(12) PATENT APPLICATION PUBLICATION (21) Application No.202541101450 A

(19) INDIA

(22) Date of filing of Application :21/10/2025 (43) Publication Date : 28/11/2025

(54) Title of the invention : RAG Based LLM Chatbot

(51) International classification	:G06F0040300000, G06F0016330000, G06N0003045000, G06F0016332000, G06N0003080000	(71)**Name of Applicant :** **1)R.V.R. & J.C. COLLEGE OF ENGINEERING** Address of Applicant :R.V.R. & J.C. COLLEGE OF ENGINEERING, CHOWDAVARAM – 522 019. Chowdavaram Andhra Pradesh India (72)**Name of Inventor :** **1)Mr. N. Srikanth** **2)N. KUSUMA BHARGAVI** **3)N. MANAS** **4)B. KUMARI**
(31) Priority Document No	:NA	
(32) Priority Date	:NA	
(33) Name of priority country	:NA	
(86) International Application No	:	
Filing Date	:01/01/1900	
(87) International Publication No	: NA	
(61) Patent of Addition to Application Number	:NA	
Filing Date	:NA	
(62) Divisional to Application Number	:NA	
Filing Date	:NA	

(57) Abstract :

ABSTRACT [0011] With the increasing volume of unstructured data available on institutional and organizational websites, there is a growing need for intelligent systems capable of retrieving accurate and relevant information in response to natural language queries. This project presents the design and implementation of a Retrieval-Augmented Generation (RAG)-based Question Answering System, which integrates web scraping, semantic embedding generation, and large language model (LLM) inference to provide context-aware responses. The system extracts textual content from general-purpose websites and processes it into semantically meaningful chunks. [0012] These chunks are then converted into vector representations using transformer-based embedding models and stored in a vector database for efficient similarity search. During query processing, user input is encoded and matched against indexed vectors to retrieve the most relevant context. A fine-tuned LLM generates human-readable answers based on the retrieved information. The backend is built using LangChain, Hugging Face Embeddings, and Pinecone, while the frontend is developed using Flask, offering an intuitive interface for end users. To enhance retrieval accuracy, multi-query expansion and metadata filtering techniques are employed. This system demonstrates how modern NLP techniques can be effectively applied to build scalable, domain- specific knowledge assistants that operate without reliance on structured databases or manual annotation. Experimental results validate the system's ability to deliver precise, contextually appropriate responses across diverse informational queries.

No. of Pages : 8 No. of Claims : 5