RAG Chatbot Project Documentation

Assignment Overview

This project enhances a Retrieval-Augmented Generation (RAG) chatbot using Streamlit, building on an initial RAG implementation. By incorporating domain-specific sources, optimizing retrieval performance, refining LLM integration, and upgrading the user interface, the chatbot aims to improve response quality, relevance, and conversation flow. The project includes integration with Llama3 and a second Mistral model to provide flexibility in response generation.

Knowledge Base Enhancement

Adding Domain-Specific Sources

To improve response relevance, we included two specialized domains:

Domain	Source URLs
Technology	TechCrunch, Wired
Health	CDC Health, Healthline

Each source provides up-to-date content, allowing the chatbot to answer domain-specific queries effectively.

Domain	Source 1	Source 2	Description
Technology	News API (TechCrunch)	Hacker News API	TechCrunch provides current technology news, while Hacker News offers community-driven tech stories.
Health	CDC API	WHO Website	CDC provides health data, and WHO offers global health information on specific topics.
General Knowledge	Wikipedia API		Provides diverse general knowledge on a variety of topics from Wikipedia.

Retrieval Performance and LLM Integration

Query Efficiency and Caching Mechanism

A caching mechanism was implemented to prevent reprocessing identical queries, reducing response time for repeated questions. This cache is stored in a query_cache.pkl file and loaded on startup.

Customizable Retrieval Parameters

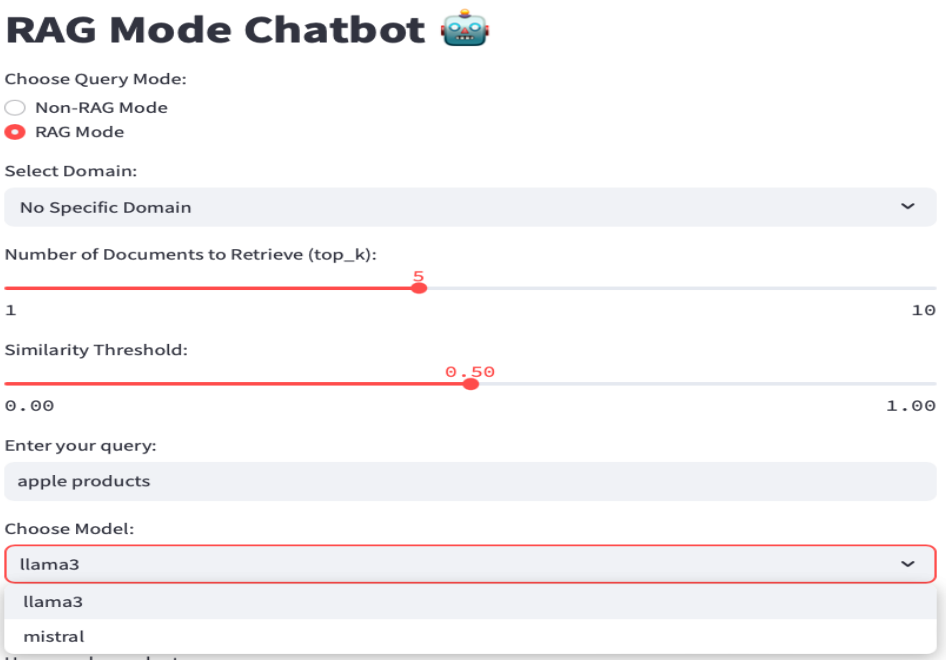
The chatbot allows users to adjust two key retrieval parameters:

- Top-K: Defines the number of documents to retrieve (default set to 5).
- Similarity Threshold: Filters retrieved documents based on relevance score (default set to 0.5).

Using these parameters, the chatbot retrieves contextually relevant information, enhancing both response quality and speed.

LLM Integration and Mode Switching

The chatbot includes two LLMs, Llama3 and a second model Mistral, integrated into both RAG and non-RAG modes. Users can toggle between modes and select an LLM based on query complexity and response requirements.



Conversation Improvements

Basic Context Maintenance

The chatbot now maintains conversational context, allowing for more natural, multi-turn interactions. Key entities and topics are stored in the session state, enabling the chatbot to reference previous queries effectively.

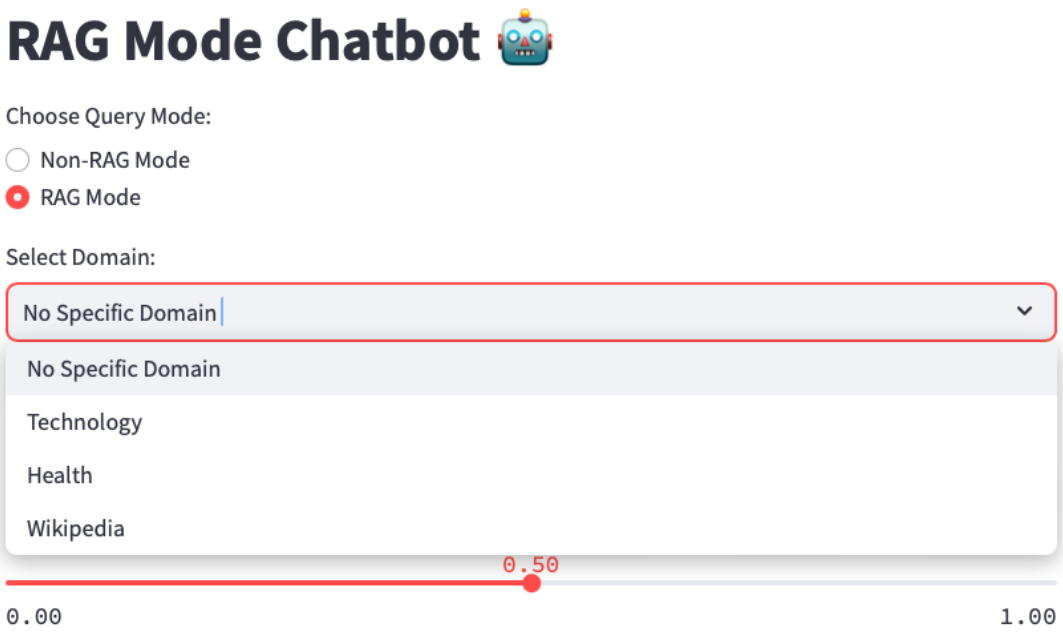
Reference System for Past Queries

With commands like "What did you say about X?", users can refer to previous answers or topics discussed earlier in the session.

User Interface Enhancements

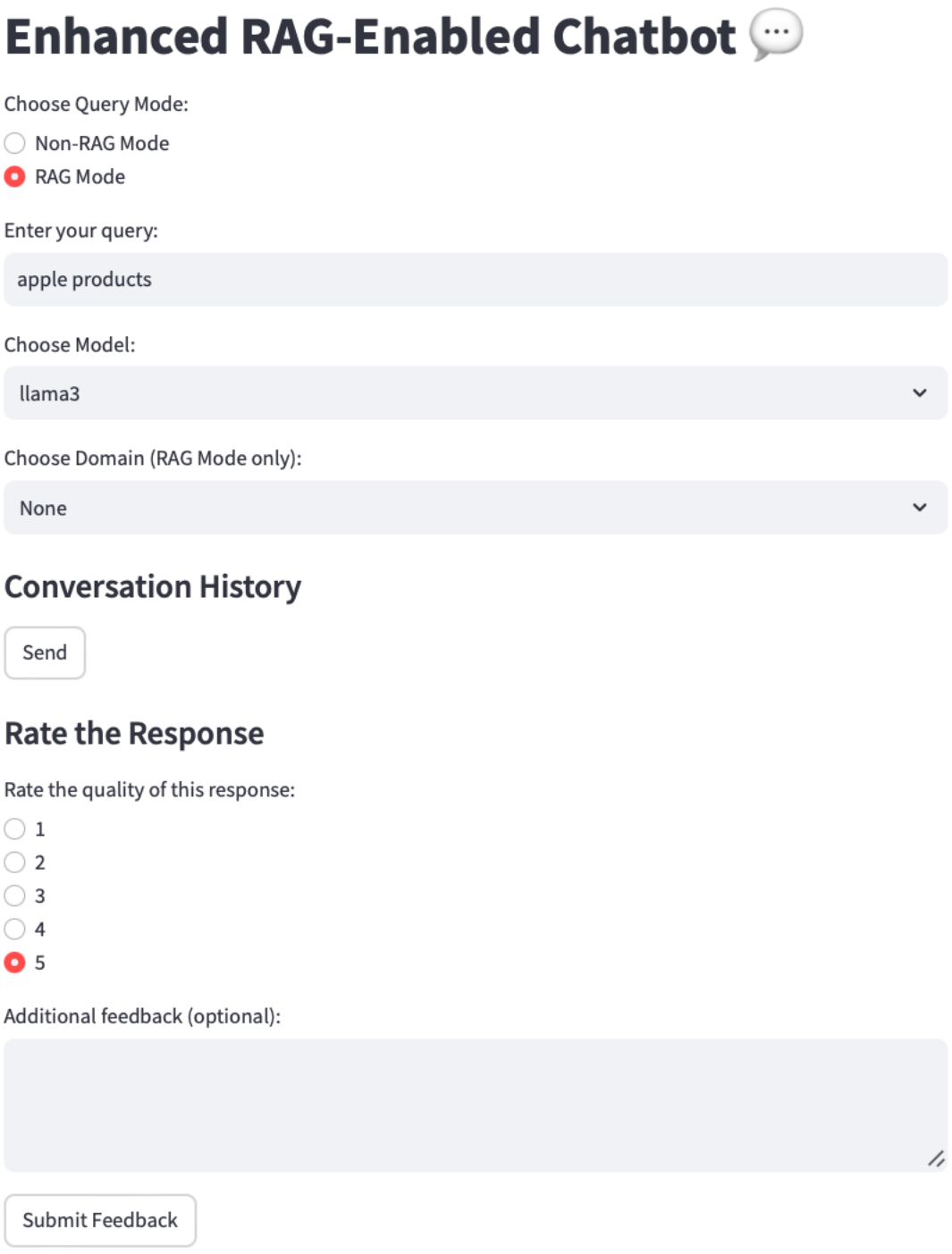
Mode Selection and Domain Selection

The UI features options for switching between RAG and non-RAG modes and selecting domains. For RAG mode, users can select between **Technology**, **Health**, or **No Specific Domain**. This customization enhances user control over the chatbot's retrieval process.



Feedback and Performance Indicators

The chatbot interface includes a feedback system that allows users to rate response quality, helping to refine the chatbot's performance. Additionally, response time indicators inform users of potential delays, particularly in RAG mode.



RAG Mode Chatbot 🤖

Choose Query Mode:

- ○ Non-RAG Mode
- ◉ RAG Mode

Select Domain:

No Specific Domain

Number of Documents to Retrieve (top_k):

5

1 10

Similarity Threshold:

0.50

0.00 1.00

Enter your query:

apple products

Choose Model:

llama3

Conversation History

User: apple products

Bot: I'm not sure how this is related to the error message, but if you're asking about Apple products, here are some popular ones:

- iPhones
- MacBooks
- iPads
- iWatches
- AirPods
- Apple TVs
- Apple Watches
- Beats headphones

Performance Analysis

Query	Domain	Non-RAG Mode Response	RAG Mode with Domain-Specific Sources	Response Quality	Relevance	Response Time
What are the latest advancements in AI technology?	Technology	*Response*: "AI is being used in various fields, including healthcare, finance, and robotics."	*Response*: "Recent AI advancements include developments in GPT models, autonomous vehicles, and discussions on ethical concerns in automation."	High (RAG) – Specific and up-to-date information	High (RAG) – Relevant to AI trends	Moderate – Retrieval adds latency
	Wikipedia	*Response*: "Artificial intelligence is the simulation of human intelligence processes by machines, especially in computer systems."	Moderate – Provides foundational understanding	Moderate – Relevant to AI basics	Fast – Wikipedia retrieval is faster	

What are the risks of using wearable health tech?	Health	*Response*: "Some wearable devices may cause skin irritation or data privacy concerns."	*Response*: "Risks include data security concerns, potential skin irritation, and dependence on device accuracy, which may impact health decisions."	High (RAG) – Comprehensive coverage of risks	High (RAG) – Directly relevant to health	Moderate – Due to health source retrieval
	Wikipedia	*Response*: "Wearable technology refers to electronic devices that can be worn on the body."	Low – Lacks health-specific risks	Low – Only general relevance to wearables	Fast	
Can you explain the principles of quantum computing?	Technology	*Response*: "Quantum computing uses principles of quantum mechanics to process information."	*Response*: "Quantum computing leverages superposition and entanglement for faster computation than classical systems."	High (RAG) – Technical depth and clarity	High (RAG) – Relevant to quantum computing	Moderate – Adds retrieval delay
	Wikipedia	*Response*: "Quantum computing is the study of how to use phenomena in quantum mechanics to perform computation."	Moderate – Foundational but lacks depth	Moderate – Relevant but not comprehensive	Fast	

Reflection and Future Improvements

Key Improvements Achieved

- Enhanced Relevance: Domain-specific sources significantly improved response quality for specialized queries.
- Contextual Conversations: The chatbot's ability to reference previous interactions enhances user experience in multi-turn conversations.
- User Control: UI enhancements, such as model and domain selection, empower users to tailor chatbot responses to their needs.

Limitations and Ethical Considerations

- Data Privacy: Web content is scrupulously selected to avoid sensitive information, adhering to ethical and legal standards.
- Context Management: Maintaining complex context remains challenging; future enhancements may address more sophisticated conversation continuity.

Future Applications

- Additional Domains: Expanding to new domains (e.g., Finance, Education) will increase the chatbot's versatility.
- Enhanced Context Tracking: Integrating advanced memory capabilities to track more complex conversational flows would further improve user experience

Reference

Github - https://github.com/Nehagopinath15neu/RAG-2-Assignment-.git

YouTube link - https://youtu.be/RV9ow1ytH5s