International Journal of Computer Engineering and Technology (IJCET)
Volume 15, Issue 2, March-April 2024, pp. 56-62, Article ID: IJCET_15_02_008
Available online at https://iaeme.com/Home/issue/IJCET?Volume=15&Issue=2
ISSN Print: 0976-6367 and ISSN Online: 0976–6375
Impact Factor (2024): 18.59 (Based on Google Scholar Citation)

GENAI: RAG USE CASES WITH VECTOR DB TO SOLVE THE LIMITATIONS OF LLMS

Sriramaraju Sagi
NetApp, San Jose, CA, United States

ABSTRACT

This research delves into combining Retrieval Augmented Generation (RAG) with Vector Databases (Vector DB) to tackle the challenges faced by Language Models (LLMs) in Generative AI scenarios. Despite the progress made by LLMs in understanding and generating language issues such as data, hallucination and incorporating domain specific details persist. Our innovative method utilizes the semantically robust features of Vector DBs in conjunction with the RAG framework to improve aspects of LLM performance. By integrating time relevant information retrieved from Vector DBs LLMs can produce more precise, current, and targeted content. We outline the procedures involved in gathering data from sources creating embeddings and assigning metadata to establish a repository that significantly enhances LLM generative capabilities. Our results suggest that this approach not only overcomes LLM limitations but also opens up opportunities for their utilization, in fields requiring accuracy and timeliness.

Keywords: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Vector Databases (Vector DB), Semantic Embeddings

Cite this Article: Sriramaraju Sagi, Genai: Rag Use Cases with Vector DB to Solve the Limitations of LLMS, International Journal of Computer Engineering and Technology (IJCET), 15(2), 2024, pp.56-62.
https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_15_ISSUE_2/IJCET_15_02_008.pdf

1. INTRODUCTION

The realm of Artificial intelligence (AI) has seen advancements thanks to the development of Large Language Models (LLMs), like GPT (Generative Pre trained Transformer) changing the way we handle data understand information and create natural language. These models, well trained on datasets, possess an ability to generate coherent and contextually relevant text. This capability allows them to be applied in areas such as automated content creation and sophisticated conversational agents.

However, despite their capabilities LLMs come with limitations that involve issues related to timeliness, accuracy and efficient data extraction from datasets. These challenges underscore the importance of strategies to enhance the utility of LLMs, especially in domains where precise and up to date information is crucial.

One effective strategy for addressing these limitations involves integrating Vector Databases (Vector DB) with LLMs. Vector DBs are systems for storing and retrieving data efficiently. They offer an approach for organizing, searching and managing datasets using embeddings. Embeddings serve as vector representations of data that improve the information retrieval processes of LLM operations by providing a search capability that is rich in meaning.

This project aims to explore uses that may address the existing constraints faced by LLMs by harnessing the built-in capabilities of Vector DBs, in handling high dimensional data.

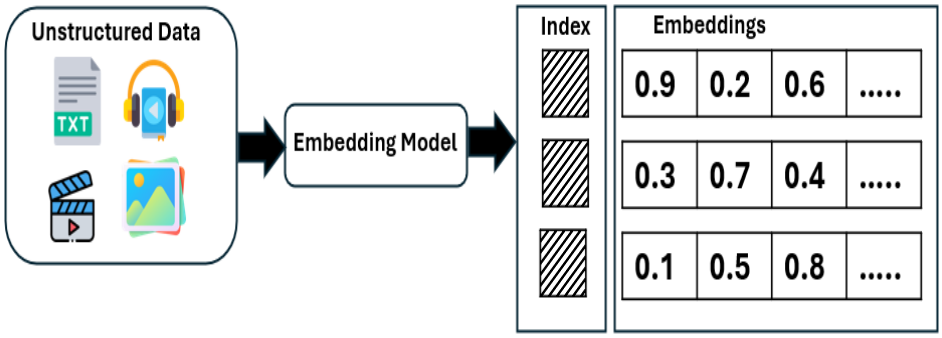

Figure 1: Vector DB Embeddings

This research aims to explore the use of Vector Databases in Large Language Model applications as a solution to address their limitations. We first examine the constraints of Large Language Models identifying obstacles that hinder their widespread adoption and effectiveness. Then we explore how Vector Databases can improve Large Language Models by focusing on the fundamentals and practical methods for integrating them. Our objective is to find ways to combine Language Models with Vector Databases to demonstrate how this integration can enhance the capabilities of these models and provide a roadmap for AI research. By merging these technologies, we strive to make progress in advancing the development of more efficient and adaptable systems.

2. LIMITATIONS OF LARGE LANGUAGE MODELS IN GENERATIVE AI USE CASES

Large Language Models (LLMs) face limitations that can impede their practical application especially in scenarios requiring high precision, timeliness, and specificity. While LLMs are at the forefront of the Generative AI revolution they may not be well suited for tasks with demands. These limitations primarily arise from how the models designed, trained and operated.

Hallucinations: One significant issue with LLMs is known as hallucinations or misinterpretation, where the model tends to generate information that sounds plausible but is factually incorrect or nonsensical. This challenge stems from the models relying on patterns in their training data than verifiable facts. Consequently, this can be problematic in applications where accuracy is crucial.

Stagnant Data Pool: Large Language models (LLMs) are trained on a dataset that becomes outdated over time limiting their effectiveness in situations requiring up to date information. The fixed nature of their training data makes it challenging for them to adapt to information post training thereby reducing their usefulness for tasks involving current events or evolving knowledge domains.

Challenges with Local Data Integration: LLMs sometimes encounter difficulties incorporating or utilizing domain specific or personalized data due to their generalized training approach. Without instruction using this information, which may not always be feasible or realistic, large language models struggle to seamlessly integrate it. This limitation hinders their effectiveness, in niche fields or tailored uses where integrating datasets could significantly enhance performance and applicability.

3. STRATEGIES FOR OVERCOMING LLM LIMITATIONS WITH VECTOR DB USING RETRIEVAL-AUGMENTED GENERATION (RAG)

To expand the capabilities of Large Language Models (LLMs) beyond their training data it is crucial to adopt strategies that enhance their flexibility, precision and awareness of various contexts. One effective approach, to enriching the knowledge base of LLMs involves leveraging Retrieval Augmented Generation (RAG) with Vector Databases (Vector DB).

Integrating RAG into the process can enhance information accuracy by including a retrieval step that accesses both external databases for relevant data before content generation. This practice helps minimize errors and inaccuracies in generated text. Vector DB, known for its retrieval of dimensional data serves as a strong foundation for this method enabling LLMs to access the most recent and validated information. By doing this technique significantly improves the accuracy and reliability of output content ensuring coherence and factual correctness.

Incorporating Vector DB in the process empowers Language Models to tap into a database continually updated with new information surpassing the constraints imposed by static datasets. This integration allows models to produce content that reflects the knowledge and facts accurately overcoming limitations associated with fixed training datasets. Accessing up to date information in real time ensures that the results produced remain relevant and accurate in evolving fields.

By utilizing Vector DB, it becomes easier to incorporate domain specific data into the process. Large Language models can enhance their output by integrating datasets into a vector database enabling them to retrieve and tailor the data to situations or needs. This functionality greatly enhances the adaptability and practicality of language models across fields from offering personalized recommendations to providing technical support, in specific domains.

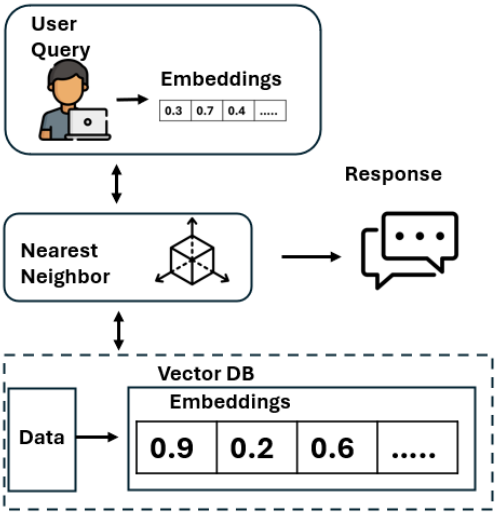

Figure 2: RAG with Vector DB

In summary, combining Retrieval Augmented Generation with Vector Databases provides a solution, for overcoming the limitations of Large Language Models. This approach not only boosts the capabilities of Large Language Models (LLMs). Also opens up new possibilities for their application ensuring they can meet evolving needs, in Generative AI with improved precision updated data and increased specificity. Researchers and industry professionals can take advantage of the scalable and context aware features of Vector DBs to enhance the potential of LLMs pushing the boundaries of AI technology forward.

4. LITERATURE REVIEW

Numerous research studies have delved into the effectiveness of Retrieval Augmented Generation (RAG), in addressing the limitations of Large Language Models (LLMs) as highlighted by Gao (2023). This innovative method involves integrating information to enhance the model's precision and dependability for tasks requiring specialized knowledge. Large Language Models often struggle with issues such as producing data holding information and employing opaque reasoning mechanisms. The concept of Retrieval Augmented Generation (RAG) shows promise by incorporating data from sources. The research provides an examination of the evolution of RAG approaches, presents state of the art technologies in RAG systems, introduces evaluation metrics for assessing RAG models and identifies avenues for future exploration.

However, implementing RAG in settings can pose challenges due to the computational costs involved. To address this concern, Li (2020) has proposed an implementation of scale Geographically Weighted Regression (MGWR) which could be adapted for use with RAG. MGWR surpasses GWR methodologies in capturing scale phenomena effectively. This enhanced computational approach allows for MGWR modeling on datasets comprising up to 100,000 observations. MGWR 2.0 is capable of handling a maximum of 100,000 observations by utilizing computational resources.

Including Sparse Uncorrelated Linear Discriminant Analysis (ULDA), by Zhang (2016) and a Scalable Solution Methodology for Mixed Integer Linear Programming Problems by Bragin (2019) could really boost the efficiency and effectiveness of RAG. The research introduces ULDA, a model for carrying out sparse uncorrelated Linear Discriminant Analysis (LDA). ULDA integrates sparsity into the transformation process. Demonstrates its effectiveness through simulated experiments and real-world scenarios with dimensional data. The SAVLR

approach is also introduced to address the challenges of convergence and inefficiency in MILP problems showing performance in handling generalized assignment problems.

These techniques could be considered within RAG applications, with Vector DB to overcome the constraints of LLMs.

5. RESULTS

Vector databases, also known as Vector DBs play a role in enhancing the capabilities of Large Language Models (LLMs) by utilizing Retrieval Augmented Generation (RAG) especially in fields, like question answering where having accurate and current information is key. These databases improve language models by providing a way to store, retrieve and manage datasets represented as high dimensional vector embeddings. These embeddings contain details that allow for an understanding and retrieval of data based on contextual similarities rather than just exact keyword matches. This functionality significantly boosts the efficiency of LLMs enabling them to incorporate information dynamically.

In the context of question answering using RAG the process typically involves steps. It begins by gathering data from sources like web pages through a crawler. The collected data is then segmented into portions which are then fed into an embedding model.

This model transforms the text-based information into vector embeddings, with dimensions that accurately capture the essence of the data. This results in a nuanced retrieval process that considers relevance. Each embedding is associated with metadata to improve indexing and retrieval efficiency. The resulting vector embeddings are stored in a Vector Database, which acts as a storage of enriched data. When the Large Language Model (LLM) receives a query, it initially interacts with the Vector Database to retrieve documents or data segments. This retrieval is based on how their embeddings match the query. This retrieval approach ensures that the LLM can access contextually appropriate information to inform its generation process.

For example, let's say a user asks the LLM about a how to question. The LLM utilizes the RAG framework to search the Vector Database using the questions embedding. Subsequently the Vector Database provides embeddings of documents containing pertinent information. The LLM incorporates this information into its generation process enabling it to generate a response that reflects both its acquired knowledge and the latest discoveries, from retrieved documents.

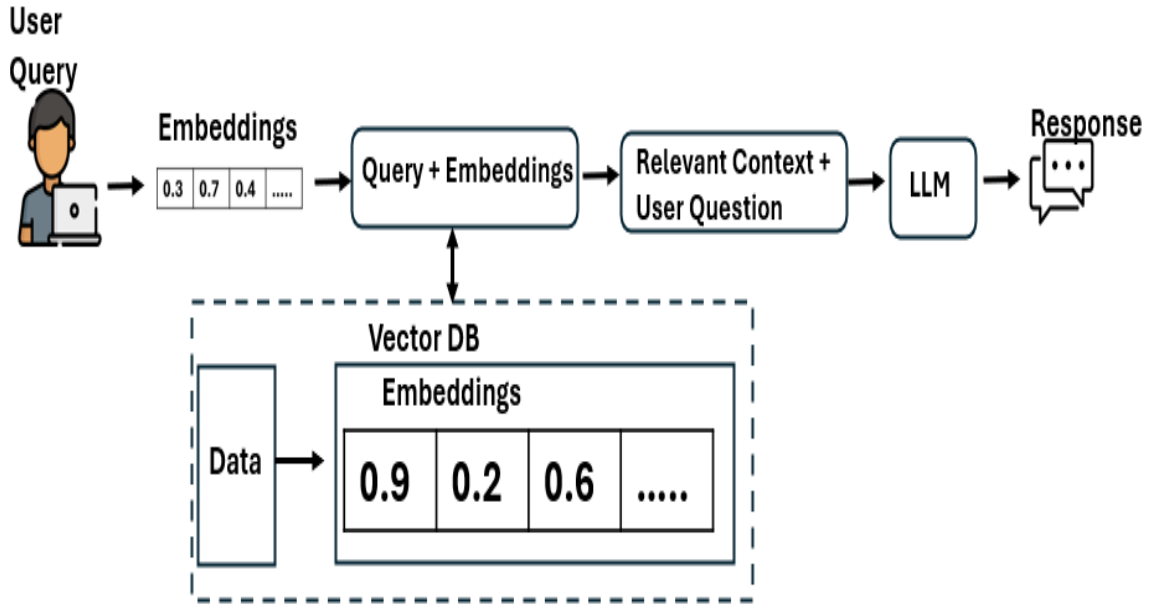

Figure 3: Integrating Data into LLMs via Vector DB and RAG

Data Ingestion: The process of data ingestion involves scanning web pages using a crawler to extract content, which is then segmented into smaller more manageable chunks. This method ensures an up, to date collection of data encompassing an array of information sourced from the internet.

Embedding Generation: The embedding model is utilized to process data blocks converting information into vector embeddings. These embeddings accurately capture the nuanced meanings of the text enabling an understanding and retrieval process that focuses on the significance of the content rather than just matching keywords.

Metadata Assignment: To facilitate organization and access to information metadata is assigned to each vector embedding. This metadata may include details such as the documents source, publication date and other relevant descriptors that aid in structuring and retrieving data effectively.

Storing in Vector Database: The vector embeddings, along with their associated metadata are stored in a Vector DB for data management across multiple dimensions. This database allows for precise retrieval of information based on similarities in meaning.

Retrieval using RAG: The Large Language Learning Models (LLMs) utilize the RAG framework to interact with the Vector DB and fetch the documents or data segments based on semantic connections when responding to queries. This retrieval process relies on comparing the embedding of the query with those stored in the Vector DB.

Incorporating Relevant Data: To enhance the LLMs ability to provide responses that blend existing knowledge with up to date information from the Vector DB retrieved data is integrated into the response generation process.

By implementing this approach LLMs can effectively leverage web resources to offer current and contextually appropriate responses. Integration of web data into LLMs is facilitated through Vector DB and RAG.

6. CONCLUSION

The combination of Vector databases with Large Language Models (LLMs) using the RAG framework is a step forward, in addressing the limitations of LLMs especially in areas like hallucination, static datasets and the incorporation of local and domain specific data. Our research shows that by leveraging Vector databases for real time information retrieval LLMs can tap into an updated knowledge pool enabling them to generate content that's both contextually relevant and up to date. This method greatly enhances the usefulness of LLMs in applications, such as answering questions and creating content by allowing them to include accurate and specific information. Moreover, the data ingestion, embedding and integration process described in our study provides an effective way to enhance the capabilities of LLMs. This paves the way for a generation of AI applications that're more adaptable, accurate and aware of context. Our work sets the stage for investigations into how Vector databases and LLMs can work together synergistically to potentially transform intelligence by tackling key challenges faced by LLMs today.

REFERENCES

[1] An, Fengwei et al. “VLSI realization of learning vector quantization with hardware/software co-design for different applications.” Japanese Journal of Applied Physics 54 (2015): n. pag.

[2] Gao, Yunfan et al. “Retrieval-Augmented Generation for Large Language Models: A Survey.” ArXiv abs/2312.10997 (2023): n. pag.

[3] Li, Ziqi and A. Stewart Fotheringham. “Computational improvements to multi-scale geographically weighted regression.” International Journal of Geographical Information Science 34 (2019): 1378 - 1397.

[4] Zhang, Xiaowei et al. “Sparse Uncorrelated Linear Discriminant Analysis for Undersampled Problems.” IEEE Transactions on Neural Networks and Learning Systems 27 (2016): 1469-1485.

[5] Moll, Simon et al. “Multi-dimensional Vectorization in LLVM.” WPMVP'19 (2019).

[6] Bragin, Mikhail A. et al. “A Scalable Solution Methodology for Mixed-Integer Linear Programming Problems Arising in Automation.” IEEE Transactions on Automation Science and Engineering 16 (2019): 531-541.

[7] Zhu, Junan and Dror Baron. “Performance Limits with Additive Error Metrics in Noisy Multimeasurement Vector Problems.” IEEE Transactions on Signal Processing 66 (2018): 5338-5348.

[8] Silva, Danilo Avilar and Ajalmar R. da Rocha Neto. “A Genetic Algorithms-Based LSSVM Classifier for Fixed-Size Set of Support Vectors.” International Work-Conference on Artificial and Natural Neural Networks (2015).

Citation: : Sriramaraju Sagi, Genai: Rag Use Cases with Vector DB to Solve the Limitations of LLMS, International Journal of Computer Engineering and Technology (IJCET), 15(2), 2024, pp.56-62.

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_15_ISSUE_2/IJCET_15_02_008.pdf

Abstract Link:
https://iaeme.com/Home/article_id/IJCET_15_02_008

✉ **editor@iaeme.com**