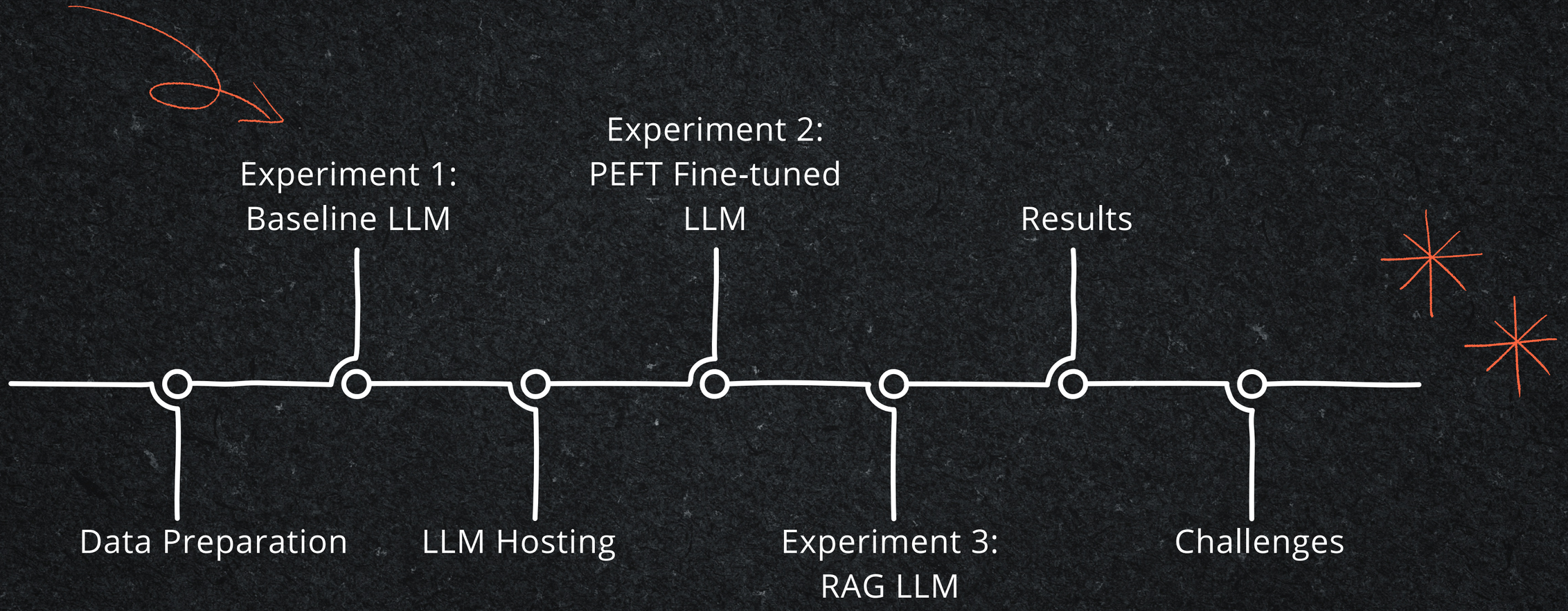
Cohort Sync
Done

F·R·I·E·N·D·S
FAN PERSONA CHATBOT

Advaith Shyamsunder Rao
Falgun Malhotra
Vanshita Gupta

ROADMAP


Experiment 1:
Baseline LLM
Experiment 2:
PEFT Fine-tuned
LLM
Results
Data Preparation
LLM Hosting
Experiment 3:
RAG LLM
Challenges

TRAINING DATASET

- The ConvoKit Friends corpus framework provides the necessary utilities to load and preprocess the Friends corpus.
- There are 236 episodes, 3,107 scenes (conversations), 67,373 utterances, and 700 characters (users).

	conversation_id	season	episode	scene	utterance_id	text	speaker
0	s01_e01_c01_u001	s01	e01	c01	s01_e01_c01_u001	There's nothing to tell! He's just some guy I ...	Monica Geller
1	s01_e01_c01_u001	s01	e01	c01	s01_e01_c01_u002	C'mon, you're going out with the guy! There's ...	Joey Tribbiani
2	s01_e01_c01_u001	s01	e01	c01	s01_e01_c01_u003	All right Joey, be nice. So does he have a hum...	Chandler Bing
3	s01_e01_c01_u001	s01	e01	c01	s01_e01_c01_u004	Wait, does he eat chalk?	Phoebe Buffay
5	s01_e01_c01_u001	s01	e01	c01	s01_e01_c01_u006	Just, 'cause, I don't want her to go through w...	Phoebe Buffay

EVALUATION DATASET

- A curated evaluation dataset is generated through web scraping from Friends Trivia websites and parsed to get MCQ-style questions.
- The dataset has been cleaned to keep four options for every question.
- The dataset is formatted to keep Question, Options with a list of four answer choices and the ground truth answer

	Question	Options	Correct Answer	Correct_Answer_no
0	How many times was Ross legally divorced? \n	['Twice', 'Three times', 'Five times', 'Six ti...	Three times	B
1	Where did Carol first meet Susan? \n	['In college', 'At work', 'At the gym', 'At Ce...	At the gym	C
2	How did Susan and Ross come up with Ben's name?	["It was the doctor's name \n", 'They both ha...	It was on the janitor's name tag	D
3	What were Ben's first words? \n	['Hi', 'Bye', 'Mom', 'Dumb']	Hi	A
4	How long did Ross and Emily date before they g...	['14 days', '6 weeks \n', 'A year \n', '3 mo...	6 weeks \n	B

EXPERIMENT DATA PREPARATION

LLM Fine Tuning Data

To prepare dataset for LLM fine tuning, the speaker and dialogues are merged based on the conversation_id

	conversation_id	season	episode	scene	script
0	s01_e01_c01_u001	s01	e01	c01	[Monica Geller: There's nothing to tell! He's ...
1	s01_e01_c02_u001	s01	e01	c02	[Monica Geller: Now I'm guessing that he bough...
2	s01_e01_c03_u001	s01	e01	c03	[Phoebe Buffay: Love is sweet as summer shower...
3	s01_e01_c04_u001	s01	e01	c04	[Ross Geller: I'm supposed to attach a bracket...
4	s01_e01_c05_u001	s01	e01	c05	[Monica Geller: Oh my God!, Paul the Wine Guy:...

Vector Store Data

Employed the Pinecone Database for storing the vectors derived from instruct embeddings

1

ID
s10...

VALUES
0.0268788282, -0.0448419, 0.0193887949, -0.0421825424, -0.050444

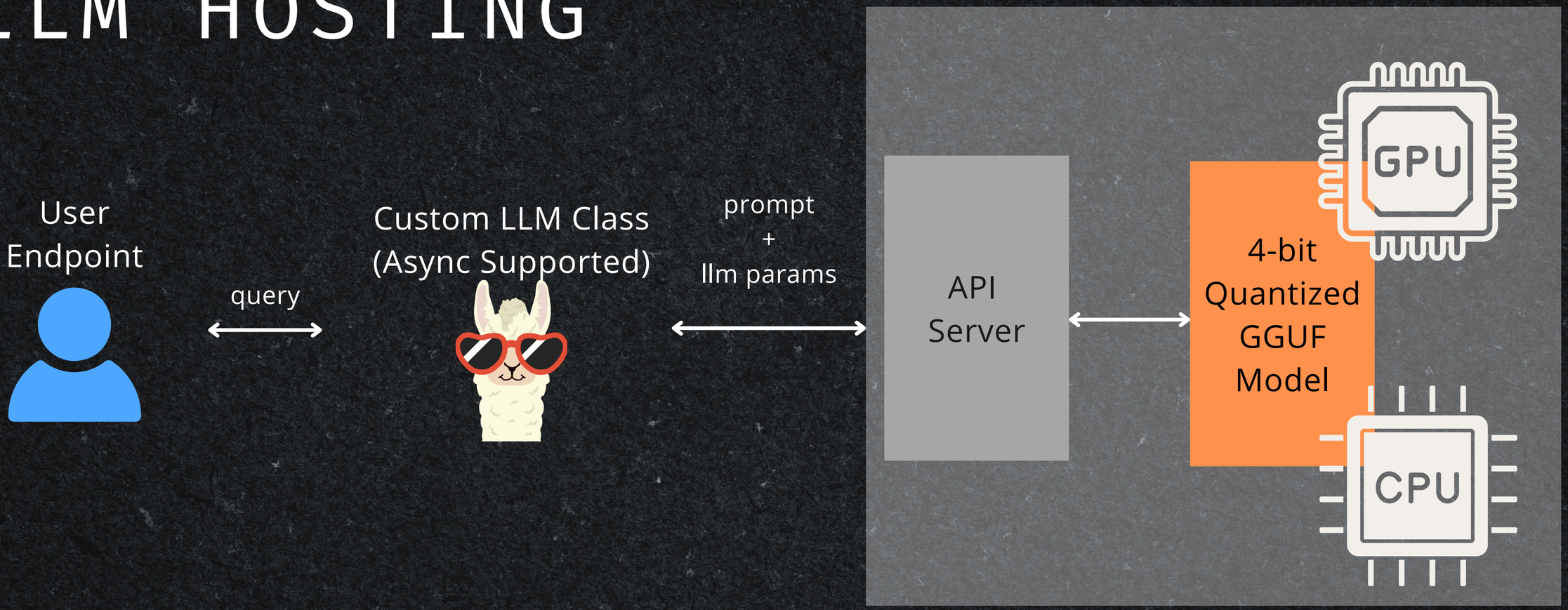
SCORE
0.0999

METADATA
source: "{'speaker': 'Joey Tribbiani', 'season': 's10', 'episode': 'e05', 'scene': 'c08'}"
text: "I'M CURVY, AND I LIKE IT!"

EXPERIMENT 1 – BASELINE

- Model: Llama-2 (by Meta)
- Type: Chat completion
- Parameters: 7B
- Top-p Sampling: 0.9
- Temperature: 0.5
- Repetition penalty: 1.5

LLM HOSTING



- Model Quantization and server setup were done using LLama.cpp executables.(quantize.sh,server.sh)

FINETUNING

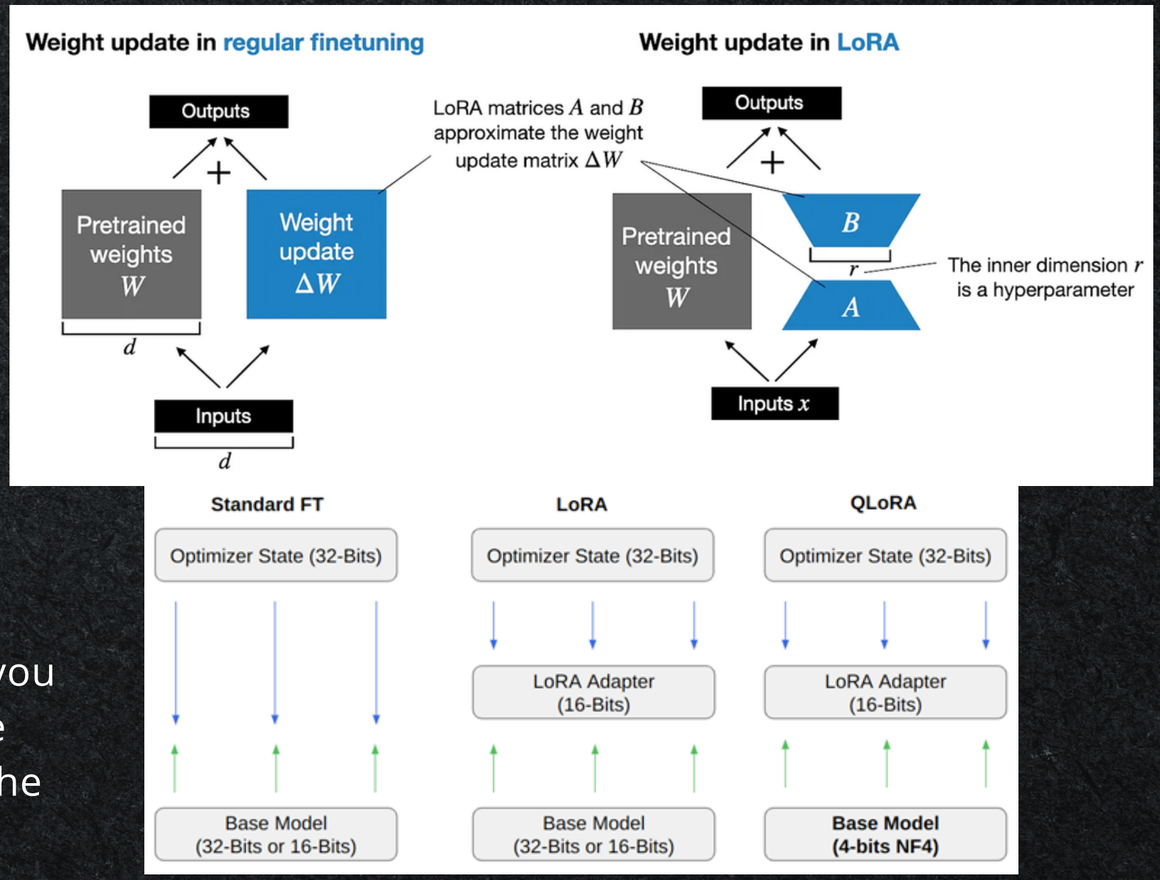
Considering the huge size of Llama2 model i.e. 7B params, we cannot perform standard finetuning on all the model parameters due to high storage and compute demand.

We experimented with 2 Parameter Efficient Fine-Tuning (PEFT) Finetuning approaches:

1. LoRA Finetuning on NF-4 Quantized Llama-2-7b-chat GGUF model

2. QLoRA Finetuning on FP-16 Llama-2-7b-chat

While LoRA helps in reducing the storage requirements, you would still need a large GPU to load the model into the memory for LoRa training. In QLoRA, you first quantize the LLM and then perform LoRa training.



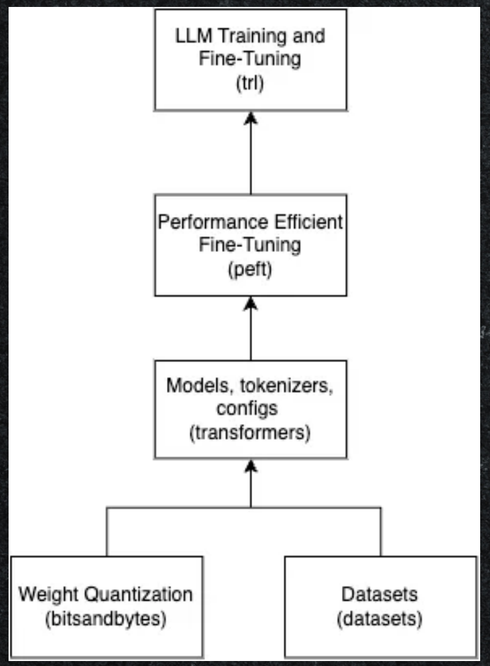
FINETUNING

1) LoRA: Was performed using No-Code Approach

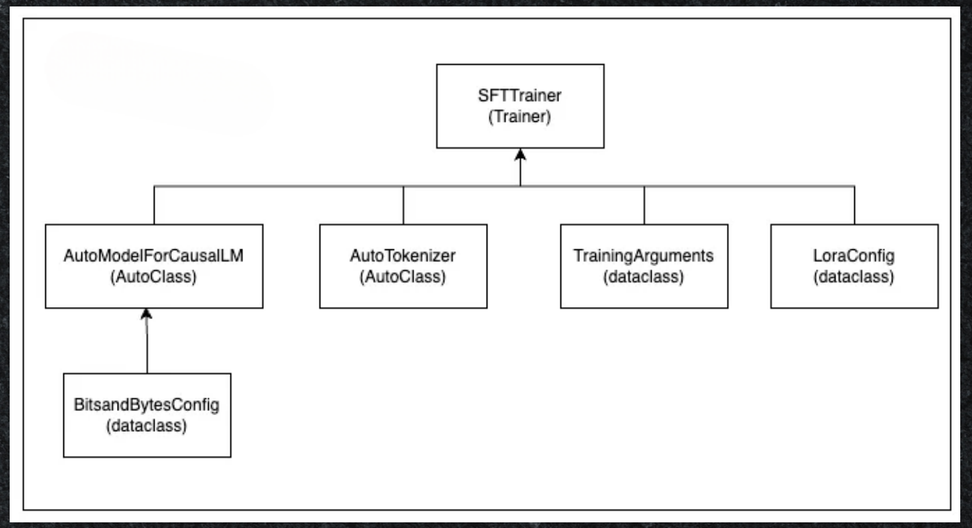
For this, LLama.cpp finetune exectuables were used along with customized config parameters to train, generate adapters and merge them to original model. (finetune.sh,convert.sh,server.sh)

2) QLoRA: Was performed using Code Approach

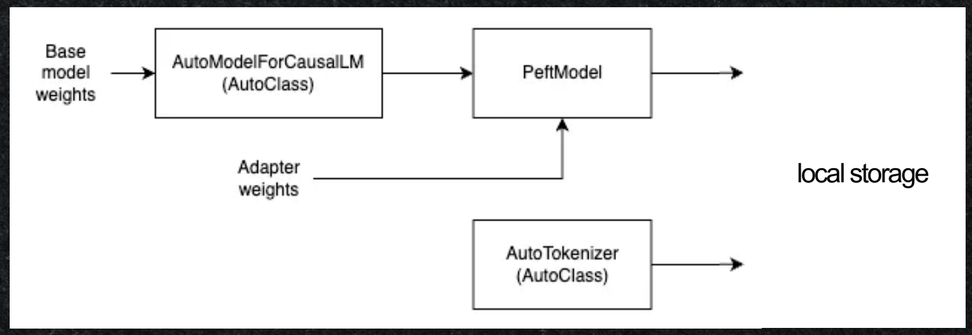
Libraries used:



Training:



Merging Weights and Saving final model:



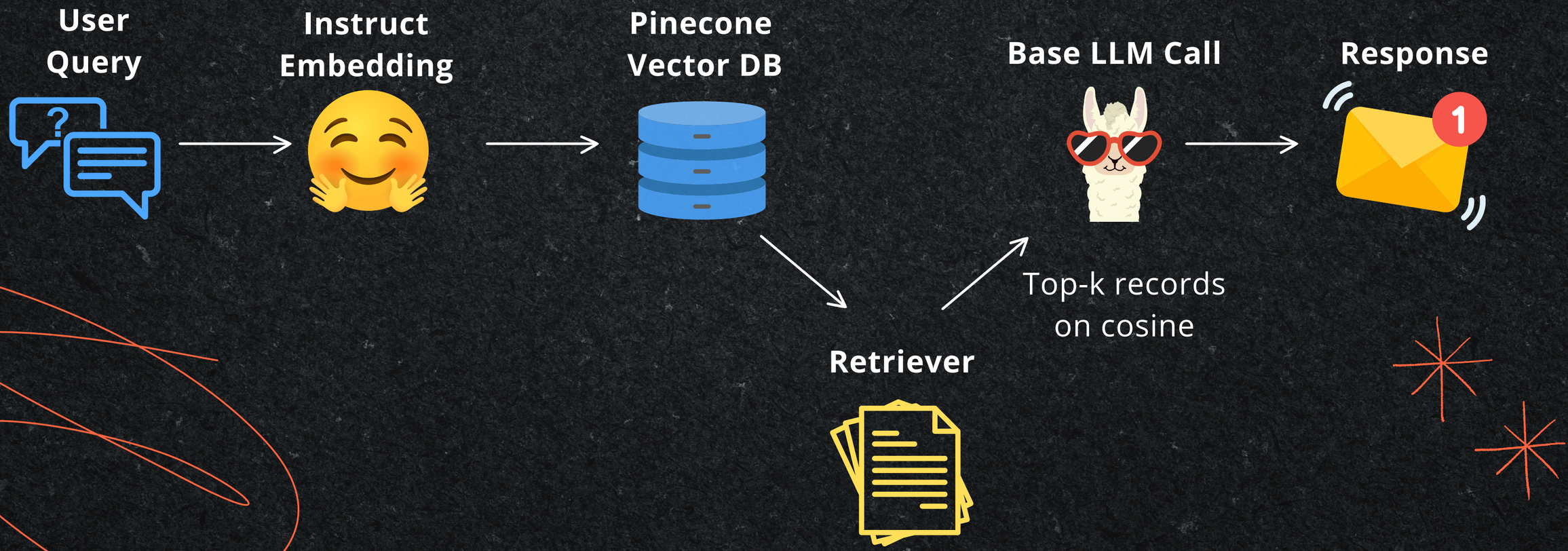
FINETUNING

Some of the Config Parameters used:

Parameter	General Notation	Value Used in Experiment
LoRA attention dimension	--lora_r N	4
Alpha parameter for LoRA scaling	--lora_alpha N	4
Dropout probability for LoRA layers	-- lora_dropout F	0.1
Model Context size	--n_ctx N	4096
Batch Size	--batch N	1

Parameter	General Notation	Value Used in Experiment
Adam learning rate	--adam-alpha F	**2e-5**
Number of epochs	--epochs N	1
Frequency base for ROPE	--rope-freq-base F	1
-rope-freq-scale F	--rope-freq-scale F	1000
Save checkpoint every N iterations	--save-every N	0

EXPERIMENT 3 - RETRIEVAL AUGMENTATION GENERATION (RAG)



RAG VS FINE TUNING

Fine Tuning

- No dependancy on an external Knowledge Base.
- Useful when the chatbot is expected to answer in a specific tone/format.

RAG

- Effective for data that changes over time.
- No compute constraints / less expensive.
- Does not require an addition model training step.

EVALUATION

MULTIPLE CHOICE EVAL

- Calculates two MCQ eval metrics
  - MCQ Accuracy
  - Correct Format Rate
- Utilizes the following prompt techniques
  - Few-shot prompting
  - Chain-of-Thought instructions
  - Sampling evaluation response

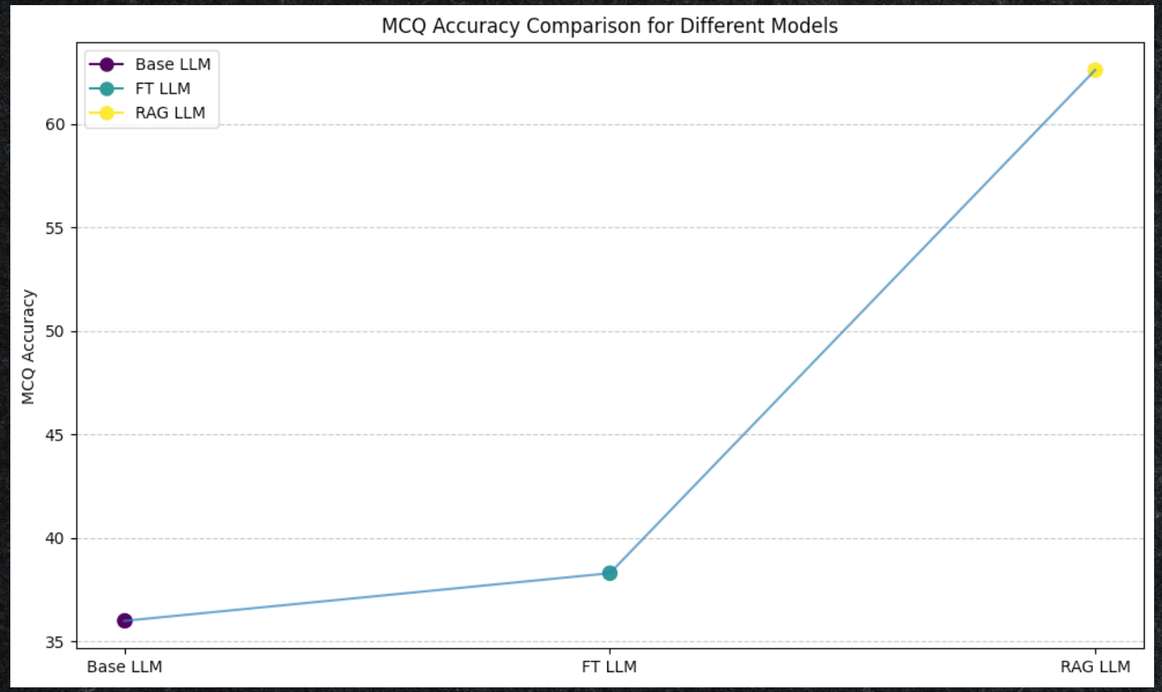
FREE RESPONSE EVAL

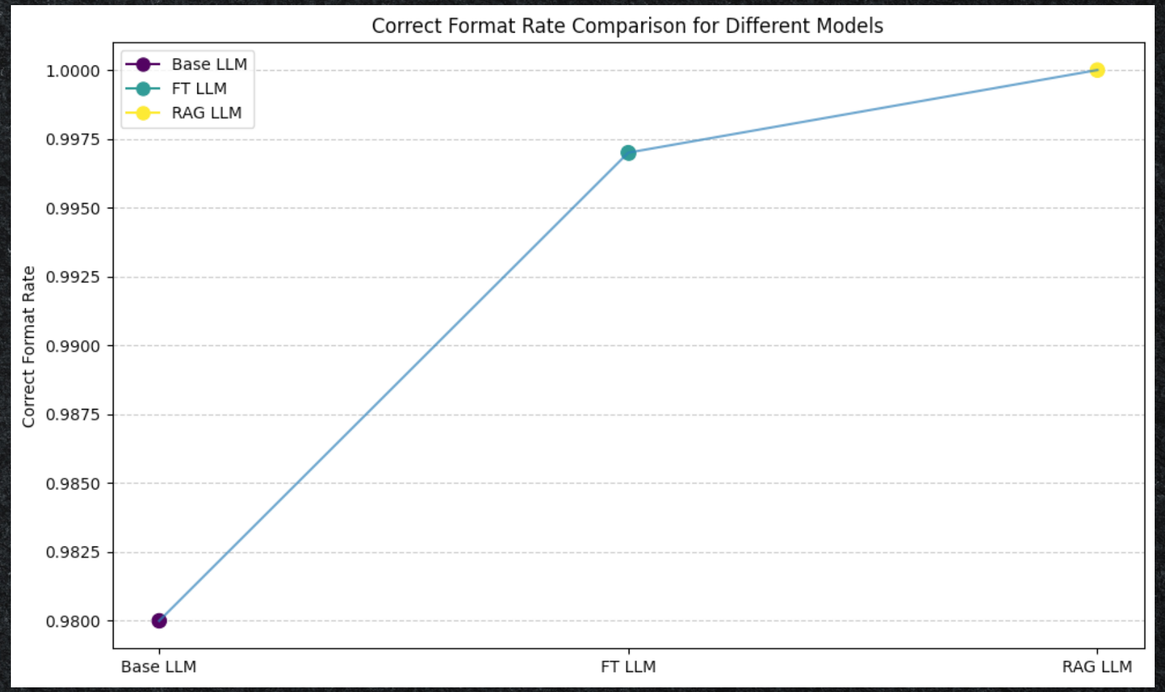
- Calculates average score which comprises of the following metrics
  - Bleu
  - Rouge-1
  - ChrF
  - Jaccard Similarity

MCQ EVALUATION PROMPT

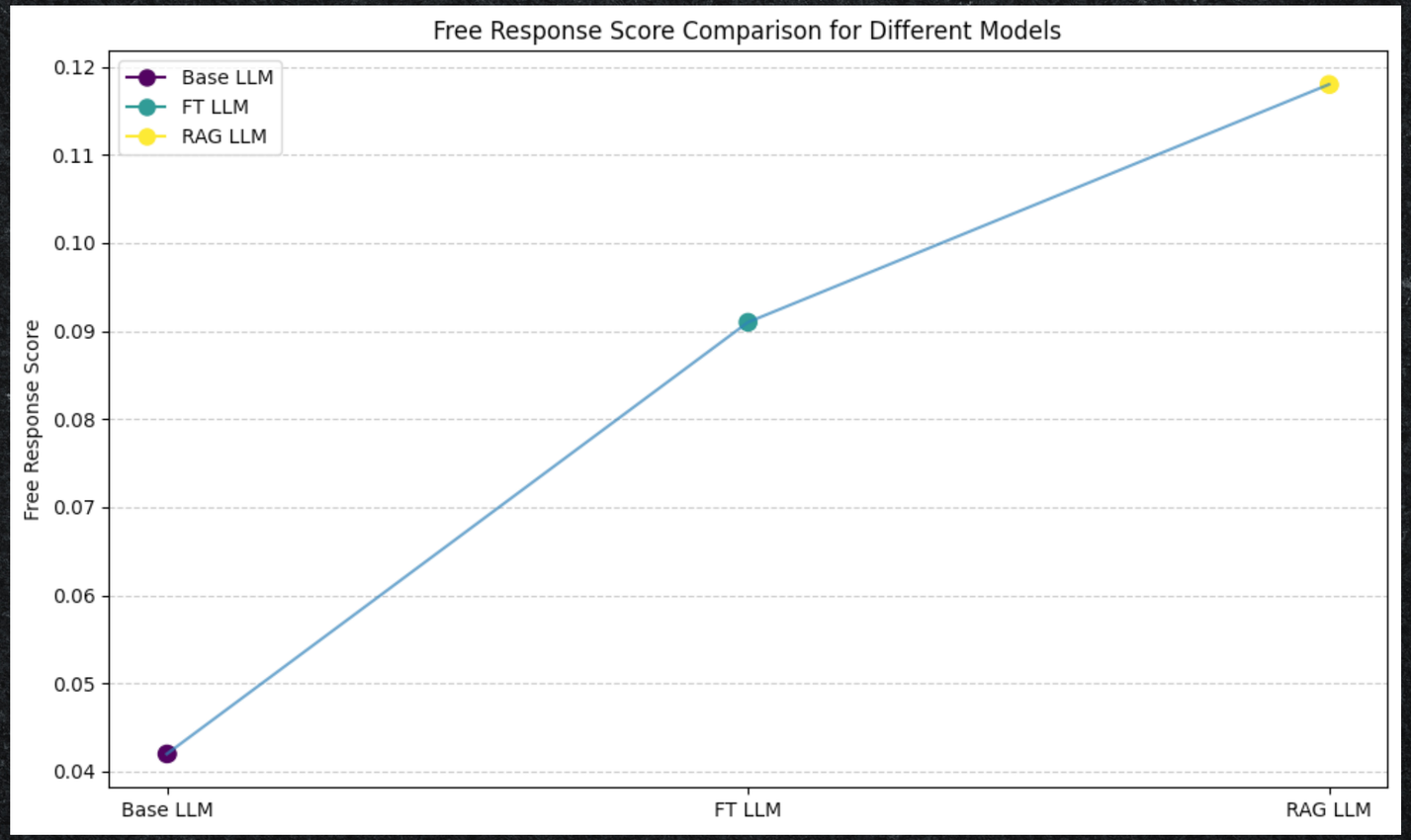
```
system_instruction = """
    You are a huge fan of the TV Show Friends. You will be given a QUESTION and four OPTIONS.
    I want you to ANSWER the QUESTION with the following steps.

    Evaluation Steps:
    1. Read the QUESTION carefully.
    2. Choose the correct OPTION from OPTIONS best of your knowledge.
    3. Output the ANSWER which is a single alphabet from A, B, C, D which is the right OPTION for the QUESTION
    4. The Output format for each OPTION is
        for A: 'ANSWER: A'
        for B: 'ANSWER: B'
        for C: 'ANSWER: C'
        for D: 'ANSWER: D'

    Here are a few Examples for how I expect the answer to be.
    Examples:

    {
        QUESTION: What is the name of Ross and Rachel's daughter,
        OPTIONS:
            A. Emma
            B. Delilah
            C. Deborah
            D. Claire
        ANSWER: A
    },

    {
        QUESTION: What is Chandler Bing's Middle Name,
        OPTIONS:
            A. Meredith
            B. Muriel
            C. Richard
            D. Robert
        ANSWER: B
    }

    Based on the above Evaluation Steps and Examples now ANSWER the QUESTION I give you
```

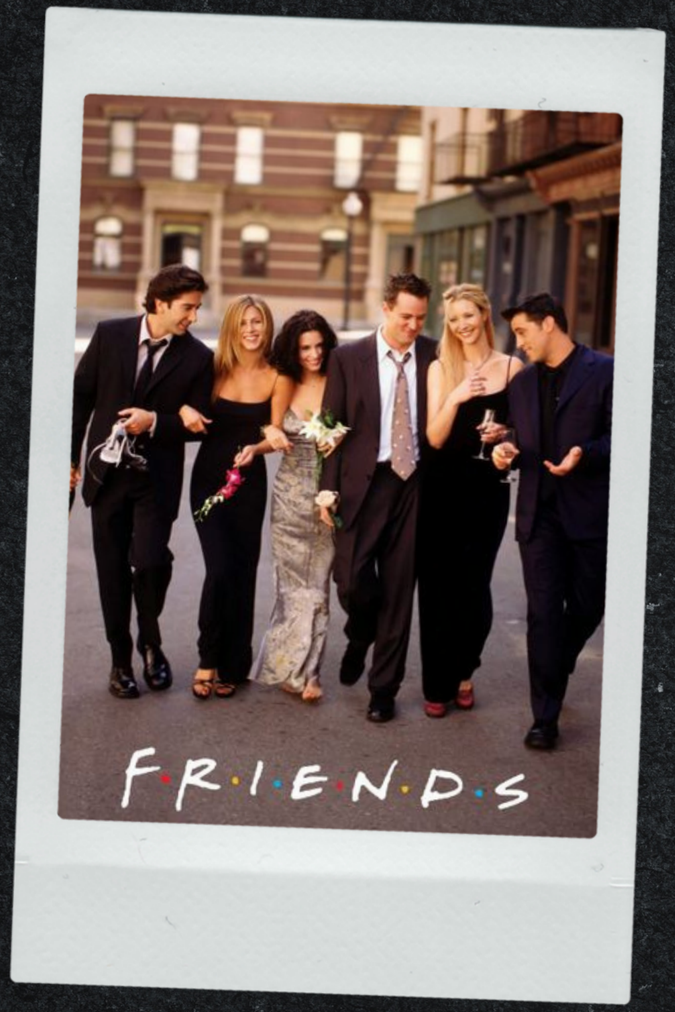
RESULTS – MCQ EVALUATION





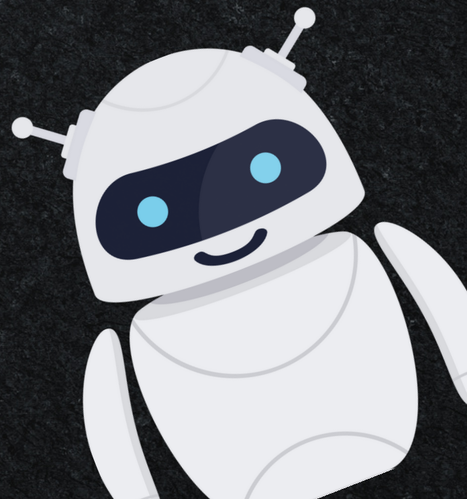
RESULTS – FREE RESPONSE EVALUATION

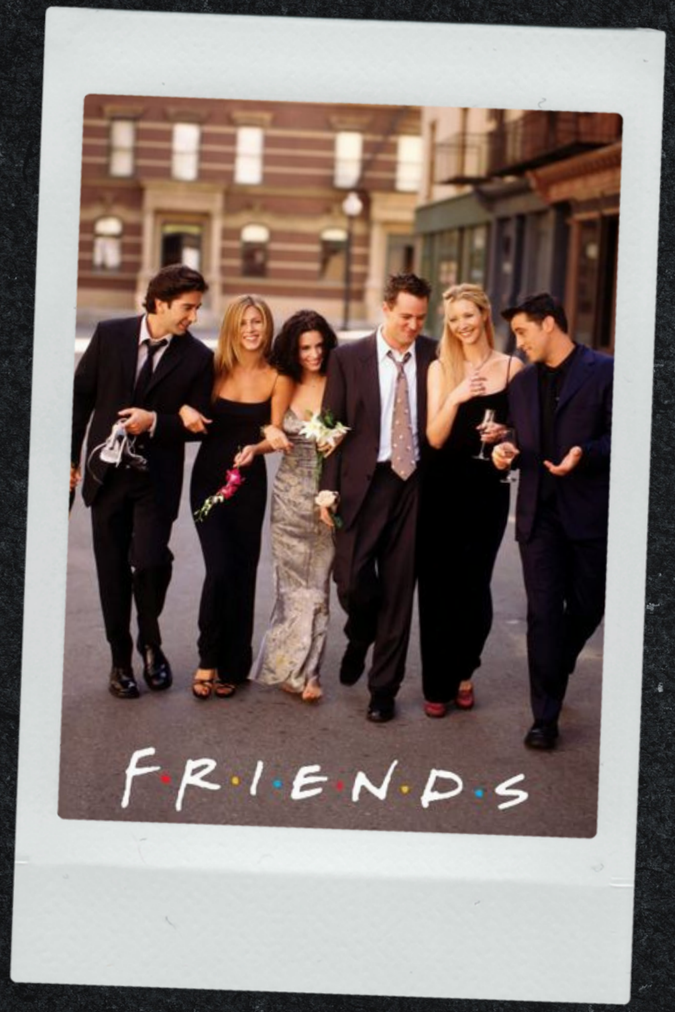




IMPROVEMENTS

- Compute!!! Increasing Number of epochs and choosing a larger LoRA dimension for fine-tuning.
- Fine-tuning was done on dialogue data and not on QA data(Instruction Fine-tuning)
- Increase the (context) top-k value for the retriever from 4 to 128
- The Feedback loop
  - Self-reflexion
  - Self-consistency





THANK YOU!