FINETUNING FUNCTION CALLING

- TUSHAR ARORA

TASK

FINE-TUNING A LARGE LANGUAGE MODEL FOR FUNCTION CALLING

MAJOR FINETUNING TYPE

Instruction Finetuning

Paramter Finetuning

✓ BECAUSE OF GPU

(Because of Multitask)

(Specialized Finetuning)

CHOOSE LLM MODEL

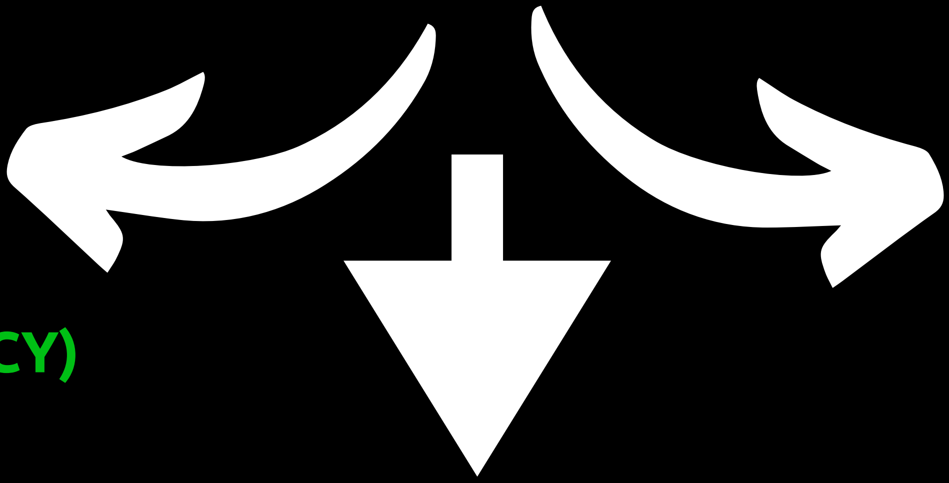
PAID SERVICES

- OPENAI
- COHERE LLM

FREE SERVICES

- OLLAMA
- ML STUDIO
- GROQ
- HUGGINGFACE

TECHNIQUES OF PARAMTER FINETUNING



LORA

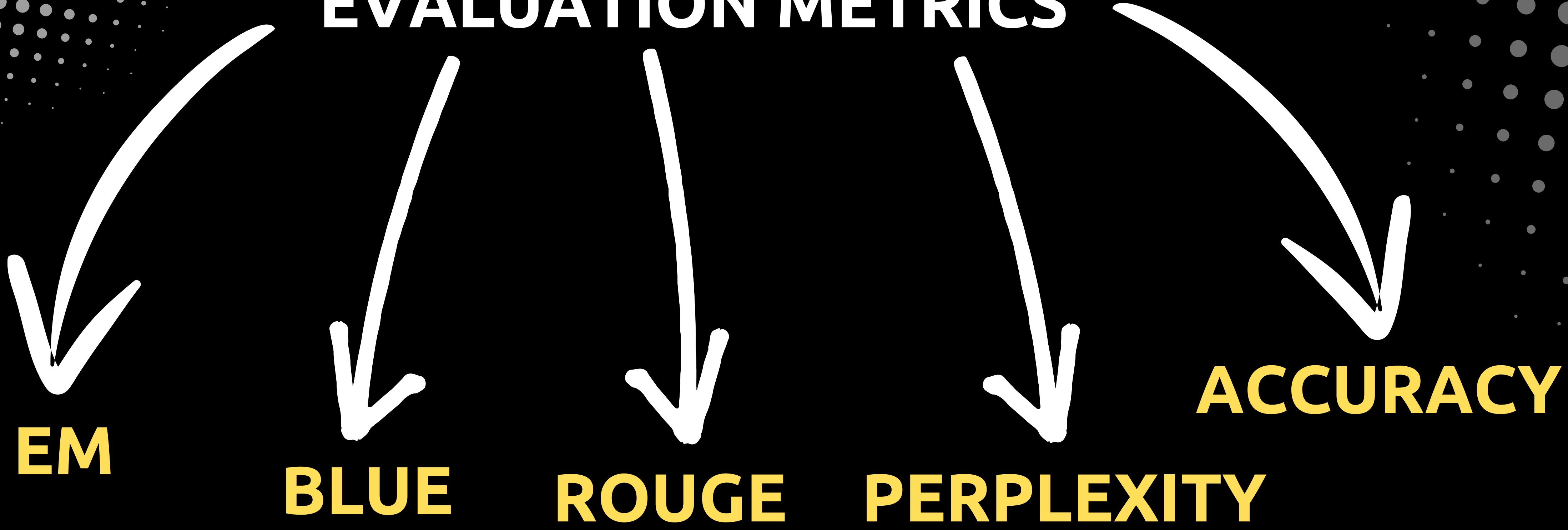
(BETTER FOR MEMORY EFFICIENCY)

QLORA

(BETTER FOR FAST TRAINING)

TRYING BOTH

EVALUATION METRICS



EM

BLUE

ROUGE

PERPLEXITY

ACCURACY

★ DIFFRENT USECASES
DIFFRENT MATRIX

CHOOSE LLM MODEL

DEPEND ON SIZE OF MODEL, PARAMETERS, COMPUTATION RESOURCES, ETC.

FEW OPTIONS :

- **LLAMA 2**
- **LLAMA 3**
- **DISTILBERT**
- **....**

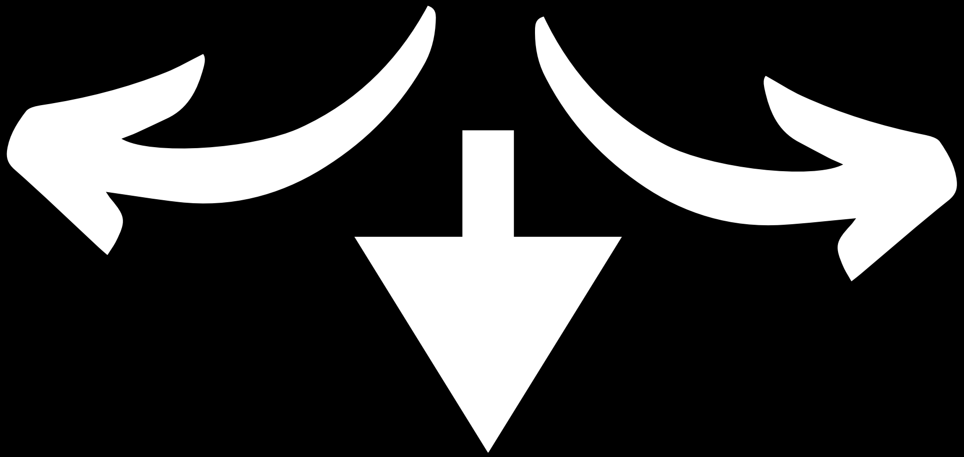
CHOOSE

- Designed for conversational AI tasks
- Balance between performance and resource
- Optimized for chat-based applications

Approach to Choosing LLM Model or Dataset for Finetuning

- Llama
- Gpt
- Distillbert
- Gemini
- ...

(Diffrent size & Parameters)



- Glaive v1
- Glaive v2
- Lilacli
- ShareGPT
- ...

(Diffrent No. of rows and size)

Problem:- Which Model and dataset should I choose?

Answer:-

- Depend on GPU
- Depend on Parameters of LLM Model

I have 8GB GPU, Google Collab Provides 15GB GPU, Kaggle Provide 16GB GPU

- Example of Llama 2 with 7Billion Parameters
- Total parameter size = 7 billion parameters * 4 bytes/parameter = 28 billion bytes
- Since 1 GB = 1 billion bytes, the total parameter size is approximately 28 / 1 = 28 GB.
- Lora or Quara Technique will Reduce /2

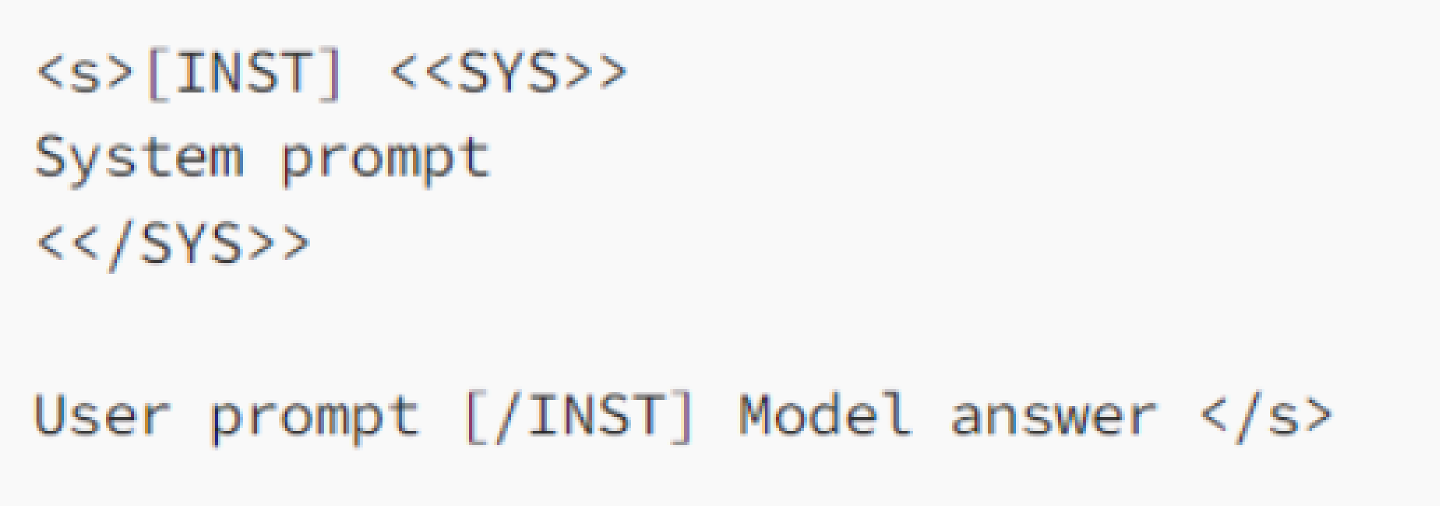
APPROACH IS:-

- Low Parameter LLM model with High sample no. of fine-tuned dataset
- High Parameter LLM model with Low sample no. of fine-tune dataset
- Buy GPU
- Paid Service like Gradient

We Also Can make Own Dataset Like this:-

```
[
  {
    "input": "Calculate the sum of 4 and 5",
    "output": "sum(4, 5)"
  },
  {
    "input": "Find the maximum of 7 and 10",
    "output": "max(7, 10)"
  }
  // Add more examples as needed
]
```

Problem:- Different LLm model Have Own Data Format. Llama 2 Formate like:

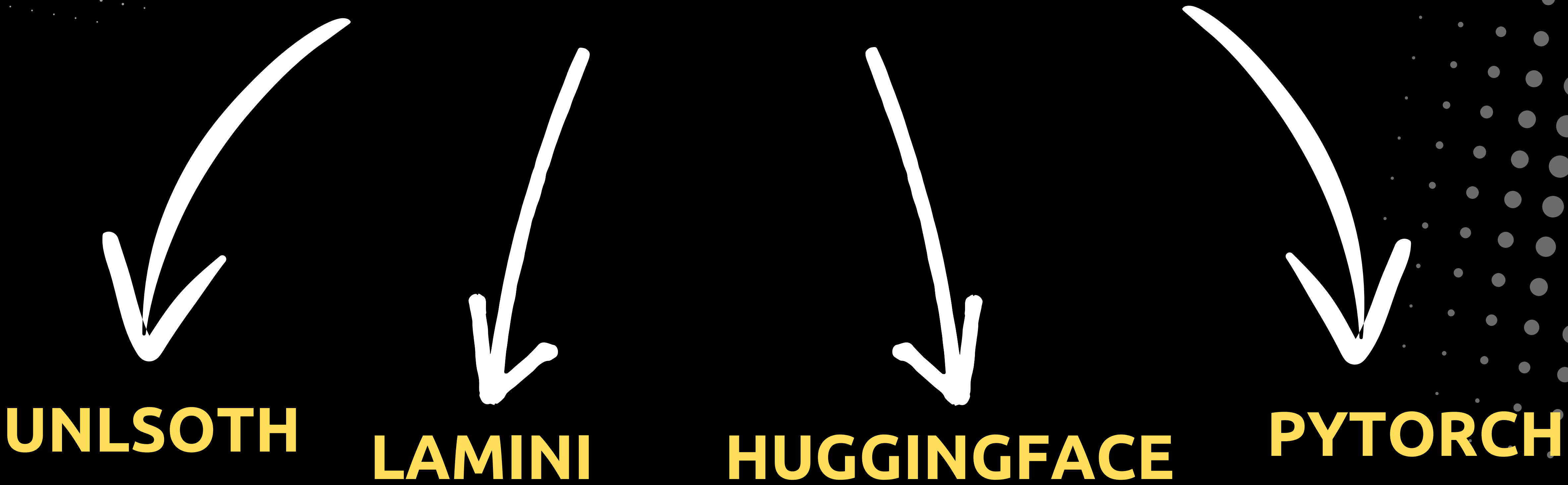


```
<s>[INST] <<SYS>>
System prompt
<</SYS>>

User prompt [/INST] Model answer </s>
```

Solution: Write a Code to Convert or using Huggingface converter

TOOLS TO DO FINETUNING



UNLSOTH

LAMINI

HUGGINGFACE

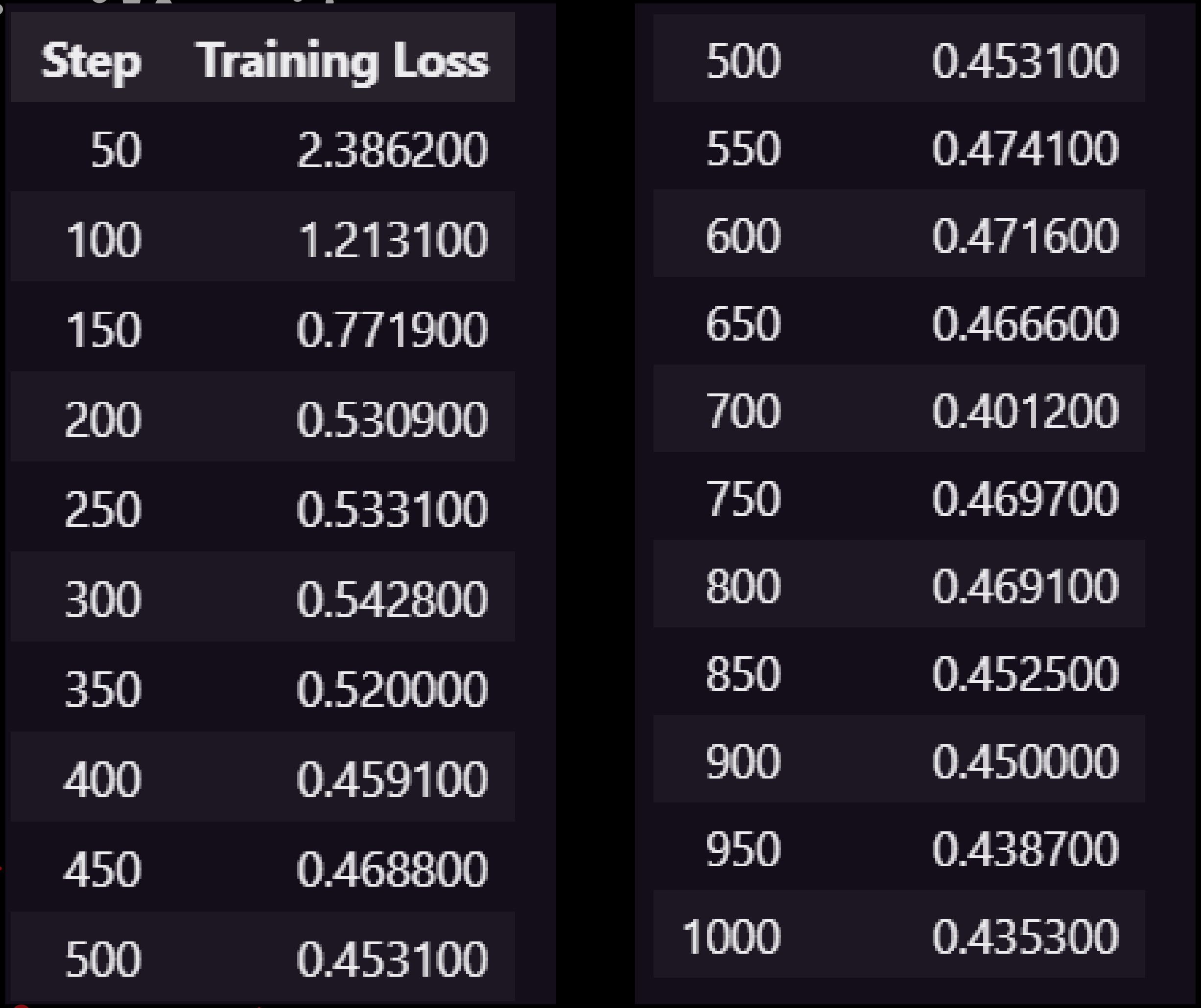
PYTORCH

LOW PARAMETER LLM MODEL WITH HIGH SAMPLE NO. OF FINE-TUNED DATASET WITH UNSLOTH

- So I choose Google Gemma 1BIllion parameter with Function calling dataset called Lilacli of 112960 rows using Unsloth tool
- https://huggingface.co/datasets/lilacai/glaive-function-calling-v2-sharegpt (why Choose this?)

FINE TUNE MODEL LINK ON HUGGING FACE

https://huggingface.co/Danjin/unsloth-gemma-glaive-function-callingv3



Step	Training Loss
50	2.386200
100	1.213100
150	0.771900
200	0.530900
250	0.533100
300	0.542800
350	0.520000
400	0.459100
450	0.468800
500	0.453100
500	0.453100
550	0.474100
600	0.471600
650	0.466600
700	0.401200
750	0.469700
800	0.469100
850	0.452500
900	0.450000
950	0.438700
1000	0.435300

Danjin

/unsloth-gemma-glaive-function-callingv3

huggingface.co

Danjin/unsloth-gemma-glaive-function-callingv3 · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open source and open science.

huggingface

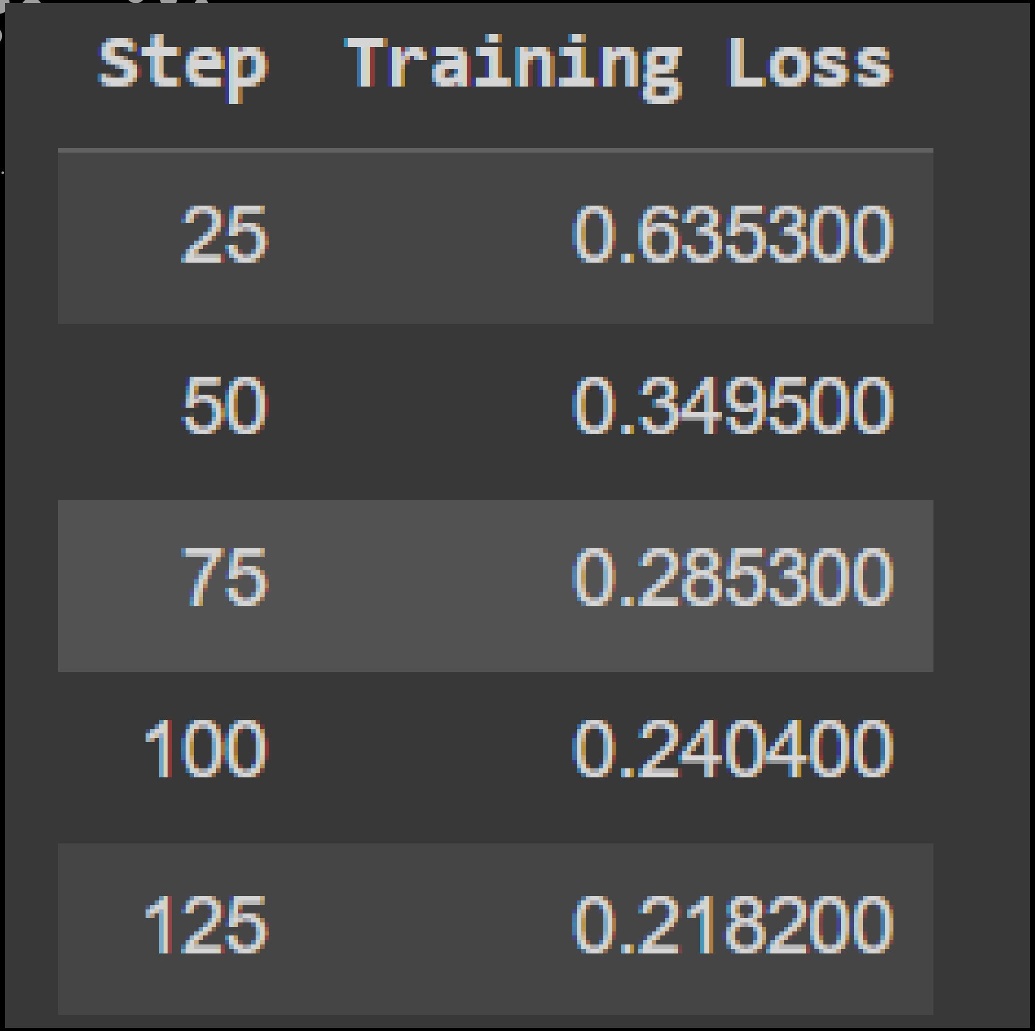
HIGH PARAMETER LLM MODEL WITH LOW SAMPLE NO. OF FINE-TUNED DATASET

So I choose Llama2 7BIllion parameter with a Function calling dataset called glaiveai of 500 rows using Hugging face and PyTorch.

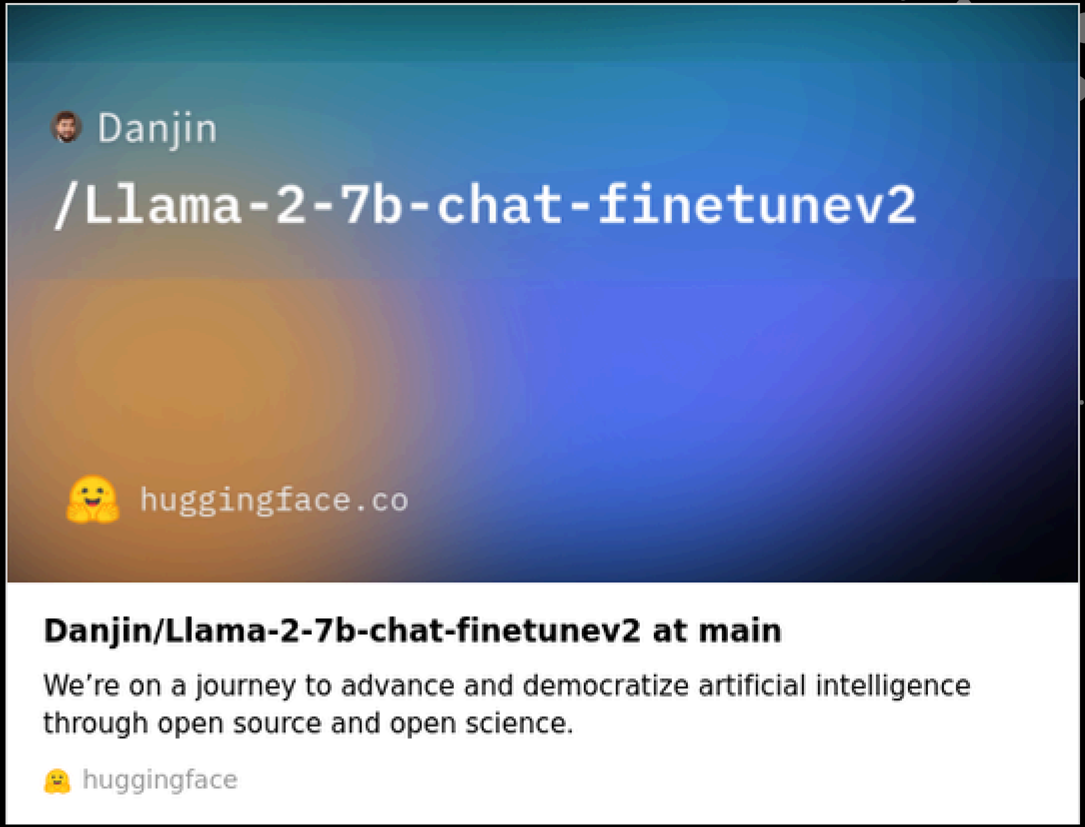
https://huggingface.co/datasets/glaiveai/glaive-function-calling-v2

FINE TUNE MODEL LINK ON HUGGING FACE

https://huggingface.co/Danjin/Llama-2-7b-chat-finetunev2/tree/main



Step	Training Loss
25	0.635300
50	0.349500
75	0.285300
100	0.240400
125	0.218200



Danjin

/Llama-2-7b-chat-finetunev2

🤗 huggingface.co

Danjin/Llama-2-7b-chat-finetunev2 at main

We’re on a journey to advance and democratize artificial intelligence through open source and open science.

🤗 huggingface