DocEase: A Next-Gen RAG-Based System for Inclusive and Accessible Document Interaction

Akanksha Palve[1], Akshata Kamerkar[2], Kanchan Tayade[3],
Shubham Murtadak[4], Prof. Mrs. Priti Malkhede[5]
Artificial Intelligence and Data Science
Progressive Education Society's Modern College of Engineering

***Abstract*—The rapid growth of unstructured digital data presents significant challenges in retrieving relevant information efficiently, often leading to information overload and reduced user productivity. To address this, we propose DocEase, a next-generation system based on Retrieval-Augmented Generation (RAG), designed to enhance digital document interaction by offering AI-powered features such as User QA, content summarization, text-to-speech conversion, multilingual translation, and image generation. Built on advanced deep learning techniques, including natural language processing and dynamic AI-driven query generation, DocEase enables users to interact with various types of documents effortlessly. This paper examines the technological advancements underpinning information access and demonstrates how DocEase Vision streamlines data retrieval, significantly improving accessibility, engagement, and retrieval efficiency across diverse use cases.**

***Index Terms*—Content Summarization, Large Language Models, Deep Learning, Generative AI, Information Access, Multimedia Generation, Multilingual Translation, Natural Language Processing, Query Response, Retrieval-Augmented Generation, Text-to-Speech.**

I. Introduction

In the age of digitalization and a fast-moving world, it is essential for an individual to keep up with the rapid advancements in technology and knowledge. Traditionally, this is achieved by reading books, newspapers and research papers. However, in an age where information overload is prevalent it is challenging to read such a vast landscape of available data in a short period of time. This challenge leads to reduced efficiency and engagement as individuals struggle to shift through and comprehend the sheer volume of available information.

To overcome these issues of information overload and inefficiency, we introduced DOCEASE, a Retrieval-Augmented Generation (RAG) based system with multiple functionalities that aims to enhance user engagement, and overall efficiency. By integrating cutting-edge technologies in generative artificial intelligence and natural language processing, DOCEASE offers advanced features such as content summarization, real-time query responses, text-to-speech conversion, multilingual translation and image generation. These features empower users to quickly retrieve and digest relevant information, saving a huge amount of reading time.

At the core of DOCEASE lies a transformer-based architecture, supported by attention mechanisms, which enables the system to efficiently handle long-range dependencies and dynamically interact with user queries. Users can upload documents in various formats, such as PDFs, PowerPoint presentations, Word files, or images, and instantly apply features like summarization, translation, or text-to-speech conversion. The system's embedded data storage ensures that users can easily access previously uploaded documents and their associated summaries or translations in the future.

One of the unique aspects of DOCEASE is its ability to generate images, based on the retrieved data, making information not only easier to access but also more interactive. By offering an intuitive, all-in-one platform, DOCEASE enhances the user experience for a wide range of individuals, from researchers seeking fast and accurate information to non-technical users who require greater accessibility and engagement.

II. Literature Survey

The foundational work by Saad-Falcon [1] lays the groundwork for document processing in RAG-based systems, focusing on organizing and prioritizing information retrieval in digital documents. This research emphasizes the need for efficient techniques to handle unstructured data, which is directly relevant to DocEase's goal of improving digital document interaction through automated summarization and query response.

Gavilanes' work [2] explores how large language models (LLMs) can enhance traditional information retrieval (IR) methods. This study provides insights into using LLMs for more dynamic query generation and response, a concept that DocEase leverages to improve the accuracy and relevance of retrieved information.

The study by Lewis et al. [16] demonstrates how RAG models can significantly enhance information retrieval tasks by combining generative and retrieval components. DocEase utilizes this approach to facilitate accurate responses to user queries and improve the quality of content summarization.

Muludi's report [3] discusses the integration of retrieval mechanisms with generative models to augment information generation with relevant contextual data. This hybrid approach aligns with DocEase's architecture, which utilizes RAG to facilitate document interaction, making content summarization and multilingual translation more efficient.

Rajpurkar et al.'s work on the SQuAD dataset [17] serves as a benchmark for evaluating the accuracy of natural language understanding models. By employing benchmarks such as SQuAD, DocEase ensures that its query response capabilities

meet high standards of performance, leading to more reliable information retrieval.

The study on BART by M. Lewis et al. [19] presents a sequence-to-sequence architecture with denoising capabilities for generating and understanding text. DocEase leverages similar techniques to enhance its summarization and translation functionalities, allowing for more accurate and coherent output across different document types.

T. Labruna's study [4] further elaborates on incorporating IR techniques into LLM-based systems, aiming to optimize the retrieval process by allowing models to selectively utilize retrieved data. This technique is employed in DocEase to improve content relevance when generating responses to user queries, ensuring that the system retrieves and presents the most pertinent information from a document.

Lozano's research on a [5] showcases the benefits of open-source frameworks in enhancing RAG's capabilities. By adapting such frameworks, DocEase can incorporate state-of-the-art techniques to streamline the integration of LLMs and retrieval mechanisms, thus allowing for rapid development and continuous improvement of document processing functionalities.

The work by Raffel et al. on a unified text-to-text transformer [18] explores the limits of transfer learning, providing insights into the power of transforming all NLP tasks into a text-to-text format. This approach is reflected in DocEase's ability to handle various tasks such as summarization, translation, and query response using a unified architecture.

Dai's [6] provides performance metrics for IR tasks augmented by LLMs. This benchmark highlights the challenges in integrating LLMs with traditional IR techniques, which are addressed in DocEase by utilizing deep learning models to bridge the gap between retrieval efficiency and generative capabilities.

The seminal paper by Devlin et al. on BERT [21] discusses pre-training deep bidirectional transformers for language understanding, forming the basis for many modern NLP tasks. DocEase incorporates BERT-inspired attention mechanisms to improve the handling of long-range dependencies in document processing.

Liu's work on [7] introduces methods for fine-tuning models to generate instructions or content based on visual input. While DocEase focuses primarily on text-based document interaction, integrating visual content generation is a future direction inspired by Liu's findings to enhance multimedia generation features.

Vaswani's [8] provides the foundational transformer architecture that underpins many modern deep learning models, including those used in DocEase. The attention mechanisms described by Vaswani are critical for handling complex dependencies in textual data, enabling features like real-time query response and content summarization.

H. Zhang et al.'s work on multimodal transformers for image captioning and visual question answering [20] showcases how cross-modal attention mechanisms can improve multimedia content generation. This informs DocEase's approach to integrating visual data with textual analysis, further enhancing document interaction and making information more accessible.

Radford et al.'s study on learning visual models from natural language supervision [23] informs DocEase's future direction in integrating visual content generation. This approach will enable the system to combine text and image data more effectively, enhancing multimedia generation.

The mathematical exploration by Ji [9] provides a theoretical basis for understanding how attention mechanisms improve model performance. DocEase leverages these insights to optimize attention mechanisms for better handling of diverse document formats and complex queries.

Shazeer's [10] and Pope's [11] focus on scaling transformer models to handle large datasets. These studies are essential for DocEase's development, as the system must efficiently process substantial amounts of text and multimedia data across various document types.

Rombach's research [12] explores advanced techniques for image generation, providing a foundation for DocEase's ability to generate images from text-based information.

Yao's [13] discusses methods for enhancing the reasoning capabilities of LLMs, which is relevant to DocEase's goal of delivering accurate and context-aware responses.

Xiao's [14] highlights the importance of efficient streaming processing for real-time applications. DocEase can adopt such techniques to improve the performance of its text-to-speech and multilingual translation features.

The work by Cho et al. on RNN encoder-decoder models for statistical machine translation [25] explores early methods for sequence-to-sequence modeling, providing a historical perspective on how current models, including those used in DocEase, have evolved to handle translation tasks more efficiently.

Hendrycks et al.'s research [22] emphasizes the robustness of models in handling unexpected or challenging inputs. This aligns with DocEase's requirement to process diverse document types while maintaining accuracy and stability.

III. Methodology

The development of DOCEASE centers on creating an efficient Retrieval-Augmented Generation (RAG) system for processing documents and user queries. The system begins by ingesting documents in various formats, which undergo preprocessing steps such as tokenization, text normalization, and stemming to prepare the content for optimal use. Extracted text is then converted into vector embeddings using an embedding model that captures the semantic essence of each document. These embeddings are stored in a Vector Database, linking each vector to its corresponding text, allowing for efficient retrieval during query processing.

When a user submits a query, it is also preprocessed and transformed into a query embedding. Using similarity measures like cosine similarity, the system searches the Vector Database to retrieve the most relevant documents based on the proximity of the query vector to the stored document vectors. The retrieved documents, along with any associated metadata, are used to form a contextualized prompt by combining

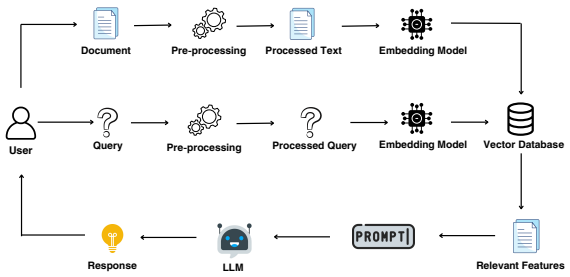

Fig. 1. System Architecture

the original query with key document excerpts. This prompt is passed to a language model (LLM), which generates a response by leveraging the augmented context, thus improving the accuracy and relevance of the output.

DOCEASE integrates this document retrieval with language model capabilities to produce high-quality, context-aware responses. The system allows for continuous improvements through the Inference Store, which tracks user interactions and feedback. This feedback loop is used to refine the system, updating embeddings and fine-tuning the LLM, ensuring adaptability to evolving document content and user preferences.

Additionally, the system offers multimodal features such as summarization, multilingual translation, text-to-speech conversion, and image generation. These features enhance accessibility and engagement by offering users varied ways to interact with content. Summarization condenses lengthy documents, while translation and text-to-speech ensure inclusivity for non-native speakers and those with visual impairments. The image generation function creates visual representations from text, enriching user comprehension.

The overall system is evaluated on performance metrics such as accuracy, response time, and user satisfaction, ensuring robust performance with real-world documents. Post-deployment, DOCEASE is continuously monitored, and user feedback guides updates, enabling the model and knowledge base to remain current with new data and emerging use cases. This iterative approach ensures that DOCEASE delivers efficient and personalized information retrieval while adapting to user needs over time.

IV. Conclusion

DOCEASE is a cutting-edge information retrieval system that integrates Retrieval-Augmented Generation (RAG) techniques with advanced natural language processing. By efficiently processing documents and user queries, DOCEASE enhances user experience through accurate, context-aware responses.

The system's multimodal features, including content summarization, multilingual translation, text-to-speech conversion, and image generation, promote engagement and accessibility for diverse users, including non-native speakers and those with visual impairments.

With a continuous feedback loop for ongoing improvement, DOCEASE adapts to evolving user needs and information trends. Overall, DOCEASE streamlines information retrieval, transforms user interaction, and empowers individuals to navigate the complexities of the digital landscape effectively.

References

[1] J. Saad-Falcon, *PDFTriage*, Stanford University, 2023.
[2] J. Gavilanes, *Use of LLM for Methods of IR*, University of Twente, 2023.
[3] Muludi, *Retrieval Augmented Generation (RAG) Report*, Darmajaya Informatics and Business Institute, 2024.
[4] T. Labruna, *Teaching LLMs to Utilize Information Retrieval*, University of Bozen-Bolzano, 2024.
[5] A. Lozano, *Open-source RAG LLM System*, Stanford University, 2024.
[6] S. Dai, *Information Retrieval Benchmark with LLM-generated Documents Integration*, Gaoling School of AI, 2024.
[7] H. Liu, *Visual Instruction Tuning*, University of Wisconsin-Madison, 2023.
[8] A. Vaswani, *Attention is All You Need*, Google Brain, 2023.
[9] S. Ji, *A Mathematical View of Attention Models in Deep Learning*, 2024.
[10] N. Shazeer, *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*, 2022.
[11] R. Pope, *Efficiently Scaling Transformer Inference*, 2022. [Online]. Available: [Insert URL].
[12] V. Madhusudhana Reddy, *Speech-to-Text and Text-to-Speech*, 2023. [Online]. Available: [Insert URL].
[13] G. Xiao, *Efficient Streaming Language Models with Attention Sniks*, 2023. [Online]. Available: [Insert URL].
[14] R. Rombach, *High-Resolution Image Synthesis with Latent Diffusion Models*, 2022. [Online]. Available: [Insert URL].
[15] S. Yao, *REACT: Synergizing Reasoning and Acting in Language Models*, 2023. [Online]. Available: [Insert URL]. Transactions on Multimedia, 2021.
[16] B. Lewis *et al.*, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 2020.
[17] P. Rajpurkar *et al.*, *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, 2016.
[18] C. Raffel *et al.*, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, 2020.
[19] M. Lewis *et al.*, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, 2019.
[20] H. Zhang *et al.*, *Multimodal Transformers for Image Captioning and Visual Question Answering*, 2020.
[21] J. Devlin *et al.*, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019.
[22] D. Hendrycks *et al.*, *Natural Adversarial Examples*, 2021.
[23] A. Radford *et al.*, *Learning Transferable Visual Models from Natural Language Supervision*, 2021.
[24] L. Dong *et al.*, *Unified Language Model Pre-training for Natural Language Understanding and Generation*, 2019.
[25] K. Cho *et al.*, *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*, 2014.