PDF QA using fine-tuned LLaMA-3 8B

Swadesh Swain

Enrollment No: 22116091

Electronics and Communication Engineering

Indian Institute of Technology Roorkee

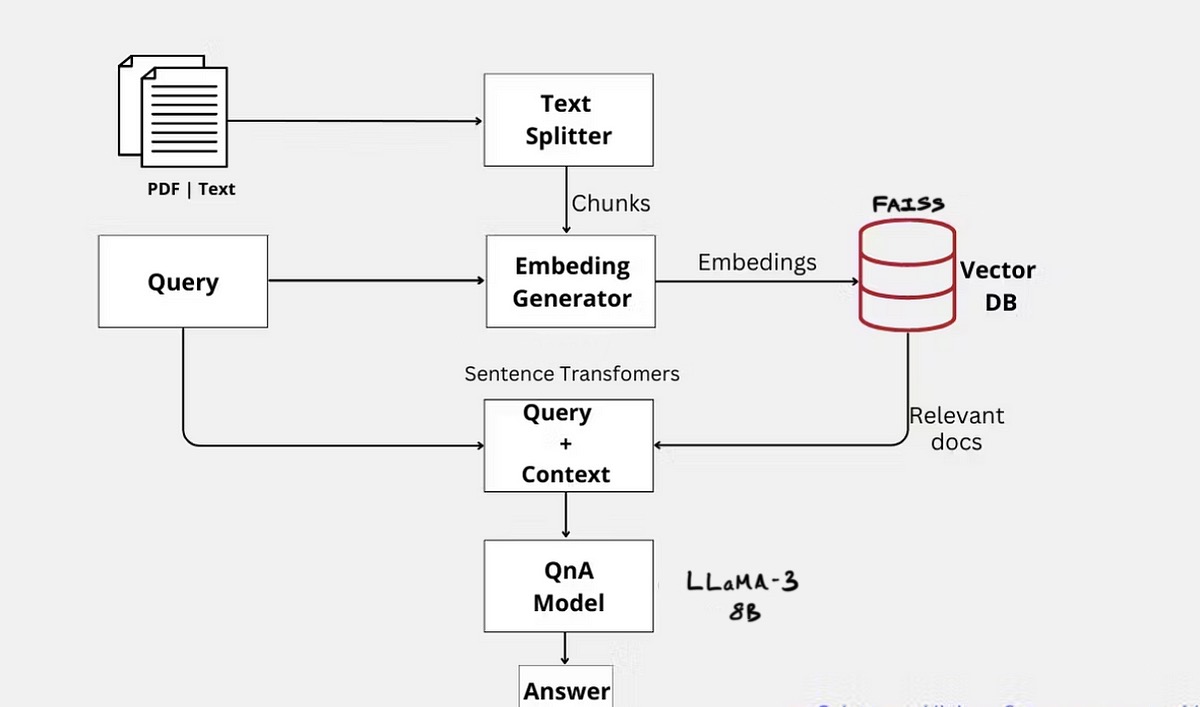
ARiES summer open project - 2024

27 May - 20 June

# Introduction

This student project was aimed at making a working pipeline where in modern AI tools and Language models can be used to perform the task of answering user queries and questions from a pdf. Please note that the entire project has been done without the use of an API and minimal use of the internet. In this particular project, Meta AI’s LLaMA 3 model, the 8 billion parameter version, has been used as a fine-tuned model to answer user questions. A popular framework for integrating Large Language models , Langchain , has also been used in this project , however , its involvement was limited to the creation of vector databases from the user’s document. To demonstrate efficiency and ease of use , I have developed the project and fine-tuned the model, entirely on Google Colab , using only the free available resources . This was possible only after careful iteration and numerous trial and error experiments to find out the best set of tools and methods to make this as compact and efficient as possible. The final code is a highly refined version of my work where I have tried to make the code as clean and concise as possible, with proper instructions along the entirety of it. The 8 billion parameter model of the LLaMA family was found to be the most effective and efficient for the task compared to other available open source models, a comprehensive list of which you will be presented with in the following sections. The entire code along with the other components of the project is available on github on the following link : <https://github.com/Swadesh06/ARiES_PDF_Q-A/tree/main>

# Approach



Highlighting my approach , it consists of the following steps:

1. Fine-tuning the LLaMA3-8B model on the alpaca\_cleaned dataset which involves the underlying steps:
   1. Using the Unsloth library to load the model in quantized 4 bit version for faster loading.
   2. Invoking RoPE (Rotary Positional Encoding) scaling (optional) to increase model context window size.
   3. Fine-tuning the model on the dataset consisting of a user query , and input to provide further context (this is fetched from the pdf as described later), and text output using Low Rank Adaptation fine-tuning for efficiency.
   4. Saving the LoRA adapters to load the model later for future use.
2. Loading a document (or a directory of Documents using PyMuPDFLoader, because upon trying numerous other loaders , for pdfs this turned out ot be the fastest. Other loaders (listed in the imports too) were PyPDFLoader, Directory Loader, and Unstructured. For loading, you must specify a file path or directory path in the code.
3. Using Recursive Character Splitting to split the raw documents extracted into smaller chunks for speed and efficiency as well as accuracy while searching for relevant information later.
4. Using an embedding function , in this case the all-MiniLM-L6-v2 embedding function, was used to embed the splitted documents into embedded vectors of 384 dimensions.
5. These embeddings were then used to create a vector database incorporating the Facebook AI Similarity Search (FAISS) , to store the chunks for retrieval. I have specifically used the gpu utilizing version (faiss-gpu) for faster creation of the database but if one’s work is resource constrained they can use faiss-cpu while installing dependencies
6. This is then stored in the path you specify in the save\_directory variable , so that there is no need to create a vector database again and again for the same document each time you run it.
7. A system prompt is designed for Retrieval Augmented Generation to be incorporated within the model with the input variable as the user query, relevant data retrieved from the pdf , and the model response. This enables us to pass the relevant information pertaining to the user’s query fetched from the database to be passed into the model along with and garner a good response from it.
8. Now finally , for the inference part

# Method (Heading Level 1)

## Participants (Heading Level 2)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan.

## Assessments and Measures

Ut blandit malesuada quam, ac varius tortor gravida eget. Vestibulum id ligula leo, ut accumsan mi. Sed tristique euismod convallis. Nulla facilisi. Etiam vestibulum est id orci interdum vitae porta enim blandit. Cras sit amet arcu dolor, at venenatis erat. Vestibulum accumsan placerat mauris. Morbi nec nibh nibh. Duis ultricies posuere nunc. Morbi at tellus quis magna vestibulum eleifend.

### Heading Level 3

Maecenas ullamcorper bibendum consequat. Pellentesque ultrices, eros eu tincidunt pretium, magna leo volutpat libero, non bibendum diam nunc eget urna. Vivamus eu tortor et dui aliquam vestibulum at vel augue. Vivamus elit dui, porttitor eget egestas at, rhoncus in justo. Curabitur tristique, elit ac venenatis volutpat, eros mauris iaculis diam, vitae rhoncus erat metus vitae eros.

**Heading level 4 is inline with the following paragraph, separated by a period.** Nulla congue egestas ante, id ultricies orci dignissim commodo. Fusce placerat, libero eu pharetra pulvinar, lorem dui pulvinar nisi, et semper orci orci vitae magna. Nullam sodales, felis id feugiat scelerisque, tortor nulla interdum mauris, ac porttitor odio dolor eget eros.

**Second level 4 heading.**Duis sit amet ipsum pretium erat accumsan iaculis vitae eget risus. Donec ut dui in lorem volutpat fermentum bibendum pulvinar libero. Nunc imperdiet eros et mi posuere pellentesque. Donec tincidunt ipsum eget nisl ullamcorper eu placerat libero ullamcorper. Maecenas id luctus ligula. Cras condimentum eleifend nibh sit amet iaculis. Suspendisse placerat sollicitudin mi, vel ornare augue hendrerit ac. Nulla sed suscipit sapien. Cras pellentesque orci lectus, eu consequat enim.

***Heading level 5 is also inline.*** Nulla congue egestas ante, id ultricies orci dignissim commodo. Fusce placerat, libero eu pharetra pulvinar, lorem dui pulvinar nisi, et semper orci orci vitae magna. Nullam sodales, felis id feugiat scelerisque, tortor nulla interdum mauris, ac porttitor odio dolor eget eros.

***Second level 5 heading.*** Duis sit amet ipsum pretium erat accumsan iaculis vitae eget risus. Donec ut dui in lorem volutpat fermentum bibendum pulvinar libero. Nunc imperdiet eros et mi posuere pellentesque. Donec tincidunt ipsum eget nisl ullamcorper eu placerat libero ullamcorper. Maecenas id luctus ligula. Cras condimentum eleifend nibh sit amet iaculis. Suspendisse placerat sollicitudin mi, vel ornare augue hendrerit ac. Nulla sed suscipit sapien. Cras pellentesque orci lectus, eu consequat enim.

# Results

Maecenas id luctus ligula. Cras condimentum eleifend nibh sit amet iaculis. Suspendisse placerat sollicitudin mi, vel ornare augue hendrerit ac. Nulla sed suscipit sapien. Cras pellentesque orci lectus, eu consequat enim.

## Outcome 1

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

## Outcome 2

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

# Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan.

# References

Lastname, C. (2008). Title of the source without caps except Proper Nouns or: First word after colon. *The Journal or Publication Italicized and Capped*, Vol#(Issue#), Page numbers.

Lastname, O. (2010). Online journal using DOI (digital object identifier). *Main Online Journal Name*, Vol#(Issue#), 159-192. <https://doi.org/10.1000/182>

Lastname, W. (2009). *Title of webpage*. Site Name. Retrieved July 3, 2019, from<http://www.example.com>