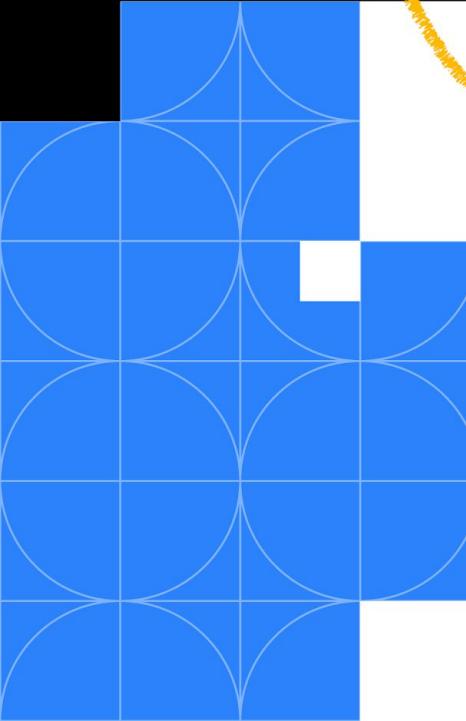


```
text  
'Section Title',  
style: TextStyle(  
  color: Colors.blue[200],  
,  
)  
,
```

# devfest

```
s.star,  
r: Colors.blue[500],
```

```
Text('23'),
```



# Fine-Tuning Open-Source LLMs: Best Practices



@aadityaura



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

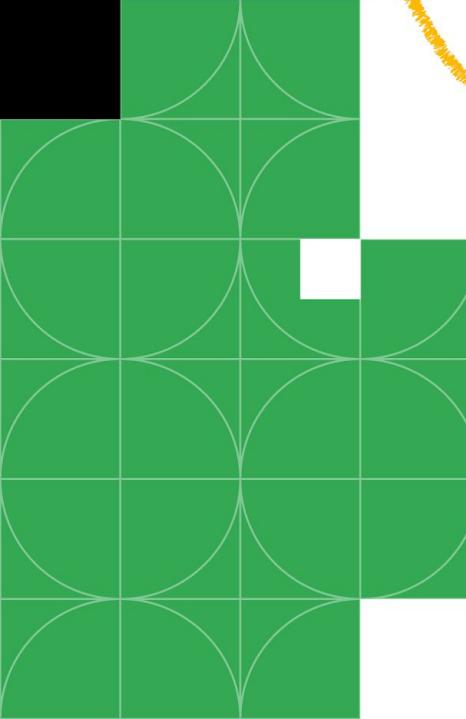
- Senior AI Research Engineer [@ Saama AI Research Lab](#)
- Research interests involve  
Representation Learning on Graphs and Manifolds
- Generative Modeling, MLOps, Signal Processing  
and their applications in Healthcare data

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],
```

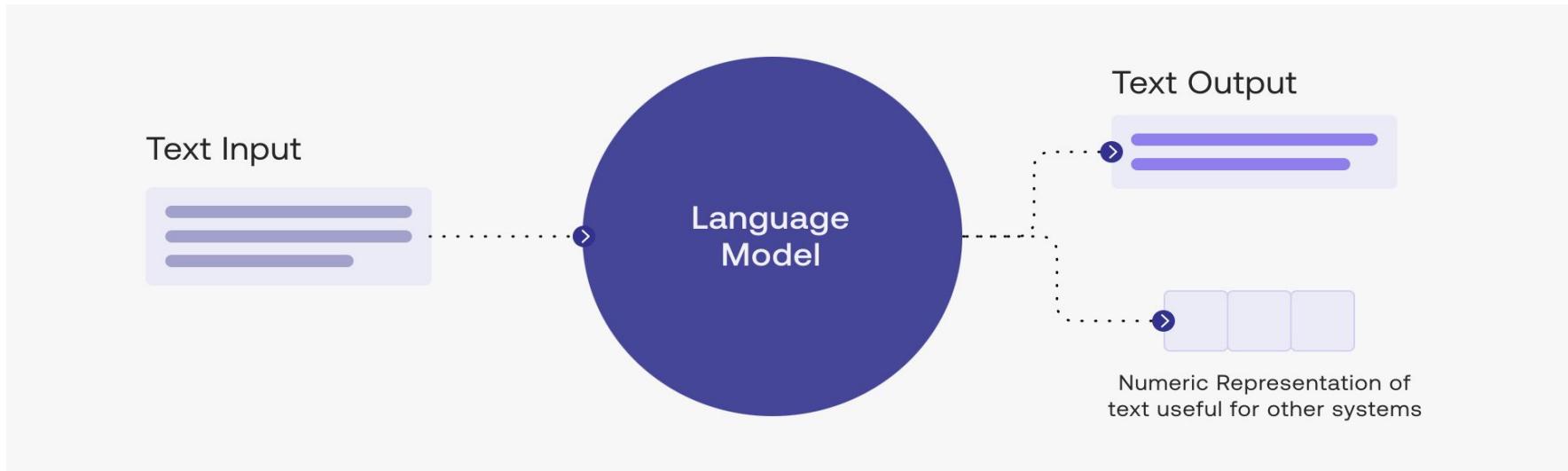
```
Text('23'),
```



## Large Language Models(LLMs)

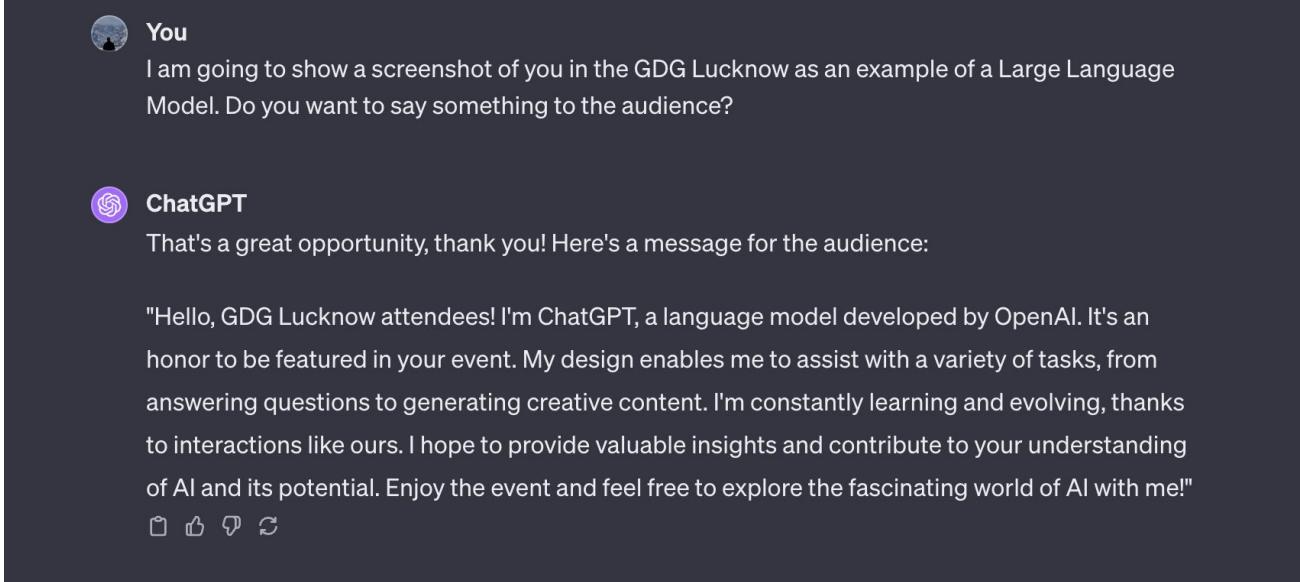
# Understanding Large Language Models

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



You

I am going to show a screenshot of you in the GDG Lucknow as an example of a Large Language Model. Do you want to say something to the audience?

ChatGPT

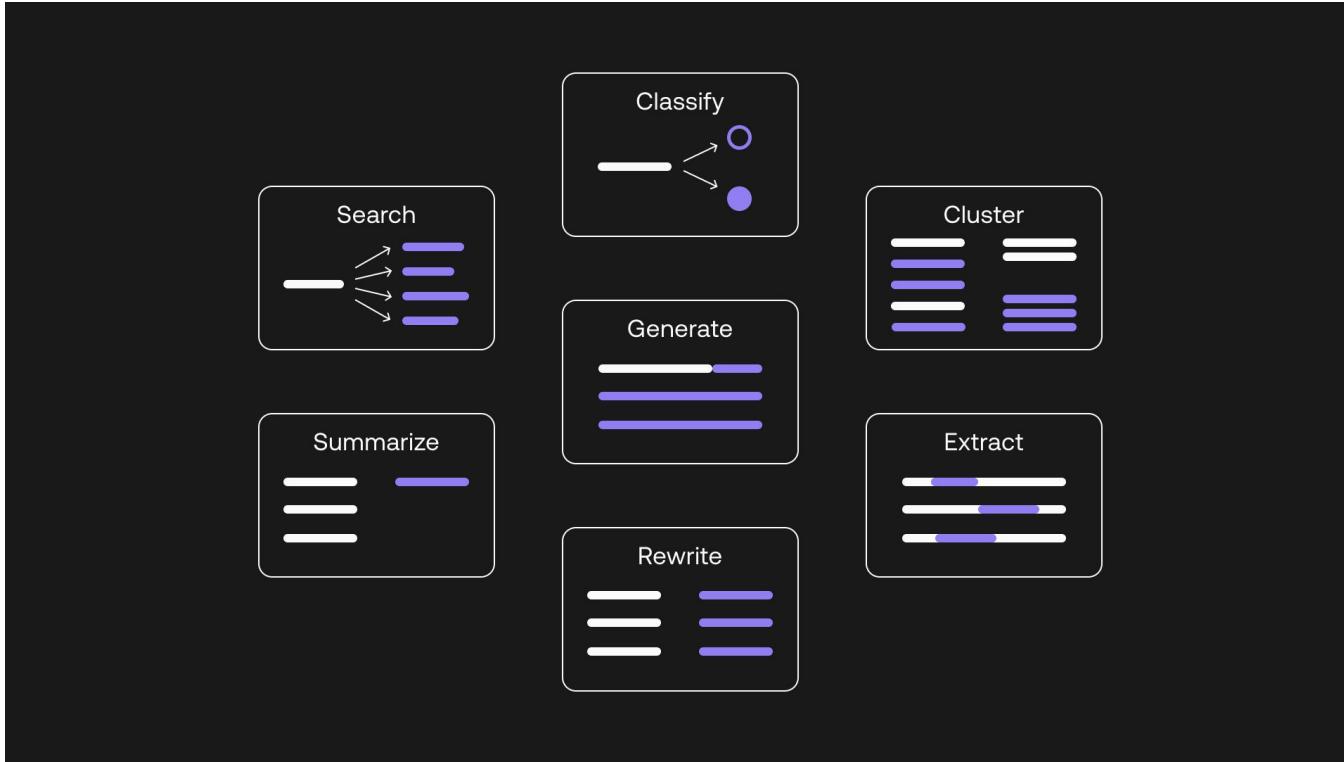
That's a great opportunity, thank you! Here's a message for the audience:

"Hello, GDG Lucknow attendees! I'm ChatGPT, a language model developed by OpenAI. It's an honor to be featured in your event. My design enables me to assist with a variety of tasks, from answering questions to generating creative content. I'm constantly learning and evolving, thanks to interactions like ours. I hope to provide valuable insights and contribute to your understanding of AI and its potential. Enjoy the event and feel free to explore the fascinating world of AI with me!"

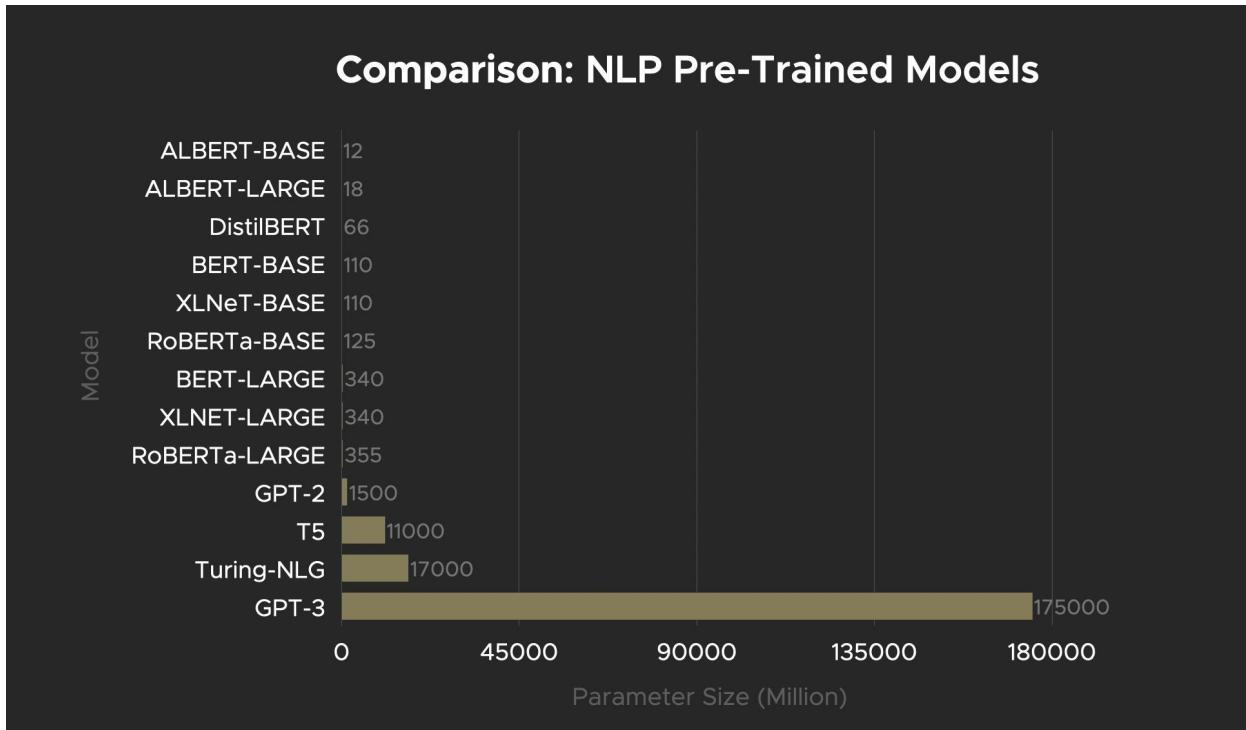
□ ↻ ⌂ ⌂

ChatGPT is one example of a Large Language Model.

# Understanding Large Language Models



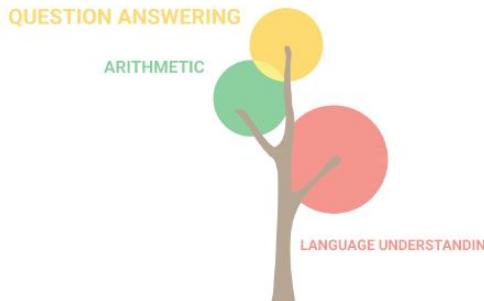
# Understanding Large Language Models



# Understanding Large Language Models



# Understanding Large Language Models

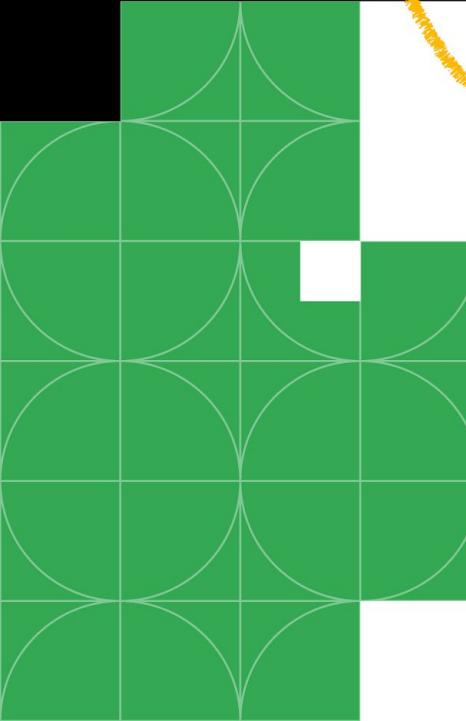


8 billion parameters

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



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



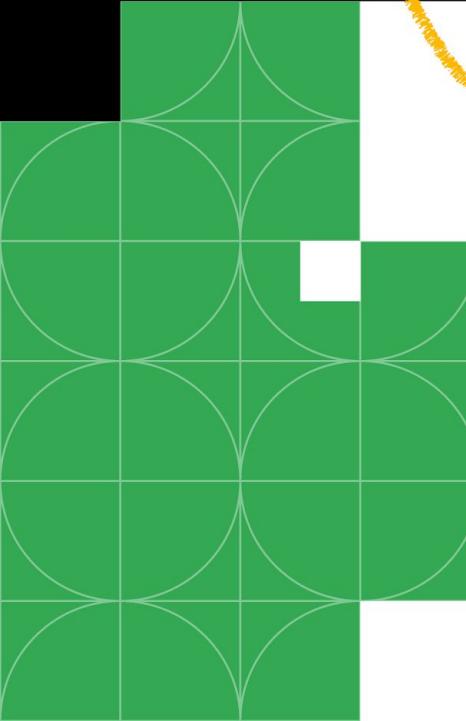
**GPT-3**

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

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

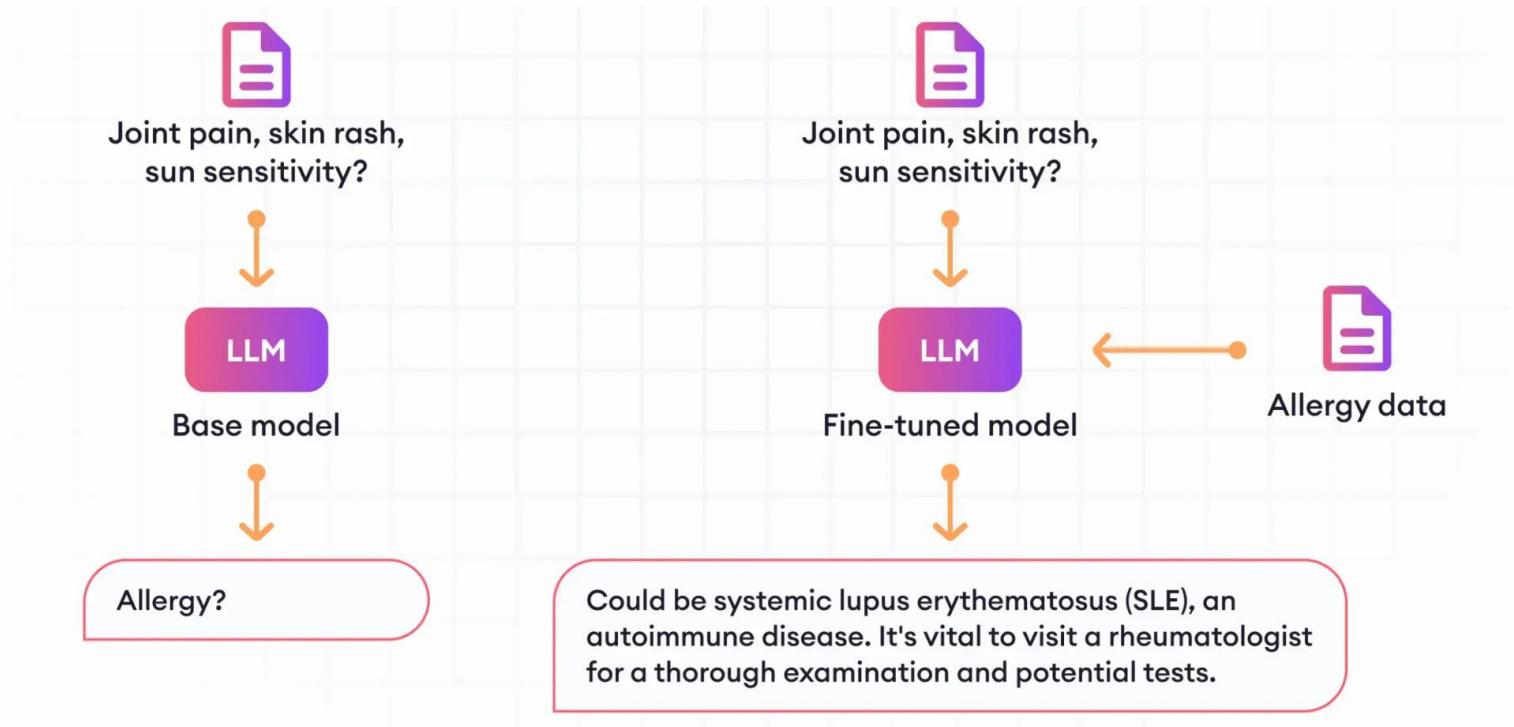
```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



Google Developer Groups

## Why We Need to Fine-Tune?

# Need of Fine-Tuning

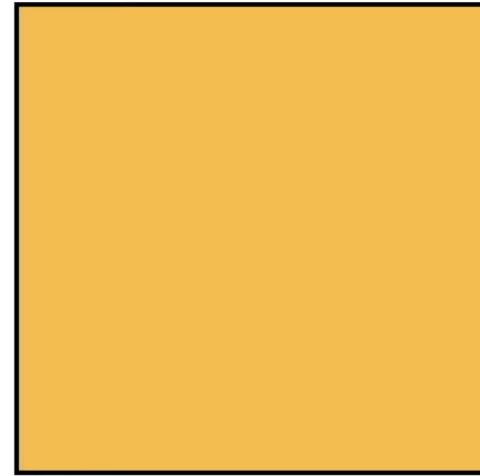


# Need of Fine-Tuning

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



**InstructGPT (1.3B)**



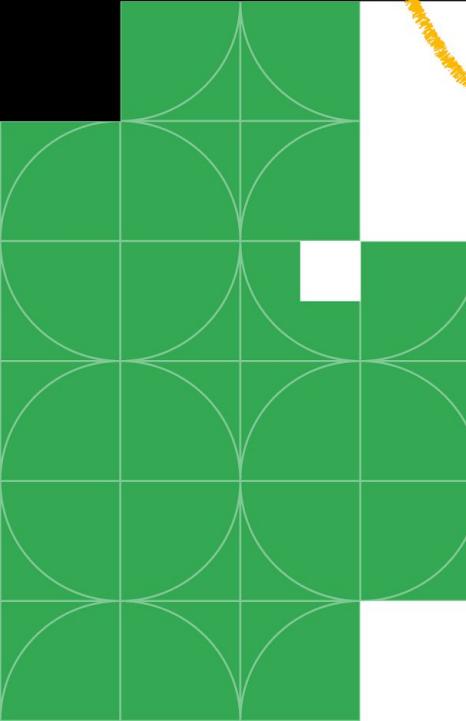
**GPT-3 (175B)**

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],
```

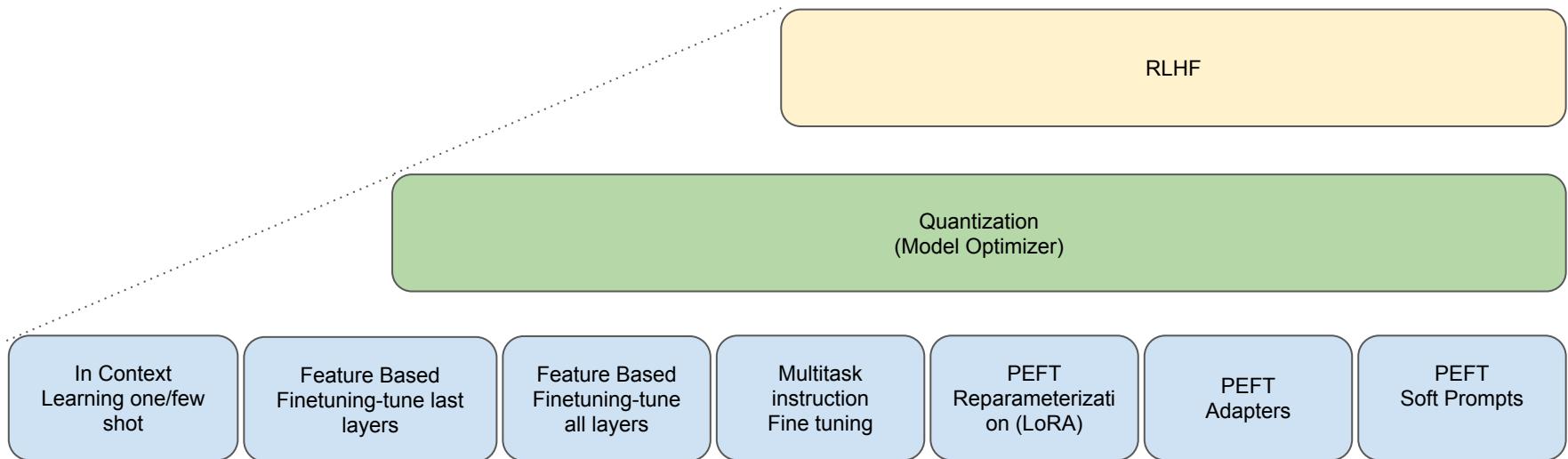
```
Text('23'),
```



Google Developer Groups

## Fine-Tuning Approaches

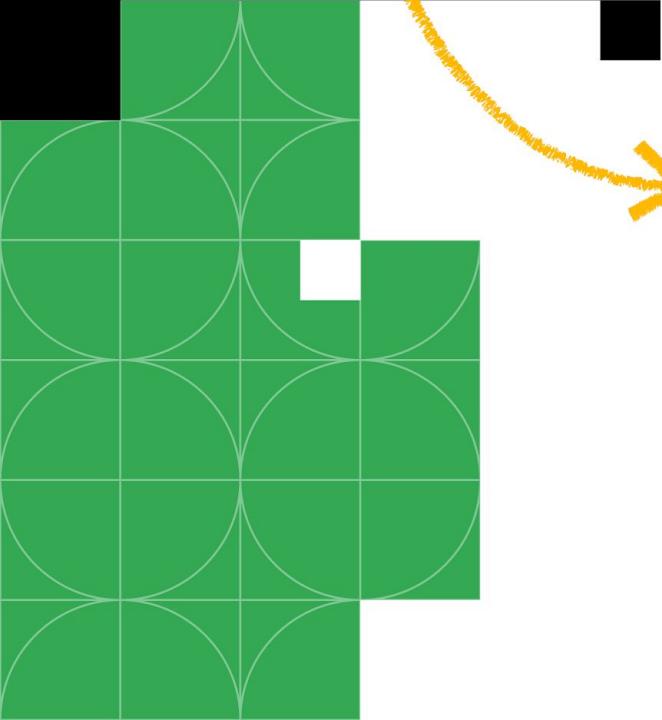
# Fine-Tuning Approaches



```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



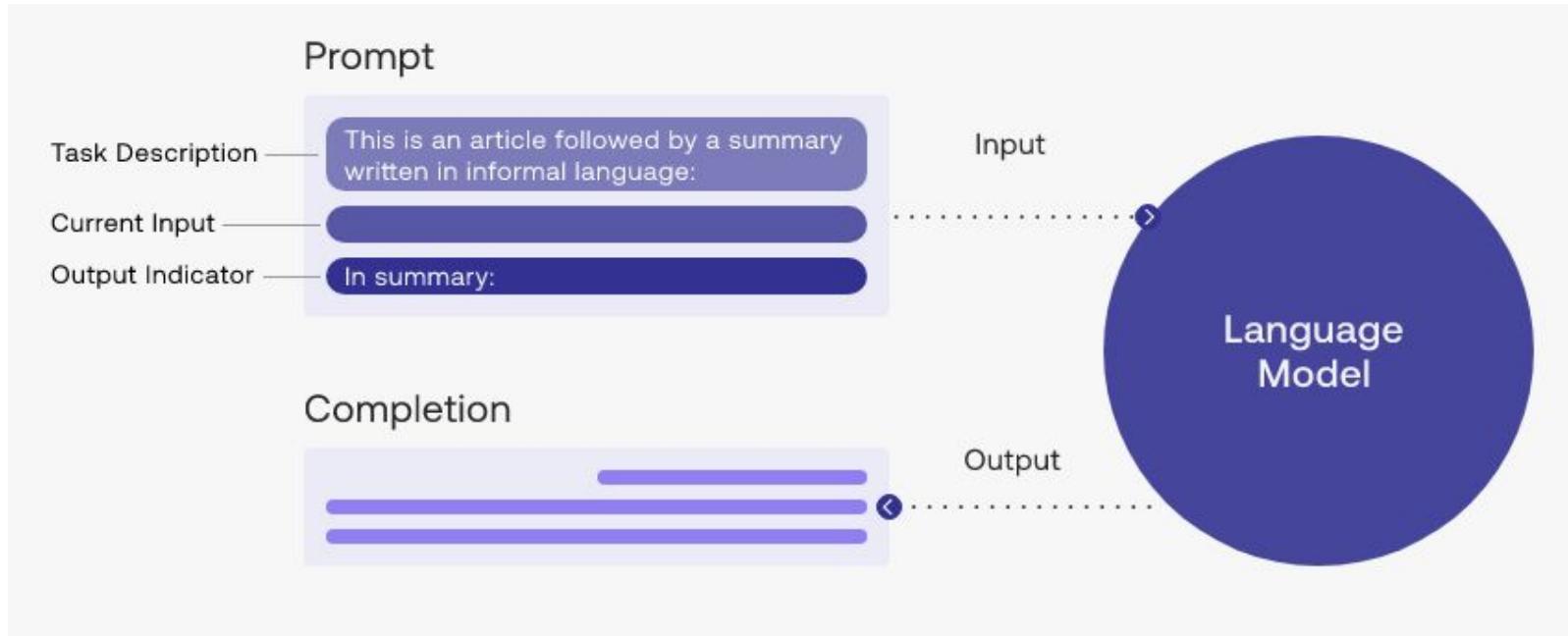
## In-Context Learning

# Fine-Tuning Approaches

Prompt  
Engineering



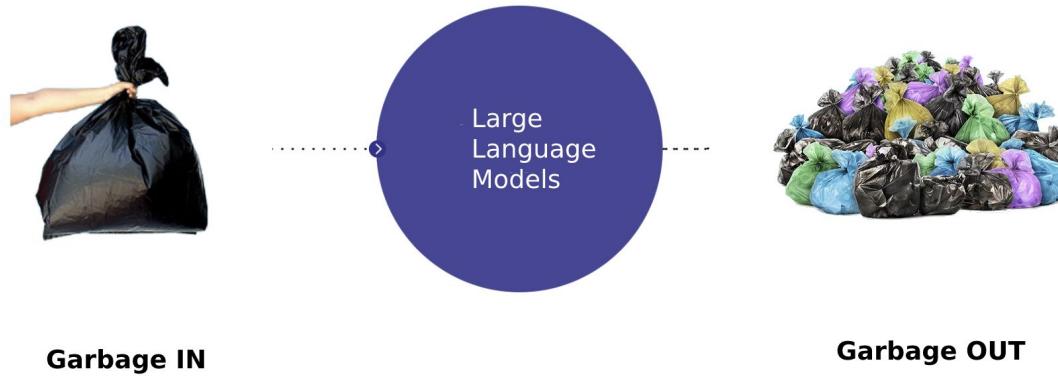
# Prompt-Engineering



Prompting is a method that allows users to interact with LLMs

# Prompt-Engineering

**Prompt plays a key role in LLMs**



# Types of prompting

The diagram shows a grey rounded rectangle containing the ChatGPT Online logo (a green circle with a white swirl) and the text "ChatGPT Online". Below it is a yellow speech bubble containing the question "What is the sum of 12 and 42 in words?". To the right is a green speech bubble containing the response "The sum of 12 and 42 in words is fifty-four." At the bottom left is a lightbulb icon with the text "Zero-shot prompting" next to it. Below the lightbulb are two bullet points: "- Instant predictions" and "- No additional training".

ChatGPT  
Online

What is the sum of 12  
and 42 in words?

The sum of 12 and 42 in  
words is fifty-four.

Zero-shot  
prompting

- Instant predictions
- No additional training

# Types of prompting

The image shows a screenshot of the ChatGPT Online interface. At the top, there is a green circular logo with a white, interlocking knot-like pattern. Below it, the text "ChatGPT" is written in a large, bold, black font, with "Online" in a smaller, regular black font underneath. A grey speech bubble shape surrounds the top portion of the interface. Inside this bubble, the AI's response to three math questions is displayed in yellow speech bubbles:

- Q: What is the sum of 2 and 3?  
A: Five
- Q: What is the sum of 12 and 10000?  
A: Ten thousand and twelve
- Q: What is the sum of 12 and 42?

Below these, a green speech bubble contains the AI's reasoning and final answer:

The sum of 12 and 42 is:  
 $12 + 42 = 54$   
So, the answer is fifty-four.

**Few-shot prompting**



- Examples or templates needed
- One to five examples

# Types of prompting

The image shows a screenshot of the ChatGPT Online interface. At the top, there is a green circular icon with a white AI-like symbol inside. Below it, the text "ChatGPT" is written in a large, bold, black font, with "Online" in a smaller, regular black font underneath. A grey speech bubble contains a question: "Q: What is the sum of 14 and 18?". Below the question, the AI's response is shown in a yellow speech bubble: "A: To sum 14 and 18, add 8 and 4 to give 12, carry over 1. Add the carried over 1 to 1 and 1. This sums to 31." Another question, "Q: What is the sum of 32 and 49?", is shown in a yellow speech bubble. The AI's detailed reasoning for this sum is provided in a green speech bubble: "To sum 32 and 49, start by adding the ones place, which gives  $2 + 9 = 11$ . Write down the 1 and carry over the 1. Then add the tens place, which gives  $3 + 4 +$  the carried over 1, for a total of 8." Below this reasoning, the AI concludes with "Therefore:  $32 + 49 = 81$ ".

**Chain-of-thought prompting**



- Breaks down problems
- More interpretable answers

# Types of prompting

## LLM Model Parameters to Control

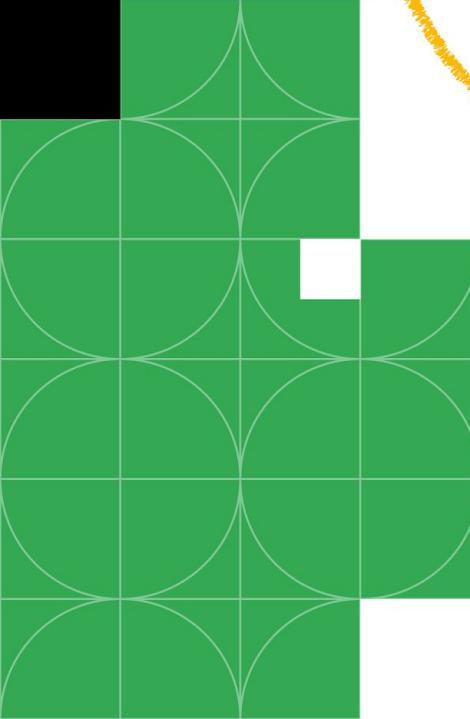
- Stopping criteria
  - Number of tokens - maximum number of tokens to generate
  - Stop sequences - character or the sequence used to separate the training examples
- Repetition Reduction
  - Frequency penalty - penalizes based on the frequency of the word in the preceding text
  - Presence penalty - penalizes based on the presence of the word

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

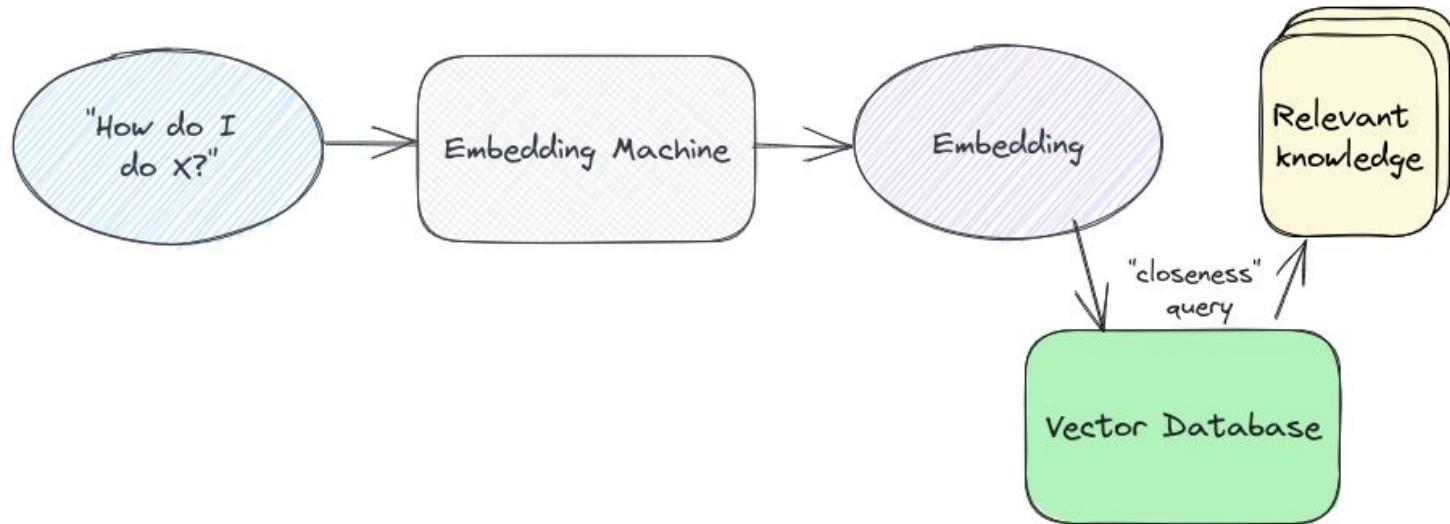
```
s.star,  
r: Colors.green[500],
```

```
Text('23'),
```

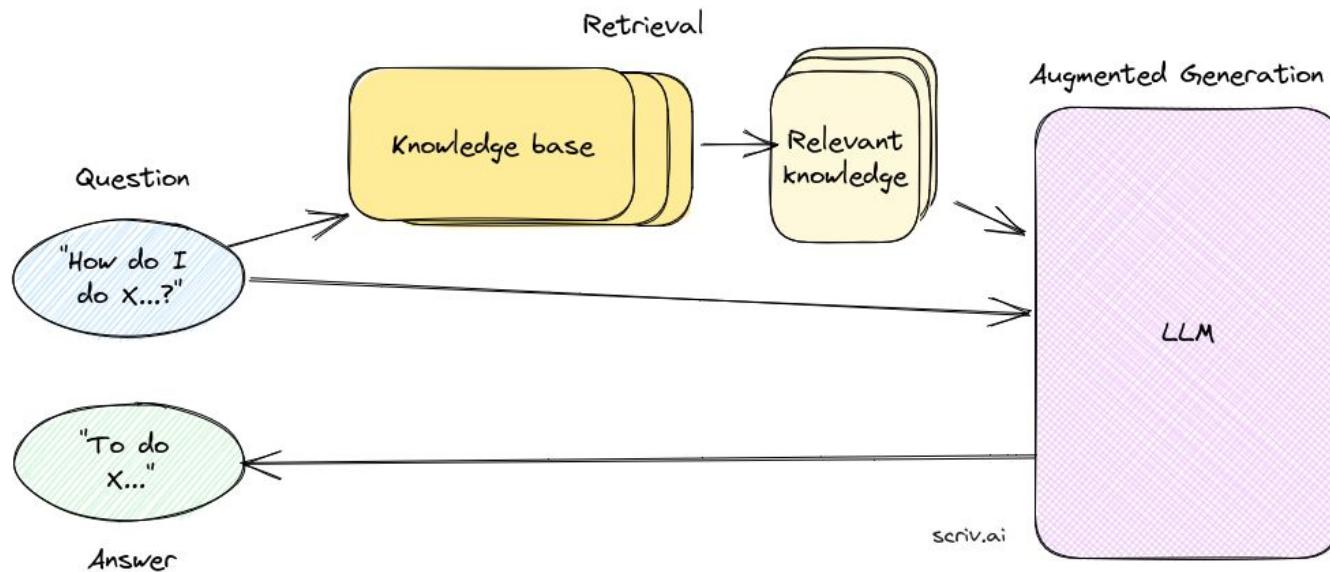


## **Retrieval Augmented Generation (RAG)**

# Retrieval Augmented Generation (RAG)

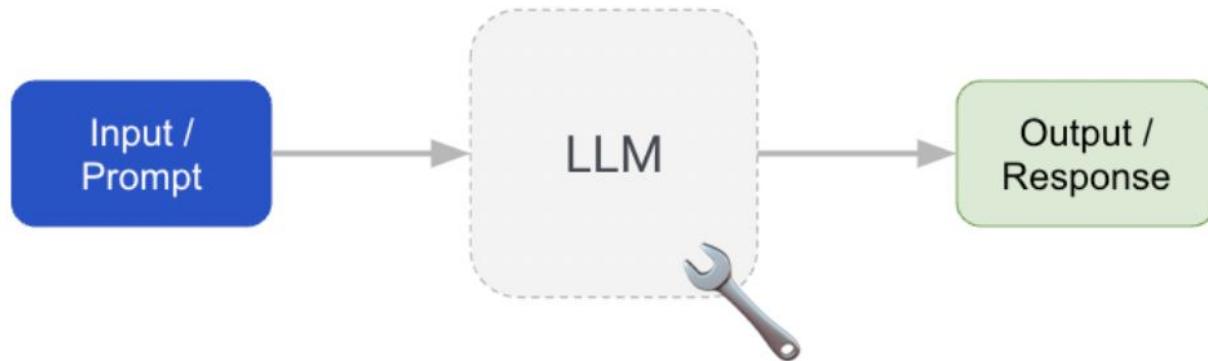


# Retrieval Augmented Generation (RAG)



# Fine-Tuning Approaches

Fine-tuning

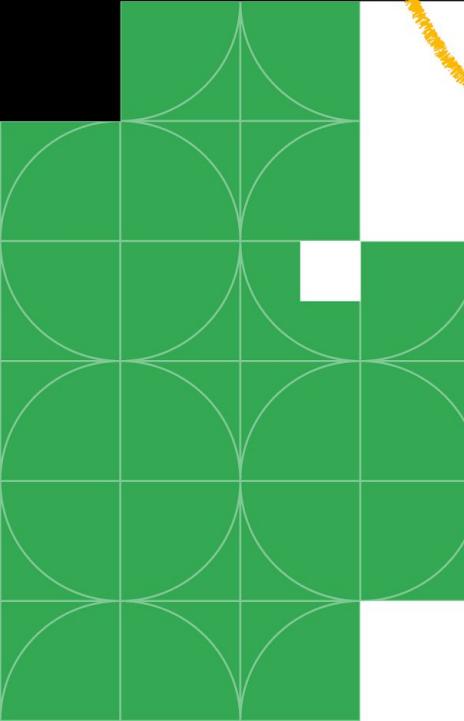


```
text:  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],
```

```
Text('23'),
```



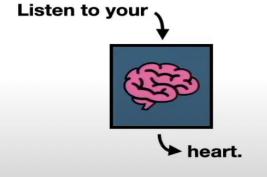
## Fine-Tuning

# Types of Fine-Tuning



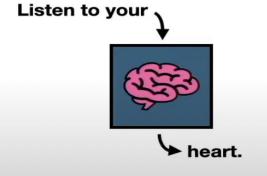
## Self-Supervised

# Types of Fine-Tuning



## Self-Supervised

# Types of Fine-Tuning



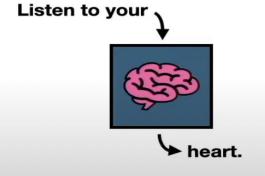
The cat

## Self-Supervised

# Types of Fine-Tuning



Input	Output



The cat

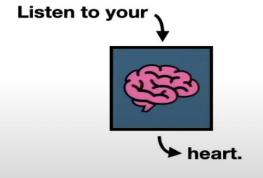
**Self-Supervised**

**Supervised**

# Types of Fine-Tuning



Input	Output



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

**Output:** John F. Kennedy

The cat

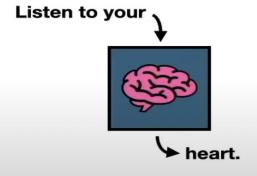
**Self-Supervised**

**Supervised**

# Types of Fine-Tuning



Input	Output



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

**Output:** John F. Kennedy

The cat

""""Please answer the following question.

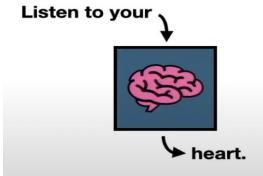
Q: {Question}

A: {Answer}""""

## Self-Supervised

## Supervised

# Types of Fine-Tuning



The cat

Input	Output

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

**Output:** John F. Kennedy

i. Supervised FT

""""Please answer the following question.

Q: {Question}

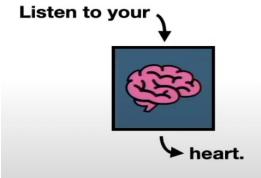
A: {Answer}""""

## 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

i. Supervised FT



""""Please answer the following question.

Q: {Question}

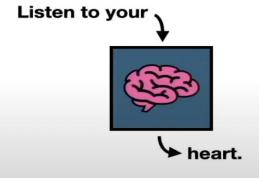
A: {Answer}""""

## 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

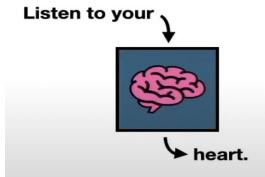
## Supervised

## Reinforcement learning

### i. Supervised FT



# 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}""""

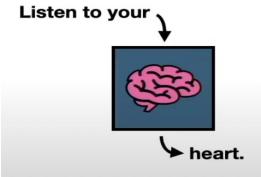
### i. Supervised FT



### ii. Train Reward Model

## 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

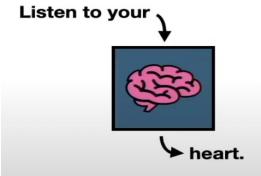
### i. Supervised FT



### ii. Train Reward Model

Prompt

# 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

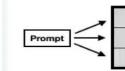
## Supervised

## Reinforcement learning

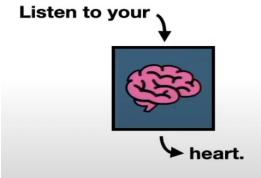
### i. Supervised FT



### ii. Train Reward Model



# 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

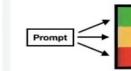
## Supervised

## Reinforcement learning

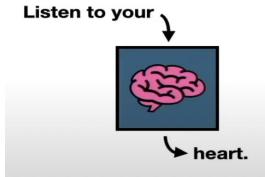
### i. Supervised FT



### ii. Train Reward Model



# 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

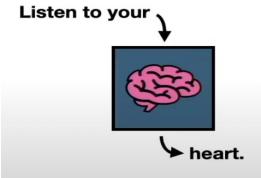
### i. Supervised FT



### ii. Train Reward Model



# 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

### i. Supervised FT



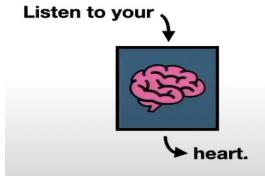
### ii. Train Reward Model



### iii. RL with PPO

Prompt

# 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

### i. Supervised FT



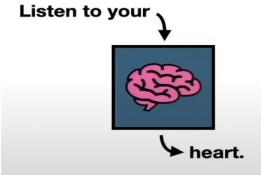
### ii. Train Reward Model



### iii. RL with PPO



# 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

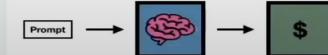
### i. Supervised FT



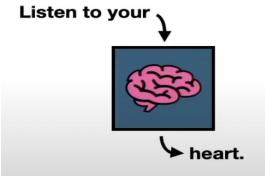
### ii. Train Reward Model



### iii. RL with PPO



# 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

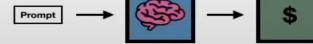
### i. Supervised FT



### ii. Train Reward Model

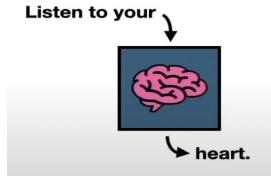


### iii. RL with PPO



## 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

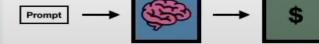
### i. Supervised FT



### ii. Train Reward Model



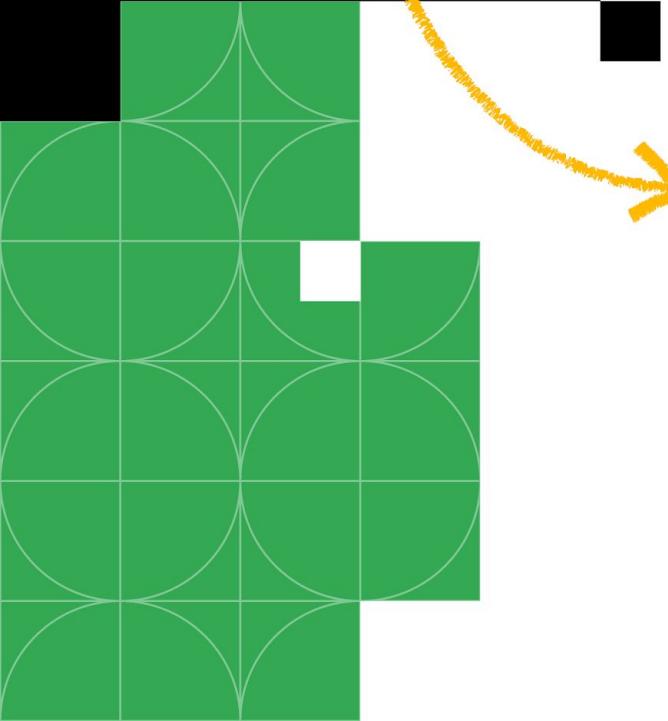
### iii. RL with PPO



```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

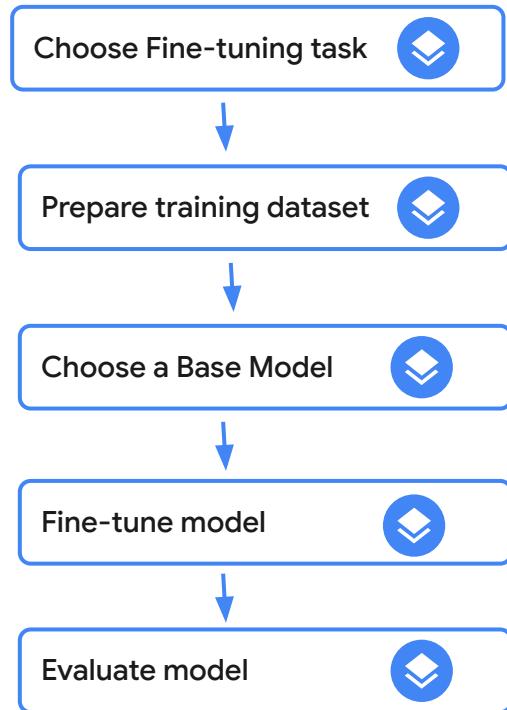
# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



## Supervised Training

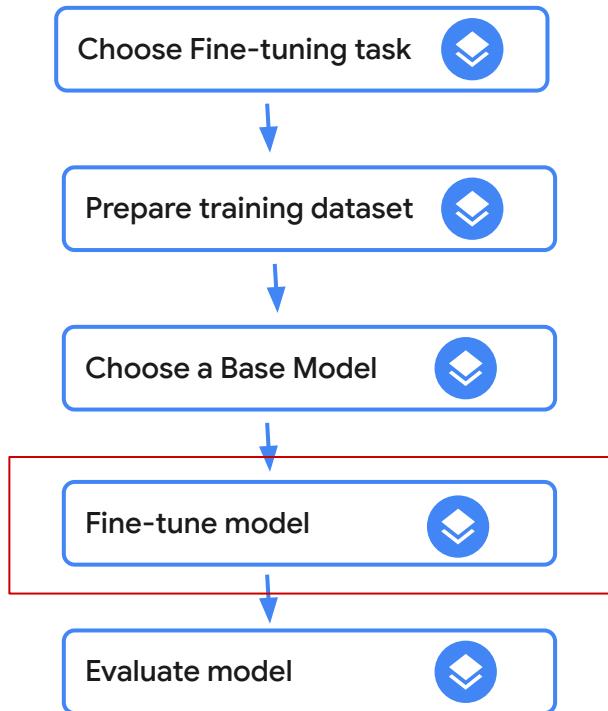
# 5 Steps



Input	Output



# 5 Steps

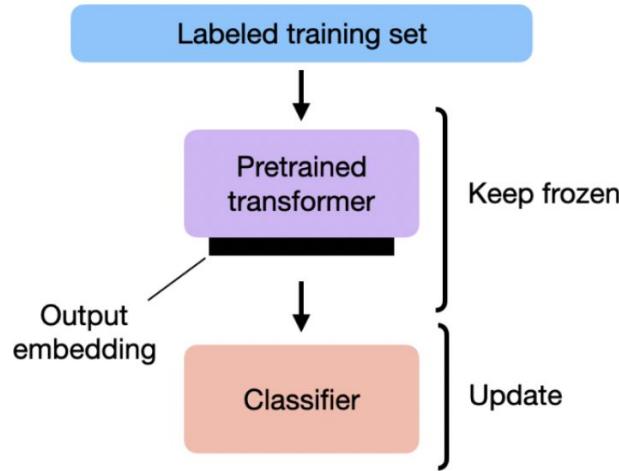


Input	Output



# Fine-tune model via supervised learning

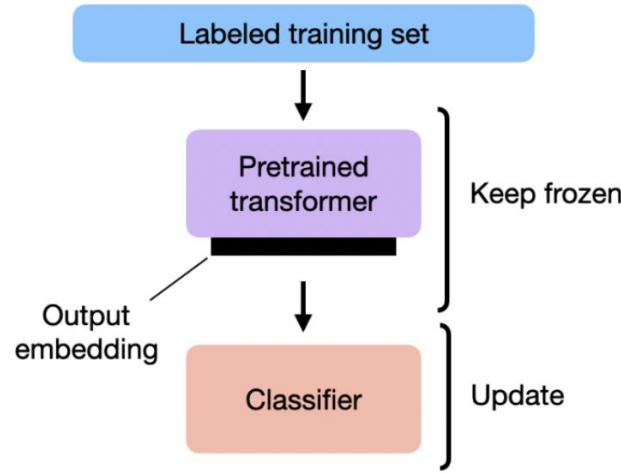
## 1) FEATURE-BASED APPROACH



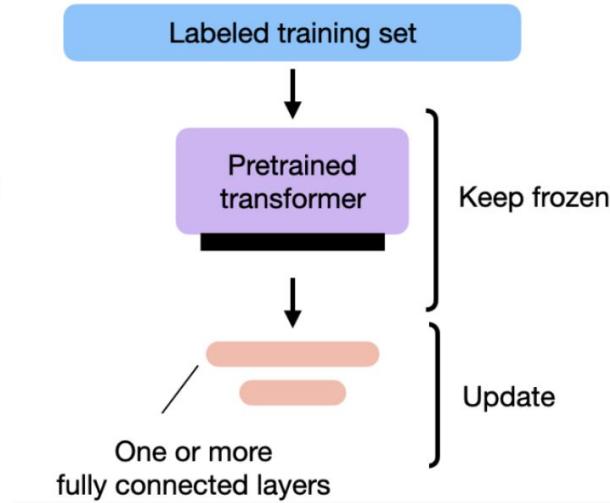
<https://magazine.sebastianraschka.com/p/finetuning-large-language-models>

# Fine-tune model via supervised learning

## 1) FEATURE-BASED APPROACH



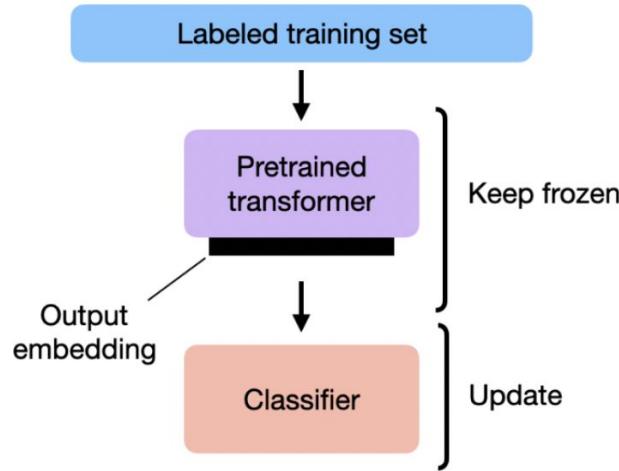
## 2) FINETUNING I



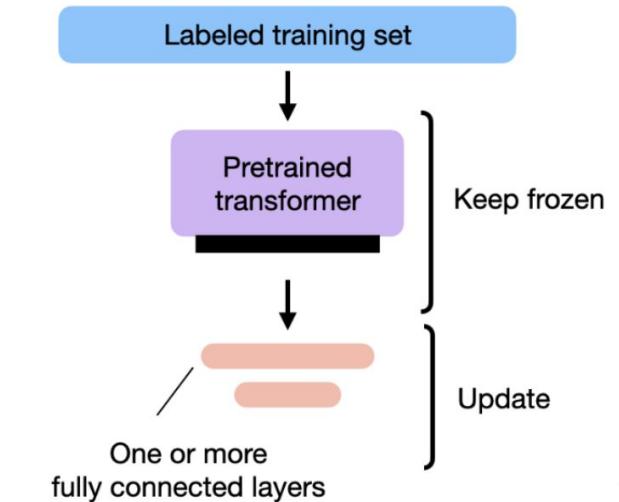
<https://magazine.sebastianraschka.com/p/finetuning-large-language-models>

# Fine-tune model via supervised learning

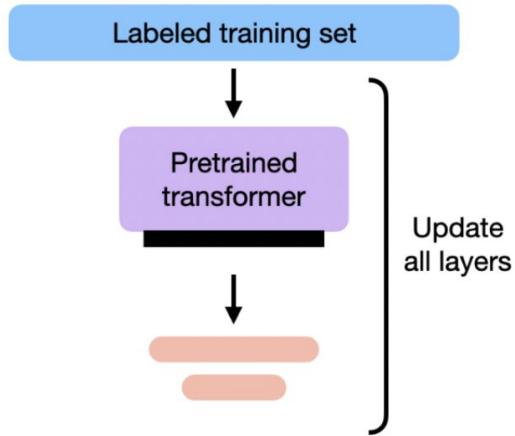
## 1) FEATURE-BASED APPROACH



## 2) FINETUNING I

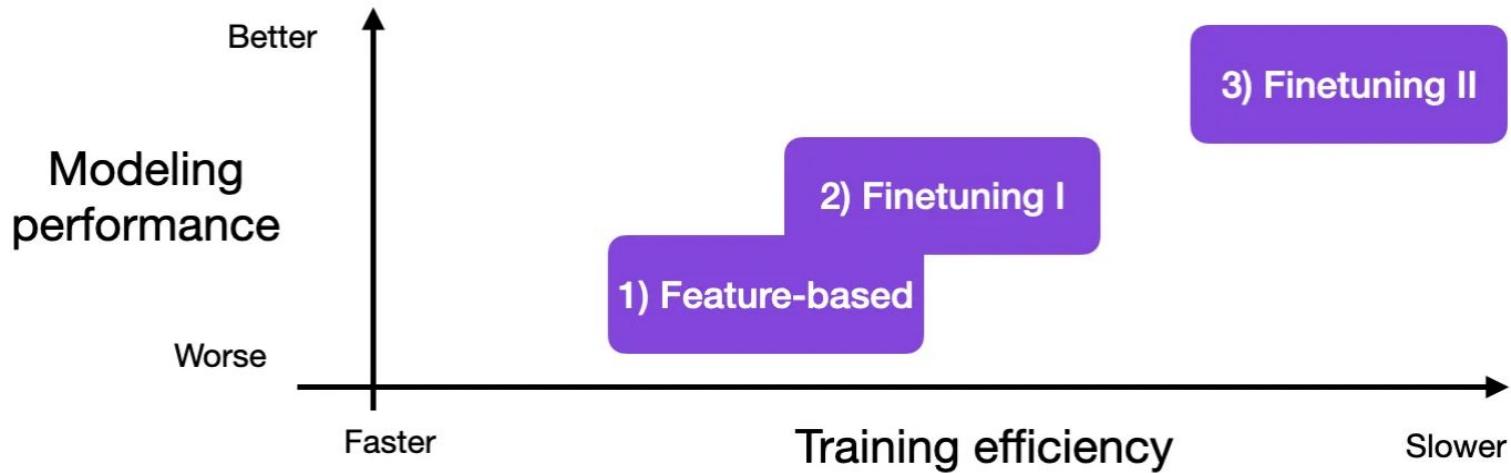


## 3) FINETUNING II



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

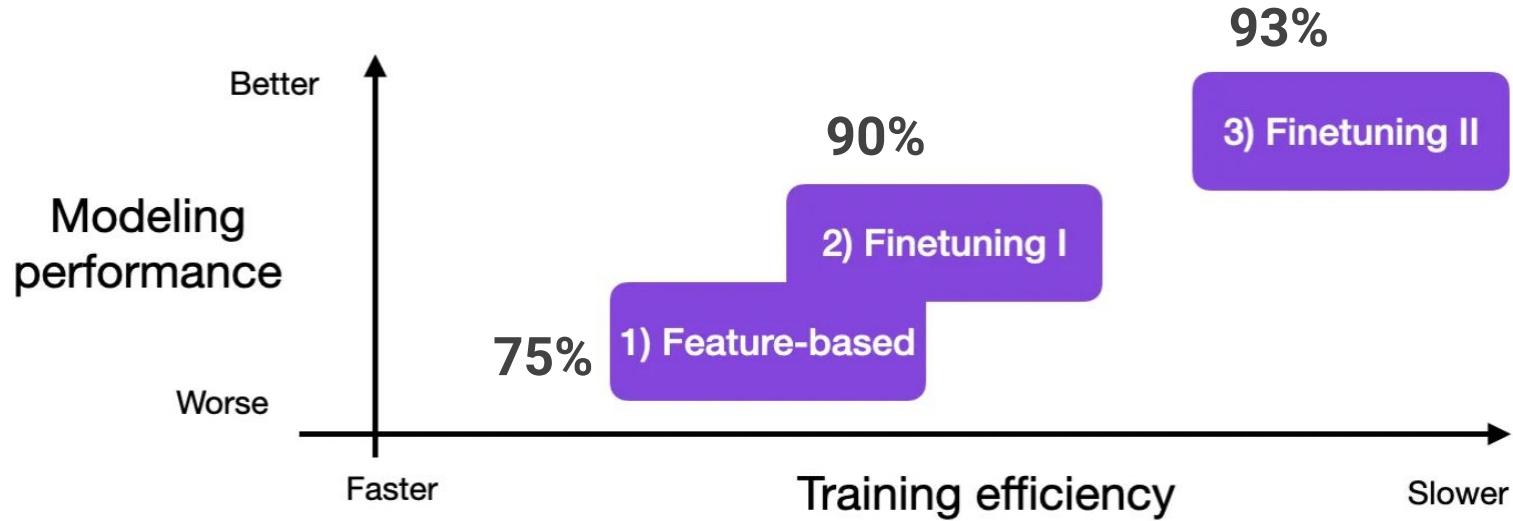
# Comparing LLM fine-tuning methods



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

<https://magazine.sebastianraschka.com/>

# Comparing LLM fine-tuning methods



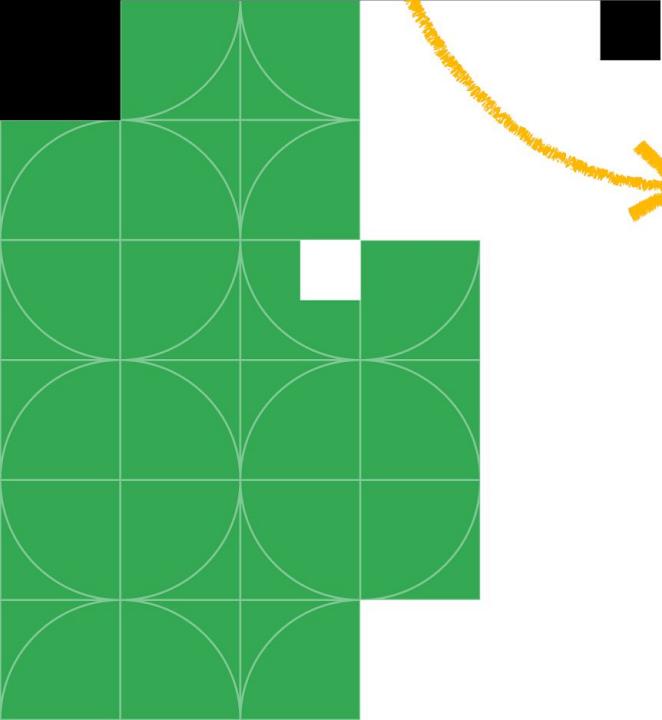
Rule-of-thumb computational and modeling performance trade-offs for various approaches.

<https://magazine.sebastianraschka.com/>

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



## Instruction Fine-Tuning

# Instruction Fine-Tuning

2

## Supervised finetuning

More next-token prediction

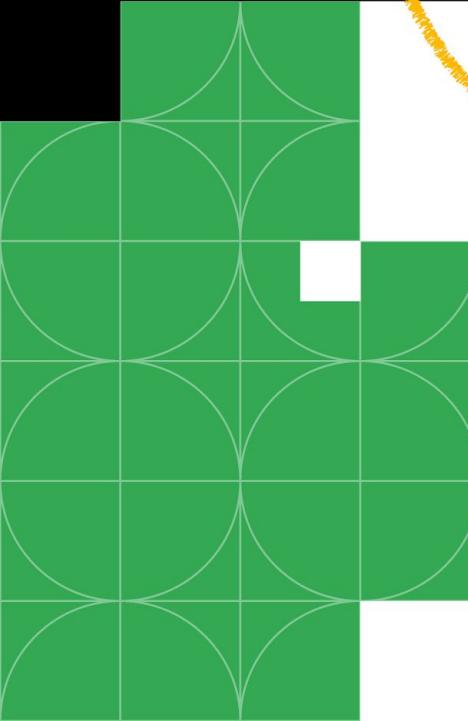
Usually 1k-50k instruction-response pairs

```
{
    "instruction": "Evaluate this sentence for spelling and grammar mistakes",
    "input": "He finnished his meal and left the restaurant",
    "output": "He finished his meal and left the restaurant.",
},
{
    "instruction": "Give three tips for staying healthy.",
    "input": "",
    "output": "\n        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.",
}
```

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

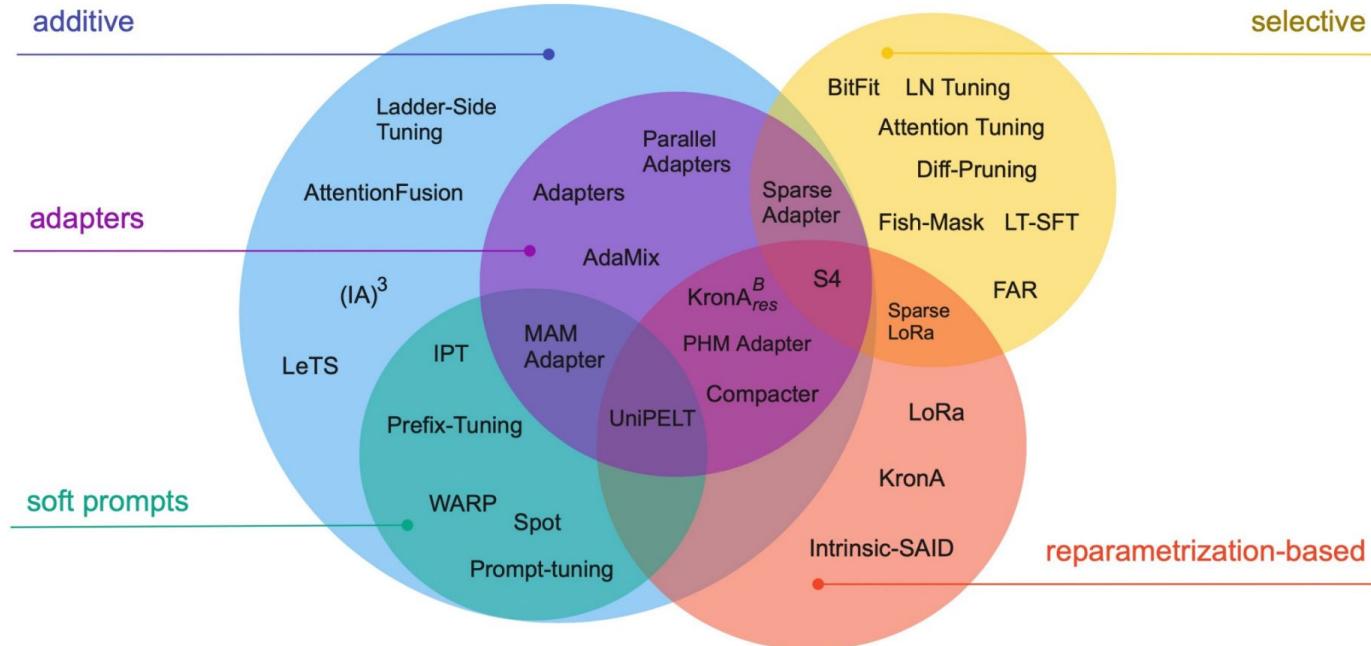
# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



## Parameter-efficient fine-tuning

# Parameter-efficient fine-tuning

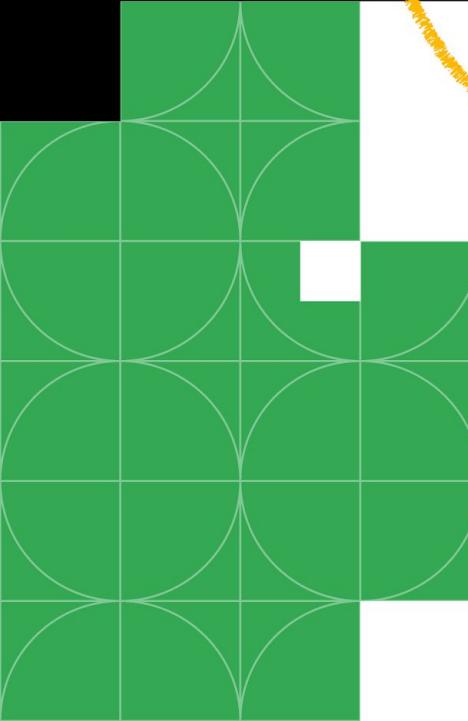


```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],
```

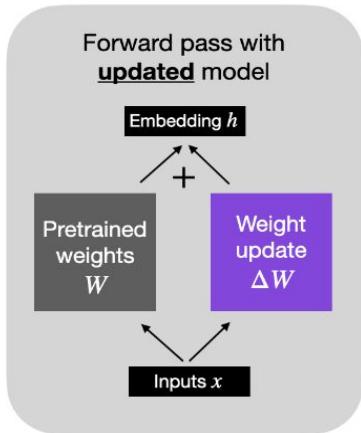
```
Text('23'),
```



## Low-rank adaptation (LoRA)

# Low-rank adaptation (LoRA)

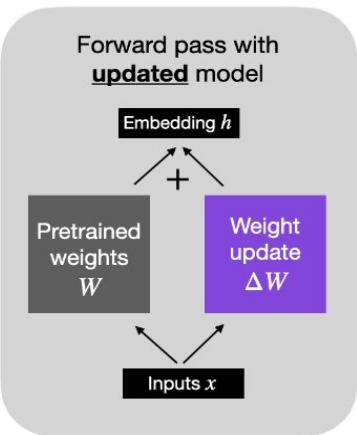
Alternative formulation (regular finetuning)



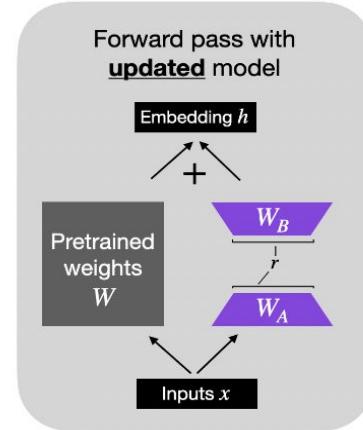
<https://magazine.sebastianraschka.com/>

# Low-rank adaptation (LoRA)

Alternative formulation (regular finetuning)



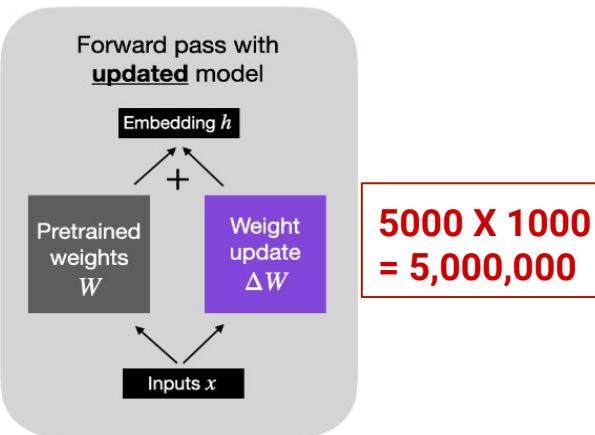
LoRA weights,  $W_A$  and  $W_B$ , represent  $\Delta W$



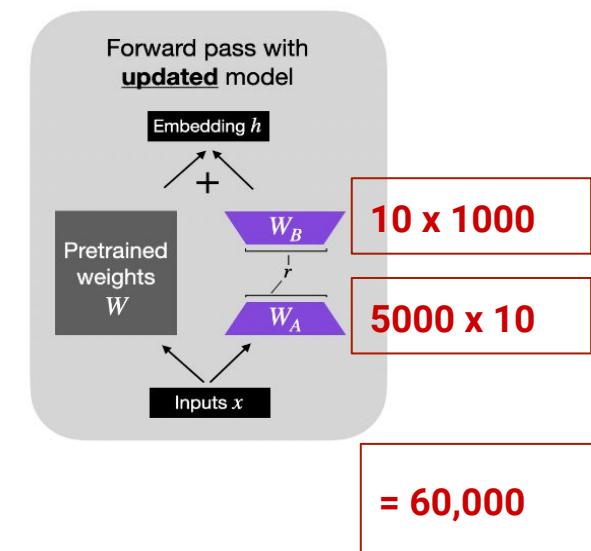
<https://magazine.sebastianraschka.com/>

# Low-rank adaptation (LoRA)

Alternative formulation (regular finetuning)



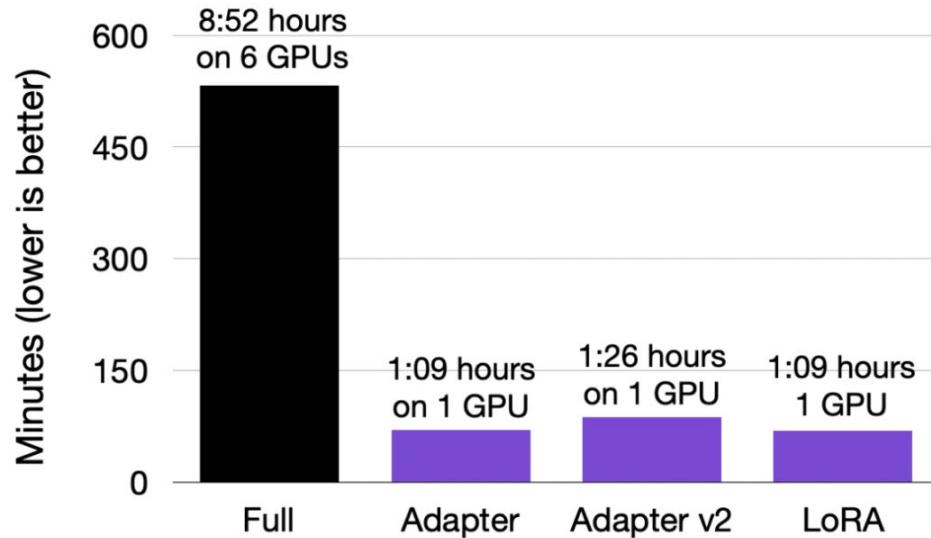
LoRA weights,  $W_A$  and  $W_B$ , represent  $\Delta W$



<https://magazine.sebastianraschka.com/>

# Low-rank adaptation (LoRA)

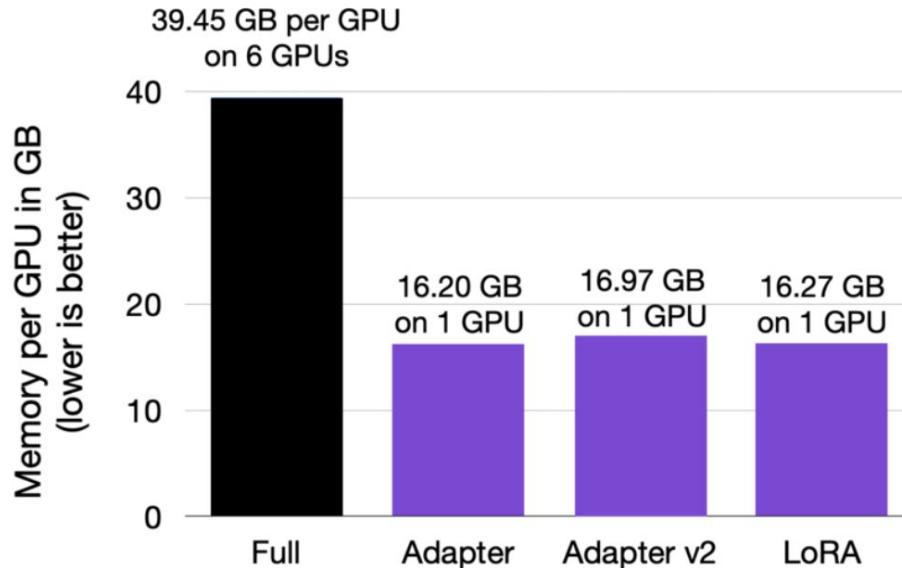
Time to complete 52k training iterations for Falcon 7B



<https://magazine.sebastianraschka.com/>

# Low-rank adaptation (LoRA)

## Memory requirements per GPU for Falcon 7B

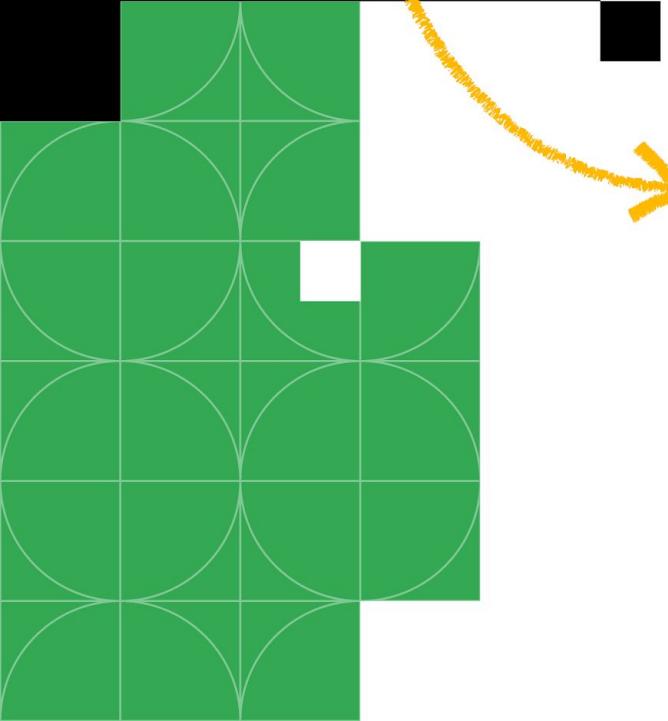


<https://magazine.sebastianraschka.com/>

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



## Open Source LLMs

# LLaMA

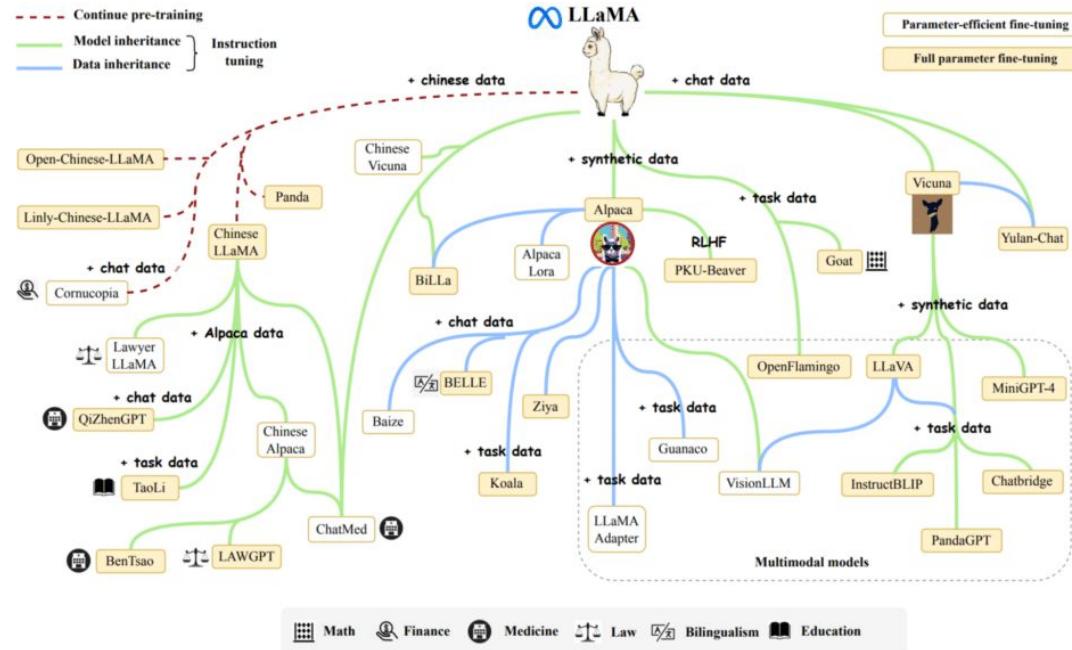
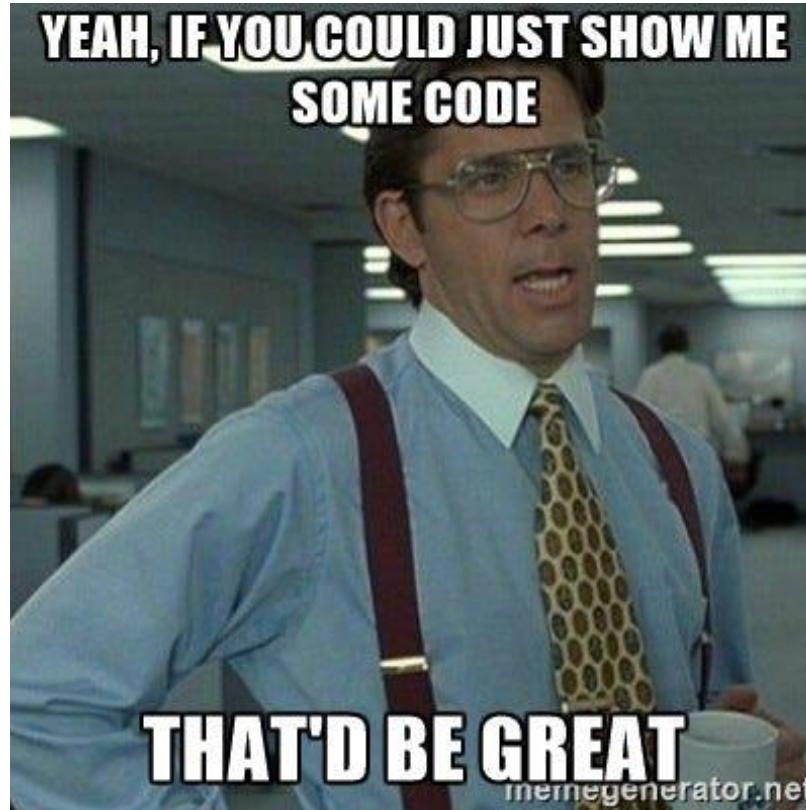


Fig. 4: An evolutionary graph of the research work conducted on LLaMA. Due to the huge number, we cannot include all the LLaMA variants in this figure, even much excellent work. To support incremental update, we share the source file of this figure, and welcome the readers to include the desired models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

# Code Please

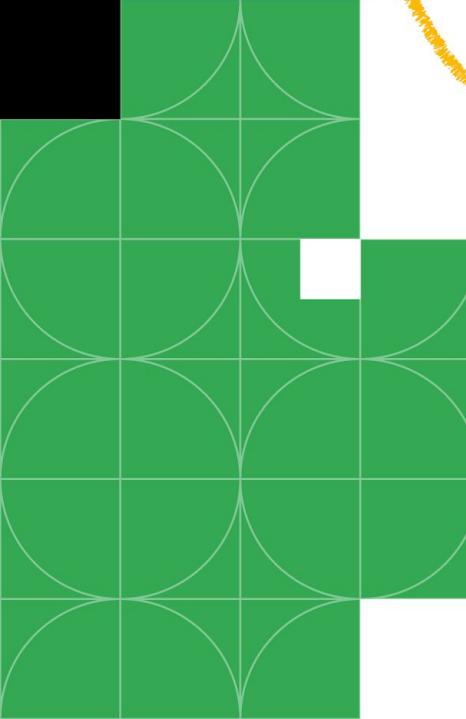


```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

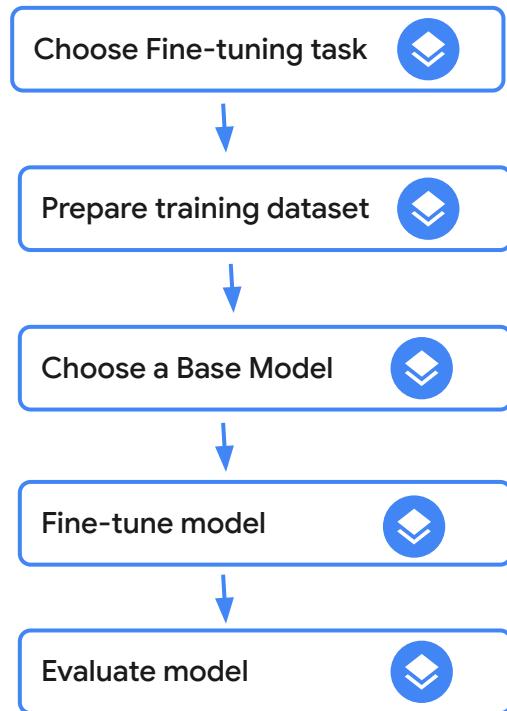
```
s.star,  
r: Colors.green[500],
```

```
Text('23'),
```



## Fine-tuning LLaMA-2 on Personal Whatsapp data

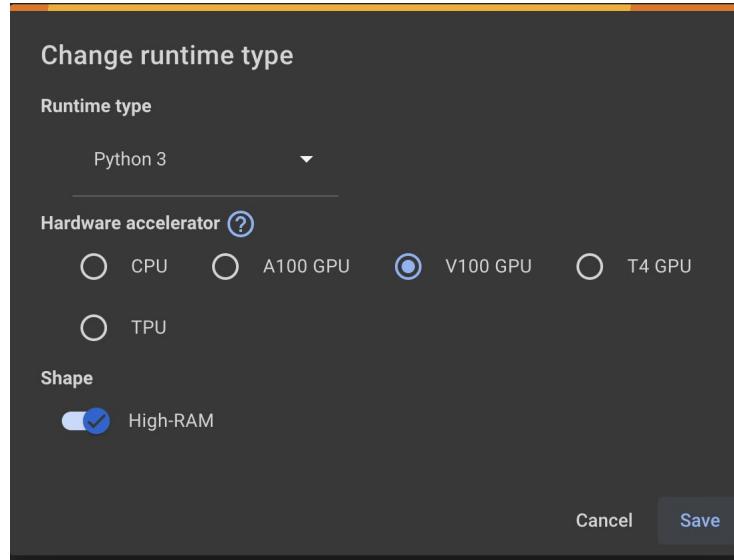
# 5 Steps



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!

A screenshot of a GitHub repository page for "README.md". At the top, there are buttons for "Edit Pins", "Unwatch" (3), "Fork" (9), and "Starred" (93). The main content shows a circular icon of a llama wearing headphones, followed by the text "LLMTuner" and a brief description: "LLMTuner: Fine-Tune Llama, Whisper, and other LLMs with best practices like LoRA, QLoRA, through a sleek, scikit-learn-inspired interface." Below this are links for "License Apache 2.0", "PRs welcome", "Discord Community", and "Open in Colab". A section titled "Installation" follows, with a "With pip" subsection and a note about Python 3.7+ compatibility.

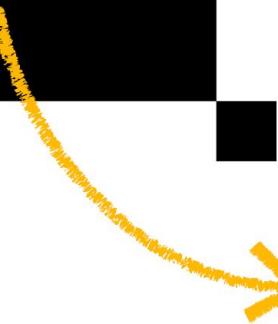
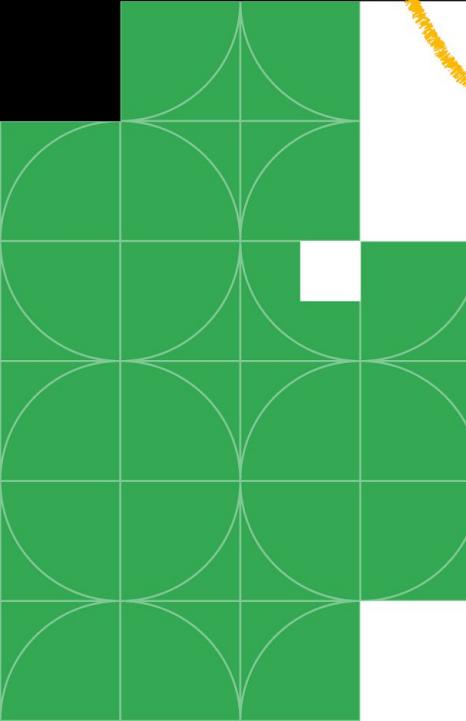
<https://github.com/promptslab/LLMtuner/>

```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

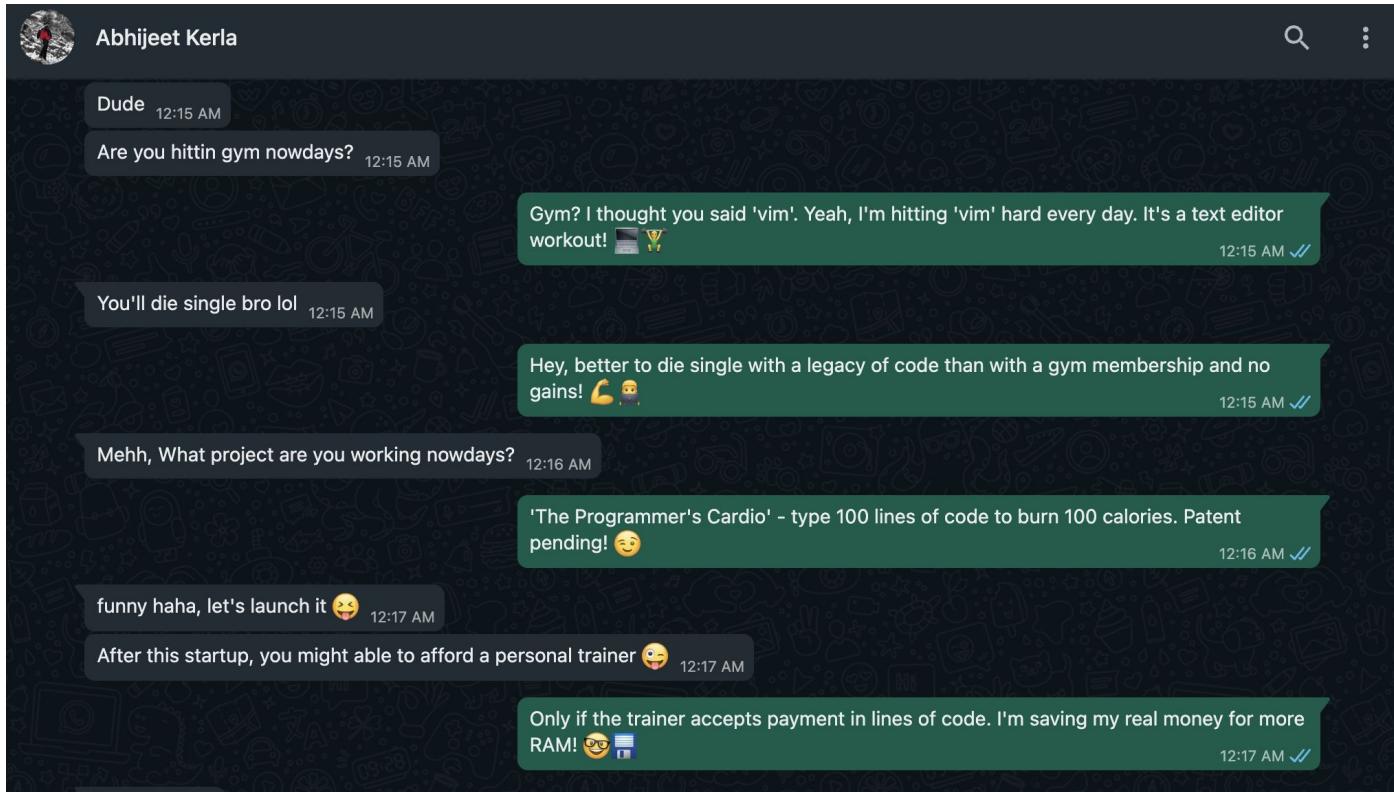
```
s.star,  
r: Colors.green[500],
```

```
Text('23'),
```



## Inference

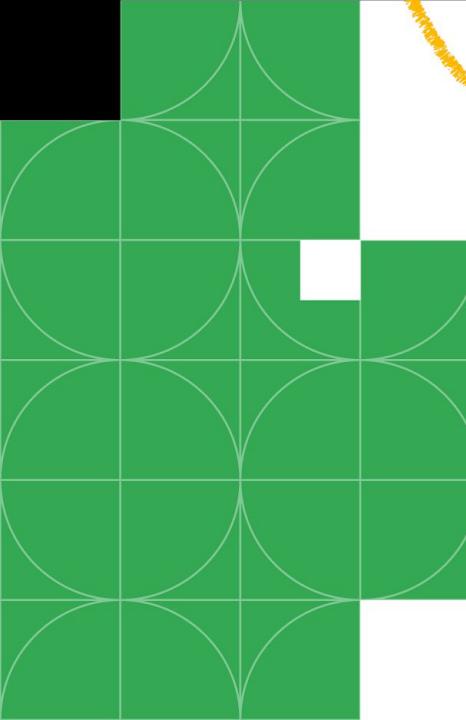
# LLaMA is talking behalf of me



```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

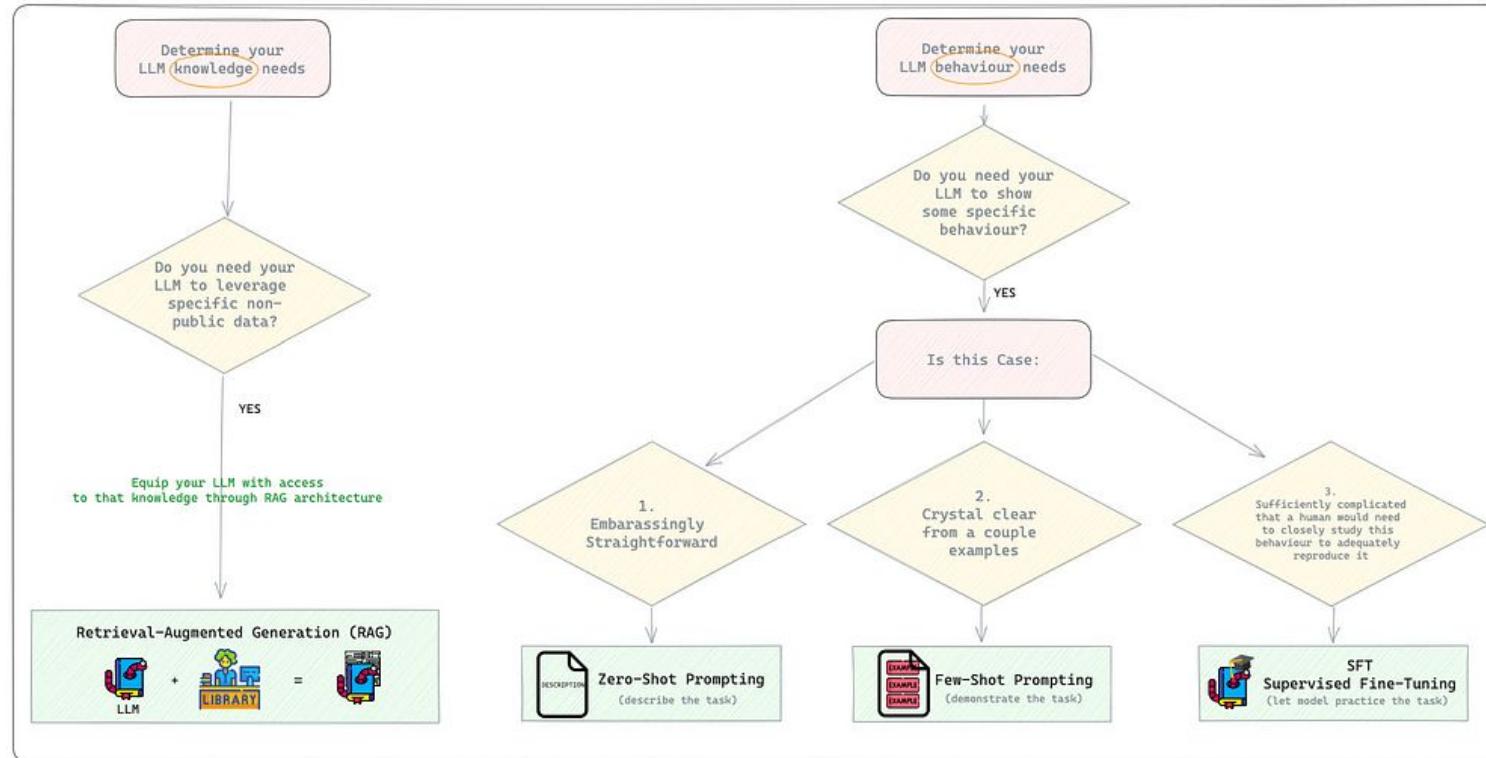
# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



## To fine-tune or Not?

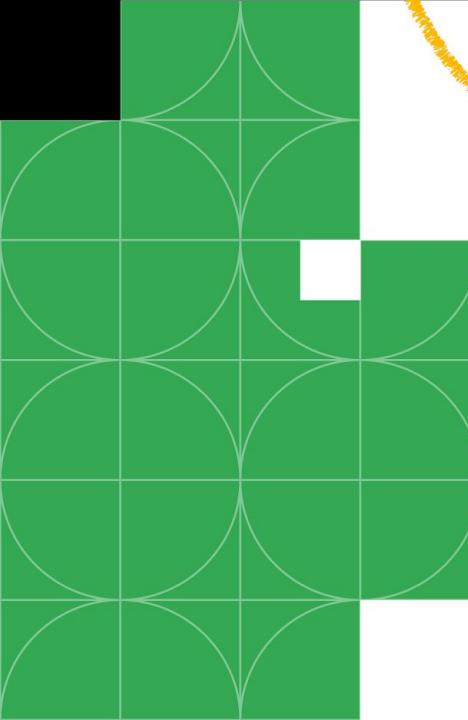
# LLaMA is talking behalf of me



```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



## 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  
[@aadityaura](https://twitter.com/aadityaura)

