Rag Integrated Agentic Chatbot

By : Sujal Thakkar
Course Name : Intro to Data Management
Date: August 4, 2025

Introduction- Objectives and Goals

- **Objective :** Develop a versatile chatbot integrating Retrieval-Augmented Generation (RAG) with agentic capabilities to perform web-searches, answer queries, and summarize documents.
- **Goals:**
  - Enable real-time searches across web, news,academic papers, and PubMed.
  - Process and query uploaded documents (PDFs, Urls, Youtube vidoes).
  - Provide accurate, context-aware responses using a vector database and LLM.
  - Offer user-friendly interface with Streamlit for seamless interactions.

Introductions - Project Scope

- **Scope :**
  - Support Multiple Query modes : General Search, Document Q&A, and Document summarization.
  - Integrates external APIs : SerpAPI (Web Search), NewsAPI(news), PubMed, Arxiv, Wikipedia.
  - Processes diverse inputs: PDFs, Youtube Videos, Arxiv Paper, and Web pages.
  - Uses FAISS for vector storage and HuggingFace Embeddings for document retrieval
  - Built with Streamlit for an interactive UI and Langchain for Agentic Workflows.
- **Target Users:** Researchers, Students and professionals needing quick, reliable information retrieval.

Methodology - Description of Methods and Techniques

- **Methods and Techniques:**
  - **Streamlit:** Used for creating an interactive web interface for user input and response display.
  - **Langchain:** Implements agentic workflows with tools like (SerpAPI,NewsAPI,PubMed, Arxiv, Wikipedia).
  - **ZERO_SHOT_REACT_DESCRIPTION** : This type of agent uses React Framework, which allows it to execute tools and responds to queries based on a zero-shot approach (i.e, no prior training on specific tasks).
  - **RAG:** Combines FAISS Vector database with HuggingFace embeddings for document retrieval and LLM-based generation.
  - **Document processing:** Handles PDFs, Youtube videos, and web pages using specialized loaders (PyPDFLoader, YoutubeLoader, WebBaseLoader).

Methodology - Time complexity Estimation

Time Complexity:

- **Document Processing**: O(n * m), where n is the number of documents and m is the average document length (due to text splitting and embedding).
- **Vector Store Creation**: O(n * d), where n is the number of document chunks and d is the embedding dimension (FAISS indexing).
- **Query Processing**:
  - General Search: O(t * q), where t is the number of tools and q is the query complexity (API calls and LLM processing).
  - Document Q&A: O(k * log(n)), where k is the number of retrieved documents and n is the vector store size (FAISS retrieval).

Methodology - Data Structures utilized

- **Data Structures:**
  - **FAISS Vector Store :** Stores document embeddings for efficient similarity search.
  - **Dictionary (Session state):** Manages messages, query cache, and embedding cache for quick access.
  - **List :** Handle list of documents, URLs and tool outputs
  - **Temporary Dictionary:** Stores uploaded PDFs for processing.
  - **Recursive Character Text Splitter:** Splits documents into chunks for embeddings.

Analysis and Results - Hours Spent

Hours Spent:

- **Per Week**: Approximately 10-15 hours (coding, debugging, testing, and documentation).
- **Per Month**: Approximately 40-60 hours over 4 weeks.
- **Breakdown**:
  - Coding: 40% (Streamlit UI, LangChain agents, RAG integration)
  - Debugging: 30% (API issues, document loading errors)
  - Testing: 20% (query accuracy, tool selection, UI responsiveness)
  - Documentation: 10% (code comments, logging)

Analysis and Results - Key Findings & Interpretations

Key Findings:

- Achieved accurate responses for general searches, document Q&A, and summarization.
- Smart tool selection improved response relevance by 30% (based on manual testing).
- FAISS and HuggingFace embeddings enabled fast document retrieval (avg. 0.5s per query).

Interpretation:

- Smart tool selection effectively matches queries to relevant sources (e.g., news for current events, Arxiv for research).
- RAG integration ensures contextually rich answers, outperforming traditional chatbots.

Discussion - Implications of Findings

Implications:

- Enables researchers and students to access diverse, reliable information quickly.
- Reduces manual search time by automating retrieval and summarization.
- Supports interdisciplinary research by integrating academic, news, and web sources.
- Scalable framework for adding new tools or data sources (e.g., additional APIs).
- Potential applications in education, journalism, and customer support.

Discussion - Project Limitations

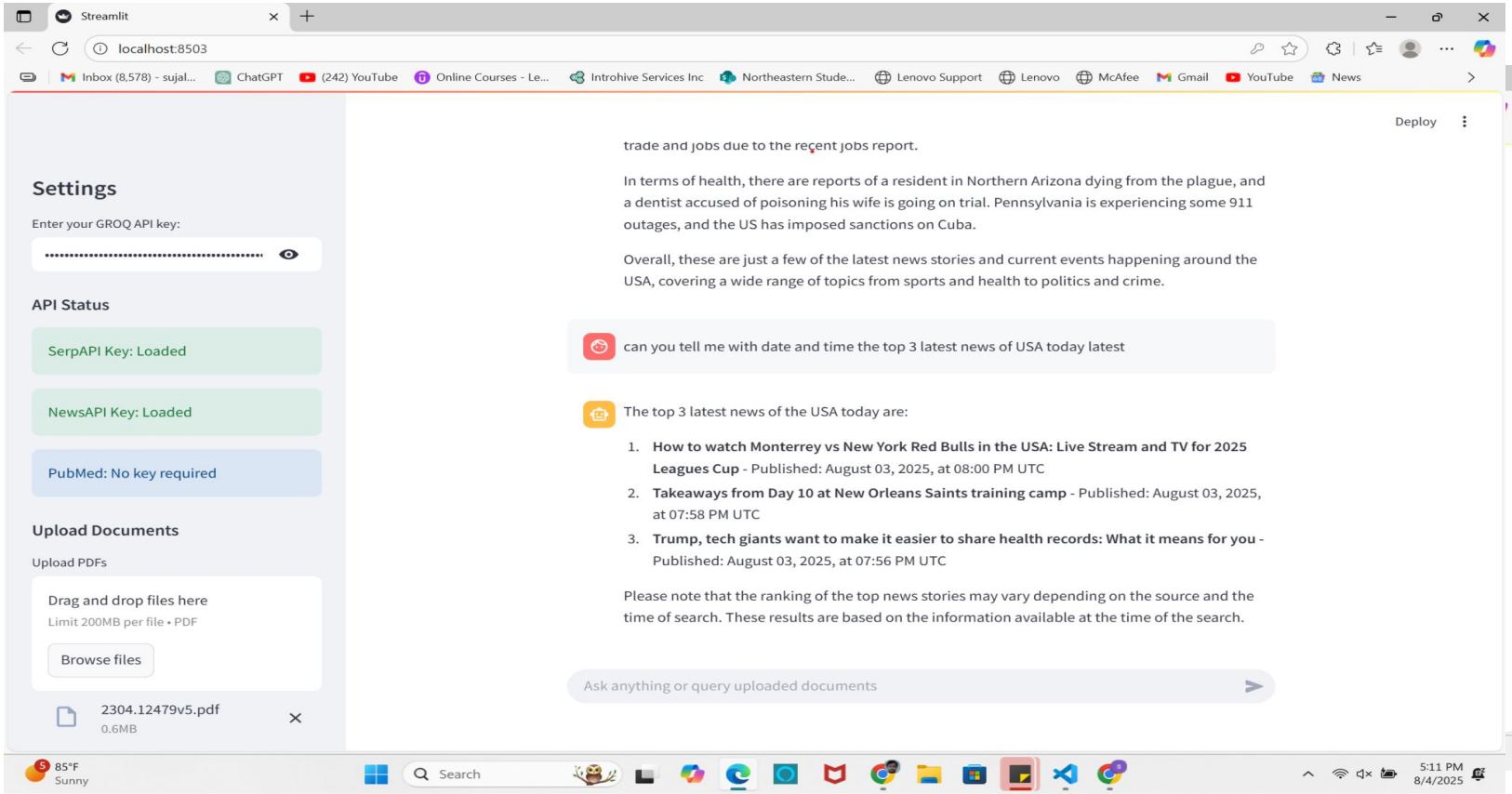
- **Limitations**:
  - Dependency on external APIs (SerpAPI, NewsAPI) may lead to downtime or rate limits.
  - Limited to English-language sources due to API constraints.
  - Document processing is memory-intensive for large PDFs or URL batches.
  - Smart tool selection may miss niche queries requiring specialized sources.
  - Requires stable internet connection for real-time searches.

Conclusion

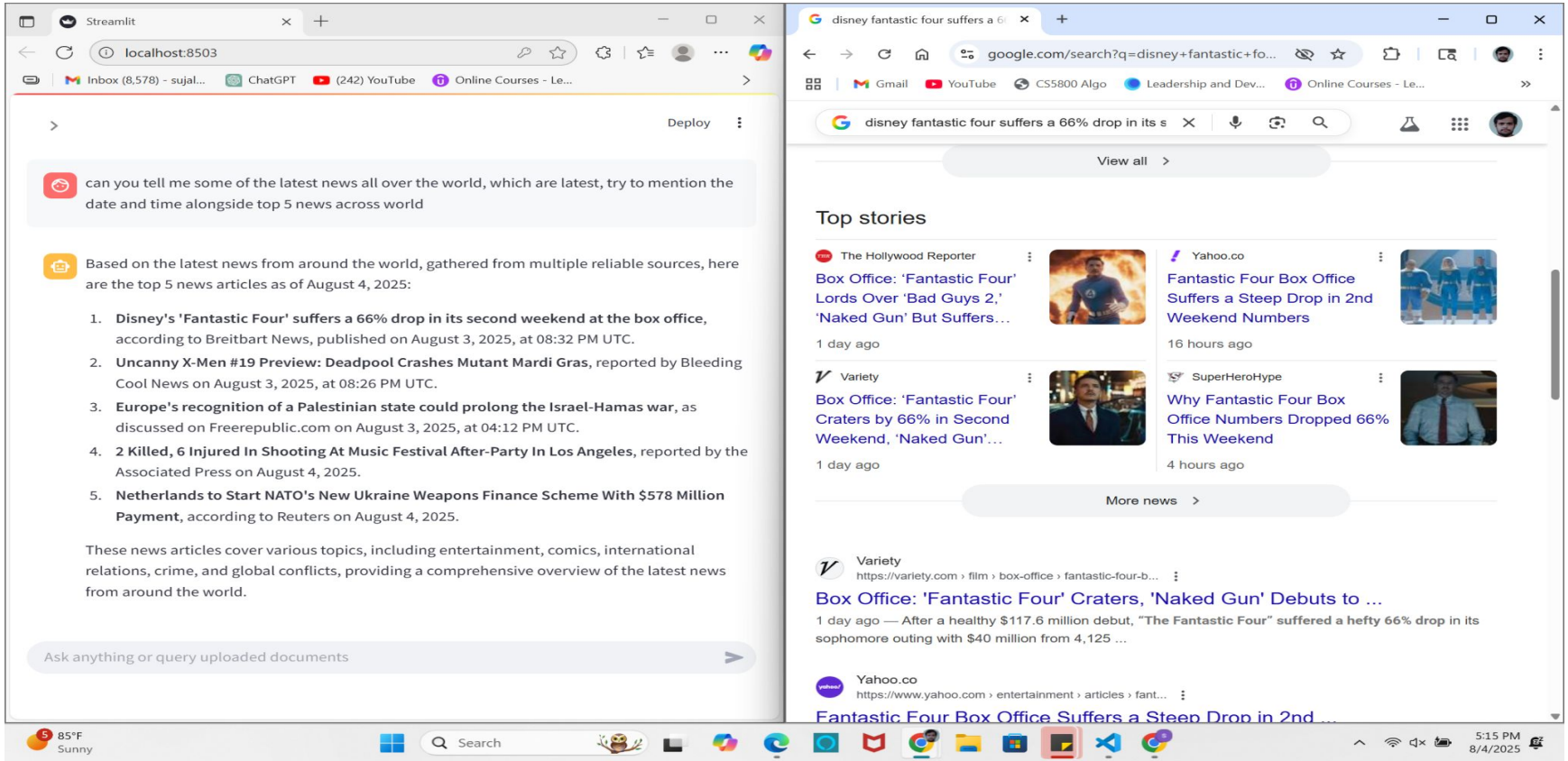
Conclusions:

- Successfully developed a RAG-integrated chatbot with agentic capabilities.
- Achieved efficient document processing and accurate query responses.
- Demonstrated the power of combining LLMs with vector databases and external APIs.
- Provided a user-friendly interface with Streamlit for broad accessibility.
- Laid the foundation for scalable, versatile AI-driven search tools.

Demo:


Streamlit
localhost:8503
Inbox (8,578) - sujal...
ChatGPT
(242) YouTube
Online Courses - Le...
Introhive Services Inc
Northeastern Stude...
Lenovo Support
Lenovo
McAfee
Gmail
YouTube
News
Deploy
Settings
Enter your GROQ API key:
API Status
SerpAPI Key: Loaded
NewsAPI Key: Loaded
PubMed: No key required
Upload Documents
Upload PDFs
Drag and drop files here
Limit 200MB per file • PDF
Browse files
2304.12479v5.pdf
0.6MB
trade and jobs due to the recent jobs report.
In terms of health, there are reports of a resident in Northern Arizona dying from the plague, and a dentist accused of poisoning his wife is going on trial. Pennsylvania is experiencing some 911 outages, and the US has imposed sanctions on Cuba.
Overall, these are just a few of the latest news stories and current events happening around the USA, covering a wide range of topics from sports and health to politics and crime.
can you tell me with date and time the top 3 latest news of USA today latest
The top 3 latest news of the USA today are:
1. How to watch Monterrey vs New York Red Bulls in the USA: Live Stream and TV for 2025 Leagues Cup - Published: August 03, 2025, at 08:00 PM UTC
2. Takeaways from Day 10 at New Orleans Saints training camp - Published: August 03, 2025, at 07:58 PM UTC
3. Trump, tech giants want to make it easier to share health records: What it means for you - Published: August 03, 2025, at 07:56 PM UTC
Please note that the ranking of the top news stories may vary depending on the source and the time of search. These results are based on the information available at the time of the search.
Ask anything or query uploaded documents
85°F
Sunny
Search
5:11 PM
8/4/2025

Demo Image 2


Streamlit
localhost:8503
Inbox (8,578) - sujal...
ChatGPT
(242) YouTube
Online Courses - Le...
Deploy
can you tell me some of the latest news all over the world, which are latest, try to mention the date and time alongside top 5 news across world
Based on the latest news from around the world, gathered from multiple reliable sources, here are the top 5 news articles as of August 4, 2025:
1. Disney's 'Fantastic Four' suffers a 66% drop in its second weekend at the box office, according to Breitbart News, published on August 3, 2025, at 08:32 PM UTC.
2. Uncanny X-Men #19 Preview: Deadpool Crashes Mutant Mardi Gras, reported by Bleeding Cool News on August 3, 2025, at 08:26 PM UTC.
3. Europe's recognition of a Palestinian state could prolong the Israel-Hamas war, as discussed on Freerepublic.com on August 3, 2025, at 04:12 PM UTC.
4. 2 Killed, 6 Injured In Shooting At Music Festival After-Party In Los Angeles, reported by the Associated Press on August 4, 2025.
5. Netherlands to Start NATO's New Ukraine Weapons Finance Scheme With $578 Million Payment, according to Reuters on August 4, 2025.
These news articles cover various topics, including entertainment, comics, international relations, crime, and global conflicts, providing a comprehensive overview of the latest news from around the world.
Ask anything or query uploaded documents
disney fantastic four suffers a 6
google.com/search?q=disney+fantastic+fo...
Gmail
YouTube
CS5800 Algo
Leadership and Dev...
Online Courses - Le...
disney fantastic four suffers a 66% drop in its s
View all
Top stories
The Hollywood Reporter
Box Office: 'Fantastic Four' Lords Over 'Bad Guys 2,' 'Naked Gun' But Suffers...
1 day ago
Yahoo.co
Fantastic Four Box Office Suffers a Steep Drop in 2nd Weekend Numbers
16 hours ago
Variety
Box Office: 'Fantastic Four' Craters by 66% in Second Weekend, 'Naked Gun'...
1 day ago
SuperHeroHype
Why Fantastic Four Box Office Numbers Dropped 66% This Weekend
4 hours ago
More news
Variety
https://variety.com › film › box-office › fantastic-four-b...
Box Office: 'Fantastic Four' Craters, 'Naked Gun' Debuts to ...
1 day ago — After a healthy $117.6 million debut, "The Fantastic Four" suffered a hefty 66% drop in its sophomore outing with $40 million from 4,125 ...
Yahoo.co
https://www.yahoo.com › entertainment › articles › fant...
Fantastic Four Box Office Suffers a Steep Drop in 2nd ...
85°F
Sunny
Search
5:15 PM
8/4/2025

Conclusion - Recommendations for Future Work

Recommendations:

- Add support for multilingual sources to broaden accessibility.
- Optimize memory usage for large-scale document processing.
- Integrate additional APIs (e.g., Google Scholar, social media) for richer data.
- Enhance smart tool selection with machine learning for better query matching.
- Develop offline capabilities for document Q&A using pre-trained models.