REAL ESTATE RESEARCH TOOL

AI-Powered Insights from Real Estate News Articles

Presented by: Vaibhav Garg

PROBLEM STATEMENT

- Exploring real estate and housing market trends through online articles is time-consuming due to the sheer volume of content and the need to manually read and extract relevant insights.
- Traditional research methods involving manual article review are inefficient, prone to oversight, and limit a researcher's ability to analyze multiple sources quickly.
- There is a critical need for an automated, intelligent system that can extract and summarize insights from real estate news articles while allowing users to ask domain-specific questions.
- A retrieval-augmented generation (RAG) approach using LLMs and vector databases enables fast, accurate, and interactive querying of real estate content, streamlining the research process and enhancing decision-making.

PROJECT OBJECTIVES

- Accept real estate article URLs as input and automatically extract relevant textual content.
- Generate semantic embeddings from the article content and store them in a vector database for efficient information retrieval.
- Allow users to ask natural language questions and receive accurate, context-aware answers based on the provided articles.
- Deliver an interactive and user-friendly Streamlit web interface that streamlines real estate research through automation and AI.

SOLUTION OVERVIEW

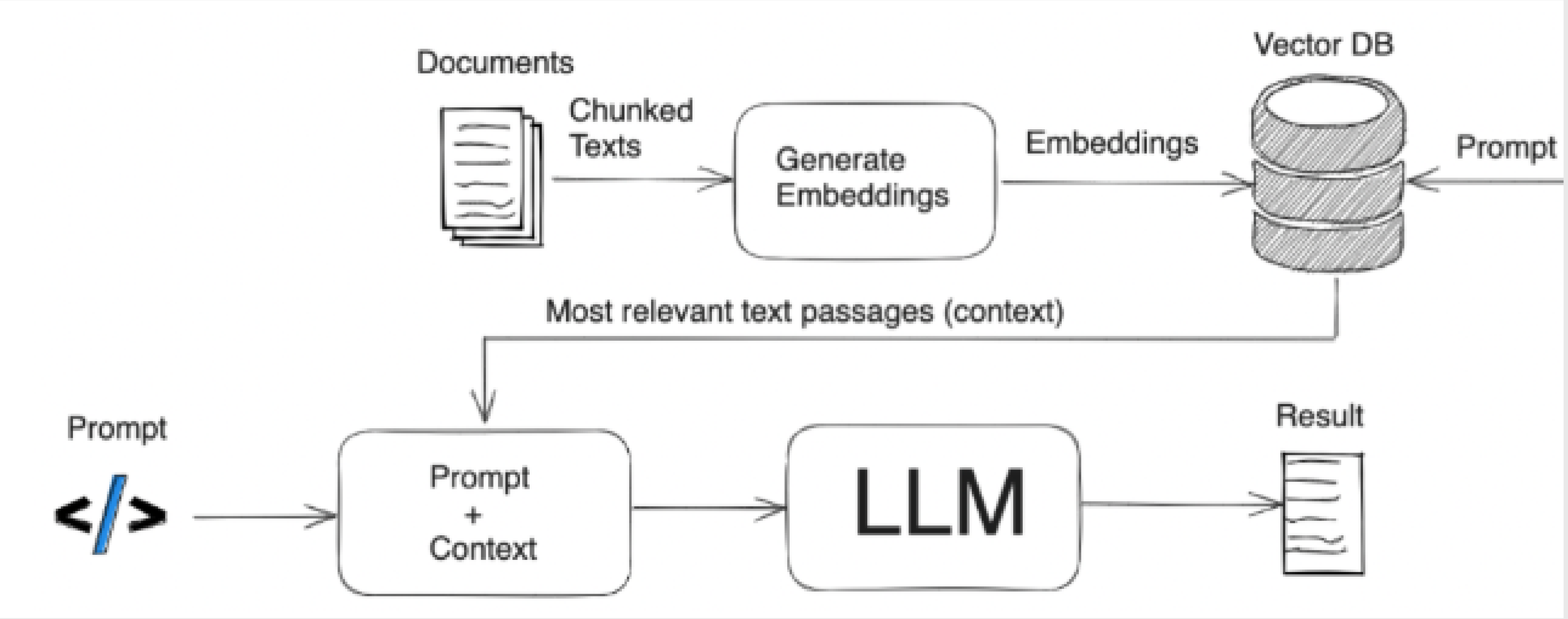
This tool simplifies real estate research by automating the process of extracting insights from news articles using AI-powered retrieval and question-answering.
Users can enter URLs of relevant articles, and the system will:

- Automatically scrape and clean the content using web scraping techniques.
- Convert article text into vector embeddings using HuggingFace NLP models.
- Store the embeddings in a Chroma vector database for fast retrieval.
- Answer user questions by retrieving the most relevant article content and passing it to a powerful LLM (LLaMA 3 via Groq API).
- Provide source-linked answers, ensuring responses are grounded in real data from the original articles.

The entire system runs in an intuitive, interactive Streamlit web app, making it easy to explore complex housing and financial trends without manually reading through every article.

SYSTEM ARCHITECTURE

The diagram below illustrates the complete flow of data through the system – from raw article URLs to insightful answers.



FUNCTIONAL WORKFLOW

This tool automates the extraction of insights from real estate news articles through an AI-powered retrieval and question-answering pipeline composed of six key steps.

1. Document Loading

- Users provide URLs of real estate news articles. The content is scraped and cleaned using web scraping tools.

2. Chunking

- The full text from each article is split into smaller, manageable chunks to preserve context and optimize retrieval accuracy.

3. Embedding and Storage

- Each chunk is transformed into a high-dimensional vector (embedding) using HuggingFace models.
- These vectors are stored in ChromaDB, which enables efficient similarity-based retrieval.

4. Query Embedding and Retrieval

- When a user enters a question, it is also embedded into a vector.
- ChromaDB retrieves the most relevant text chunks by comparing similarity scores with stored vectors.

5. Context + Prompt Construction

- The retrieved chunks are compiled with a carefully designed prompt and sent to the LLM (LLaMA 3 via Groq) to ensure the answer is grounded in article content.

6. Response Generation

- The LLM processes the input and generates a human-readable response, along with the sources used to answer the question.

TECHNOLOGY STACK

This project integrates modern tools from the AI and web development ecosystem to enable fast, accurate, and interactive real estate research:

- **Frontend**
  - **Streamlit:** Used to create an intuitive, responsive web interface for URL input and question-answering.
- **Web Scraping**
  - **Requests + BeautifulSoup:** Fetches and parses article content from user-provided URLs.
- **Data Processing**
  - **LangChain:**
    - **RecursiveCharacterTextSplitter** for chunking text into manageable units.
    - **RetrievalQAWithSourcesChain** for combining LLM outputs with document retrieval.

- **Embeddings & Storage**
  - **HuggingFace Embeddings:** Uses the "all-MiniLM-L6-v2" model for semantic vector generation.
  - **ChromaDB:** Serves as the vector store for efficient similarity search and document retrieval.

- **Language Model**
  - **Groq API + LLaMA 3 (70B):** Delivers fast and grounded responses using a powerful open-weight LLM hosted on Groq’s inference engine.

- **Environment Management**
  - **Python-dotenv:** Loads API keys and configuration variables from a .env file securely.

KEY FEATURES

- **End-to-End Automation**
  - From scraping real estate articles to answering questions, the tool handles the entire research pipeline with minimal user input.

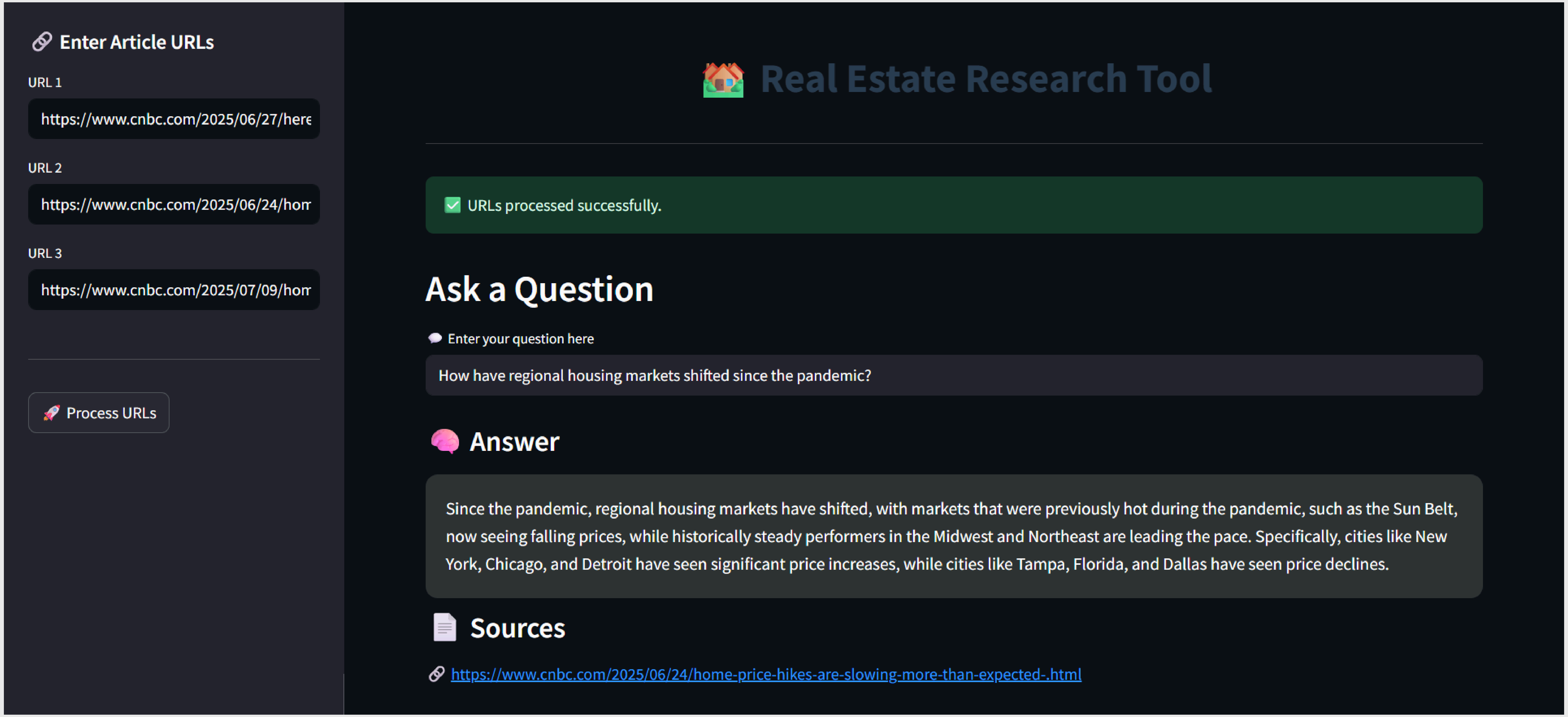
- **Natural Language Q&A**
  - Users can ask real estate-related questions in plain English and receive grounded answers with source references using LLaMA 3 via Groq API.

- **AI-Powered Search & Retrieval**
  - Uses advanced NLP embeddings (HuggingFace) and a vector database (ChromaDB) to find the most relevant article content quickly and accurately.

- **Streamlit-Based Interface**
  - A simple, intuitive web UI makes the tool accessible for both domain experts and casual users – no technical expertise required.

- **Reusable and Scalable Design**
  - Modular code structure built on LangChain allows easy extension to other domains like finance, policy, or education.

USER INTERACTION PREVIEW



URL 1: https://www.cnbc.com/2025/06/27/here

URL 2: https://www.cnbc.com/2025/06/24/hom

URL 3: https://www.cnbc.com/2025/07/09/hom

✅ URLs processed successfully.

Ask a Question

💬 Enter your question here

How have regional housing markets shifted since the pandemic?

🧠 Answer

Since the pandemic, regional housing markets have shifted, with markets that were previously hot during the pandemic, such as the Sun Belt, now seeing falling prices, while historically steady performers in the Midwest and Northeast are leading the pace. Specifically, cities like New York, Chicago, and Detroit have seen significant price increases, while cities like Tampa, Florida, and Dallas have seen price declines.

📄 Sources

🔗 https://www.cnbc.com/2025/06/24/home-price-hikes-are-slowing-more-than-expected-.html

PROJECT SUMMARY

- Developed an AI-powered research assistant that extracts insights from real estate news articles using automated retrieval and question-answering.
- Implemented web scraping logic to fetch and clean article content from user-provided URLs using BeautifulSoup.
- Converted article text into vector embeddings using HuggingFace’s MiniLM model and stored them in ChromaDB for fast semantic search.
- Integrated Retrieval-Augmented Generation (RAG) pipeline using LangChain and Groq’s LLaMA 3 (70B) to generate accurate, source-linked answers from context.
- Designed an interactive Streamlit app with a sidebar-based interface to input URLs and ask questions with real-time response display.
- Live App: https://vaibhav-project-real-estate-research-tool.streamlit.app/
- GitHub: https://github.com/vaibhavgarg2004/Real-Estate-Research-Tool

THANK YOU