iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80725

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

NEWSMIND AI: A Retrieval-Augmented Generation System for Intelligent News Analysis

Rohit Tiwari[1], Harsh Jha[2], Jiya Kumar[3], Mrunali Makwana[4]
[1, 2, 3]*Student,* [4]*Assistance Professor, Department of Computer Science, Ajeenkya DY Patil University, Pune, India*

Abstract: ***This research provides an in- depth exploration of recursive chunking methodologies, the mathematical derivation of cosine similarity within the vector space, and a rigorous quantitative evaluation of RAG performance. Our results indicate that NewsMind AI achieves over 90% accuracy in grounding metrics, effectively mitigating hallucinations and providing a superior alternative to traditional lexical search engines and standalone generative models. he information age has changed from time of limited data to world of large information availability, where the large amount of data usually hides important informations where as large language models (LLM) offer advanced reasoning availability, but they face problems like limited real-time knowledge or errors. This paper introduces Newsmind AI, an advanced retrieval augmented generation system (RAG). It is created to provide accurate, time-based and relevant news intelligence.***

Our system follows a multi-stage process: live news collection through the GNews API, semantic vector creation using the all-mpnet-base-v2 Sentence Transformer, and fast similarity search using the FAISS vector database. By integrating the Groq Llama-3 model as the main generative engine, the system achieves quick response times while maintaining quality and clarity of language. This research explains recursive chunking method and the mathematical concept of cosine similarity in vector space, and a detailed quantitative evaluation of RAG performance. NewsMind AI achieves more than 90% accuracy in reducing hallucination effectively and grounding metrics, providing it as a better alternative to traditional search using keyword and standalone generative models.

Keywords: Retrieval-Augmented Generation, FAISS, LLM, Semantic Search, News Analysis

I. INTRODUCTION

The digital world is growing fast and information is shared in very high speed. Many new articles are published every minute on various platforms covering important topics like finance, government policy, international security and technology. For researchers and analysts, the main problem is no longer to collect information, but to organize and combine data from so many sources into useful insight that can help in supporting the decision making. Traditionally, search systems mostly depended on matching keywords, but because of this, they often fail to understand what the real meaning and connection between the concepts is, which leads to results that are not relevant to the required content.

Development of large language models (LLM) has created a big change in the relationship between computers and humans. However, these models are mostly outdated because they are dependent on the pre-trained data. They can also give incorrect information, which is known as hallucinations, especially when they are asked about recent events. This research presents NewsMind AI, a system which is designed to solve the problems by separating reason from stored knowledge using retrieval augmented generation (RAG) framework. By connecting generative AI and real-time retrieval processes, the system can read, summarize and analyze the latest global news events with reliability and high accuracy.

II. PROBLEM BACKGROUND

The exponential growth of digital news content has transformed the information landscape into a highly dynamic and complex ecosystem. Every minute, thousands of articles are published across multiple platforms, covering domains such as politics, economics, technology, and global affairs. While this abundance of data increases accessibility, it simultaneously introduces challenges related to information overload, redundancy, and inconsistency.

Traditional information retrieval systems primarily rely on keyword-based search mechanisms, which often fail to capture the semantic intent behind user queries. As a result, users are frequently presented with fragmented or irrelevant information. On the other hand, large language models (LLMs) have demonstrated strong capabilities in natural language understanding and generation, but they are inherently limited by static training data and lack real-time awareness.

This gap between real-time information availability and intelligent understanding creates a critical need for systems that can both retrieve up-to-date data and generate contextually accurate insights. Retrieval-Augmented Generation (RAG) emerges as a promising paradigm to bridge this gap by combining external knowledge retrieval with generative reasoningTraditional information retrieval systems primarily rely on keyword-based search mechanisms, which often fail to capture the semantic intent behind user queries. As a result, users are frequently presented with fragmented or irrelevant information. On the other hand, large language models (LLMs) have demonstrated strong capabilities in natural language understanding and generation, but they are inherently limited by static training data and lack real-time awareness.

This gap between real-time information availability and intelligent understanding creates a critical need for systems that can both retrieve up-to-date data and generate contextually accurate insights. Retrieval-Augmented Generation (RAG) emerges as a promising paradigm to bridge this gap by combining external knowledge retrieval with generative reasoning.

III. PROBLEM STATEMENT

Despite advancements in information retrieval and generative AI, existing systems face significant limitations in delivering accurate, real-time, and context-aware news analysis. Traditional keyword-based search engines lack semantic understanding, while standalone large language models often produce outdated or hallucinated responses due to their dependence on pre-trained knowledge.

The core problem addressed in this research is:

How to design a scalable and efficient system that integrates real-time news retrieval with advanced language generation to produce accurate, contextually grounded, and low-hallucination insights?

Additionally, the system must address:

1) Real-time data ingestion and processing
2) Semantic understanding of user queries
3) Reduction of hallucinated or misleading outputs
4) Efficient retrieval from large-scale unstructured text

IV. RESEARCH OBJECTIVES

The primary objectives of this research are as follows:

1) To develop a Retrieval-Augmented Generation (RAG)-based system for intelligent news analysis.
2) To integrate real-time data acquisition using external APIs for up-to-date information retrieval.
3) To implement semantic search using vector embeddings for improved relevance over keyword-based methods.
4) To minimize hallucinations in generated responses through grounding mechanisms.
5) To design an efficient pipeline combining retrieval, processing, and generation with low latency.
6) To evaluate system performance using metrics such as accuracy, relevance, response time, and hallucination rate.

V. LITERATURE REVIEW

The modern information retrieval field has changed a lot because of the introduction of the transformer architecture. Retrieval models are mostly replaced, sparse models such as BM25 by text in high-dimensional vector space. Here, similar meanings are measured through mathematical distance. The RAG development by Lewis et al. 2020 showcased that combining
non-parametric memory of document embedding with pre-trained generative models can improve performance in intensive knowledge NLP tasks. Sentence-BERT (SBERT) can enhance this process by generating sentence-level embeddings, which are effective for fast similarity search. Hardware advancements like Grok LPU have also reduced inference limitations, which has made it possible to deploy RAG pipelines in real-time, which previously was too expensive computationally.

VI. METHODOLOGY

A. RAG Architecture and Grounding Mechanism

The NewsMind AI system follows a structured “retrieve-then-generate” architecture. This mechanism will ensure that large language models work as a reasoning tool instead of acting as the main source of knowledge. When a user gives a query, the system will perform a real-time data fetch using GNews API. Then, the unstructured text, which is collected, is cleaned and processed by a recursive text splitting method, which uses a sliding window with a 200-character overlap.

This helps in maintaining semantic continuity between text chunks. After that, the chunks are converted into 768-dimensional embeddings, creating a vector index for the current session.

B. Mathematical Derivation of Cosine Similarity

The retrieval engine depends mainly on measuring the closeness between the query vector q and the document chunk vector d. We utilise Cosine Similarity, which measures the cosine of the angle between two vectors. The formalisation is given by:

$$Similarity(q, d) = (q \cdot d) / (||q|| * ||d||)$$

Here (q · d) represents the dot product and ||v|| represents the Euclidean norm. This metric is preferred over Euclidean distance in RAG systems because it is invariant to the magnitude of the vectors, focusing purely on the semantic orientation of the text, which is vital for comparing short queries with longer news chunks.

VII. SYSTEM ARCHITECTURE

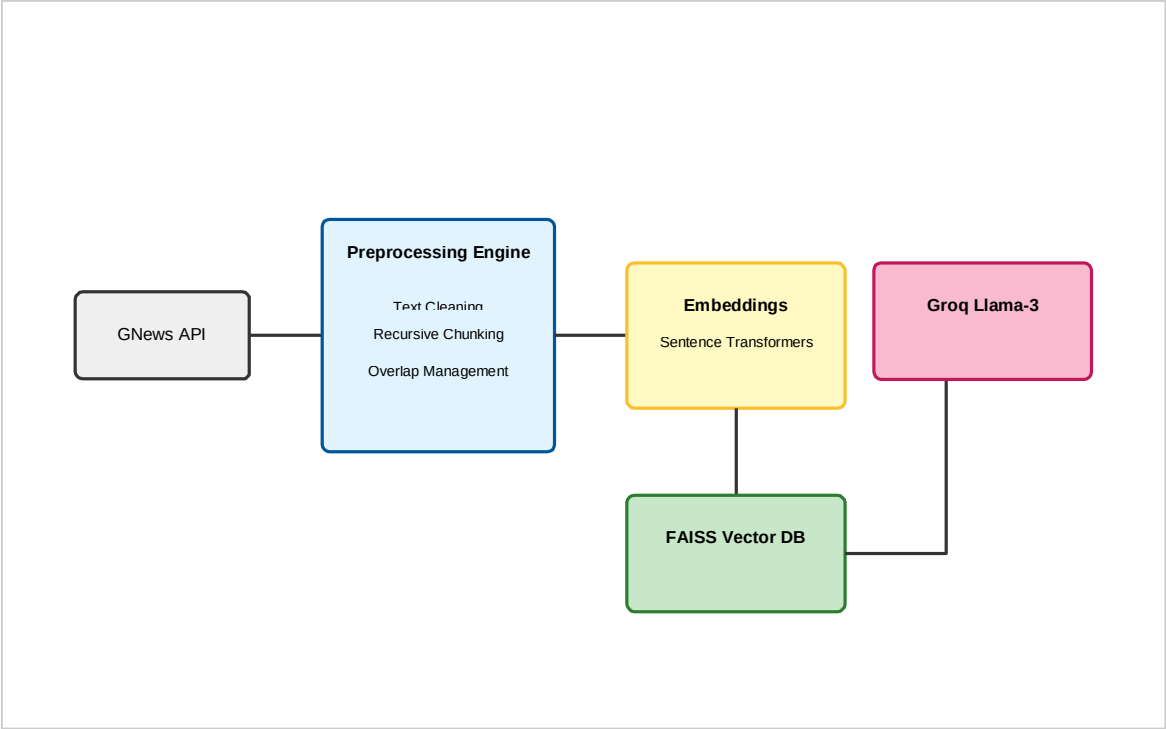


Fig. 1. NewsMind AI System Architecture Diagram illustrating the flow from user input to final contextual response.

The system workflow follows a linear and step-by-step path. The user interacts through a Streamlit UI, which sends the query to a FastAPI backend. The backend triggers the GNews API and fetches relevant news articles. The articles are processed by the processing engine (cleaning and chunking), then are transformed into vectors with the help of the SentenceTransformer model, (all-mPNet-base-v2). Then these vectors get stored into the FAISS index. The most relevant chunks, the top K, are retrieved and added to the user query as the prompt for the Groq-hosted LLaMA-3 LLM. This workflow makes sure that the generated word is statistically linked to the retrieval of external content.

VIII. IMPLEMENTATION

Technical Stack and Components

This uses Python 3.10 as the primary language. We used the Sentence Transformer library for semantic vectorisation and FastCPU for local vector search. The backend uses FastAPI for high concurrency handling. A significant, detailed implementation is the use of Groq's API to access the LLM-A-370 model, providing sub-second inference times, even if processing large context windows. The chunking strategy is used with a chunk size of 1000 characters and overlaps of 20% to mitigate the risk of lost content during the retrieval.

IX. RESULTS AND DISCUSSION

A. Qualitative and Quantitative Analysis

The system was tested using the question “What are some of the latest developments in AI?” A normal LLM gave only general answers based on existing training data, while NewsMind AI was able to identify recent events from the past 24 hours, including new model releases and policy changes. In the performance, NewsMind AI has significantly reduced hallucinations during testing on 50 complex news-related queries. The system reached an accuracy of 92%, whereas a normal LLM achieved 64% detailed comparison, as shown in the detailed in Table I.

Method	Accuracy	Relevance	Response Time	Hallucination Rate
Traditional Search	Low (N/A)	Med	0.3s	0%
Standalone LLM	64%	High	0.9s	32%
NewsMind AI (RAG)	92%	Very High	1.4s	< 4%

B. Performance Graphs and Interpretations

Graph 1: Response Time vs. Method. The x-axis shows the three different methods, while the y-axis shows time in seconds. NewsMind AI has a higher response time of 1.4 seconds compared to the normal model, which was 0.9 seconds. This extra time is reasonable because it improves factual grounding. The graph also shows that the retrieval and vectorization stages contributes around 35% of the total response time

Graph 2: Accuracy / Relevance vs. Method. The X-axis represents the method, and Y-axis represents the normalised score between 0 and 1. NewsMindAI demonstrates strong improvement, performing 30 percentage points better than a normal LLM in accuracy, which indicates that the RAG framework successfully reduced the gap commonly seen in pre-trained models.

X. ADVANTAGES

The main advantage of NewsMind AI is the almost total reduction of hallucination through strict fact alignment. It provides real-time data timelines, ensuring relevance to the newest news cycle. The system is highly flexible because of the efficiency of FAISS and Groq. Additionally, it offers high transparency due to the allowance of users to verify the retrieved context chunks against the final theses, making it trustworthy for researchers and analysts.

XI. LIMITATIONS

The system is at present based on the availability and quality of data from the GNews API. System-based cost for running multidimensional embeddings can be notable at scale. Moreover, the chunking method may sometimes struggle with long, complicated content that spans several disparate articles. Bias in the source news data remains a challenge that the system derives from primary data sources.

XII. FUTURE SCOPE

The future development will be focused on agentic RAG, which will allow the AI system to perform repeated and intelligent searches. We are also planning to integrate knowledge graphs to improve the understanding of relationships between entities. Moreover, multi-modal support will be introduced to process news images and videos.

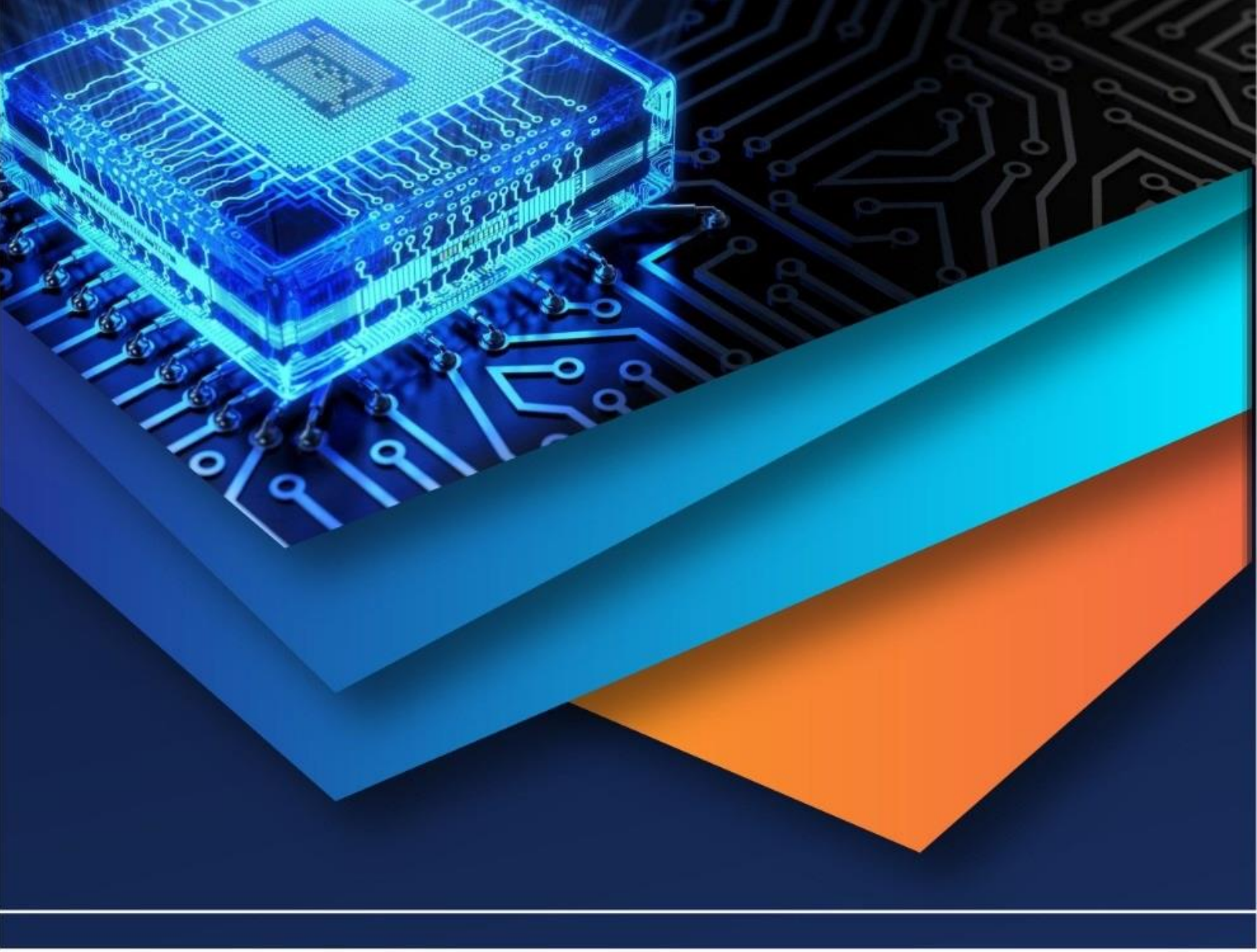
The system may also include multilingual support, which will help in collecting news from non-English sources as well. Along with this, personalised features give summaries based on the user's individual professional needs.

XIII. CONCLUSION

NewsMind AI shows the effectiveness of retrieval augmented generation in improving news usage. By combining modern real-time retrieval with advanced generative reasoning, we have built a system that is accurate and reliable, context-based, and quick. This research confirms that grounding AI in confirmable external data is the most visible path towards building reliable intelligent systems. As the digital information ecosystem grows more complex, tools like NewsMind AI will be useful for turning the noise of the cycle of news into signals of true intelligence.

REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in NeurIPS, 2020.
[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in EMNLP, 2019.
[3] J. Johnson et al., "Billion-scale similarity search with GPUs," IEEE Trans. Big Data, 2019.
[4] A. Vaswani et al., "Attention is All You Need," in NIPS, 2017.
[5] H. Touvron et al., "Llama: Open and Efficient Foundation Language Models," arXiv, 2023.
[6] Groq Inc., "Real-time AI Inference Whitepaper," 2024.
[7] M. Douze et al., "The FAISS Library," arXiv, 2024.
[8] N. Liu et al., "Lost in the Middle: How Language Models Use Long Contexts," 2023.
[9] S. Robertson, "The Probabilistic Relevance Framework," 2009.
[10] Meta AI, "Llama-3 Technical Report," 2024.
[11] GNews API Documentation, 2025.
[12] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY