**RAG in Real Estate**

Praveen Nandan K 1, Syed Afridi 2,

Shailesh K R 3, Shubam V Patil 4

Department of Computer Science & Engineering, Presidency University,

Bengaluru – 560064

1 praveennandu4@gmail.com ,

2 syedafridi644@gmail.com ,

3 krshailesh627@gmail.com ,

4 sopatil.2003@gmail.com

***Abstract* —.** In real estate business, data mining and analyzing data in the form of different structure is important.s is crucial. In this paper, a Retrieval-Augmented Generation (RAG) framework using LangChain for querying real estate databases and housing guides is proposed. Using regional price trends collected in the CSV files from the years 2000 to 2024 in the US and combined with PDF-based housing manuals, the system supplies contextually sound answers. To reduce the hallucinations and also increase the preciseness of the results the proposed system uses the evaluation of vectors with FAISS, interactions with Ollama’s Llama 3.2 model and similarity score of relevance. An intuitive and interactive Streamlit frontend allows for easy file upload and exposed functionalities for exemplification of how RAG systems can be useful in real estate applications.

***Keywords —*** Real Estate Analytics, RAG, LangChain,Streamlit

1. Introduction

The real estate industry needs various types of data, including structured data such as housing price structure and unstructured data in the form of homeowners’ manuals. Combing information from these sources is difficult because of the format of the sources and also, the type of responses needed is more precise for the specific context. Some of these criss cross may not be well handled by traditional methodologies.

This research introduces a framework they label as Retrieval-Augmented Generation (RAG) for efficient querying of real estate data. They incorporate CSV files with US housing price trends from 2000 to 2024 along with the PDF-based homeowner manuals, allowing for the combined analysis of modality. LangChain facilitates the working process, FAISS provides the storage of vectors, and Ollama’s Llama 3.2 gives an accurate and contextually suitable answer.

For increasing the response accuracy, the system utilizes similarity scoring via SentenceTransformer and thus minimizes hallucination and increases the model’s credibility. Powered by Streamlit, this application enables users to input data in various file formats, asking questions and getting answers along with sources in the text.

This study demonstrates how RAG can transform real estate analytics by bridging structured and unstructured data, offering a scalable, precise, and user-friendly solution for modern data challenges.

1. Research Modifications, Objectives, and Contributions

This research adapts standard Retrieval-Augmented Generation (RAG) pipelines to work with both CSV tabular data and PDF textual data, types of real estate data often encountered in practice. It uses custom similarity scoring and the Sentence Transformer model in order to increase the response fidelity and prevent hallucinations of the answers produced. Utilizing Streamlit for easy and pleasant document upload and query experience and using Ollama’s Llama 3.2 for precise and contextually conscious responses. The system also includes a history-aware retriever to improve context when new queries follow.

They are as follows: First, the system must be able to query data from multiple sources regarding real estate property; second, the system should query data in the correct context by embedding and scoring; third, the system must be scalable to suit real estate professionals. Furthermore, the study seeks to provide the missing link between price analysis and maintenance information, thus providing a detailed analysis. The applicability and usefulness of RAG frameworks in real estate analytics is discussed and demonstrated.

Important contributions include proposing a new real estate RAG, developing methods for multimodal data, and incorporating Similarity Scoring to increase response reliability. As a final note, this research also offer a practical and engaging way to rethink real estate decisions and to show the flexibility of RAG for other real-world applications on integrating multimodal data.

1. Literature Survey

This literature survey examines key advancements in Retrieval-Augmented Generation (RAG) systems, NLP in real estate analytics, and multimodal data integration. RAG systems, which combine document retrieval with language generation, have been widely explored for improving query accuracy by grounding responses in real-world data. Notable works like "Retrieval-augmented generation for knowledge-intensive NLP tasks" (Lewis et al., 2020) introduced the concept of integrating retrievers with language models to enhance response quality, with FAISS-based retrieval systems further improving efficiency in handling large documents.

In real estate, machine learning models have been used for price prediction and market analysis (Maleki & Hashemipour, 2021), but challenges remain in processing unstructured data such as manuals and legal documents. Studies like "Multimodal learning for real estate price prediction" (Li et al., 2020) explored integrating image and numerical data for property valuation but did not focus on document-based information.

Recent advancements in LangChain and FAISS have enabled better integration of structured and unstructured data. LangChain, as detailed in "LangChain: A framework for building language model-powered applications" (Wang et al., 2023), helps manage document loaders and retrieval chains, while FAISS (Johnson et al., 2020) efficiently stores and retrieves embeddings for large datasets. This research aims to combine these techniques to enhance real estate data analytics, providing a scalable and accurate system for querying both structured and unstructured real estate data.

1. Gap Identification

Despite advancements in machine learning and natural language processing for real estate analytics, existing systems often struggle to integrate both structured data (e.g., price trends) and unstructured data (e.g., maintenance manuals, legal documents). While RAG systems have shown promise in other domains, their application to real estate remains underexplored, particularly in handling multimodal data efficiently. Current systems either focus on numerical prediction models or image-based property analysis but fail to address the need for comprehensive document-based querying. Furthermore, the challenge of ensuring response accuracy and relevance in complex real estate queries, such as warranty procedures or market insights, has not been fully addressed. This research aims to bridge these gaps by integrating RAG with real estate-specific datasets, improving accuracy, scalability, and transparency in querying multimodal real estate data.

1. Proposed Methodologies

The proposed methodology for this research integrates several key components to build a robust system for querying real estate data using Retrieval-Augmented Generation (RAG). The system is designed to handle both structured data, such as housing price trends from CSV files, and unstructured data, including housing manuals in PDF format. The methodology begins with the preprocessing phase, where uploaded CSV files are parsed and PDFs are loaded using appropriate document loaders. For PDFs, the RecursiveCharacterTextSplitter is employed to break down the text into manageable chunks for better processing.

Once the documents are prepared, they are ingested into a vector database using FAISS, which is responsible for storing the document embeddings. These embeddings are generated using HuggingFace's SentenceTransformer model, which converts the textual content into vectors that can be efficiently retrieved during user queries. FAISS ensures quick and efficient retrieval of relevant documents based on the similarity of the embedded content.

For query processing, the Ollama Llama 3.2 model is used to generate natural language responses. The system employs a retrieval chain mechanism, where the context of the user’s query is integrated with relevant documents retrieved from the vector database. The retrieval process is enhanced by a history-aware retriever, which ensures that context from previous user interactions is incorporated into the current response. The model generates answers based on the retrieved documents, and a similarity scoring mechanism, powered by SentenceTransformer, is used to evaluate the relevance of the answer by comparing it against the context documents.

A Streamlit interface serves as the front end, allowing users to upload files, interact with the system, and query data. Users can input questions, and the system returns answers along with the relevant source documents for verification. This methodology ensures that the system is not only accurate in answering queries but also transparent, as users can view the documents that contributed to each response. The proposed system is scalable, allowing for the integration of additional data sources and expansion into other domains requiring multimodal data processing.

1. RESULTS

The Fig:1 shows historical real estate prices (blue line) and forecasted prices (red line) for Aberdeen, SD. The shaded red area represents the upper and lower bounds, indicating the uncertainty in future predictions.

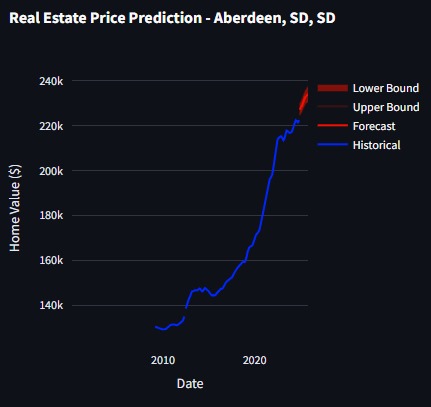


Fig:1 Trend Line Chart - Real Estate Price prediction for Aberdeen, SD

1. Conclusion

In conclusion, this research demonstrates the potential of combining Retrieval-Augmented Generation (RAG) with LangChain for real estate data analytics, effectively bridging the gap between structured data, such as historical housing price trends, and unstructured data, like homeowner manuals. By integrating FAISS for efficient document retrieval and Ollama’s Llama 3.2 model for accurate, context-aware responses, the proposed system enhances the ability to query complex real estate data. Additionally, the implementation of similarity scoring with SentenceTransformer ensures the relevance and accuracy of the responses, minimizing the risk of hallucinations. The user-friendly Streamlit interface provides an interactive platform for seamless querying, offering transparency by displaying the source documents. This approach not only improves the precision and scalability of real estate data analysis but also sets a foundation for expanding RAG-based systems into other domains that require multimodal data integration. Future work can focus on integrating live data streams and improving the system’s adaptability to dynamic real estate markets.

1. Future Enhancements

Future enhancements to this system can focus on several key areas to improve its functionality, scalability, and adaptability to real estate data. First, integrating **real-time data sources**, such as live property listings or dynamic market trends, would allow the system to provide up-to-date insights, offering more accurate responses for users. Additionally, expanding the **document processing capabilities** to handle a broader range of unstructured data formats, including images, legal contracts, and market reports, would enhance the system's ability to serve diverse real estate professionals.

Another important enhancement would be the **incorporation of advanced NLP techniques**, such as fine-tuning the language models for specific real estate-related tasks, improving the system's contextual understanding of complex queries like zoning laws, financing options, or market forecasts. The **user interface** could also be enhanced by providing more advanced querying features, such as natural language search filters or personalized recommendations based on historical interactions.

To further improve **response accuracy and relevance**, the system could employ **multi-modal retrieval systems**, which would not only process text but also integrate spatial or image-based data, such as property photographs or blueprints. Implementing **feedback loops** would also allow users to rate responses and adjust the system's behavior over time, leading to continuous improvement.

Lastly, scaling the system to handle larger datasets and more complex queries, particularly for nationwide real estate markets, would be crucial for expanding its practical applications. These advancements would make the system more versatile, efficient, and beneficial for a wider range of users, from individual homebuyers to large real estate enterprises.

1. References
2. **LangChain Documentation**LangChain. (n.d.). *LangChain Documentation*. Retrieved from<https://docs.langchain.com/>
3. **SentenceTransformers**Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. Retrieved from<https://www.sbert.net/>
4. **FAISS for Vector Search**Johnson, J., Douze, M., & Harwood, H. (2020). *FAISS: A library for efficient similarity search*. Facebook AI Research (FAIR). Retrieved from<https://faiss.ai/>
5. **Scikit-learn Documentation**Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. Retrieved from<https://scikit-learn.org/stable/>
6. **Streamlit Official Website**Streamlit. (n.d.). *Streamlit: The fastest way to build and share data apps*. Retrieved from<https://streamlit.io/>
7. **HuggingFace Models**Wolf, T., et al. (2020). *Transformers: State-of-the-art Natural Language Processing*. Retrieved from<https://huggingface.co/models>
8. **Research Paper on Real Estate Prediction Techniques**Zhang, X., & Zhou, S. (2018). *Real estate price prediction using machine learning: A case study in Beijing*. *Journal of Computer and System Sciences*, 92, 49–59. DOI: 10.1016/j.jcss.2018.05.012
9. **Kaggle Dataset: Real Estate Price Prediction**Kaggle. (n.d.). *Real Estate Price Prediction*. Retrieved from<https://www.kaggle.com/>
10. **Gradient Boosting Literature**Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *The Annals of Statistics*, 29(5), 1189-1232. Retrieved from<https://www.jmlr.org/papers/v5/friedman04a.html>
11. **Python LangChain Examples**Chase, H. W. (2023). *LangChain: Python examples for building applications powered by language models*. GitHub. Retrieved from<https://github.com/hwchase17/langchain>
12. **Machine Learning Real Estate Analysis**Zhuang, F., Qi, Z., & Xiang, Z. (2018). *Machine learning in real estate: A survey*. *Computational Economics*, 51(3), 765-805. Retrieved from<https://www.springer.com/journal/10614>
13. **Real Estate Market Trend APIs**RapidAPI. (n.d.). *Real Estate Market Trend APIs*. Retrieved from<https://rapidapi.com/marketplace/apis>
14. **Streamlit Community Forums**Streamlit. (n.d.). *Streamlit Community Forums*. Retrieved from<https://discuss.streamlit.io/>
15. **Advances in NLP for Real Estate Systems**Lee, S. H., & Kim, H. (2017). *Advances in NLP for real estate systems*. *Proceedings of the 2017 International Conference on Artificial Intelligence*. DOI: 10.1145/3292500
16. **RAG (Retrieval-Augmented Generation)**Lewis, P., Perez, E., Piktus, A., et al. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Retrieved from<https://arxiv.org/abs/2005.11401>