FINANCE CHATBOT FOR INDIAN STOCK MARKET

IE643 COURSE PROJECT

TEAM NAME : DEEP DIVERS
ROHIT KOURAV : 22B0720
YASH SHAH : 22B0717

OUTLINE

DESCRIPTION OF THE PROBLEM

- WE WERE ASSIGNED THE TASK OF MAKING A QUESTION-ANSWERING CHATBOT BASED ON EXISTING MODELS BY FINE-TUNING THEM ON OUR CUSTOM DATASET
- THE CHATBOT SHOULD ANSWER GENERAL QUERIES REGARDING THE INDIAN STOCK MARKET, GIVE INSIGHTS INTO NEWS RELATED TO THE INDUSTRY OR SPECIFIC COMPANY AND PROVIDE SUGGESTIONS ABOUT QUESTIONS IF ASKED
- EXPLORING MODELS AND PAIRING THEM WITH DATASETS THAT ALIGN WITH THE NEEDS OF TRAINING AND FINE TUNING
- PREPARING EVALUATION METRICS FOR CHECKING THE PERFORMANCE OF THE MODEL AND DESIGNING AN INTERACTIVE INTERFACE THAT IS EASILY UNDERSTANDABLE AND USABLE BY THE END USER.

PROJECT WORKFLOW

DATA COLLECTION → DATA CLEANING → DATA PREPROCESSING

DATASET CREATION

MODEL SELECTION AND RESEARCH → MODEL FINE-TUNING → MODEL EVALUATION AND TESTING

MODEL CREATION

INTERFACE CREATION

WORK DONE BEFORE PREP PRESENTATION

01 EXPERIMENTING AND EXPLORING THE DIFFERENT TECHNIQUES FOR DATA COLLECTION

02 CONVERTING THE DATA INTO THE RIGHT JSON FORMAT FOR FEEDING INTO THE MODEL

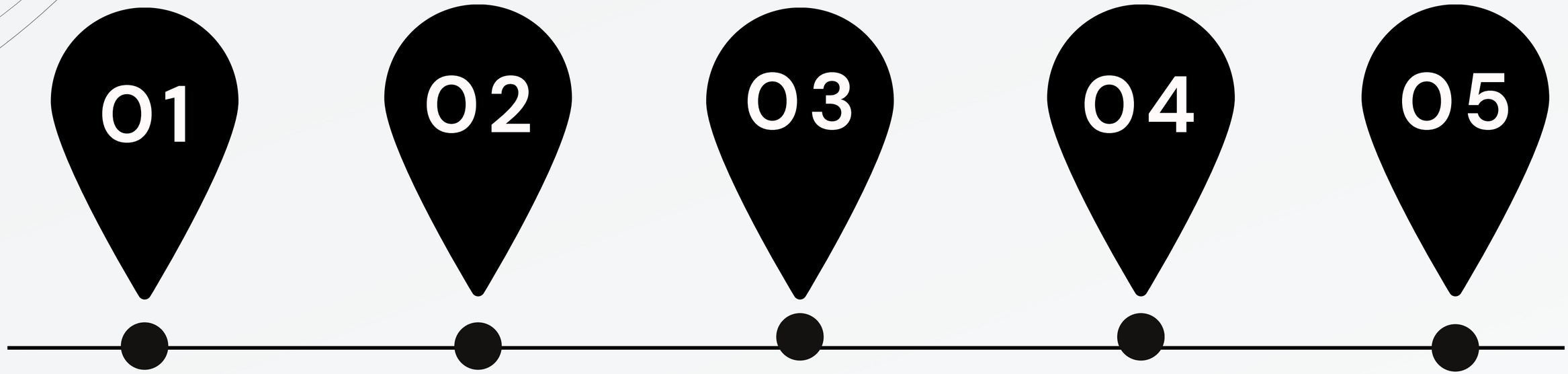
03 EXPLORING DIFFERENT MODELS WHICH CAN BE USED FOR THE TASK – GPT, BERT, FINBERT, ETC.

04 LEARNING ABOUT THE FUNDAMENTALS AND WORKING OF A CHATBOT

05 LEARNING TOOLS LIKE BEAUTIFULSOUP AND HUGGINGFACE FOR FUTURE USE

06 CREATED A ROADMAP FOR THE WEEKS TO FOLLOW

WORK DONE AFTER PREP PRESENTATION



01 DATA COLLECTION AND CLEANING

- USED MONEY CONTROL TO SCRAPE FINANCIAL NEWS DATA
- USED BEAUTIFULSOUP AND CREATED THE CODE TO EXTRACT DATA AND CLEANED IT TO FEED INTO THE CHATBOT

02 DATA FORMATING AND SELECTION

- DATASET WAS SCATTERED AND SEPARATED
- CONVERTED INTO READABLE FORMATS INTO CSV

03 DATA PRE-PROCESSING

- MODELS FOR QUERY UNDERSTANDING AND CLASSIFICATION
- MODEL FOR SENTIMENT ANALYSIS
- MODELS FOR STOCK PRICE PREDICTION

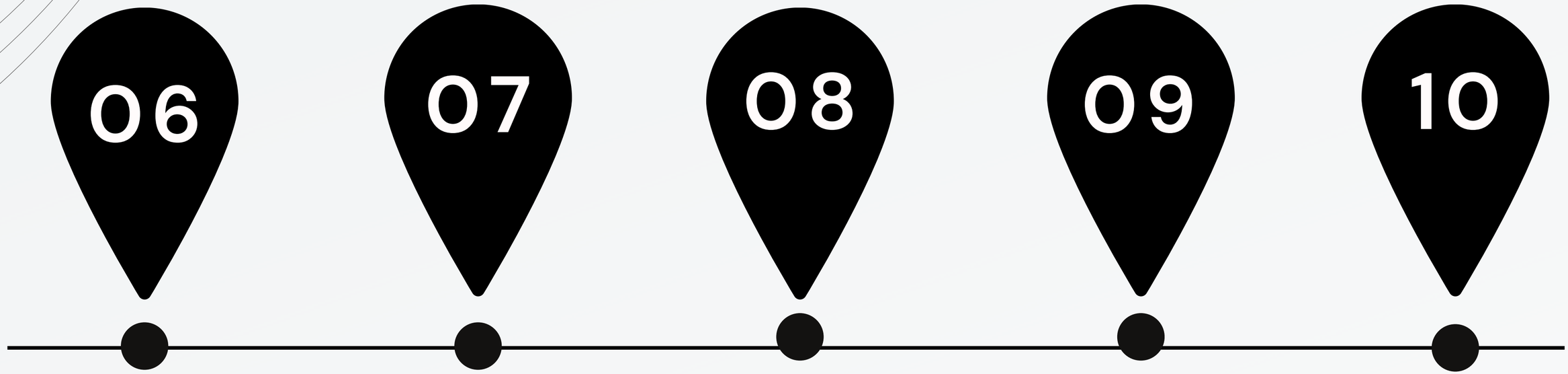
04 MODEL DESIGN AND TRAINING

- BUILDING MODEL PIPELINES
- FINE-TUNING MODELS ON COLLECTED DATA

05 PARAMETER SETTING

- DECIDING ON THE MODEL PARAMETERS
- ADEQUATE DATA TO BE GIVEN
- NUMBER OF EPOCHS AND ALL

WORK DONE AFTER PREP PRESENTATION



06 INTRODUCING PEFT

- FULL MODEL COULD NOT BE TRAINED AGAIN
- USED LORA UNDER PEFT LIBRARY

07 EVALUATION AND TESTING

- DEFINING PERFORMANCE METRICS
- TEST THE MODEL ON MULTIPLE QUERIES
- IMPLEMENT FEEDBACK LOOP FOR CONTINUOUS LEARNING

08 ASSESSING THE RESULTS

- CHECKED RESULTS BASED ON DIFFERENT SETTINGS
- USED THE MOST EFFICIENT FOR INTERFACE BUILDING AND FURTHER WORKS

09 CREATION OF INTERFACE

- CREATING A BASIC CHATBOT INTERFACE
- ADDING DESIGN ELEMENTS IF REQUIRED
- MAKING THE INTERFACE USER FRIENDLY AND EASY TO UNDERSTAND

10 MODIFICATIONS IN THE INTERFACE

- ADDING REVIEW FUNCTION
- REMOVING THE SCORE FUNCTION DUE TO INADEQUACY OF THE OUTPUT.

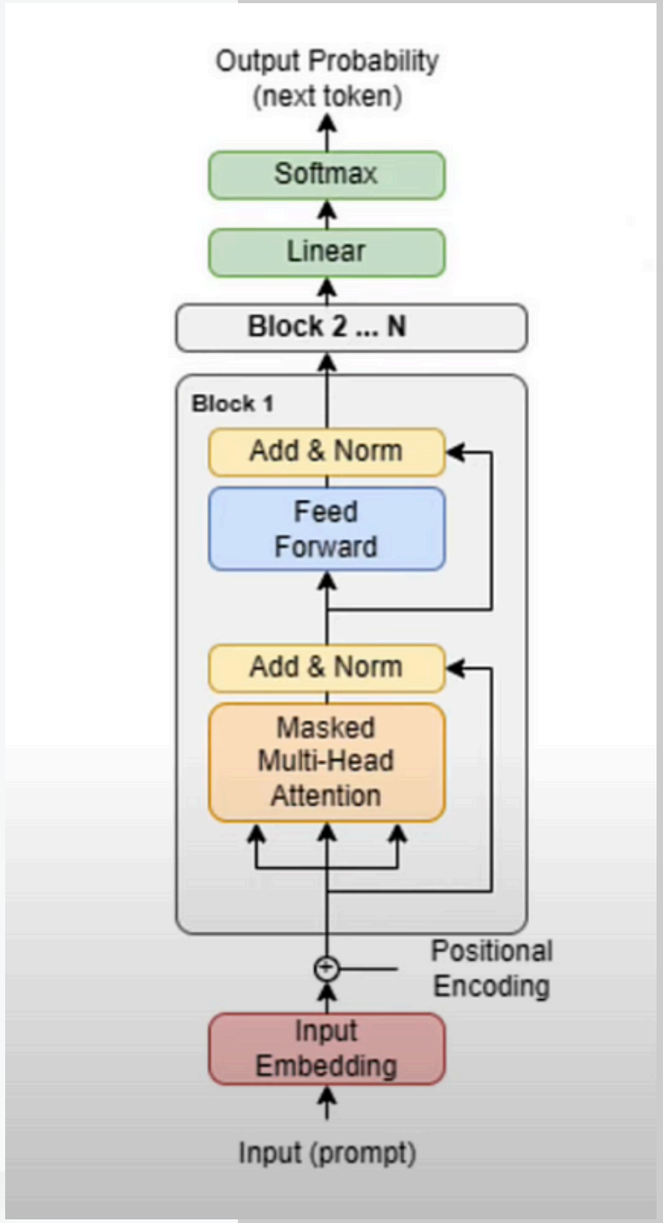
APPROACHES

Initially, we were going with the Bert and associated model, but due to a change in the dataset and not getting the required results we switched to LLama 2.

Llama 2 is a suite of LLMs with 7 to 70 billion parameters, including versions fine-tuned for chat (Llama 2-Chat) and optimized for dialogue.

Fine-tuning and safety enhancements make these models competitive with open-source benchmarks, offering an alternative to closed-source models.

The models are large and resource-intensive to fully train, requiring significant time and memory. Techniques like LoRA help efficiently reduce both memory usage and training time.



ADAPTATIONS

Updates all model parameters, requiring substantial memory and computational resources, making it challenging for very large models.

FULL MODEL

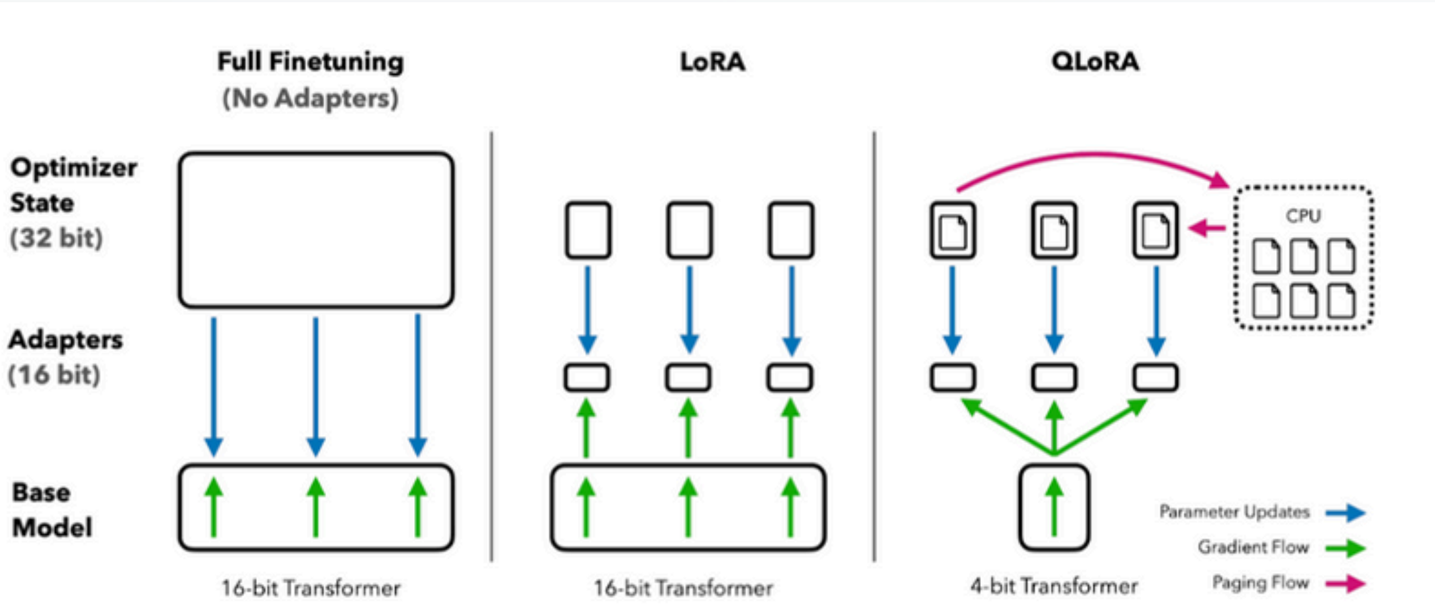
Adds trainable low-rank adapters, keeping the base model fixed, reducing memory and computational demands during fine-tuning.

LORA

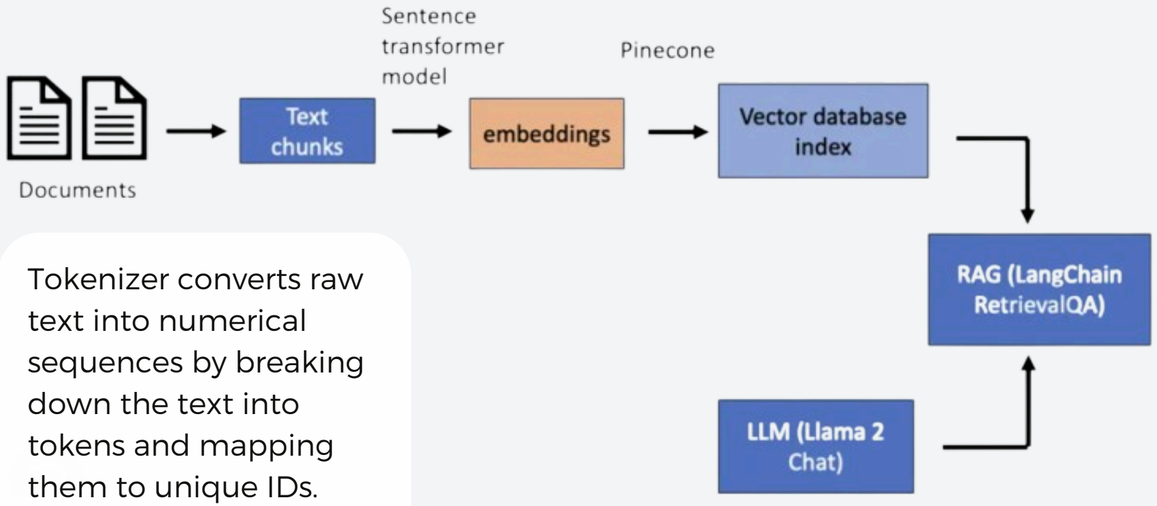
Combines LoRA with 4-bit quantization and CPU offloading, enabling efficient fine-tuning of large models within limited memory.

QLORA

- LoRA and QLoRA enhance large language model fine-tuning by freezing pre-trained weights and injecting trainable low-rank matrices, reducing trainable parameters and GPU memory usage.
- QLoRA quantizes models to 4 bits, enabling fine-tuning of 65B parameter models on a single 48GB GPU while maintaining performance, achieving significant memory reduction and outperforming existing models



TOKENIZER AND TRAINING SETUP



Tokenizer converts raw text into numerical sequences by breaking down the text into tokens and mapping them to unique IDs.

ENCODING

Tokenizer transforms model-generated numerical IDs back into human-readable text by mapping IDs to tokens and reconstructing the output string.

DECODING

SETUP

- Model Selection: The NousResearch/Llama-2-7b-chat-hf model from Hugging Face is utilized for its conversational capabilities.
- Tokenizer Initialization: A tokenizer for the LLaMA model is loaded for accurate text encoding.
- QLoRA Configuration: The BitsAndBytes configuration enables 4-bit quantization, optimizing memory usage and computational efficiency.
- LoRA Parameters: Low-Rank Adaptation (LoRA) settings are defined to facilitate effective fine-tuning with minimal resource consumption.
- Training Arguments: Hyperparameters such as batch size, learning rate, and epochs are established for structured training.
- Gradient Accumulation: Gradient accumulation allows efficient training with smaller batch sizes, enhancing performance.
- Dataset Preparation: Input text is converted into Hugging Face Dataset format for seamless integration.
- Training Execution: The SFTTrainer is used to execute training based on the defined configurations.

DATA DESCRIPTION

- Our dataset consists of Financial News data for the Indian Stock Market scraped from the ‘moneycontrol’ website.
- We have data from 3 categories of news: Stock Market related news, IPO-related news, and Company business related news.
- Our dataset consists of 4 columns - URL of the data, News Headline, Summary of the news, and Main news body. There are 1500+ rows of such data, which was used to fine-tune the model.
- This data has been scraped using the BeautifulSoup Library. It has also been cleaned to convert the raw textual data into clean data which can be fed into the model.
- The initial data had URLs, Titles, Summary and body of the news stored in CSV later converted to different formats for training.

CLEANING AND PROCESSING

Cleaning

- News had unwanted theoretical components like links, also read, and disclaimers which would unnecessarily train the data on wrong inputs, thus they were removed.
- Data was fed into the cleaning function that removed these irregularities and some other unnecessary presence of unwanted texts.

- Data was in CSV format which was not supported for use thus it was converted into a format shown beside which is understood by the model and is part of huggingface.
- The final data was in the format of question and answer which was stored between the syntaxes as shown in the format beside.

Understandable Format

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

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

EXPERIMENTATION SETUP FOR IMPROVEMENT OF RESULTS

- The model once trained requires prompt to be given by user which it responds by giving text outputs.
- To reduce loss we tried to change the number of epochs and also tried to overfit the model on our custom dataset.
- For that purpose, we copied a bunch of news 5-10 times to make a repeating dataset that could overfit the model.
- We tried to change datasets in particular ways to check on different datasets.
- Tried on the same dataset but reduced the length of answers in the question-answering format.

Step	Training Loss
25	1.408600
50	1.662100
75	1.215200
100	1.444100
125	1.176900
150	1.367300
175	1.174000
200	1.468100
225	1.158200
250	1.542700

EXPERIMENTAL OUTCOMES FOR THE ADJUSTMENTS

- On the epoch part, we noticed that on any dataset if we even increase epochs beyond 3 then also we have a very small decrease in loss and it becomes constant after some iterations.
- Changing the dataset has led to a decrease in loss only when the dataset was the one on which the model was originally trained.
- Copying large news interactions again and again in the dataset thus helps to increase accuracy and reduce loss significantly but leads to overfitting.
- Overfitting in this case is good if we only want to get news insights but is bad if we will go with general queries.
- The best BLEU score is shown here and the loss related to it with the good nature will be discussed in the next slide.

Step	Training Loss
25	1.971400
50	1.656400
75	1.457100
100	1.453800
125	1.326000
150	1.345400

```
Generated Text: <s>[INST] Provide
BLEU Score: 0.4984588154936228
```

EXPERIMENTAL OUTCOMES FOR THE ADJUSTMENTS

- This loss was observed on a dataset where the answers were pretty short to train. The word corpuses in the answer were around 80 only.
- This dataset also had around 120 news which were repeated around 5-8 times.
- The BLEU score was also around 0.5 which in itself was not very good but is significantly better than other cases.
- Epochs don't seem to affect much the outcome and the training generally reaches its saturated loss and accuracy on 2 to 3 itself, the major effect is due to change of dataset
- The average rating for various use cases was better when the data was large and general, although not performing well on the training set but answering general queries better.

Step	Training Loss
25	1.788400
50	0.875900
75	0.498200
100	0.661100
125	0.542300
150	0.408700
175	0.567600
200	0.537400
225	0.348900

CONCLUSIONS

Model

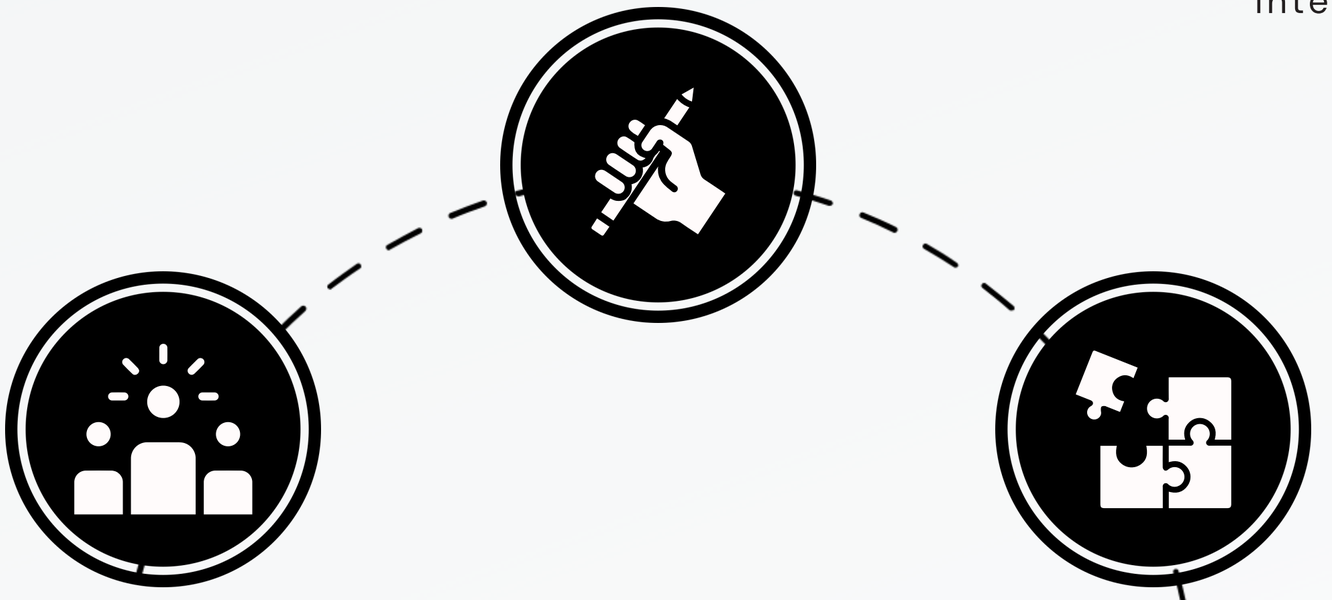
Using LoRA and QLoRA for the model can reduce efficiency but significantly improve time and memory management. It's important to choose a model that aligns with the available memory and GPU capacity.

Dataset

The dataset should be clean and compatible with the model. Datasets with short, repetitive question-answer pairs may lead to overfitting, which can be beneficial for specific, narrow applications but may limit generalizability. For broader queries, a more diverse dataset is recommended.

Results

The model performs well with repetitive datasets, but increasing the number of epochs does not yield significant improvements, as both loss and accuracy tend to plateau at higher levels. Overall, a balanced, intermediate approach works best.



PLAN FOR NOVELTY ASSESSMENT

- In the future, we aim to add dual answerability to the model, allowing the chatbot to respond to multiple aspects of a single query.
- This feature will require an updated interface to display comprehensive responses, enhancing the user experience by addressing diverse query facets.
- Implementing dual answerability also demands a specialized, cleaned dataset with well-structured formatting to ensure accurate and relevant responses.

RESEARCH PAPERS AND REFERENCES

01

AUTHORS: NEIL HOULSBY, ANDREI GIURGIU, STANISŁAW JASTRZE¸BSKI, BRUNA MORRONE, QUENTIN DE LAROUSSILHE, ANDREA GESMUNDO, MONA ATTARIYAN, AND SYLVAIN GELLY
TITLE: PARAMETER-EFFICIENT TRANSFER LEARNING FOR NLP
PUBLISHED JUN 2019 ACCESSED OCT 2024
LINK: HTTPS://ARXIV.ORG/PDF/1902.00751

02

AUTHORS: TIM DETTMERS, ARTIDORO PAGNONI, ARI HOLTZMAN, LUKE ZETTLEMOYER FROM UNIVERSITY OF WASHINGTON
TITLE: QLORA: EFFICIENT FINETUNING OF QUANTIZED LLMS
PUBLISHED MAY 2023 ACCESSED OCT 2024
LINK: HTTPS://ARXIV.ORG/PDF/2305.14314

03

AUTHORS: HUGO TOUVRON, THIBAUT LAVRIL, GAUTIER IZACARD, XAVIER MARTINET MARIE-ANNE LACHAUX, TIMOTHEE LACROIX, BAPTISTE ROZIÈRE, NAMAN GOYAL ERIC HAMBRO, FAISAL AZHAR, AURELIEN RODRIGUEZ, ARMAND JOULIN EDOUARD GRAVE, GUILLAUME LAMPLE.
TITLE: LLAMA: OPEN AND EFFICIENT FOUNDATION LANGUAGE MODELS
PUBLISHED FEB 2023 ACCESSED OCT 2024
LINK: HTTPS://ARXIV.ORG/PDF/1902.00751

CODE REPO AND WEBSITES

Abid Ali

Link: “https://www.datacamp.com/tutorial/fine-tuning-llama-2”

Published:Oct, 2023 Accessed: Oct, 2024

Krish Naik

Link: “https://colab.research.google.com/drive/1Bd7c5rioBOmtJbDEax83vAHEPru-r06l”

Published:Feb, 2024 Accessed: Oct, 2024

Published:Jun, 2020
Accessed: Oct, 2024

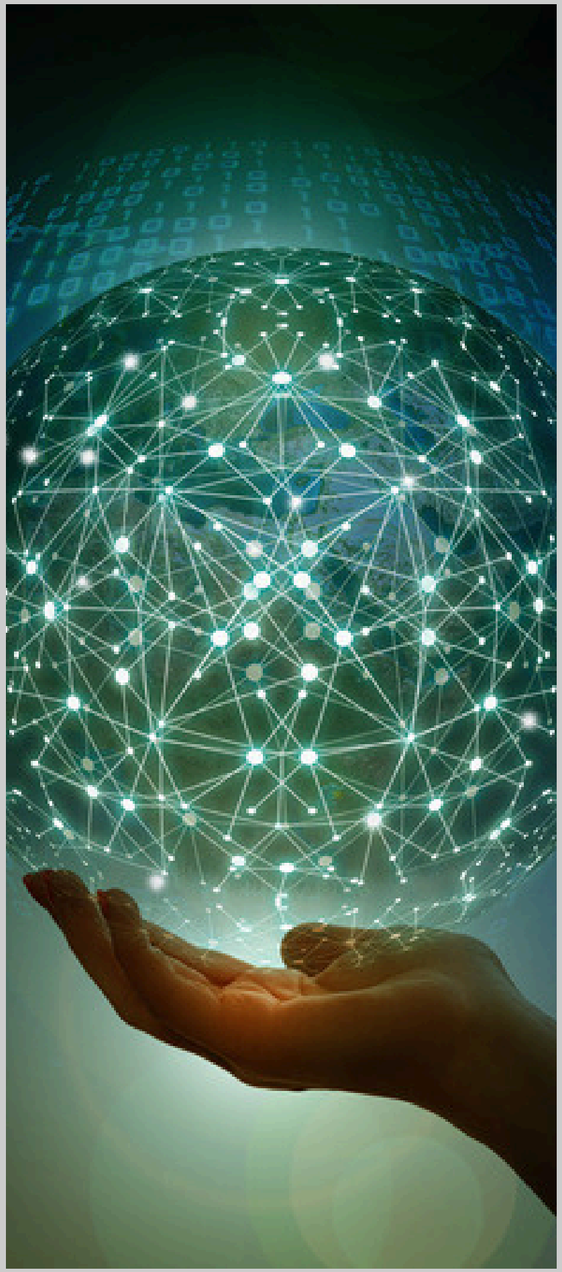
Link: https://www.connectedpapers.com/main/32ac52069e562d4f900afee70bdca63f53461481/graph

Alex Tarnavsky Eitan

Eddie Smolyansky

Itay Knaan Harpaz

Sahar Perets



SIGNING OFF

TEAM: DEEP DIVERS