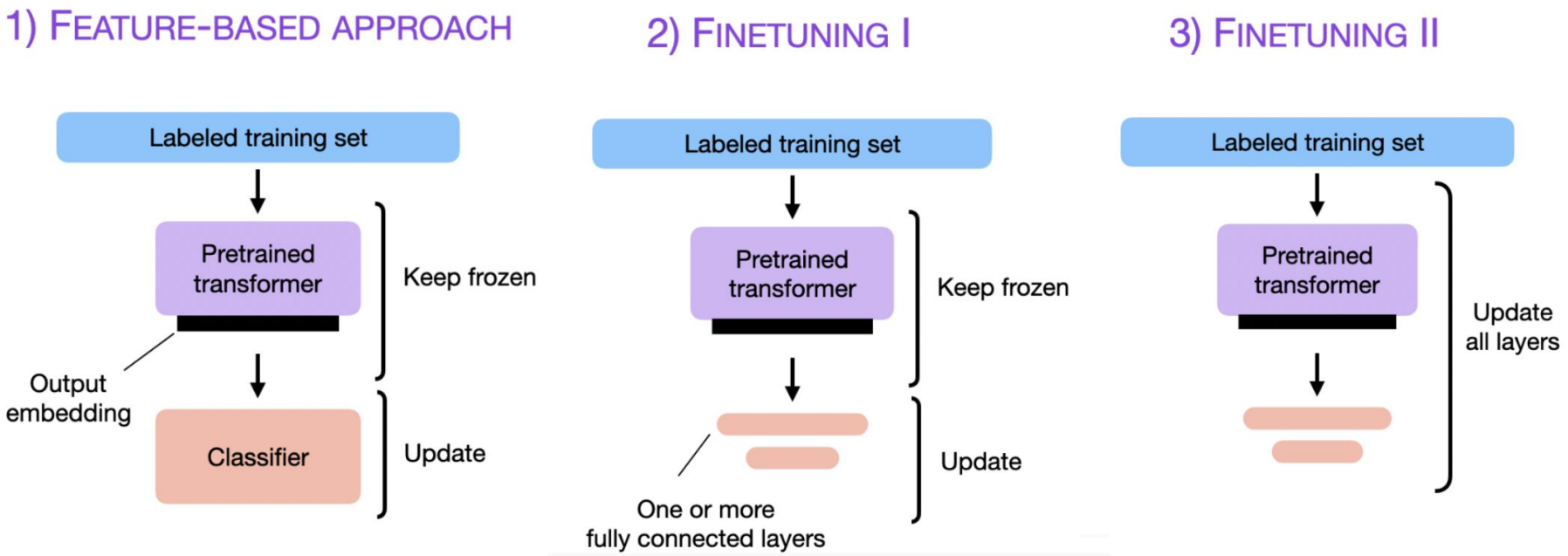
https://magazine.sebastianraschka.com/p/finetuning-large-language-models

Fine-tune model via supervised learning



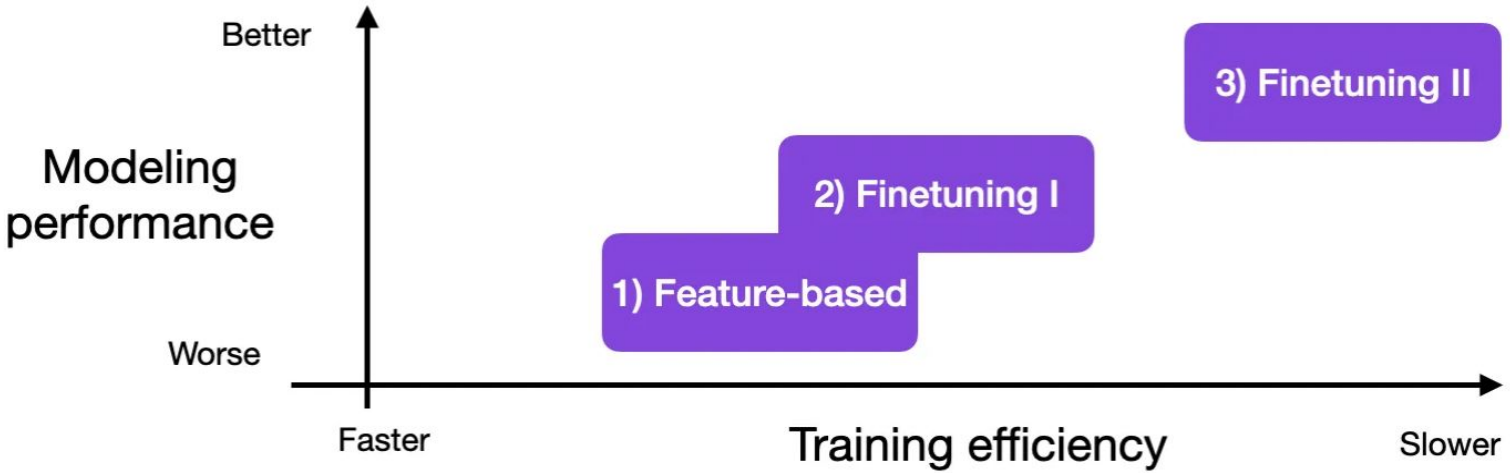
https://magazine.sebastianraschka.com/p/finetuning-large-language-models

Fine-tune model via supervised learning



https://magazine.sebastianraschka.com/p/finetuning-large-language-models

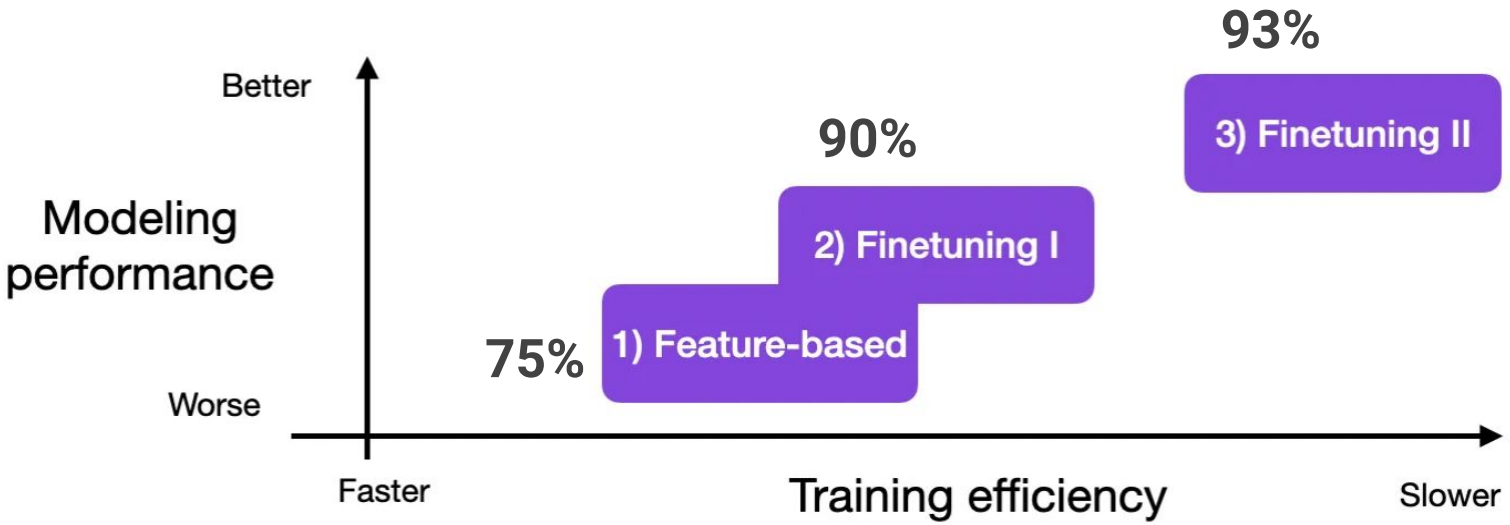
Comparing LLM fine-tuning methods



Rule-of-thumb computational and modeling performance trade-offs for various approaches.

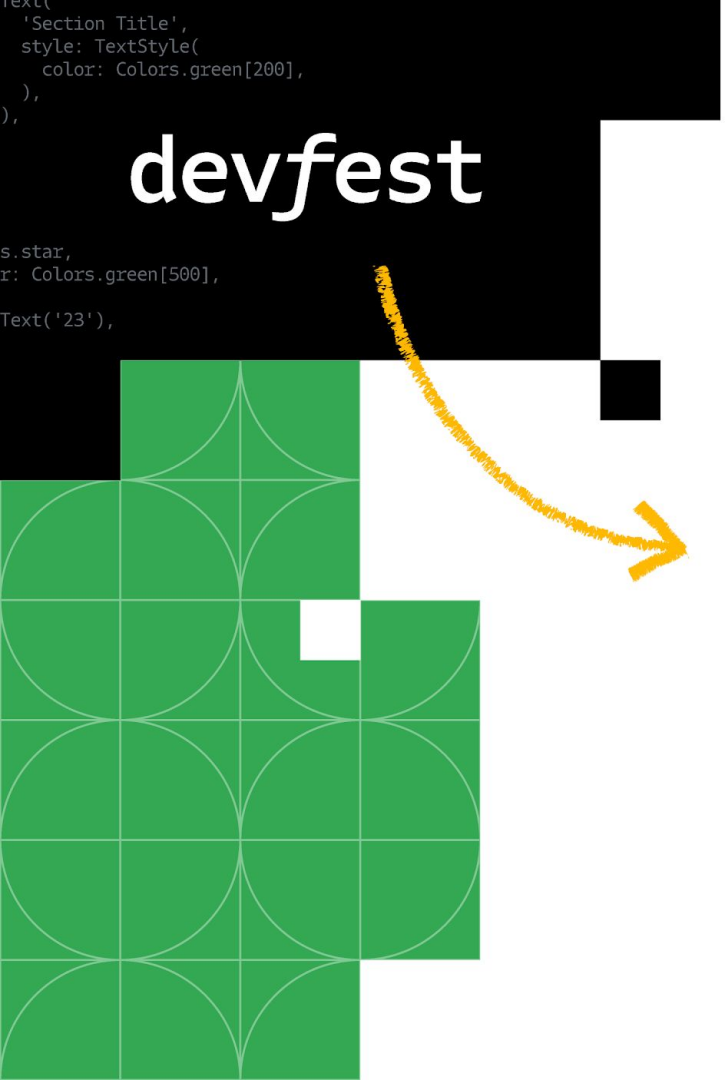
https://magazine.sebastianraschka.com/

Comparing LLM fine-tuning methods



Rule-of-thumb computational and modeling performance trade-offs for various approaches.

https://magazine.sebastianraschka.com/



Google Developer Groups

Instruction Fine-Tuning

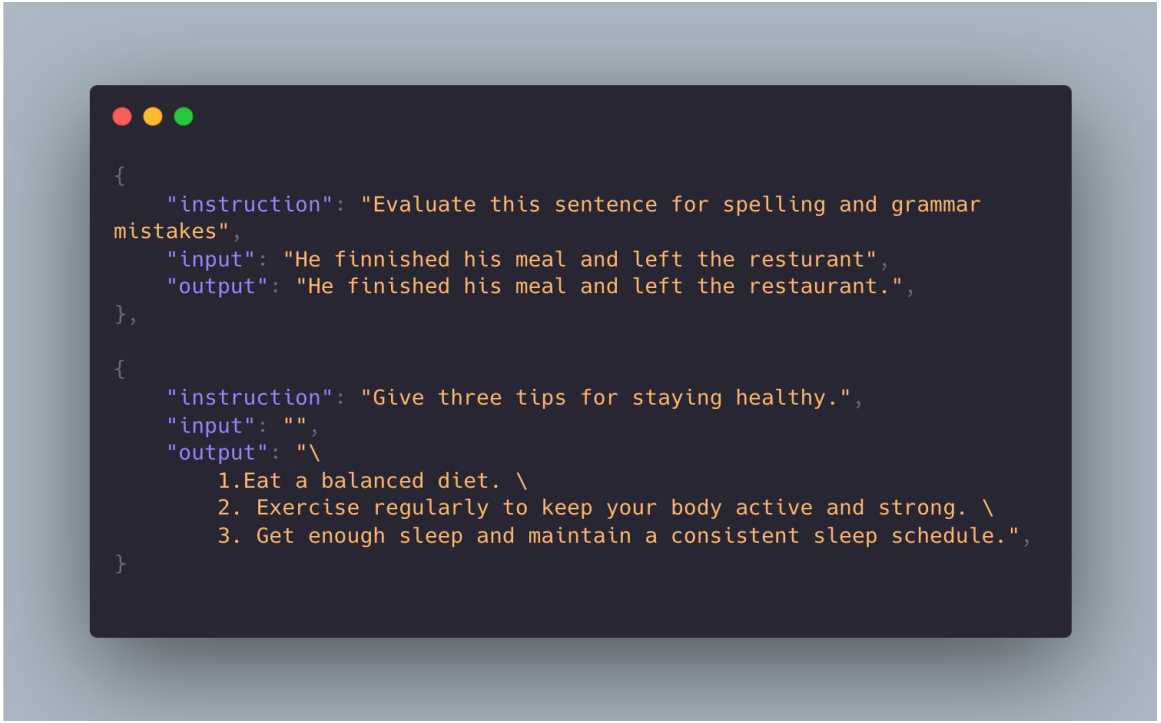
Instruction Fine-Tuning

2

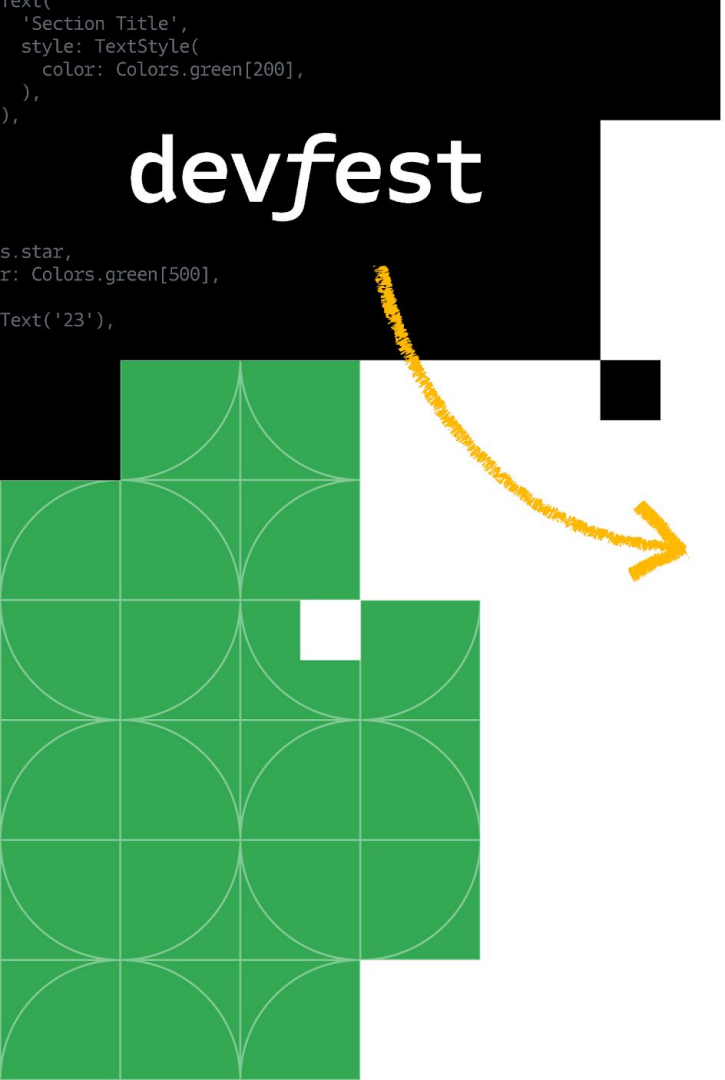
Supervised finetuning

More next-token prediction

Usually 1k-50k instruction-response pairs



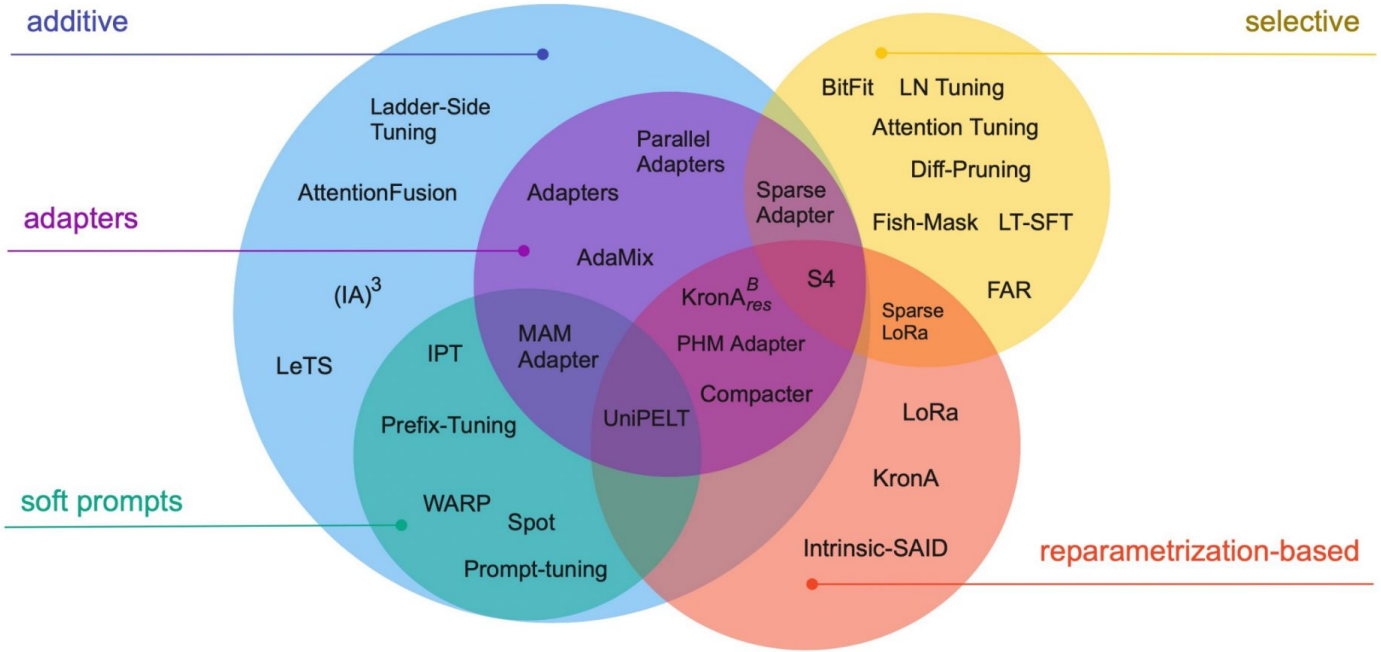
```
{
    "instruction": "Evaluate this sentence for spelling and grammar
mistakes",
    "input": "He finnished his meal and left the resturant",
    "output": "He finished his meal and left the restaurant.",
},

{
    "instruction": "Give three tips for staying healthy.",
    "input": "",
    "output": "\
        1.Eat a balanced diet. \
        2. Exercise regularly to keep your body active and strong. \
        3. Get enough sleep and maintain a consistent sleep schedule.",
}
```

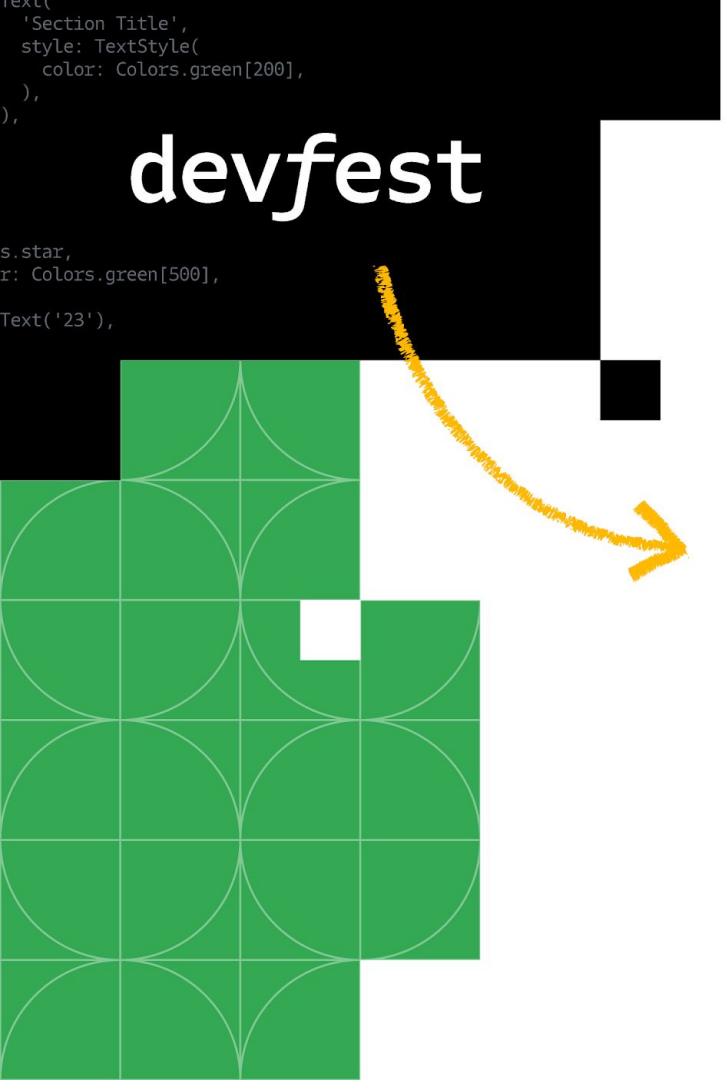


Google Developer Groups

Parameter-efficient fine-tuning

Parameter-efficient fine-tuning



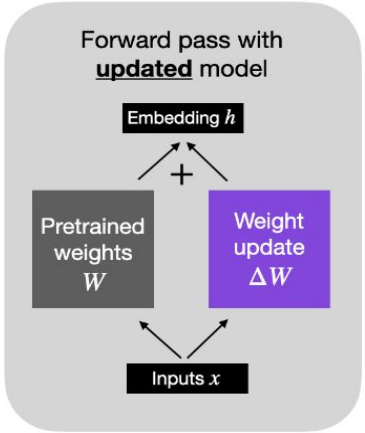


Google Developer Groups

Low-rank adaptation (LoRA)

Low-rank adaptation (LoRA)

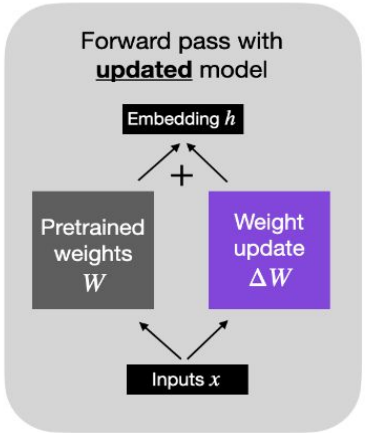
Alternative formulation (regular finetuning)



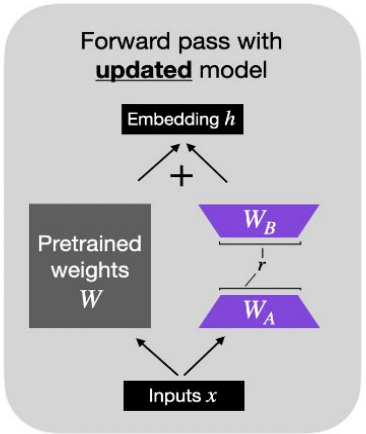
https://magazine.sebastianraschka.com/

Low-rank adaptation (LoRA)

Alternative formulation (regular finetuning)



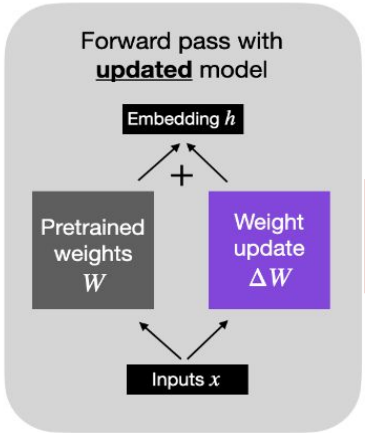
LoRA weights, W_A and W_B, represent ΔW



https://magazine.sebastianraschka.com/

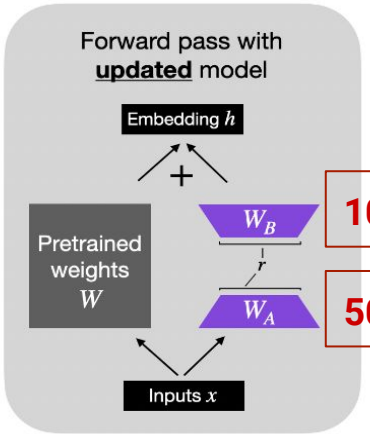
Low-rank adaptation (LoRA)

Alternative formulation (regular finetuning)



5000 X 1000
= 5,000,000

LoRA weights, W_A and W_B, represent ΔW



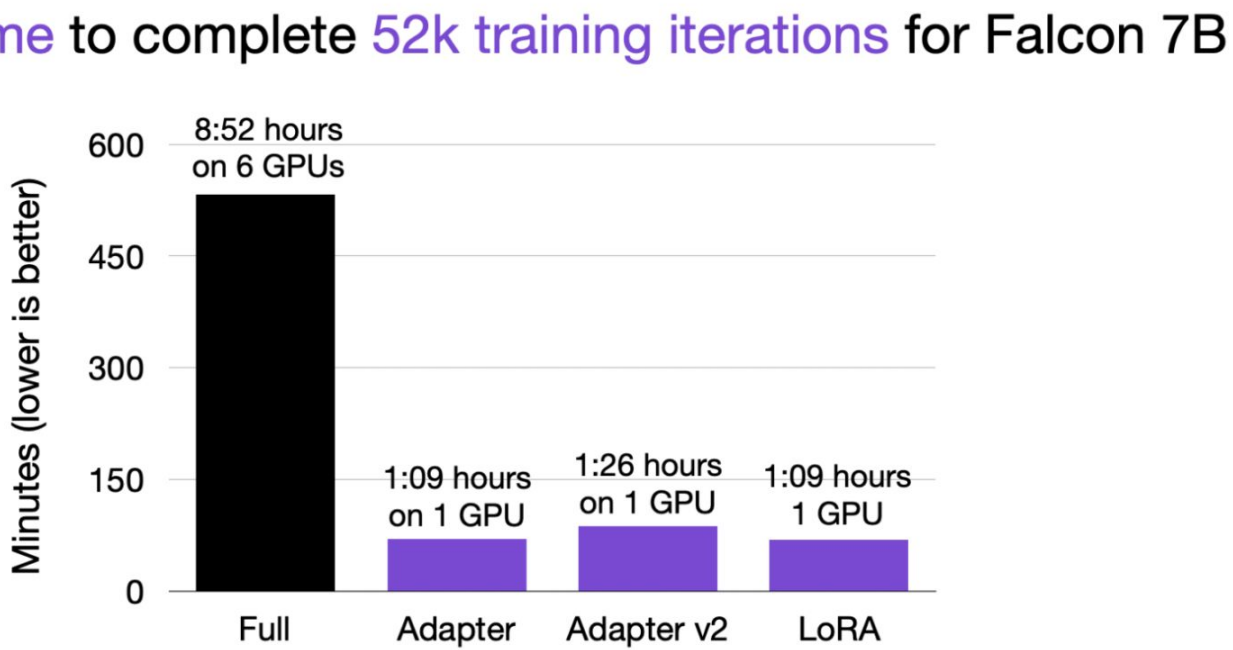
10 x 1000

5000 x 10

= 60,000

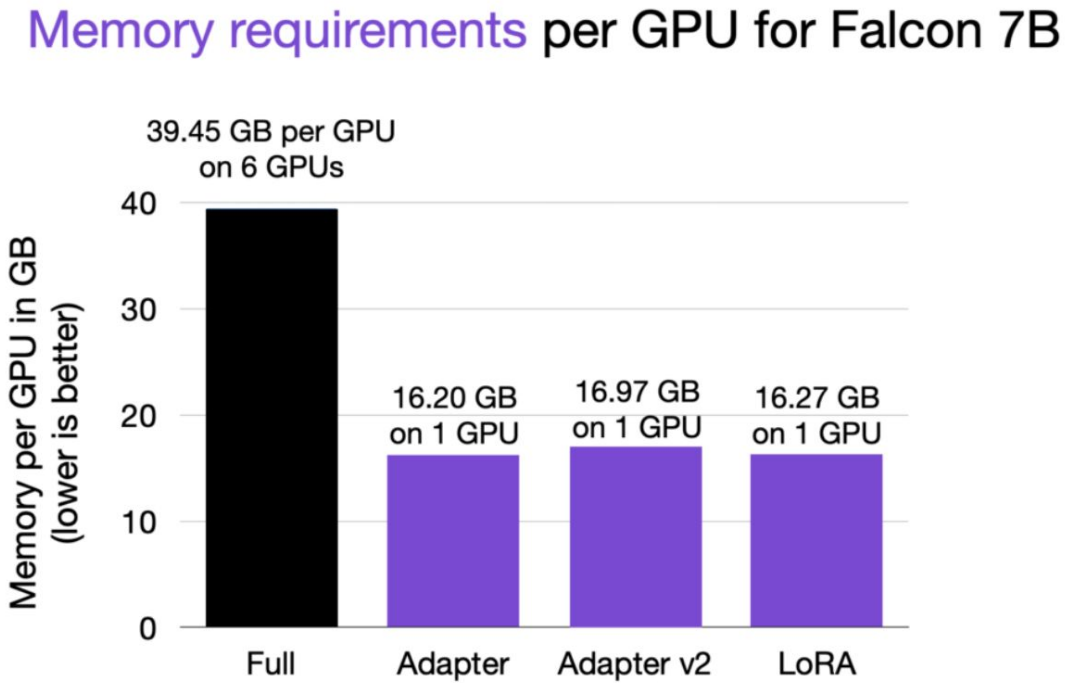
https://magazine.sebastianraschka.com/

Low-rank adaptation (LoRA)



https://magazine.sebastianraschka.com/

Low-rank adaptation (LoRA)



https://magazine.sebastianraschka.com/

Google Developer Groups

Open Source LLMs

LLaMA

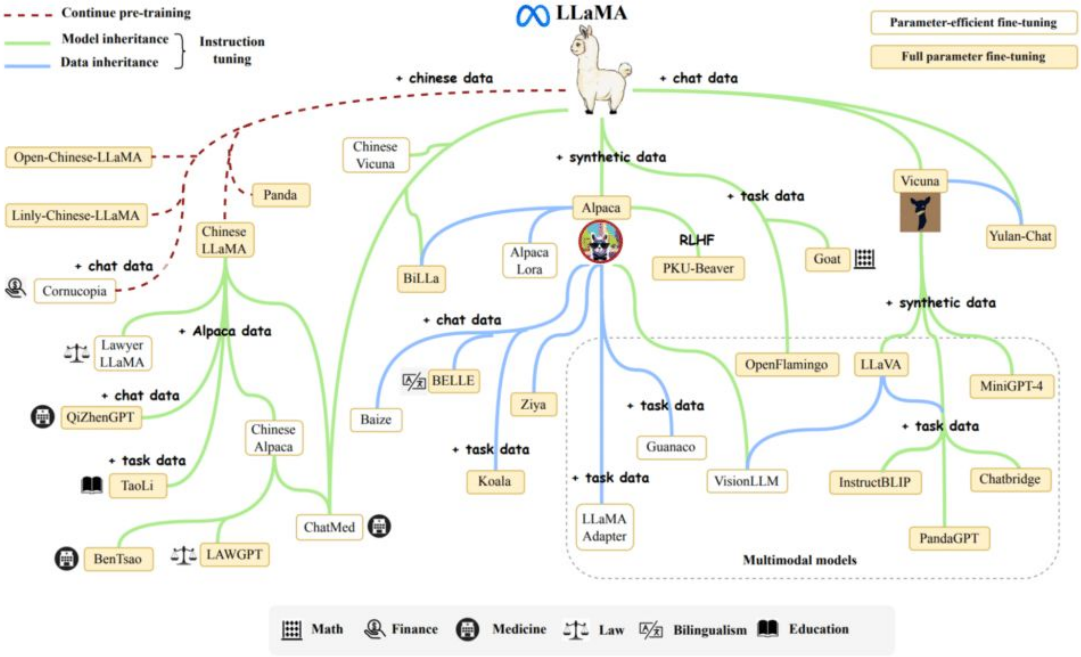
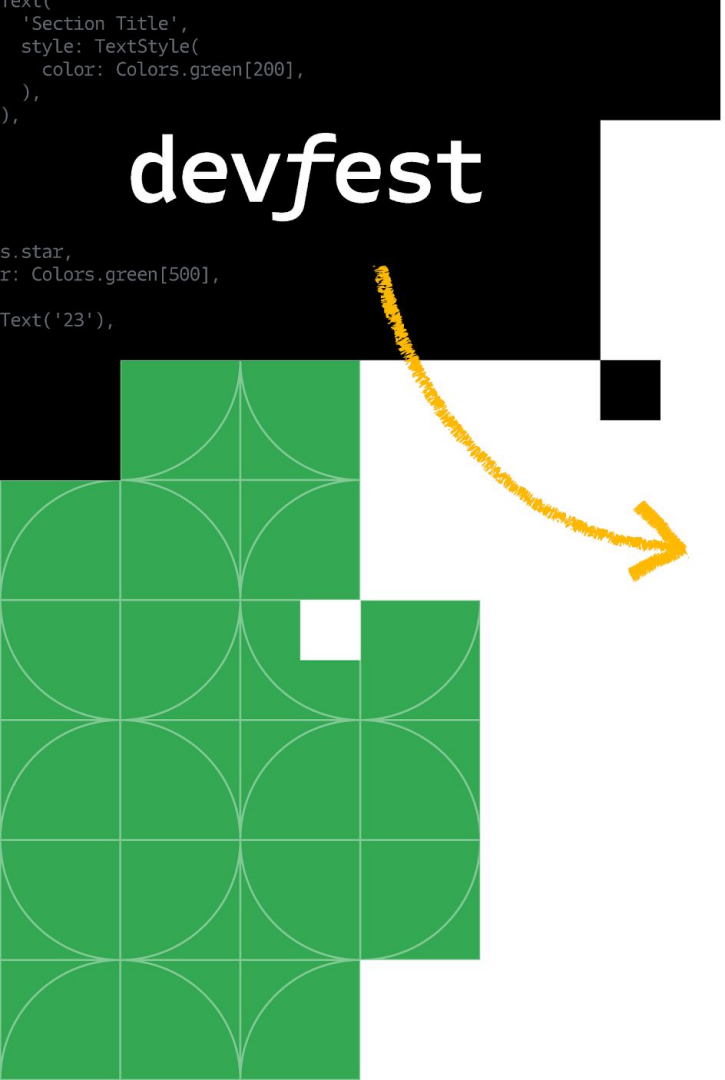


Fig. 4: An evolutionary graph of the research work conducted on LLaMA. Due to the huge number, we cannot include all the LLaMA variants in this figure, even much excellent work. To support incremental update, we share the source file of this figure, and welcome the readers to include the desired models by submitting the pull requests on our GitHub page.

Code Please

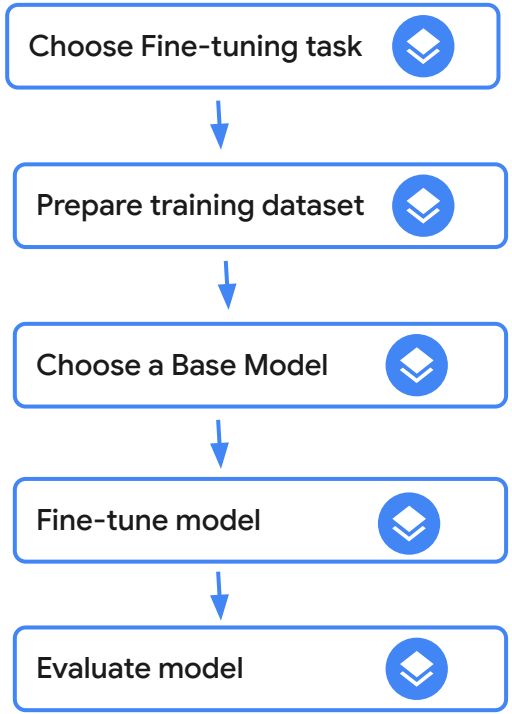

YEAH, IF YOU COULD JUST SHOW ME SOME CODE
THAT'D BE GREAT
memegenerator.net



Google Developer Groups

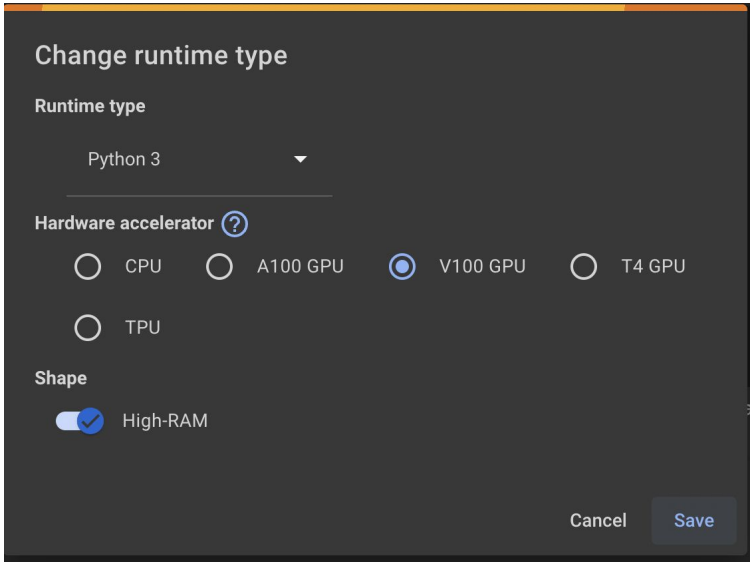
Fine-tuning LLaMA-2 on Personal Whatsapp data

5 Steps



Input	Output

Setup




```
[1]  pip3 install accelerate peft bitsandbytes transformers trl
```

Import Required Packages

```
import os
import torch
from datasets import load_dataset
from transformers import (
    AutoModelForCausalLM,
    AutoTokenizer,
    BitsAndBytesConfig,
    HfArgumentParser,
    TrainingArguments,
    pipeline,
    logging,
)
from peft import LoraConfig, PeftModel
from trl import SFTTrainer
```

Prepare dataset

```
[ ]  formatter = ChatDatasetFormatter('chat_dataset.json', 'Aadi', 'chat_dataset.csv')
     formatter.prepare_dataset()
```


Context	Reply
Abhijeet Kerla: What are you working on as res...	Working on Frozen adapters in Bert paper How's...
Abhijeet Kerla: What are you working on as res...	What was the issue?
Abhijeet Kerla: What are you working on as res...	Yup About to message you

Model setup

```
# Load base model

model = AutoModelForCausalLM.from_pretrained(
    pretrained_model_name_or_path="NousResearch/Llama-2-7b-chat-hf",
    quantization_config=quant_config,
    device_map={"": 0},
)
model.config.pretraining_tp = 1
```

Trainer Setup

```
[ ]  # Load LoRA configuration
     peft_args = LoraConfig(
         lora_alpha=16,
         lora_dropout=0.1,
         r=64,
         bias="none",
         task_type="CAUSAL_LM",
     )
```

```
# Set supervised fine-tuning parameters
trainer = SFTTrainer(
    model=model,
    train_dataset=dataset,
    peft_config=peft_args,
    dataset_text_field="text",
    max_seq_length=None,
    tokenizer=tokenizer,
    args=training_params,
    packing=False,
)
```

Train

```
[ ]  # Train model
     trainer.train()
```

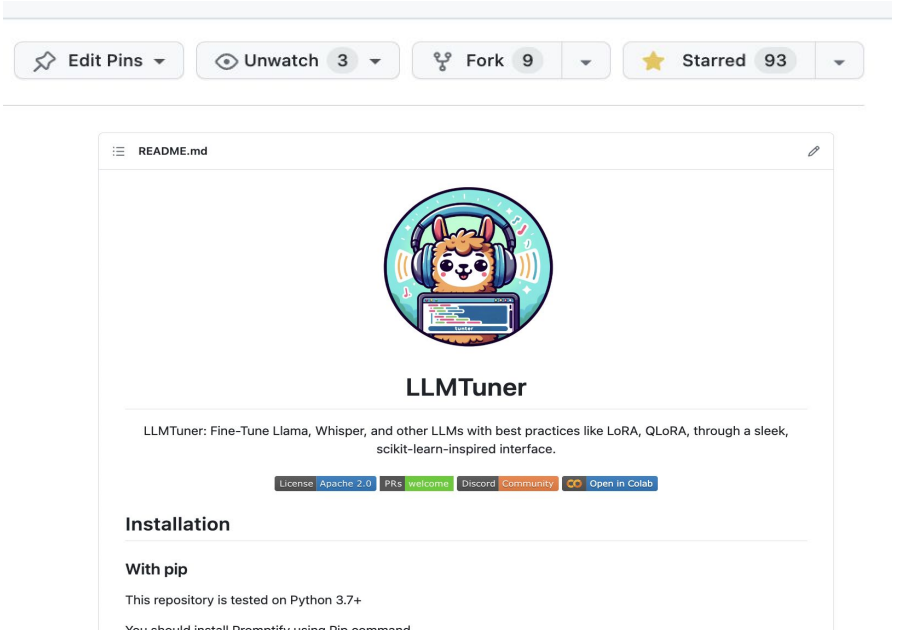
```
warnings.warn(
```

[187/250 05:43 < 01:56, 0.54 it/s, Epoch 0.74/1]

Step	Training Loss
25	1.408300
50	1.656600
75	1.213100
100	1.443900
125	1.176500
150	1.366400
175	1.173500

LLMtuner

Scan me!



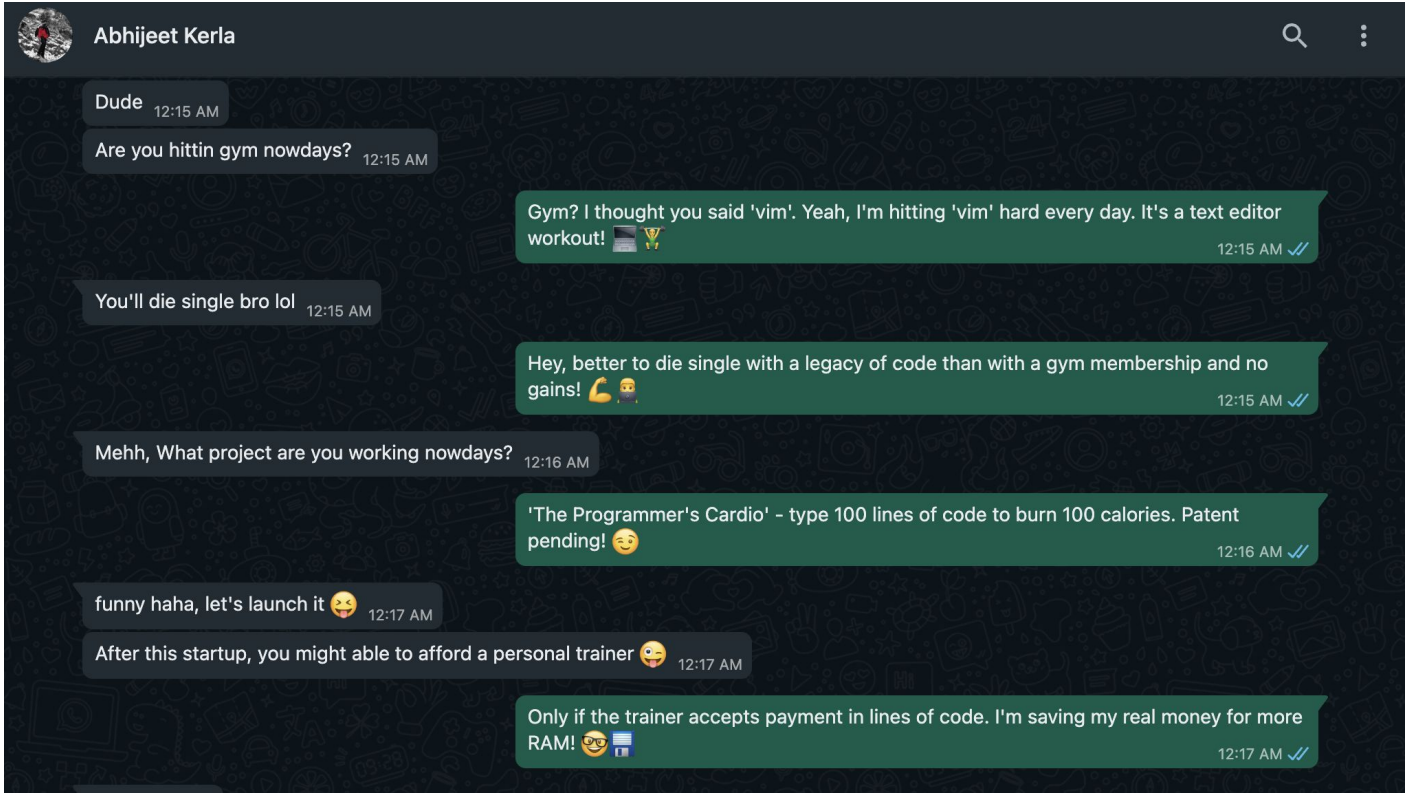
https://github.com/promptslab/LLMtuner/

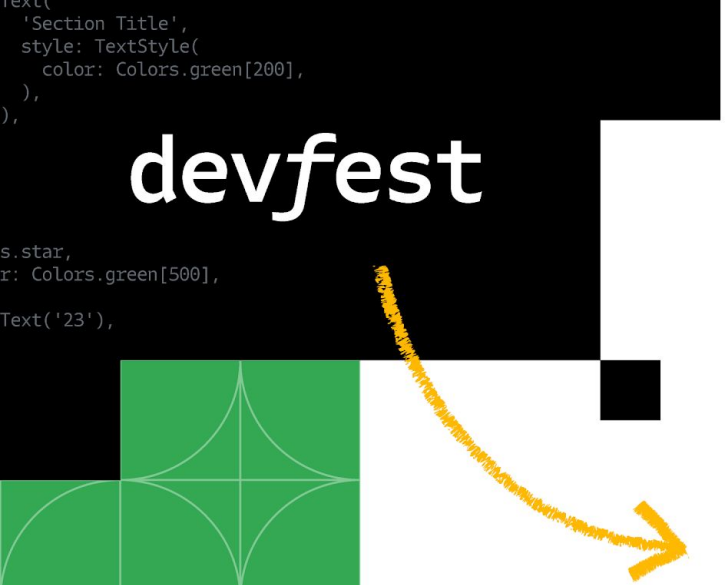
Google Developer Groups

Google Developer Groups

Inference

LLaMA is talking behalf of me

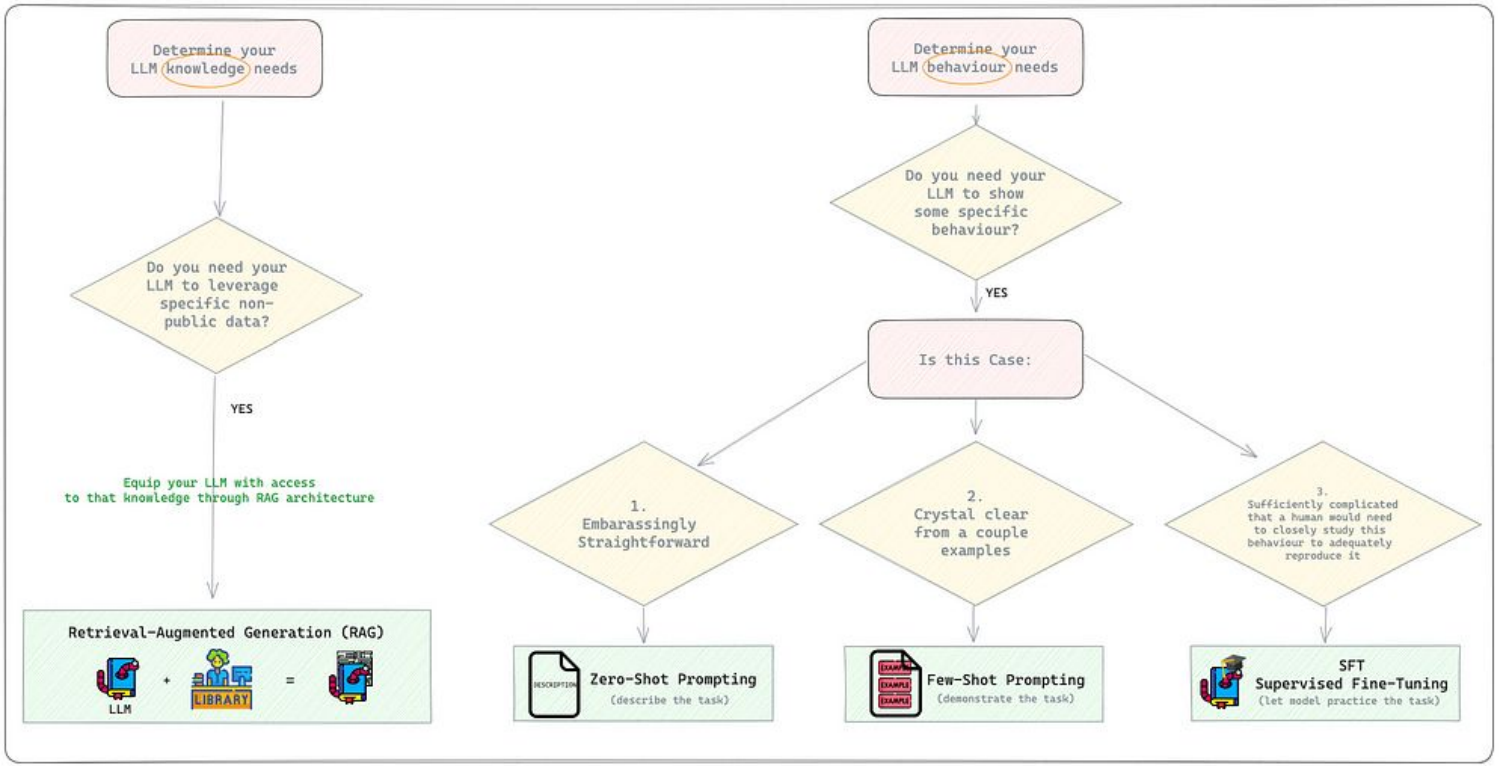


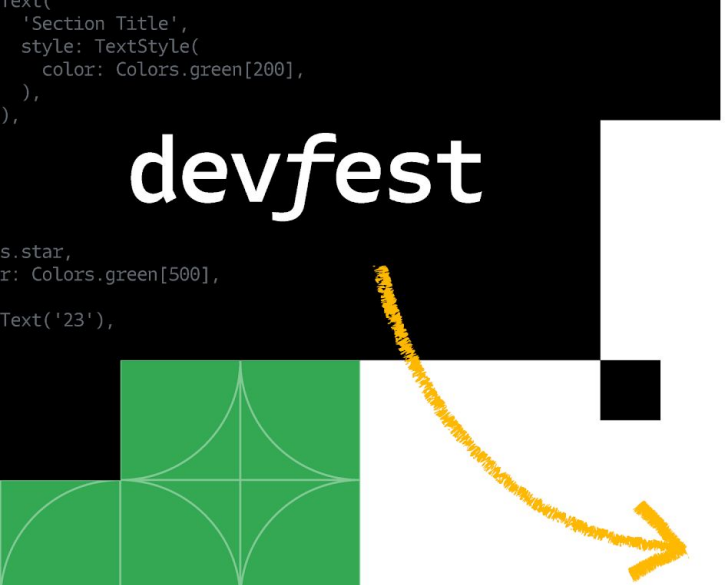


Google Developer Groups

To fine-tune or Not?

LLaMA is talking behalf of me





Google Developer Groups

Best Practices

Best Practices

Instead of fine-tuning an LLM as a first approach, try prompt Engg instead

Start with a pre-trained & Compact Model

Work more on the data

Utilize the appropriate evaluation metric

Use Low Ranking Adaptation (LoRA) & Quantized LoRA (QLoRA) etc

Use deepspeed for Distributed, Effective, and Efficient Training

Thank you!

Let's get connected on Twitter! I am @adityaura