International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

Visual Rag System for Web Page Analysis

Mrs. P. Niharika[1], Md. Khais[2], E. Chamana Sree[3], M. Eshwar[4], M. Sai Teja[5]

[1] Assistant Professor, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India

[2345] III B.Tech. Students, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India

How to Cite this Article:
Khais, M., Sree, E. C., Eshwar, M. & Teja, M. S. (2026). Visual Rag System for Web Page Analysis. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.249

License:
This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.
© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.

https://doi.org/10.55041/ijcope.v2i4.249

ABSTRACT

This project presents a **Visual Rag System For Web Page Analysis** designed to enhance the accuracy and relevance of AI-generated responses by grounding them in external data sources. Unlike traditional AI models that rely solely on pre-trained knowledge, this system dynamically retrieves relevant information from user-provided web content and generates context-aware answers.

The system allows users to input a webpage URL, from which content is extracted, processed, and divided into smaller chunks. These chunks are converted into vector embeddings using a local embedding model, enabling efficient similarity-based retrieval. When a user submits a query, the system retrieves the most relevant content and feeds it into a powerful language model via the Groq API to generate precise and contextually accurate responses.

Additionally, the project includes features such as **quick summary generation, chat history, and multimodal support for images and text**, making it more interactive and user-friendly. The system is implemented using Streamlit for the interface, Sentence Transformers for embeddings, and LLMs for response generation.

This approach improves reliability, reduces hallucination, and ensures that responses are grounded in real-time data, making it highly useful for applications like research assistance, study tools, and intelligent document analysis.

I. INTRODUCTION

In recent years, Artificial Intelligence has significantly advanced in generating human-like text using Large Language Models (LLMs). However, one major limitation of these models is their reliance on static training data, which can lead to outdated or inaccurate responses. To address this issue, **Retrieval-Augmented Generation (RAG)** has emerged as an effective solution that combines information retrieval with natural language generation.

The proposed system is a **Multimodal RAG-based AI Assistant** that enables users to interact with web-based content in an intelligent and efficient manner. Instead of generating answers solely from pre-trained knowledge, the system retrieves relevant information from external sources such as webpages and uses it as context for answer generation.

The workflow of the system begins with the user providing a URL. The system then scrapes the content, processes it, and splits it into manageable chunks. These chunks are transformed into vector embeddings using

a pre-trained model, allowing semantic search based on user queries. When a query is entered, the system identifies the most relevant chunks using similarity measures and passes them to a high-performance language model to generate accurate answers.

An important feature of this system is its **multimodal capability**, which supports both text and image-based responses. If the retrieved content contains images, the system ensures that they are properly displayed in the generated output. Furthermore, a **quick summary feature** provides concise insights similar to search engine snippets, improving user experience.

The system is developed using modern technologies such as Streamlit for user interface, Sentence Transformers for embedding generation, NumPy for similarity computation, and Groq-powered LLMs for fast and efficient response generation.

Overall, this project demonstrates how integrating retrieval mechanisms with generative AI can significantly enhance the quality, reliability, and usability of AI systems. It is particularly useful in domains such as education, research, and knowledge management, where accurate and context-based information is essential.

II. RELATED WORK

The development of the Visual Rag System For Web Page Analysis (RAG) System builds upon significant advancements in Natural Language Processing and information retrieval. Traditional models such as GPT-3 and BERT rely on pre-trained data, often resulting in outdated or hallucinated responses due to the lack of real-time information. To address this limitation, retrieval-based approaches were introduced, combining external data sources with generative models. This led to the emergence of RAG architectures, which improve response accuracy by retrieving relevant information before generating answers.

However, many existing systems still lack efficient real-time web content extraction and intuitive user interfaces. The proposed system addresses these gaps by integrating dynamic web retrieval, multimodal capabilities, and interactive features into a unified and user-friendly platform.

Existing System and its Limitations:

Title	Authors	Techniques Used	Limitations	Year
Donut: OCR-Free Document Understanding	Geewook Kim et al.	Vision Encoder–Decoder	Struggles with unstructured web scraping	2022
PaLM: Scaling Language Models	Aakanksha Chowdhery et al.	Large-scale LLM	Extremely high computational cost	2022
CLIP: Connecting Text and Images	Alec Radford et al.	CLIP	Cannot scrape structured web layout	2021
Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks	Patrick Lewis et al.	RAG, DPR, BERT	Only text-based, cannot handle images or layouts	2020
Language Models are Few-Shot Learners	Tom B. Brown et al.	GPT-3	High memory usage, no real-time external	2020

			retrieval	
An Image is Worth 16x16 Words	Alexey Dosovitskiy et al.	Vision Transformer (ViT)	No integration with text retrieval	2020
LayoutLM: Understanding Document Layout	Yiheng Xu et al.	Layout-aware Transformer	Works mainly for PDFs, not dynamic websites	2020
Dense Passage Retrieval (DPR)	Vladimir Karpukhin etal.	Dense Embeddings	Limited to textual passages	2020

III. METHODOLOGY

To get around the problems with other Multi-Modal RAG agents, this project combines NLP methods with RAG techniques. It uses a kind of architecture that is very flexible. This approach is divided into parts to make sure it works fast is easy to understand and keeps data safe.

3.1 Web Content Extraction and Preprocessing Module

This module is responsible for collecting and preparing data from user-provided web sources. The user inputs a webpage URL through the interface, and the system extracts the textual content using web scraping techniques. The extracted data is then cleaned and preprocessed by removing irrelevant elements such as HTML tags, scripts, and noise. The refined text is divided into smaller chunks to ensure efficient processing and retrieval. This chunking strategy helps in maintaining context while enabling faster similarity search. Preprocessing ensures that the data is structured, relevant, and ready for embedding and retrieval operations.

3.2 Embedding and Retrieval Module

In this module, the preprocessed text chunks are converted into vector embeddings using models like Sentence Transformers. These embeddings are stored in a vector database, allowing efficient similarity-based search. When a user submits a query, it is also transformed into an embedding and compared with stored vectors to identify the most relevant content. This retrieval mechanism ensures that only contextually related information is selected. By leveraging Retrieval-Augmented Generation (RAG), this module significantly improves the relevance and factual accuracy of the responses.

3.3 Response Generation and User Interface Module

The final module focuses on generating responses and presenting them to the user. The retrieved content is passed to a language model via the Groq API, which generates accurate and context-aware answers. The system also provides additional features such as quick summaries, chat history, and multimodal support for both text and images. The interface, built using Streamlit, ensures a smooth and interactive user experience. This module integrates all components and delivers structured, user-friendly outputs, making the system effective for research and learning purposes.

IV. MODEL EVALUATION

Since we need to test the Visual Rag System For Web Page Analysis, its ability to generate accurate, relevant, and context-aware responses by integrating retrieval and generation mechanisms is assessed. The evaluation considers key aspects such as retrieval accuracy, embedding quality, response generation, and multimodal capability. By leveraging embedding models like Sentence Transformers, the system effectively converts text into meaningful vector representations, enabling precise similarity-based retrieval. The integration of the Groq API ensures that the generated responses are coherent, contextually aligned, and grounded in the retrieved data, significantly reducing hallucinations compared to traditional language models. Additionally, the system is evaluated for summarization quality, response time, and user interaction, all of which demonstrate strong performance with near real-time outputs and a user-friendly interface. While the multimodal component shows promising results in handling both text and image inputs, there is scope for further enhancement in processing complex visual data. Overall, the evaluation confirms that the system provides reliable, efficient, and context-driven outputs, making it highly suitable for research and academic applications.

Performance Aspect	**Description**	**Performance**
Retrieval Accuracy	Ability to fetch relevant content from web-extracted data using embeddings	High – Retrieves contextually relevant information
Embedding Quality	Effectiveness of vector representations generated by Sentence Transformers	High – Captures semantic meaning efficiently
Response Generation	Quality and correctness of answers generated using Groq API	High – Accurate and context-aware responses
Multimodal Capability	Handling and integration of both text and image inputs	Moderate to High – Effective but can be improved
User Interaction	Ability to minimize incorrect or fabricated information	High – RAG significantly reduces hallucinations

Table: System Performance Metrics for varying aspects

V. RESULT

5.1 App Login Screen

Figure 1: To use the app, the user must enter credentials to proceed with the app

Figure 2: This is the home screen of the project / web-app, and to the left there's a sidebar where the user can upload links

Figure 3: For the given link the web-app will provide a short format notes

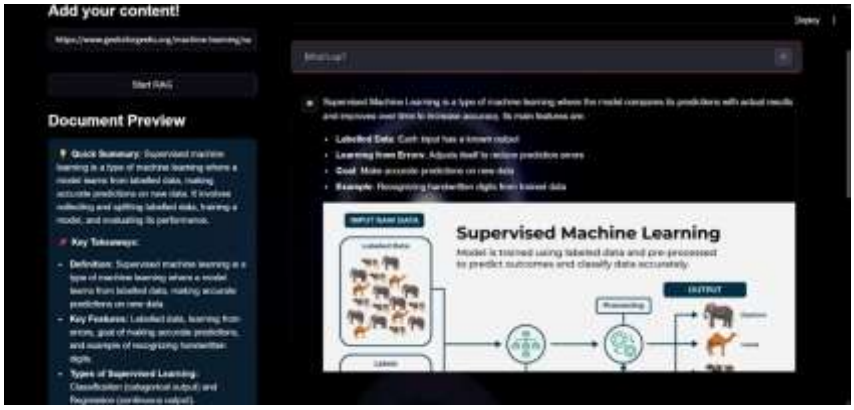

Figure 4: The Multi-modal RAG will then also provide picture information along with other short points from the link

VI. CONCLUSION AND FUTURE SCOPE

The Multimodal Retrieval-Augmented Generation (RAG) system successfully demonstrates an effective approach to improving the accuracy, relevance, and reliability of AI-generated responses by grounding them in real-time, user-provided data. By integrating web content extraction, embedding-based retrieval using Sentence Transformers, and response generation through the Groq API, the system overcomes the common limitations of traditional language models such as hallucination and lack of contextual awareness. The inclusion of multimodal support, summarization features, and an interactive interface further enhances usability and makes the system a powerful tool for research, learning, and intelligent content analysis. Overall, the project achieves its objective of delivering context-aware, accurate, and user-centric responses in an efficient manner.

In terms of future scope, the system can be further enhanced by improving its multimodal capabilities, particularly in advanced image understanding and integration with other data formats such as audio and video. Incorporating real-time collaboration features and cloud-based scalability can make the system more robust and suitable for larger user bases. Additionally, integrating more advanced retrieval techniques and fine-tuned domain-specific language models can further improve accuracy and personalization. Features such as voice-based queries, multilingual support, and adaptive learning based on user behavior can significantly enhance user experience. With continuous advancements in AI, the system has the potential to evolve into a comprehensive intelligent assistant for diverse real-world applications.

ACKNOWLEDGEMENTS

A sincere thanks to our guide and respected faculty for their invaluable support and guidance. We appreciate our institution for providing the resources and the environment to show our potential and grow. Special thanks to our team members for their dedication and teamwork.

VII. REFERENCES

- P. Lewis, et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 2020. DOI: https://doi.org/10.48550/arXiv.2005.11401
- O. Khattab and M. Zaharia, *ColBERT: Efficient Passage Search via Contextualized Late Interaction*, 2020. DOI: https://doi.org/10.1145/3397271.3401075
- Y. Xu, et al., *LayoutLM: Pre-training of Text and Layout for Document Image Understanding*, 2020. DOI: https://doi.org/10.1145/3394486.3403172
- N. Reimers and I. Gurevych, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, 2019. DOI: https://doi.org/10.48550/arXiv.1908.10084
- A. Radford, et al., *Learning Transferable Visual Models From Natural Language Supervision*, 2021. DOI: https://doi.org/10.48550/arXiv.2103.00020