Q. Write a detailed explanation of what is Retrieval-Augmented Generation (RAG) ?

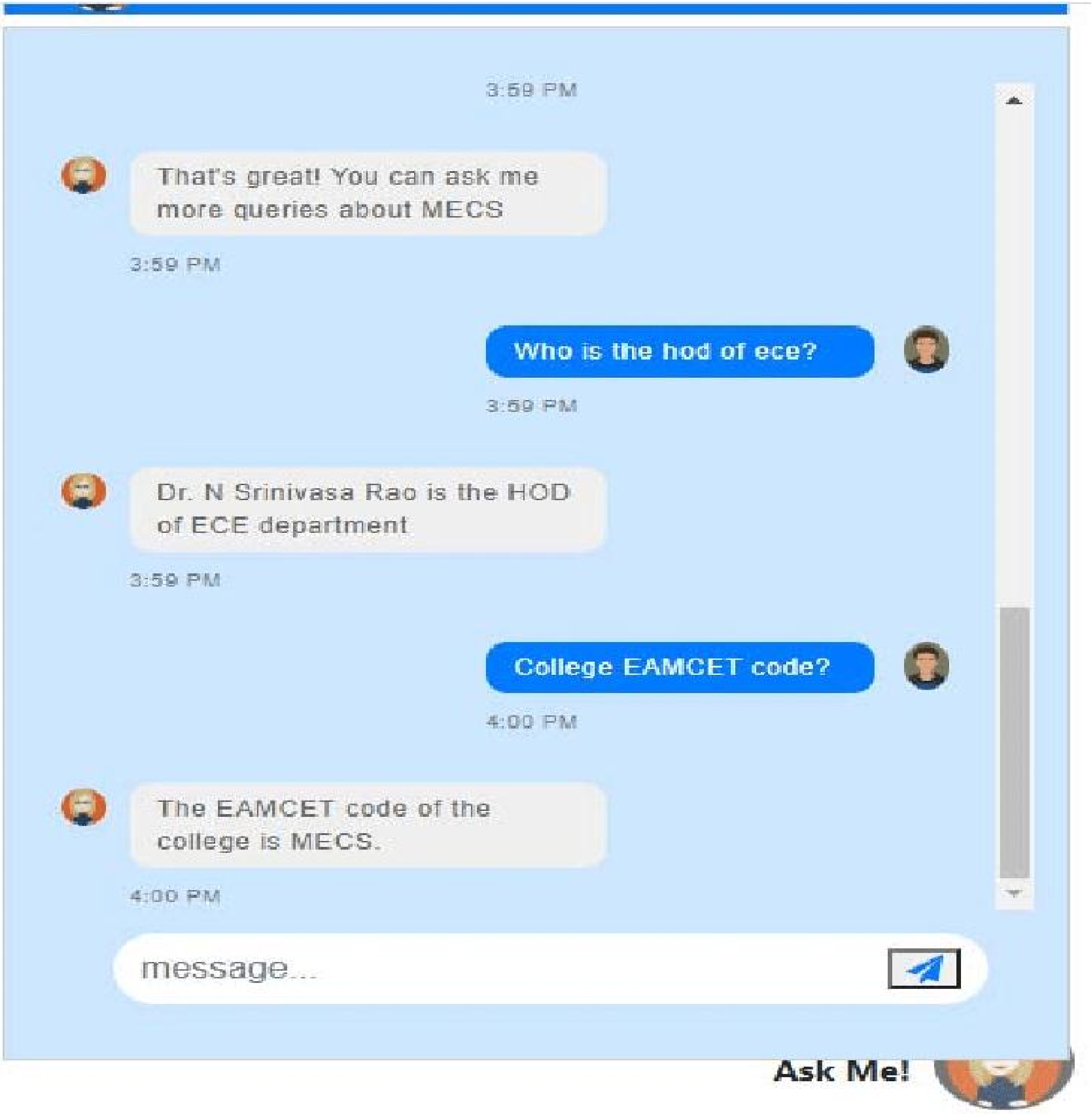
A. This is a technique that gives the answers from an external source of information that was provided to it and it will generate answer from the data that is provided to it and retrieval data by using models.

This RAG mechanism consisted of 6 major steps:

1. Ingestion
2. Chunking
3. embedding
4. indexing
5. retrieving
6. Generating

Q.Why is it used? What problem does it solve?

A. This is used for answering the user query on based of information that is provided to it and also what the information that is know of it about the query. This is used to make chart Bots so that it will be helpful to answer the user querys.



3:59 PM

That's great! You can ask me more queries about MECS

3:59 PM

Who is the hod of ece?

3:59 PM

Dr. N Srinivasa Rao is the HOD of ECE department

3:59 PM

College EAMCET code?

4:00 PM

The EAMCET code of the college is MECS.

4:00 PM

Q. What are the 6 important stages of a RAG system? Explain each stage of RAG clearly ?

A. Ingestion

In this stage we will be providing complete data to it the data may be of any form like PDFs, Word, Audio Files etc… which is required to answer the user query.

Chunking

In this stage it will chunk the data that we had provided it may be chunk them into words or letters or pages.(It will make the large amount of data into small small pieces so that it can answer easily).

Embedding

In this stage it will convert our chunks which is in natural language to numerical language for its understanding and make the chunks into vector.

We will be using embedding modals for doing this work

Indexing

In this stage their will be three major works happening by using FAISS library.

Storing

It will store the vector had we got by embedding the chunks into vector database

Indexing

It will assign the index for each and every vector that was stored in vector database

Searching

It will be searching the relevant vector for the user query from the vector database by using Cousine Similarity and Euclidean Distance

Retrieval

In this stage it will get the most relavent chunk of vector database.

Generation

In this stage we will be giving user query and most relevant chunk to model by using both it will generate the answer for the users query.

Q. Mention the Flowchart of those stages of RAG

A. Step 1:Ingestion

Step 2:Chunking

Step 3:Embedding

Step 4:Indexing

Step 5:Retervial

Step 6:Generation

Q.What is the importance of RAG in GEN AI ?

A.Gen AI also works by this mechanism only first it will be trained on a huge data and by using that data it will answers the users query so RAG in mechanism we will provide data the data will be converted to chunk and chunks will be converted to vectors and indexes are provided to them and most relevant chunk is found for the user query and it will be given to modal and from it we will be getting the answer this will be applicable for Gen Ai. RAG plays a major role in Gen AI to generate the answer.

Q.List at least 5 real-world applications where RAG is more suitable than standalone LLMs

A. Amazon helpline chatbot(Ecommers Field)

In this by using RAG mechanism it will provide correct in information about Amazon that may includes policy and Other as the data of it is available in it but in LLMs it may give false information sometimes.

Medical Field

In this field it is helpful for providing the correct information about any medicines or any other suggestion but By using LLMs it may give some false information

Education Field

In this field it is will provide correct information of knowledge and guidance to the student but In LLMs it will give the information but if we ask for some thing it may not the 100% correct about it. As their may be any change in data after the LLM been developed

Legal Assistance

We will use RAG over LLMs for legal assistance as it may risky to take advise from LLMs because if it gives wrong information it may leads to Sevier damage

Stock Marketing

LLMs will not give the accurate information about current stocks But in place of LLMs If we use RAG it will provide the accurate information.