INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

EXPLAINABLE HATE SPEECH DETECTION USING GENERATIVE AI

Mini Project - 1

DR.SUNIL SAUMYA

Our Team

Ch Srinivasa Sai

21BDS012

K Sai Kartheek Reddy

21BDS027

K Abhiram

21BDS029

R Vinay Kumar

21BDS056

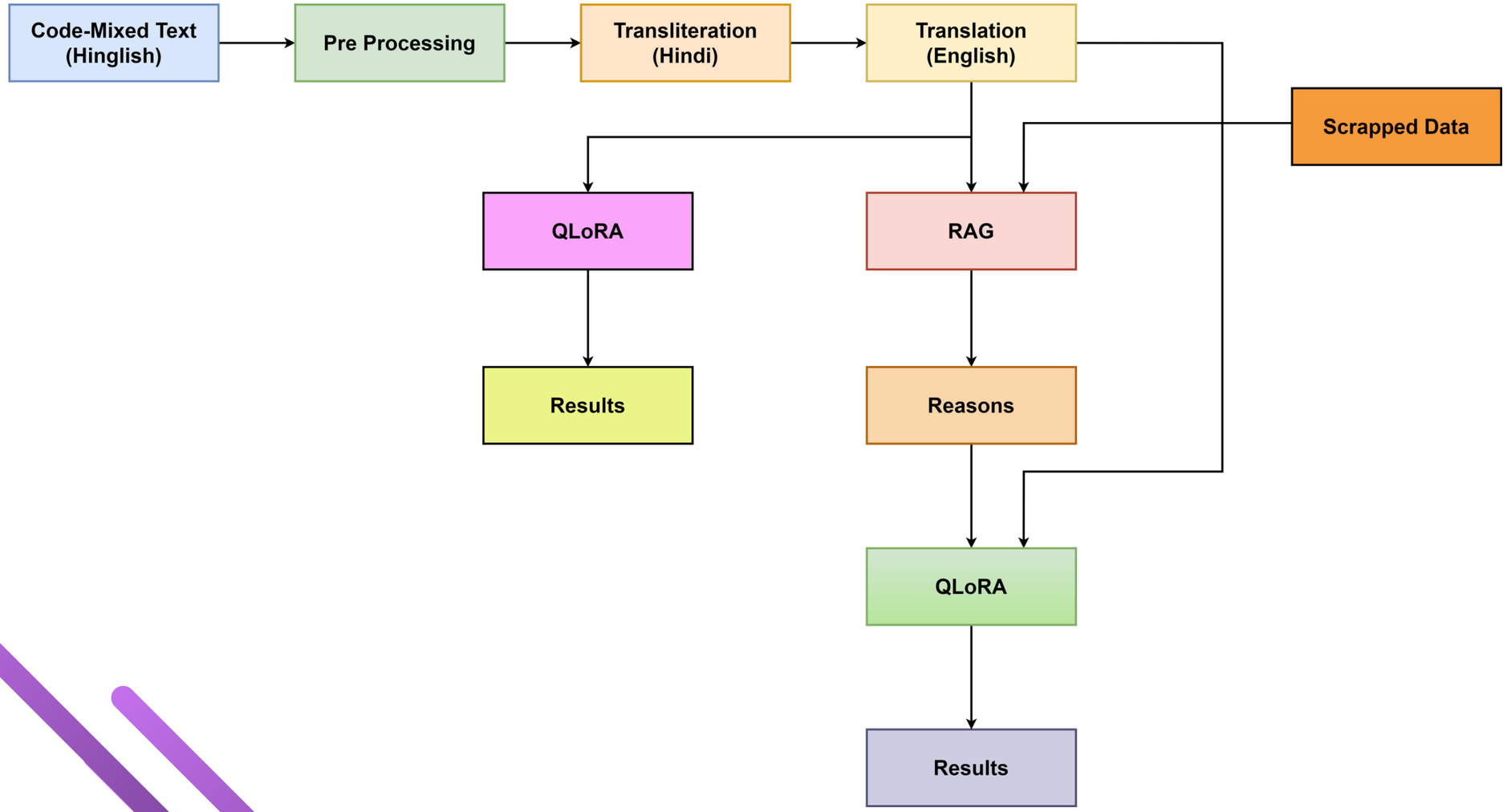
Problem Statement

Online platforms are facing a growing challenge: how to tackle hate speech and fake news effectively. With more people sharing content online, it's become harder to spot harmful messages and false information. Traditional methods of moderation take too long and can't keep up with the flood of posts. To address this, our project aims to use advanced technology, like QLoRA and RAG, to automatically detect and classify hate speech and fake news with accuracy and speed.

Introduction

The internet is a wonderful place to connect and share ideas, but it also has its dark side. Hate speech and fake news are spreading fast, causing problems like division and misinformation. Our project wants to fight back using ML techiques. We're using tools like QLoRA and RAG to quickly find and sort out the bad stuff from the good, making the internet a safer and more trustworthy place for everyone.

Flow Chart



This flowchart unveils the magic behind text summarizers, those tools that automatically condense lengthy texts. First, you provide the text, which is then prepped for analysis by transforming it into a computer-friendly format and breaking it down into manageable chunks. To guide the summarization, a prompt is created, specifying things like desired length or focus areas. The system then delves into a vast database, searching for similar text chunks to grasp the context and identify important information. Next, it analyzes your text to determine its type, like a news article or recipe. Finally, armed with knowledge from similar texts, the text's category, and its overall meaning, the system crafts a concise summary that captures the essence of the original content.

EXPERIMENTS WITH TRANSILERATION AND TRANSLATION

1 . Transliteration

- Indic Xlit is an open-source transliteration tool that can convert text between Roman script and multiple Indic languages.
- IndicXlit can be used to transliterate Romanized languages to normal languages, making it easier to access and understand the content. It uses 25 indic Romanized languages to translate
- IndicXlit has successfully converted our romanized Hindi dataset into standard Hindi frames, enabling seamless access to Hindi content

AI4BHĀRAT

HTTPS://AI4BHARAT.IITM.AC.IN/INDICXLIT-MODEL/

EXPERIMENTS WITH TRANSILERATION AND TRANSLATION

2 . Translation

- IndicTrans2 is a multilingual machine translation model that can translate between 26 Indic languages, including Hindi and English.

- IndicTrans2 is an open-source model, making it freely available for use and modification.

- IndicTrans2 has successfully converted our Hindi dataset into standard English frames, making it easier to communicate with English speakers.

AI4BHĀRAT

HTTPS://AI4BHARAT.IITM.AC.IN/INDIC-TRANS2//

After Pre-Processing

Original Records	After Trasliterated	After Translation
China Tumhari Bhen ki Chut Saale Puri Duniya ko De diye Corona Virus Hum ab pareshan Ho gye Madarchod tumhare Wajah se Joining rukhi hui hai Meri Sena ki Ab india ke Tour pe aana mat Tum warna Zinda bach ke nae ja payega tum Tum toh kisi Social site pe bhi nae ho ki Galli de Tmko	चीना तुम्हारी भें की चुट साले पूरी दुनिया को दे दिये कोरोना वायरस हम अब परेशन हो गये मदारचोद तुम्हरे वजह से जॉइनिंग रूखी हुई है मेरी सेना की अब इंडिया के तौर पे आना मत तुम वारना जिंदा बच के ने जा पाएगा तुम तुम तोह किसी सोशल साइट पे भी ने हो की गल्ली दे तमको	China hands over your buffalo calves to the whole world We are now ready for the corona virus Mother Chod joining has been withheld due to you Now do not come as India of my army you will be able to escape alive You are not even on any social site give a hug to me
China ki setting se ravish kumar ko "ramon magsaysay" philipins award mila tha Kyun ye desh virodh propaganda desh main chalata hai hai duniya main ye akela reporter tha jo corona virus par china ko support kr raha tha ye congress ka mouthpiece hai but ab No more Khabish kumar https://t.co/BOBYblzewA	चीना की सेटिंग से रवीश कुमार को रामों मैगसेसे फिलिपिंस अवार्ड मिला था क्यून ये देश विरोध प्रोपगंडा देश मैं चलता है है दुनिया मैं ये अकेला रिपोर्टर था जो कोरोना वायरस पर चीना को सपोर्ट केआर रहा था ये कांग्रेस का माउथपीस है बट अब नो मोरे खबीश कुमार एचटीटीपीएस टी सीओ बॉबीब्ल्जेवा	Ravish Kumar was awarded the Ramon Magsaysay Philippines Award for his work on the setting of China, because he was the only reporter in the world who was supporting China on coronavirus.
India has reached the bhaad me jaaye stage in respect of corona virus.Jitna freely apne desh me logg ghum rahe hai utna to shayad china vaale bhi nahi ghum rahe hongeÃ˚Â▯Â▯Â▯#CoronavirusIndia #coronavirus	इंडिया हस रीच्ड थे भाड़ मे जाए स्टेज इन रिस्पेक्ट ओएफ कोरोना वायरस जितना फ्रीली अपने देश मे लॉग घुम रहे है उतना तो शायद चीना वाले भी नही घुम रहे होंगे कोरोनावायरसइंडिया कोरोनावायरस	India was so touched by the corona virus that it is likely that the Chinese may not be able to travel as much as the corona virus in their country

Low Rank Adaptation (LoRA)

LoRA (Low Rank Adaptation) revolutionizes deep learning models by redefining how weight updates are handled. Instead of directly updating weights, LoRA tracks changes in separate, smaller matrices. These matrices are then multiplied together to form a matrix equivalent in size to the model's weight matrix.

Explaining the Different uses of Low Rank Adaptation (LoRA):

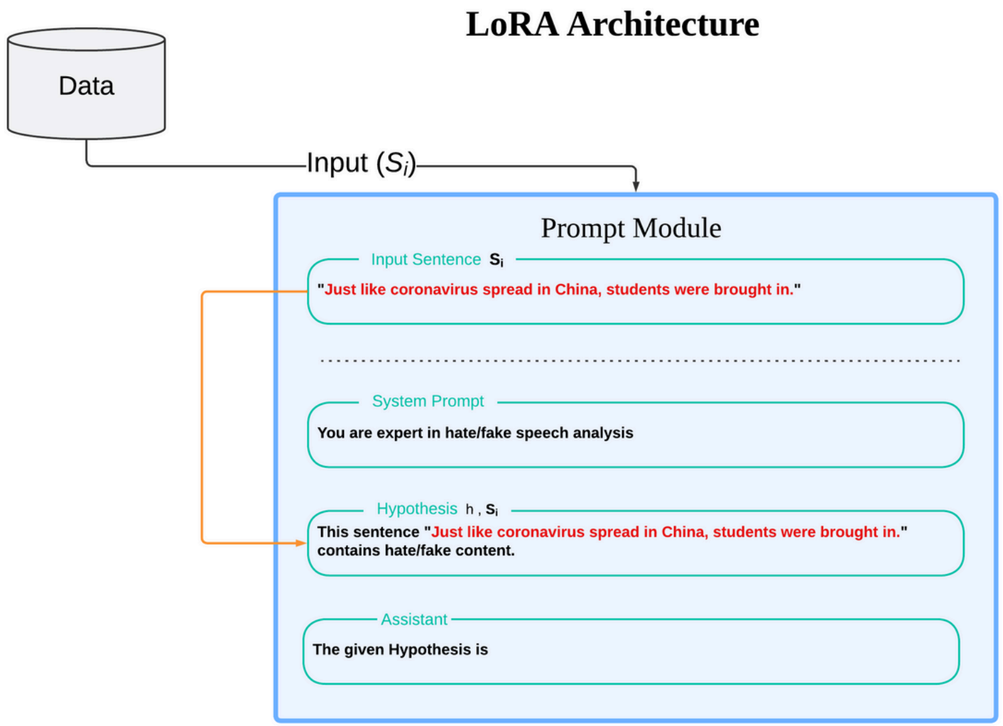
01

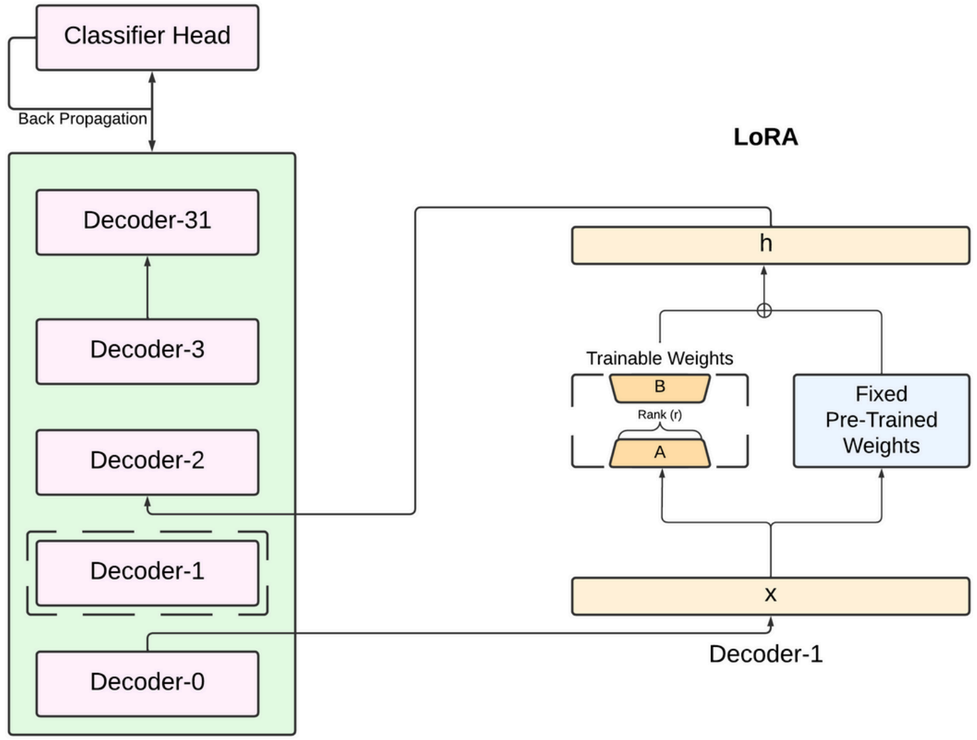
Weight Update Tracking: LoRA refrains from directly updating weights. Instead, it monitors changes in separate, smaller matrices.

02

Matrix Multiplication: These tracked weight changes are stored in two distinct matrices, which are multiplied together. This multiplication yields a matrix matching the size of the model's weight matrix.

LoRA Architecture





Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) is a state-of-the-art approach in natural language processing, combining retrieval-based techniques with generative models to enhance the output of Large Language Models (LLMs). LLMs, such as ChatGPT 3.5, are trained up to a specific point in time, potentially lacking recent information. RAG addresses this limitation by integrating fresh data into the generation process.

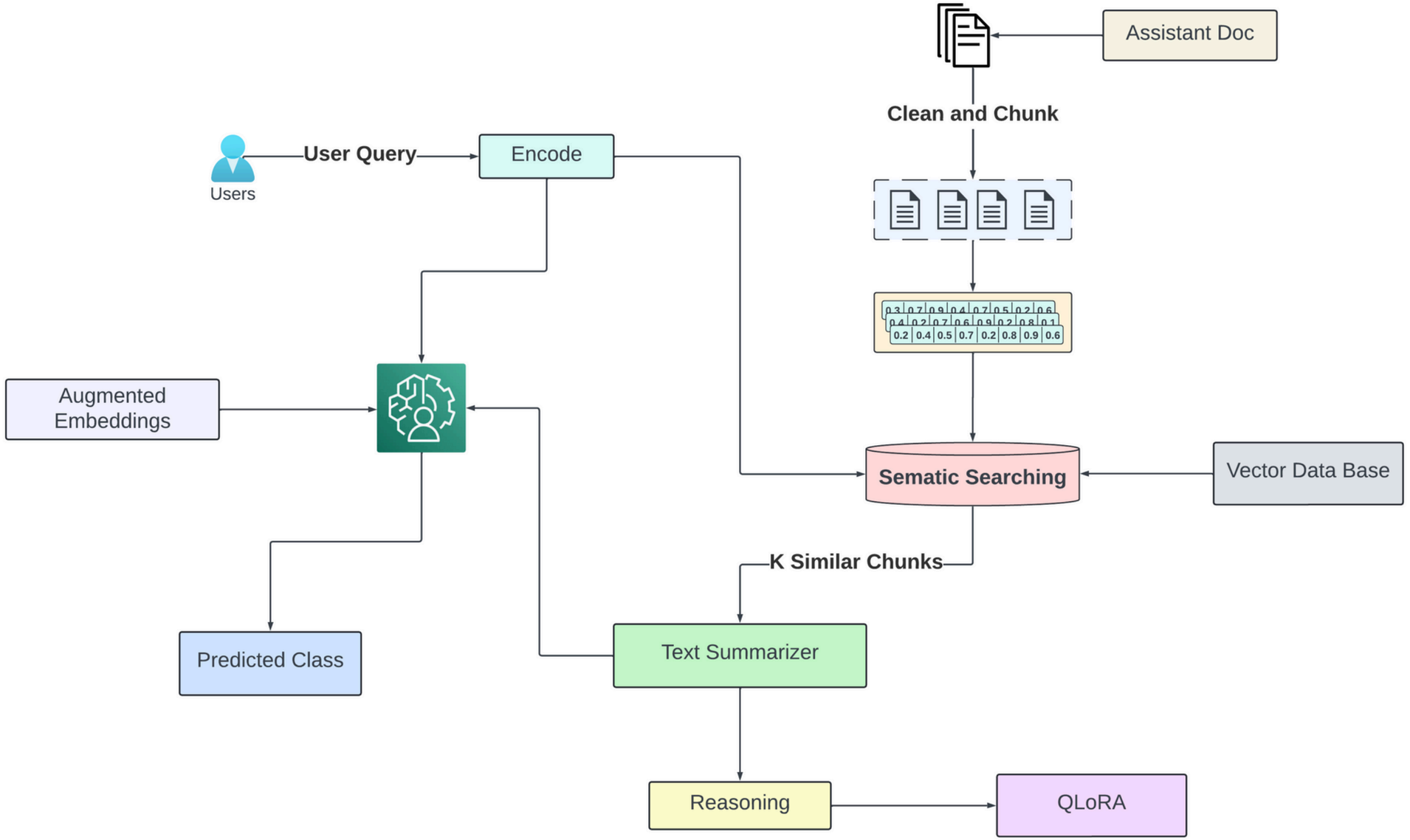
Explaining the Different Parts of RAG (Retrieval Augmented Generation):

01 Retrieval: This phase involves fetching relevant information related to the given query from a vector database or other additional data sources. The vector database stores text embeddings for efficient retrieval of pertinent data.

02 Augmentation: In RAG, the retrieved information enriches the knowledge of the LLM, offering it updated insights about the query. This augmentation empowers the LLM to provide more comprehensive responses by incorporating additional contextual information.

03 Generation: During this stage, the LLM produces new text by utilizing both the query information and the retrieved data. By leveraging this combined knowledge, the generated response aims to be more accurate and informative, as the LLM is equipped with up-to-date information through the retrieval process.

RAG Architecture



Results

Comparison of Hate Data Results: QLoRA vs. RAG with QLoRA

Model Name	Hate Macro F1 with QLoRA	Hate Macro F1 with RAG and QLoRA
Mistral 7B	72.3	72.8
DeepSeek 7B	72.3	70.9
Zephyr Beta 7B	69.6	70.8
Zephyr Alpha 7B	67.1	69.7

Comparison of Fake Data Results: QLoRA vs. RAG with QLoRA

Model Name	Fake Macro F1 with QLoRA	Fake Macro F1 with RAG and QLoRA
Mistral 7B	77.3	78.2
DeepSeek 7B	74.7	78.4
Zephyr Beta 7B	77.3	78.2
Zephyr Alpha 7B	74.7	78.4

Limitation and Future Scope

- In order to expand the capabilities of this project, we plan to enhance the model by refining a Large Language Model (LLM) for generating text. The model will be provided with the query, reason, and label, and will then produce text that clarifies the classification of the query as either Hate or Non-Hate.
- However, at present, we are unable to accomplish this due to constraints with the GPU. Particularly, when the input sequence is more than 50 tokens, it becomes difficult to feed these extensive no of tokens to the LLM given the limited GPU resources that are accessible.

THANK YOU