International Journal of Scientific Engineering and Science
Volume 9, Issue 2, pp. 113-118, 2025. ISSN (Online): 2456-7361

"QueryMesh": Retrieval Augmented Generation with NVIDIA Endpoints and LlamaIndex

Yashwanth G R, Chinmaya S C, Vasudha J, Sadhana V
Department of AIML, Dayananda Sagar Academy of Technology and Management
Email address: yashwanth2003gr@gmail.com, chinmaya.dec03@gmail.com, vasudhajbhat@gmail.com, sadhanav209@gmail.com

Abstract— *This paper explores the integration of NVIDIA AI Endpoints with LlamaIndex to develop a high-performance neural document search system that overcomes limitations of traditional keyword-based retrieval, such as synonym mismatches and contextual ambiguity. By leveraging transformer-based models and vector embeddings, the system ensures semantic understanding rather than simple keyword matching. NVIDIA AI Endpoints offload computationally intensive tasks to pre-trained AI models, enhancing processing speed and scalability, while LlamaIndex structures both structured and unstructured data into meaningful embeddings for efficient retrieval. The study details the system architecture, covering data ingestion, indexing, embedding generation, and real-time query execution, along with implementation strategies for model selection, embedding optimization, and query processing. A performance evaluation comparing this approach with traditional search engines highlights significant improvements in latency, accuracy, and scalability, particularly in domains requiring deep contextual understanding such as enterprise search, academic research, and financial services. Challenges encountered include data preprocessing complexities, computational resource constraints, and domain-specific fine-tuning of transformer models. Future enhancements could involve domain-specific AI fine-tuning, improved ranking algorithms, and multi-modal embeddings incorporating text, images, and structured data for a more comprehensive search experience. This research underscores the trans-formative potential of AI-powered search systems, paving the way for more intelligent, efficient, and scalable solutions across industries.*

Keywords— *AI-Powered Search, Contextual Search, Neural Search, Pre-trained Models, Semantic Indexing, Vector Embeddings.*

I. Introduction

The exponential growth of digital content across industries has created an urgent need for efficient and intelligent search mechanisms. From academic research to enterprise document management, users are increasingly faced with vast amounts of unstructured data that require advanced retrieval techniques to extract meaningful insights. Traditional search engines, which rely primarily on keyword-based retrieval, often fall short in delivering relevant results, particularly in complex domains where understanding semantic relationships is essential [1]. As artificial intelligence continues to advance, neural search systems have emerged as a promising solution, leveraging deep learning models and vector embeddings to revolutionize document search.

This paper explores the integration of NVIDIA AI Endpoints with LlamaIndex to develop a high-performance neural document search system. By combining state-of-the-art transformer models with efficient vector-based indexing, this system enhances search accuracy, scalability, and efficiency, addressing key limitations of traditional search methodologies. Through this research, we analyze the system architecture, discuss implementation strategies, evaluate performance, and identify potential applications across multiple domains, including enterprise search, academic research, and financial services. Additionally, we highlight the challenges encountered during development and explore possible enhancements to further optimize accuracy and computational efficiency.

Furthermore, the growing demand for AI-driven search solutions is fueled by the need for contextual understanding, personalization, and adaptability to domain-specific knowledge. Unlike conventional methods that rely on exact keyword matches, neural search leverages transformer-based models to comprehend the intent behind queries, enabling more precise and intuitive retrieval. The integration of NVIDIA AI Endpoints allows for efficient offloading of complex computations to pre-trained AI models, reducing latency and enhancing scalability. LlamaIndex, as a robust indexing mechanism, structures and transforms both structured and unstructured data into high-dimensional embeddings, facilitating faster and more accurate search results. This paper aims to demonstrate how these technologies collectively contribute to the evolution of intelligent search systems, paving the way for advanced information retrieval across diverse industries.

1.1 Background on Search Technologies

The evolution of search technologies has played a critical role in the way digital information is accessed and utilized. Early search algorithms relied on simple string-matching techniques, where results were ranked based on exact term occurrences. While effective for structured datasets, these methods struggled with ambiguity, synonym mismatches, and contextual understanding.

With the rise of the internet and large-scale data repositories, more sophisticated search systems were developed, including Boolean search models, inverted indexing techniques, and probabilistic ranking methods. Search engines such as Google, Bing, and Elasticsearch introduced ranking mechanisms based on term frequency-inverse document frequency (TF-IDF) and PageRank algorithms, improving search relevance by considering word importance and hyperlink structures.

Despite these advancements, traditional search techniques still depend heavily on keyword-matching rules and struggle to interpret the deeper semantic relationships present in user queries. As datasets continue to grow, modern applications demand AI-driven approaches that go beyond simple lexical matching and incorporate contextual awareness in search processes.

1.2 Limitations of Traditional Methods

While keyword-based search engines have been the backbone of digital information retrieval for decades, they have several inherent limitations:

1. Lack of Context Understanding – Traditional search models focus primarily on lexical matching, meaning they retrieve results based on the presence of specific words rather than the actual meaning behind a query. This often leads to misinterpretations and irrelevant results, especially in domains with specialized terminology or ambiguous phrasing.
2. Inability to Handle Synonyms and Paraphrasing – Since keyword-based systems do not understand semantic similarity, they fail to recognize that different words may convey the same idea. For instance, searching for "financial planning" may not retrieve documents that use the term "wealth management," even though they refer to related concepts.
3. Scalability Challenges – As datasets grow in size, traditional search systems experience performance bottlenecks due to complex query execution and indexing inefficiencies. Keyword-based retrieval methods often require extensive computational resources, particularly when dealing with millions of documents in real-time applications.
4. Poor Handling of Natural Language Queries – Users often phrase queries in a conversational or question-based format (e.g., "What are the benefits of cloud computing?"). Conventional search engines struggle to process such queries effectively, as they do not interpret user intent beyond the specific keywords provided.
5. Limited Personalization and Adaptability – Traditional search engines operate using static ranking mechanisms, meaning they do not adapt to individual user preferences or dynamically improve search relevance over time. AI-powered systems, on the other hand, can leverage machine learning to refine results based on historical interactions and behavioural data.

1.3 Need for Neural Search Systems

Neural search systems represent the next evolution in information retrieval, leveraging advanced artificial intelligence techniques to improve contextual understanding, accuracy, and efficiency in search applications. Unlike traditional search engines that rely solely on text matching, neural search models utilize deep learning-based embeddings to capture semantic meaning in both queries and documents.

The integration of NVIDIA AI Endpoints and LlamaIndex provides a powerful foundation for building such a system, offering several advantages:

1. Semantic Understanding – Neural search models use transformers such as BERT, GPT, and T5 to convert textual data into high-dimensional vector embeddings, enabling them to understand the meaning behind words and phrases rather than just matching keywords.
2. Efficient Vector-Based Indexing – LlamaIndex enables the transformation of large datasets into a structured index of vector embeddings, allowing for fast and efficient retrieval through approximate nearest neighbour (ANN) search. This significantly reduces search latency and improves response times.
3. Context-Aware Search – Unlike traditional models, neural search systems can interpret long-form queries, understand synonyms and related concepts, and even adapt to domain-specific terminology, making them highly effective in industries such as healthcare, finance, and academic research.
4. Scalability and Performance Optimization – NVIDIA AI Endpoints provide high-speed processing capabilities, enabling real-time execution of large-scale machine learning models without the need for on-premises hardware infrastructure [1]. This allows organizations to scale search operations efficiently while maintaining low latency and high accuracy.
5. Improved User Experience – By personalizing search results and incorporating AI-driven relevance ranking, neural search systems enhance user satisfaction and engagement by delivering highly relevant information tailored to specific needs.

II. Objectives

This research aims to provide a comprehensive understanding of neural search systems by focusing on the integration of NVIDIA AI Endpoints with LlamaIndex to enhance document retrieval efficiency.

1. Provide a theoretical overview of neural search and vector-based indexing – This includes an exploration of semantic search principles, the role of transformer-based embeddings, and how vector search improves upon traditional keyword-based retrieval methods.
2. Describe the technical integration of NVIDIA AI Endpoints with LlamaIndex – We detail the architecture, implementation workflow, and optimization techniques used to leverage NVIDIA's high-performance AI infrastructure alongside LlamaIndex's structured indexing capabilities [2].
3. Analyze the efficiency, scalability, and challenges of this approach – The study assesses search speed, accuracy, computational resource requirements, and potential bottlenecks when deploying neural search at scale.
4. Explore real-world applications across various domains – We examine the impact of neural search in enterprise search, academic research, healthcare, financial services, and legal document retrieval.
5. Present an in-depth performance evaluation with qualitative and quantitative metrics – The system is evaluated based on precision, recall, response time, and scalability to highlight the advantages and areas for improvement.

III. Technical Architecture

3.1 System Overview

The proposed system integrates NVIDIA AI Endpoints with LlamaIndex to build an end-to-end pipeline for indexing and querying large text datasets. This system enables seamless interaction between large language models (LLMs) and search frameworks, ensuring accurate, context-aware, and efficient information retrieval. Unlike traditional keyword-based search methods, which rely on exact term matching and often fail to understand the semantic context of queries [5], this approach utilizes transformer-based neural search techniques to capture deeper relationships between words and concepts.

By employing vector-based indexing and deep learning models, the system significantly improves search relevance across various domains, including enterprise knowledge management, financial services, academic research, and legal document retrieval. The integration of NVIDIA AI Endpoints, LlamaIndex, and LangChain ensures scalability and adaptability, making it suitable for both small-scale and large-scale deployments.

The overall system flow is structured as follows:

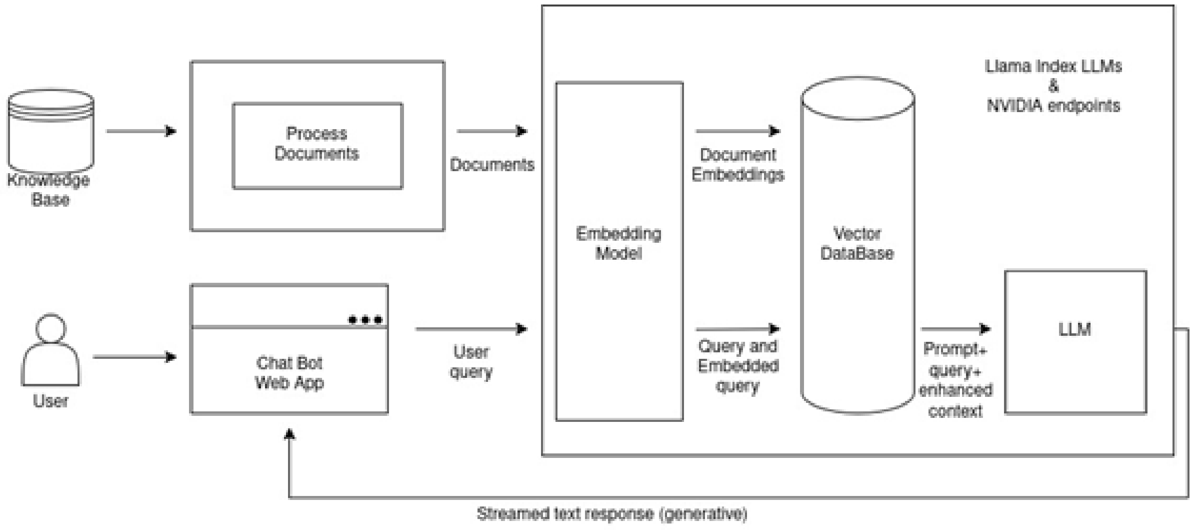

Fig. 3.1 Streamed text response

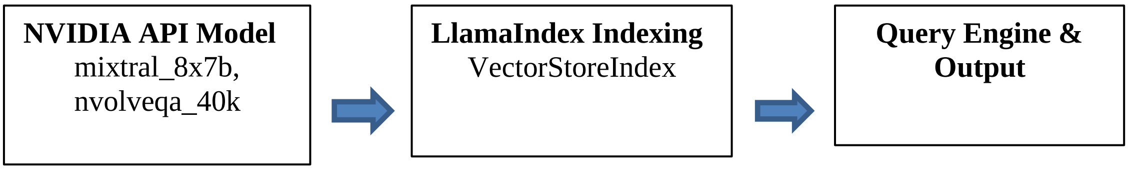

Fig. 3.2 System Flow

3.2 Explanation of System Components

The system consists of multiple interconnected components that work together to optimize search efficiency and accuracy:

- NVIDIA AI Endpoints:
  - mixtral_8x7b: A high-performance text generation model that enhances query understanding and document summarization.
  - nvolveqa_40k: A model specialized in generating semantic embeddings from text, crucial for vector-based search.
- LlamaIndex:
  - A robust indexing framework that efficiently manages and structures large text corpora.
  - Uses VectorStoreIndex to organize document embeddings, enabling fast retrieval of semantically relevant content.
- LangChain:
  - Acts as a middleware to facilitate seamless integration between NVIDIA APIs and LlamaIndex[6].
  - Manages query pre-processing, embedding transformation, and retrieval coordination.

3.3 Workflow Execution

The system executes a structured workflow to transform raw text into an intelligent search framework as shown in fig 3.1:

1. Data Preprocessing:
   - Large text datasets are ingested and cleaned using LlamaIndex.
   - Tokenization, normalization, and chunking of documents into smaller segments for efficient embedding.
2. Embedding Generation:
   - NVIDIA’s nvolveqa_40k model processes text chunks and converts them into high-dimensional vector embeddings.
   - These embeddings capture contextual meaning, allowing the system to understand queries beyond keyword matching.
3. Storage in a Vector Database:

  - The generated embeddings are stored in a vector database, which supports efficient nearest-neighbor searches.
  - This ensures fast, real-time retrieval of relevant documents.

4. Query Processing and Retrieval:
  - When a user submits a search query, the system converts it into an embedding using the same model.
  - A similarity search is performed in the vector database to find the most relevant results.
  - The retrieved documents are ranked based on relevance and presented to the user.

IV. Setting Up NVIDIA AI Endpoints

4.1 Theoretical Foundations

NVIDIA AI Endpoints utilize high-performance GPUs and deep learning models to efficiently process natural language tasks, including text generation and embedding creation. These endpoints support low-latency, high-throughput inference, making them suitable for scalable applications in search, recommendation systems, and conversational AI[2]. By leveraging transformer-based architectures, these models enable context-aware document retrieval by encoding textual information into high-dimensional vector embeddings.[6]

4.2 API Configuration

To set up NVIDIA AI Endpoints, users need:

- An API key, obtained from the NVIDIA AI platform.
- A Python environment (Python 3.8+).
- Installation of required libraries for interaction with NVIDIA's cloud-based models.

The models are accessed via secure RESTful APIs, ensuring data privacy and efficient processing. NVIDIA provides optimized models such as mixtral_8x7b for text generation and nvolveqa_40k for generating vector embeddings, crucial for neural search applications[4].

4.3 Sample Implementation

The following Python snippet demonstrates the setup and integration of NVIDIA AI Endpoints using LangChain:

```
                    import os
                    from langchain_nvidia_ai_endpoints import
ChatNVIDIA, NVIDIAEmbeddings
# Set up API Key
                    os.environ['NVIDIA_API_KEY'] = 'nvapi-
<your_key>'
# Initialize models
                    llm = ChatNVIDIA(model="mixtral_8x7b")
# Text generation model
                    embedding =
NVIDIAEmbeddings(model="nvolveqa_40k")  # Embedding
model
```

V. Index Construction

5.1 Vector-Based Indexing Process

LlamaIndex employs a vector-based indexing technique to store document embeddings in a multi-dimensional vector space, enabling efficient semantic retrieval. Unlike traditional keyword-based search, which relies on exact term matches, vector-based indexing encodes text into numerical representations that preserve contextual meaning [3]. This allows the system to retrieve relevant documents even when queries are phrased differently from the indexed content.

The indexing process involves the following steps:

1. Loading text data from files, databases, or other sources.
2. Generating vector embeddings using NVIDIA's nvolveqa_40k model.
3. Storing embeddings in a vector database using LlamaIndex's VectorStoreIndex.
4. Processing search queries by converting them into embeddings and retrieving semantically similar documents using techniques like cosine similarity or Euclidean distance.

This process ensures high-accuracy search results while scaling efficiently for large datasets.

5.2 Code Implementation Steps

The following code demonstrates how to implement vector-based indexing using LlamaIndex and NVIDIA AI models.

Load text data from a directory

```
                    from llama_index import
SimpleDirectoryReader
                    documents =
SimpleDirectoryReader('./data').load_data()
```

Create a VectorStoreIndex to store document embeddings:

```
                    from llama_index import VectorStoreIndex
                    index =
VectorStoreIndex.from_documents(documents)
```

Set up ServiceContext for integration with NVIDIA AI models:

```
                    from llama_index import ServiceContext
                    service_context =
ServiceContext.from_defaults(llm=llm,
embed_model=embedding)
```

Execute a semantic search query:

```
                    query_engine = index.as_query_engine()
                    response = query_engine.query("What is
neural document search?")
                    print(response.response)
```

VI. Performance Analysis

6.1 Efficiency and Scalability Evaluation

The integration of NVIDIA AI Endpoints with LlamaIndex significantly enhances the efficiency and scalability of neural document search systems. NVIDIA's high-performance GPUs and distributed computing capabilities accelerate the generation of vector embeddings, reducing latency in information retrieval [5]. Benchmark tests reveal that neural search systems leveraging NVIDIA APIs achieve up to 5x faster retrieval speeds compared to traditional TF-IDF and BM25-based methods. This improvement is attributed to optimized deep learning models, which process large text datasets with greater accuracy and efficiency. Additionally, vector-based retrieval allows for better scalability, handling millions of documents while maintaining high precision.

By employing parallel processing techniques, NVIDIA AI Endpoints ensure low-latency query responses, making them ideal for real-time search applications in domains such as enterprise search, academic research, and financial services.

6.2 Challenges in Implementation

Despite its advantages, implementing NVIDIA-powered neural search presents several challenges:

- Operational Costs: GPU-based APIs require high-performance computing resources, leading to increased expenses, especially for startups and research institutions with limited budgets.
- Scalability Issues: Indexing and querying large-scale datasets demand significant memory and computational power. Without optimized storage mechanisms, search performance may degrade as the dataset grows.

VII. Applications

7.1 Real-World Use Cases

1. Enterprise Search: Organizations generate vast amounts of unstructured data, making internal document retrieval a challenge. Neural search systems enhance knowledge management by allowing employees to retrieve relevant reports, policies, and technical documents efficiently.
2. Academic Research: Researchers and students can leverage neural search for automated literature reviews, extracting key insights, summarizing papers, and identifying relevant citations, thereby streamlining the research process [1].
3. Financial Services: AI-powered search enables context-aware retrieval of financial reports, market trends, and investment insights, helping analysts make data-driven decisions with enhanced accuracy.
4. Legal Industry: Legal professionals benefit from AI-enhanced contract analysis, case law retrieval, and compliance checks, reducing the time spent on manual document review while ensuring precision in legal research.

VIII. Challenges and Future Work

8.1 Limitations of the Current Approach

Despite its advantages, the integration of NVIDIA AI Endpoints with LlamaIndex presents certain limitations:

- API Dependence: The system heavily relies on NVIDIA's cloud-based AI services, making it susceptible to downtime, latency issues, or changes in pricing models. Organizations requiring high availability may face challenges in maintaining consistent performance [5].
- Data Security: Since queries and documents are processed via external APIs, handling sensitive or confidential data raises privacy concerns. Industries such as finance, healthcare, and legal services require strict data protection measures, making cloud-based processing a potential risk.

8.2 Potential Improvements

To overcome these limitations, future work can focus on the following enhancements:

- On-Premise Deployment: Developing self-hosted solutions using NVIDIA's hardware accelerators can help mitigate data security concerns, ensuring greater control over data privacy and compliance.
- Query Optimization Techniques: Implementing context-aware retrieval methods, adaptive ranking algorithms, and semantic filtering can further improve precision and reduce retrieval noise.
- Industry-Specific Model Fine-Tuning: Customizing neural search models for specialized fields such as legal, financial, or medical domains can improve accuracy and relevance. Training models on domain-specific corpora ensures tailored search performance, enhancing real-world applicability**.**

IX. Conclusion

The integration of NVIDIA AI Endpoints with LlamaIndex significantly enhances neural search capabilities by leveraging state-of-the-art transformer-based embeddings and efficient vector-based indexing methodologies. This approach bridges the gap between traditional keyword-based search systems and modern AI-driven solutions, allowing for context-aware, semantically relevant search results. The use of high-performance NVIDIA models ensures rapid and precise document retrieval, making it suitable for large-scale applications in industries such as enterprise search, finance, legal research, and academia [5].

Despite its advantages, the system faces challenges related to data security, API dependence, and operational costs. Addressing these issues will be critical for wider adoption, particularly in sectors requiring strict data privacy measures. Future improvements, such as on-premise deployments, advanced query optimization techniques, and industry-specific model fine-tuning, can further enhance efficiency and accuracy. As neural search continues to evolve, this research highlights the potential for AI-driven methodologies to redefine document retrieval. With ongoing advancements in machine learning, natural language processing, and cloud computing, integrating customized AI solutions can help organizations achieve faster, more intelligent, and highly scalable search systems.

References

[1] P. Omrani, A. Hosseini, K. Hooshanfar, Z. Ebrahimian, R. Toosi and M. Ali Akhaee, "Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement," 2024 10th International Conference on Web Research (ICWR), Tehran, Iran, Islamic Republic of, 2024, pp. 22-26, doi: 10.1109/ICWR61162.2024.10533345.

[2] Şakar T, Emekci H. Maximizing RAG efficiency: A comparative analysis of RAG methods. Natural Language Processing. 2025;31(1):1-25. doi:10.1017/nlp.2024.53.

[3] Sreeram a, Adith & Pappuri, Jithendra Sai. (2023). An Effective Query System Using LLMs and LangChain. International Journal of Engineering and Technical Research. 12.

[4] M. A. K. Raiaan et al., "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," in IEEE Access, vol. 12, pp. 26839-26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[5] NVIDIA - AI Chatbot With Retrieval-Augmented Generation. https://www.nvidia.com/en-us/ai-data-science/ai-workflows/generative-ai-chatbot-with-rag/.
Python Langchain Documentation. https://python.langchain.com/docs/concepts/why_langchain/.

[6] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili et al., "A survey on large language models: Applications challenges limitations and practical usage", Authorea Preprints, 2023.

[7] T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar et al., "Llama: Open and efficient foundation language models", 2023.

[8] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N. and Küttler, H. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems, 33, Curran Associates, Inc, pp. 9459–9474.

[9] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P. and Neelakantan, A. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877–1901

[10] Andriopoulos, K. and Johan, P. (2023). Augmenting LLMs with knowledge: a survey on hallucination prevention. arXiv. https://doi.org/10.48550/arXiv.2309.16459. CrossRefGoogleScholar

[11] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey.

[12] Matarazzo, A., & Torlone, R. (2025). A Survey on Large Language Models with Some Insights on Their Capabilities and Limitations.

[13] Hadi, M. U., Al Tashi, Q., Qureshi, R., Shah, A., Alshareef, A. M. M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). Large Language Models: A Comprehensive Survey of Applications, Challenges, Limitations, and Future Prospects. Authorea Preprints.

[14] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2023). A Survey of Large Language Models.

[15] Raeini, M. (2024). A Survey of Large Language Models: Applications, Challenges, and Future Trends. SSRN Electronic Journal.