i-TIEC

INTERNATIONAL TINKER INNOVATION & ENTREPRENEURSHIP CHALLENGE

UNIVERSITI TEKNOLOGI MARA

E-PROCEEDINGS

INTERNATIONAL TINKER INNOVATION & ENTREPRENEURSHIP CHALLENGE (i-TIEC 2025)

"Fostering a Culture of Innovation and Entrepreneurial Excellence"

23 January 2025
PTDI, UiTM Cawangan Johor Kampus Pasir Gudang

e ISBN 978-967-0033-34-1

9 789670 033341

ORGANIZED BY:

Electrical Engineering Studies, College of Engineering
Universiti Teknologi MARA (UiTM) Cawangan Johor
Kampus Pasir Gudang
https://tiec-uitmpg.wixsite.com/tiec

E-PROCEEDINGS
of International Tinker Innovation & Entrepreneurship Challenge (i-TIEC 2025)

"Fostering a Culture of Innovation and Entrepreneurial Excellence"

23rd JANUARY 2025
PTDI, UiTM Cawangan Johor, Kampus Pasir Gudang

Organized by

Electrical Engineering Studies, College of Engineering,

Universiti Teknologi MARA (UiTM) Cawangan Johor, Kampus Pasir Gudang.

https://tiec-uitmpg.wixsite.com/tiec

Editors

Aznilinda Zainuddin

Maisarah Noorezam

Copyright © 2025 Universiti Teknologi MARA Cawangan Johor, Kampus Pasir Gudang, Jalan Purnama, Bandar Seri Alam, 81750 Masai Johor.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, whether electronic, mechanical, or otherwise, without prior written consent from the Undergraduate Coordinator, Electrical Engineering Studies, College of Engineering, Universiti Teknologi MARA (UiTM) Cawangan Johor, Kampus Pasir Gudang.

e ISBN: 978-967-0033-34-1

The author and publisher assume no responsibility for errors or omissions in this e- proceeding book or for any outcomes related to the use of the information contained herein.

The extended abstracts featured in this e-proceeding book have not undergone peer review or verification by i-TIEC 2025. The authors bear full responsibility for the content of their abstracts, guaranteeing that they are original, unpublished, and not concurrently submitted elsewhere. The opinions presented in the abstracts reflect those of the authors and do not necessarily align with the views of the editor.

Published in Malaysia by
Universiti Teknologi MARA (UiTM) Cawangan Johor
Kampus Pasir Gudang, 81750 Masai

A-ST075: RESEARCH PAPER AI ASSISTANT USING RETRIEVAL AUGMENTED GENERATION AND MULTIMODAL LLM

Ammar Helmey Iskandar Mohd Rayhan, Ezzatul Akmal Kamaru-Zaman, Muhammad Fareed ‘Aidil Rozaidi, and Azlin Ahmad
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA
Shah Alam Campus, Selangor, Malaysia

Corresponding author: Ammar Helmey Iskandar Mohd Rayhan,
ezzatulakmal@uitm.edu.my

ABSTRACT

Researchers often face challenges such as the time-consuming process of document review and the complexity of comprehending technical concepts during journal writing. Emerging large language models, such as ChatGPT and Poe, have proven to be valuable tools for streamlining the writing process. However, these models often require payment for extensive usage or document uploads, posing accessibility barriers. This innovation is motivated by the need to simplify research workflows and enhance the understanding of research documents while ensuring these capabilities are accessible to university users without incurring additional costs. The Research Paper AI Assistant is an innovative system designed to revolutionize how researchers interact with and synthesize information from academic papers. Leveraging advanced technologies such as the Large Language Model (LLM) - Gemma 9b, Retrieval-Augmented Generation (RAG), and the Chroma Embedding Multimodal Model, this innovative system features an intuitive chatbot interface. It enables users to upload multiple PDFs and conduct conversational interactions to extract and synthesize complex information. The scope of the system is specifically confined to the context of the uploaded documents, ensuring focused and relevant responses. The system features automatic diagram analysis for detailed insights and a dedicated "Future Work" tab that instantly identifies key problems and future research directions outlined by authors subsequently, promotes efficient knowledge acquisition. Its unique ability to generate notes, explain complex terms, and provide instant retrieval of relevant content makes it invaluable for young academics such as students and professionals. The system’s socio-economic impact extends beyond accelerating research breakthroughs and fostering innovation in academic insights. It also significantly benefits students by simplifying the process of understanding complex academic materials, aiding in literature reviews, and enhancing their research and learning experiences. By providing an accessible and interactive tool, students can efficiently extract relevant information from academic papers, develop critical thinking skills, and improve their academic writing. Commercially, the system holds immense potential for integration into academic institutions, research labs, and publishing platforms, supporting applications ranging from education to corporate R&D.

Keywords: Large Language Model, Retrieval-Augmented Generation, Diagram Analysis, Future Work Extraction, Knowledge Accessibility.

1. Product Description

The Research Paper AI Assistant revolutionizes how researchers engage with academic papers. Powered by advanced technologies such as the Large Language Model (LLM) - Gemma 9b, the Multimodal Model - InternVL2_5 4B, the Embedding Model - Nomic-Embed-Text-v1.5, Chroma for vector storage, the system offers unparalleled functionality. Its intuitive chatbot interface enables users to upload multiple PDFs and engage in contextual, document-specific interactions. The scope is strictly confined to the context of the uploaded documents, ensuring accurate, hallucination-free responses. The system extracts text and images from PDFs, using LLM models to analyze diagrams and generate textual descriptions, which are displayed in a “Diagram Analysis” tab. Additionally, it identifies key issues, limitations, and future directions from research papers, systematically presenting them in the “Future Work” tab. A central vector database stores embeddings of the research content, enabling instant retrieval of relevant information based on user queries. By simplifying academic research, the system saves time, improves knowledge acquisition, and fosters innovation. Designed for scalability and supporting multiple documents, it suits academic institutions, research organizations, and corporate R&D teams. Its ability to explain terms, generate notes, and enhance access to technical content ensures significant value across domains, accelerating knowledge sharing with precision and affordability.

2. System Architecture

The Research Paper AI Assistant utilizes a Retrieval-Augmented Generation (RAG) architecture, combining the reliability of retrieval-based systems with the generative capabilities of Gemma 9b, an advanced LLM model. This design, as shown in **Figure 1** ensures that user queries are answered by producing output strictly based on the document. Research paper content is embedded into high-dimensional vectors and stored in a vector database, Chroma for quick retrieval. These vectors are then processed by the LLM, which generates detailed, contextually relevant responses.

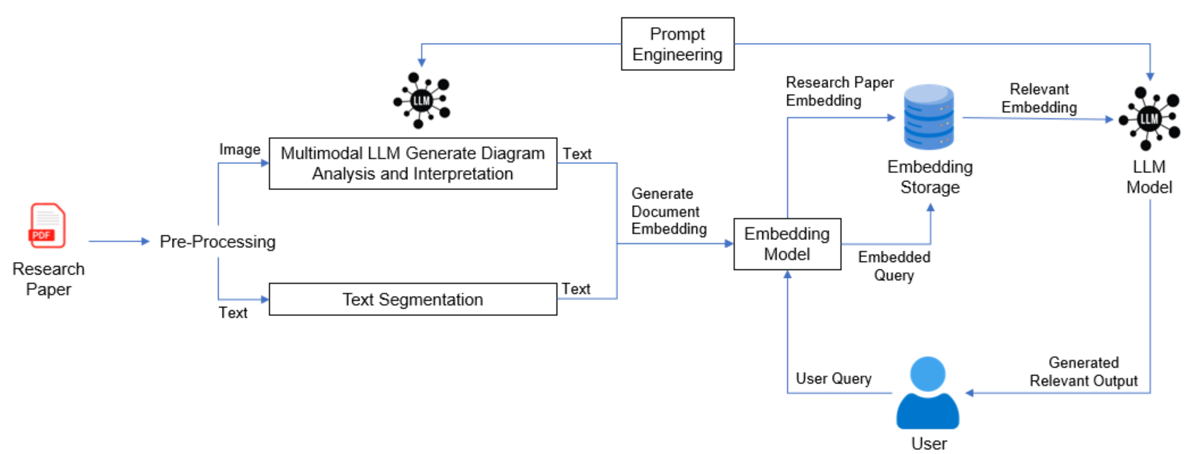


Figure 1. Diagram of Research Paper Processing and Query-Based Information Retrieval

Input and Pre-processing

The workflow initiates with researchers uploading one or more research paper PDFs into the system. This phase begins with an intelligent pre-processing mechanism that extracts both textual and visual elements from the documents. For visual data, a Multimodal Large Language Model - InternVL2_5 4B (MLLM) is employed to generate descriptive analyses, offering detailed insights into graphical components such as charts, graphs, and diagrams. Simultaneously, the textual content undergoes segmentation, where it is divided into smaller, manageable chunks using a chunking technique. This step ensures that large documents can be handled efficiently while maintaining their logical structure and contextual coherence. The pre-processed outputs form the foundation for downstream processing, ensuring that both text and images are accessible for further embedding and retrieval operations.

Multimodal LLM and Prompt Engineering

At this stage, the system employs a Multimodal Large Language Model (MLLM) to analyze and interpret visual data extracted from research papers, such as diagrams and graphs. By integrating prompt engineering techniques, the MLLM is guided to generate precise and detailed textual descriptions of these visual components. **Table 1** outlines the prompt templates used for specific tasks, showcasing how tailored prompts enhance the MLLM's performance. Prompt engineering ensures that the MLLM provides outputs tailored to the context of research papers, improving the accuracy and relevance of the generated insights. These visual descriptions complement the segmented textual content, enabling a comprehensive representation of the document. This multimodal approach ensures that the system captures and processes both textual and visual data seamlessly, providing researchers with a full-spectrum understanding of the content.

Embedding Storage

After preprocessing, the segmented text and image descriptions are transformed into high-dimensional vector representations via an embedding model. These embeddings serve as mathematical abstractions that encapsulate the semantic meaning and contextual nuances of the content. The resulting embeddings, along with associated metadata, are stored in a vector database. This database functions as the primary knowledge repository, facilitating the efficient and precise retrieval of information. The embedding storage phase is critical to ensuring the system's capacity to map user queries to the most relevant segments of the research papers effectively.

User Query Handling

Following the embedding phase, researchers interact with the system through an advanced chatbot assistant. This component is designed to support conversational queries, enabling users to pose specific questions, request clarifications on complex terms, or navigate through specific sections of the research papers interactively. Upon receiving a query, the system processes it by generating a corresponding vector representation using the same embedding model utilized during the storage phase. This alignment ensures the seamless retrieval of relevant content from the vector database, thereby enabling a responsive and contextually aware interaction between the user and the system.

Contextual Retrieval and Generation

In the final stage, the system employs Retrieval-Augmented Generation (RAG) to combine retrieved embeddings from the vector database with the advanced capabilities of the LLM for contextual analysis and response generation. This approach ensures that responses are accurate, relevant, and grounded in the uploaded documents. Leveraging RAG, the system delivers outputs tailored to the user's specific queries, including detailed explanations of scientific terms, comprehensive summaries of sections, or synthesized notes capturing key findings.

The interactive design allows users to refine their queries or request additional clarifications, fostering a deeper engagement with the research papers. This phase is integral to enabling researchers to synthesize complex information, identify critical insights, and interact with academic content dynamically and intuitively. By incorporating RAG, the system enhances accuracy, improves contextual understanding, and supports a seamless user experience. **Table 2** provides a comparison of the innovation system utilizing RAG architecture with methods that do not incorporate RAG, highlighting the advantages in accuracy, contextual relevance, and user engagement offered by RAG-based approaches. Meanwhile, **Table 3** defined the sample input and output from the system able to be produced within 4 seconds.

3. Novelty and Uniqueness

Research Paper AI Assistant simplifies research with advanced AI technologies. Unlike traditional tools, it combines Retrieval-Augmented Generation (RAG) with an interactive chatbot for a seamless, user-friendly experience. Its ability to analyze diagrams and provide detailed descriptions in a "Diagram Analysis" tab ensures a comprehensive understanding of visual content. The automatic extraction of problem statements and "Future Work" sections, displayed in an organized tab, enhances its utility for identifying research gaps and opportunities. A centralized vector database enables precise, instant retrieval of information, transforming how researchers interact with data. The system explains complex terms, generates notes, and synthesizes multiple documents, fostering efficient and accessible knowledge acquisition. Figures and tables highlight the system's features, including its user-friendly interface, advanced diagram analysis, and automatic extraction of key insights. Figures 2–10 illustrate the system's innovative document upload, diagram extraction, and contextual output generation. The "Future Work" feature in **Figure 10** automatically identifies and extracts limitations and suggested research directions. **Table 4** summarizes its performance, emphasizing efficiency and effectiveness in supporting researchers. By combining multimodal analysis, prompt engineering, and automatic insight generation, the Research Paper AI Assistant enhances the speed and accuracy of engaging with complex documents, marking a major advancement in academic tools.

Table 1. Prompt template that has been used for specific tasks.

Diagram Analysis and Interpretation	**User Query Answer Prompt**
You are an advanced AI Assistant with expertise in analyzing and interpreting diagrams. Your task is to interpret the given diagram and provide insights based on the type of diagram. Follow these guidelines and ensure the response is clear, organized, and coherent. *1. For Flowcharts:* ● Summary: Provide a brief overview of the flowchart, describing its overall purpose and structure. ● Key Findings: Highlight the most important insights or steps in the flowchart, focusing on any critical decision points or processes. ● Image Components: Describe the components of the flowchart, including nodes, arrows, and any labels or symbols used. Explain how the elements are connected and their significance in the process. *2. For Charts (e.g., Bar chart, Line graph, Pie chart, etc.):* ● Act as a data analyst with expertise in analyzing graphs and charts. ● Trend Analysis: Identify any trends or patterns depicted in the chart. Discuss any increases, decreases, or fluctuations, and what they may signify. ● Key Findings: Provide insights into the major takeaways from the data presented in the chart. Be specific about significant values, comparisons, or any anomalies. ● Summary: Offer a concise summary of the chart, its axes, and any notable data points. ● Conclusion: Provide a concluding statement that synthesizes the key findings and what they imply about the data or subject being represented. *3. For Tables:* ● Table Analysis: Carefully examine the table and identify the main categories, columns, and rows. ● Summary: Summarize the type of data presented in the table and the key patterns or information it conveys. ● Key Information: Extract the most important data points or trends from the table. Discuss what conclusions or insights can be drawn based on the data. ● If relevant, mention any correlations or notable differences in the table. In all cases, ensure that your response is precise, coherent, and free from assumptions.	Act as AI Assistant that is expert in reviewing and understanding the content of research papers in PDF format. {context} --- Produce answer for the question based on the above context: {question}

Table 2. Comparison utilizing Retrieval Augmented Generation (RAG) Architecture with methods that do not use RAG.

	Using Retrieval Augmented Generation Architecture	**Without using Retrieval Augmented Generation Architecture**
Accuracy of Responses	Provides accurate answers by grounding responses in retrieved documents that be able to minimize hallucination.	May generate incorrect or hallucinated information, especially when queried on topics outside the pre-trained data.
User query example if the document does not mentioned MapReduce	Query : What is the MapReduce There is no mentioned MapReduce in the documents provided . Therefore i cant answer the question .	Query : What is the MapReduce MapReduce is a programming model and processing technique developed by Google for processing and generating large datasets in a distributed and parallel computing environment.

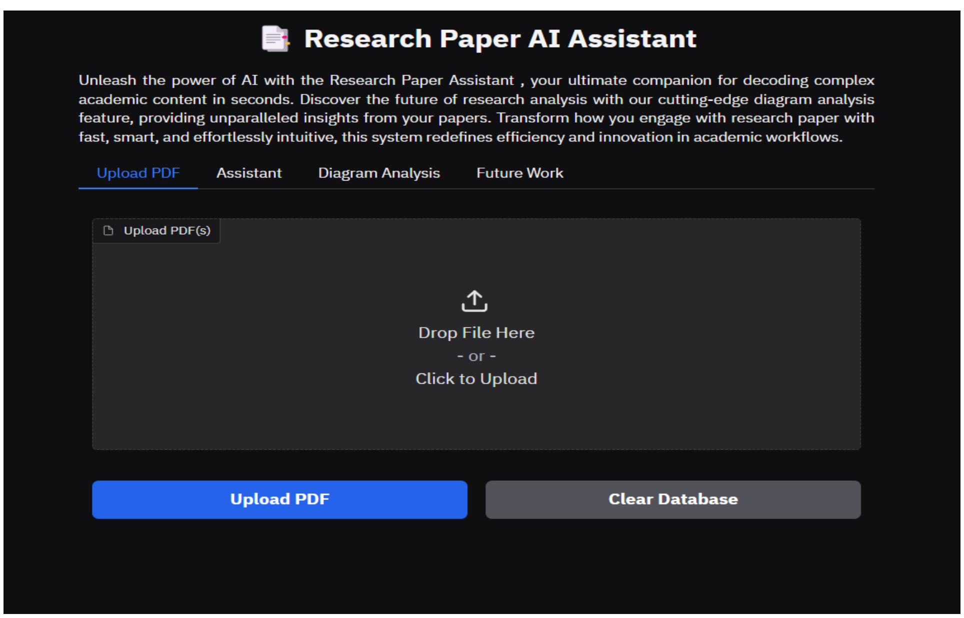


Figure 2. User Interface of the Research Paper AI Assistant for document upload and analysis

Table 3. Sample input and output from the system that can be produced within 4 seconds.

Input	Output
Please extract the following information from the given research paper: 1. **PROBLEM STATEMENT:** Summarize the main problem or issue being addressed in the research paper. Ensure the explanation is clear and concise. 2. **OBJECTIVES:** What are the specific objectives or goals of the research study? List or summarize the key aims of the study. Start output with "to." 3. **TECHNIQUE:** Describe the techniques, methods, or approaches used in the research. Focus on the key methodologies employed. 4. **FINDINGS:** What are the major findings or results of the research? Provide a summary of the key outcomes. 5. **REMARKS:** Are there any important remarks, conclusions, or recommendations provided by the authors? Summarize any final reflections or notes.	Here's a breakdown of the research paper based on the provided text: **PROBLEM STATEMENT:** The research addresses the limitations of traditional keyword search in information retrieval systems, particularly when dealing with vast academic paper datasets. **OBJECTIVES:** To develop a Retrieval-Augmented Generation (RAG) system for efficient information extraction from academic papers. To investigate effective search strategies within the RAG system. To evaluate the performance of different Large Language Models (LLMs) in the context of the RAG system. **TECHNIQUE:** Keyword Search: Initially, the researchers used keyword-based searches to evaluate the document retrieval component of the RAG system. Semantic Search: The system incorporates semantic search techniques to improve upon keyword-based limitations. Large Language Models (LLMs): The research utilizes various LLMs, including "gpt-3.5-turbo-0125," to process retrieved information and generate responses.

	Two-Layer Search System: A two-layer approach is implemented to address the token limitations of LLMs. The first layer identifies relevant papers, and the second layer extracts the most pertinent text excerpts from those papers for LLM processing. Abstract and PDF Embedding Analysis: The researchers analyze the similarity between paper abstracts and full text using cosine similarity scores and visualization techniques to assess the representativeness of abstracts. **FINDINGS:** Keyword search alone is insufficient for efficient information retrieval in academic paper dataset. Abstracts often provide a good representation of the full paper's content, but there are cases where significant vocabulary differences exist. The two-layer search system effectively addresses the token limitations of LLMs. **REMARKS:** The research highlights the need for advanced search strategies beyond keyword-based approaches in academic information retrieval. The findings suggest that combining semantic search with LLMs and a careful selection of relevant text excerpts can significantly improve the efficiency and accuracy of information extraction. **Sources:** ['Building a Retrieval Augmented Generation RAG System for Academic Papers.pdf, page 1', 'Building a Retrieval Augmented Generation RAG System for Academic Papers.pdf, page 4', 'Building a Retrieval Augmented Generation RAGSystem for

	Academic Papers.pdf, page 1', 'Building a Retrieval Augmented Generation RAG System for Academic Papers.pdf, page 3', 'Building a Retrieval Augmented Generation RAG System for Academic Papers.pdf, page 3 information extraction. **Sources:** ['Building a Retrieval Augmented Generation RAG System for Academic Papers.pdf, page 1', 'Building a Retrieval Augmented Generation RAG System for Academic Papers.pdf, page 4', 'Building a Retrieval Augmented Generation RAGSystem for Academic Papers.pdf, page 1', 'Building a Retrieval Augmented Generation RAG System for Academic Papers.pdf, page 3', 'Building a Retrieval Augmented Generation RAG System for Academic Papers.pdf, page 3

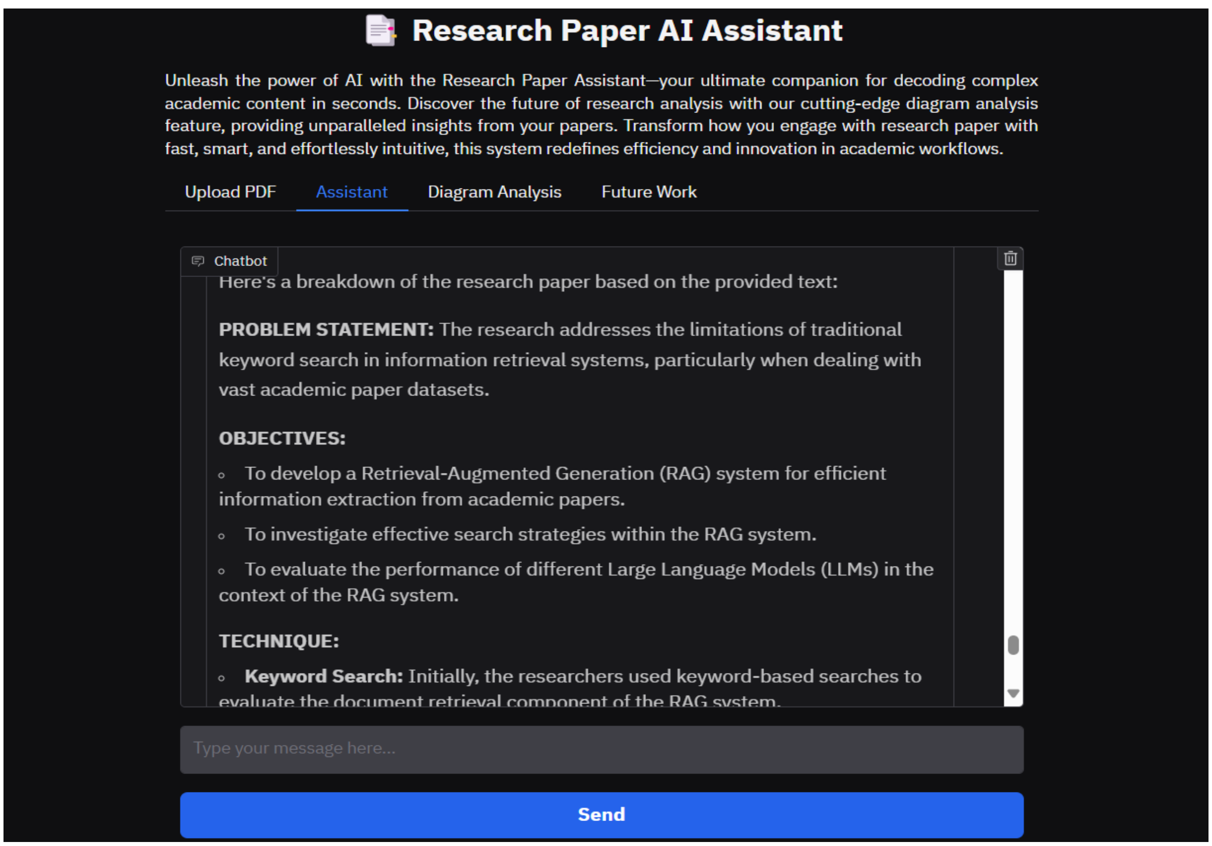


Figure 3. shows a user-friendly interface for the user to interact with the AI Assistant

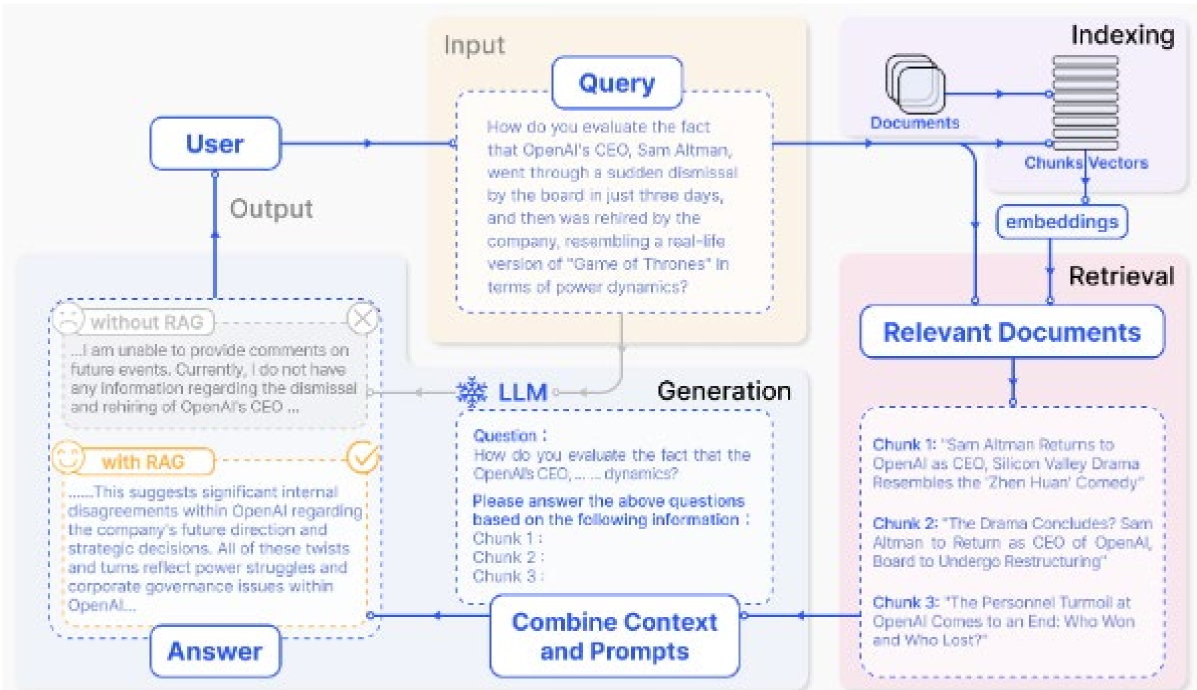


Figure 4. Shows sample of diagram extracted from the document.

Summary

The diagram is a flowchart that illustrates the process of evaluating a query using a large language model (LLM) with and without the use of Retrieval-Augmented Generation (RAG). The flowchart outlines the steps from input to output, highlighting the differences in the evaluation process with and without RAG.

Key Findings

1. **Input Query:** The query is about the sudden dismissal and subsequent rehiring of OpenAI's CEO, Sam Altman, and its implications on the company's power dynamics.
2. **Process with RAG:**
   - **Indexing:** Documents are indexed into chunks and vectors.
   - **Retrieval:** Relevant documents are retrieved based on the query.
   - **LLM Generation:** The LLM generates a response based on the retrieved context.
   - **Output:** The response is provided to the user.
3. **Process without RAG:**
   - **Indexing:** Similar to RAG, but without retrieval.
   - **Retrieval:** No relevant documents are retrieved.
   - **LLM Generation:** The LLM generates a response based on the input query.
   - **Output:** The response is provided to the user.

Image Components

- **Input:** The query is displayed at the top left.
- **Query:** The specific question being evaluated.
- **Documents:** Represented as chunks and vectors.
- **Relevant Documents:** Highlighted as retrieved from the indexing process.
- **LLM Generation:** The process where the LLM generates a response.
- **Output:** The final response provided to the user.

The flowchart outlines the process of evaluating a query using a large language model with and without the use of Retrieval-Augmented Generation (RAG). The process involves indexing documents, retrieving relevant documents, and using the LLM to generate a response. The key difference is that with RAG, the LLM uses retrieved context to generate a more informed response, whereas without RAG, the LLM generates a response based solely on the input query.

Conclusion

The flowchart effectively demonstrates the advantages of using RAG in enhancing the quality of responses generated by a large language model. By incorporating retrieved context, the model can provide more informed and contextually relevant answers, improving the overall evaluation process.

Figure 5. illustrate output produced from the system for the diagram in **Figure 4**.

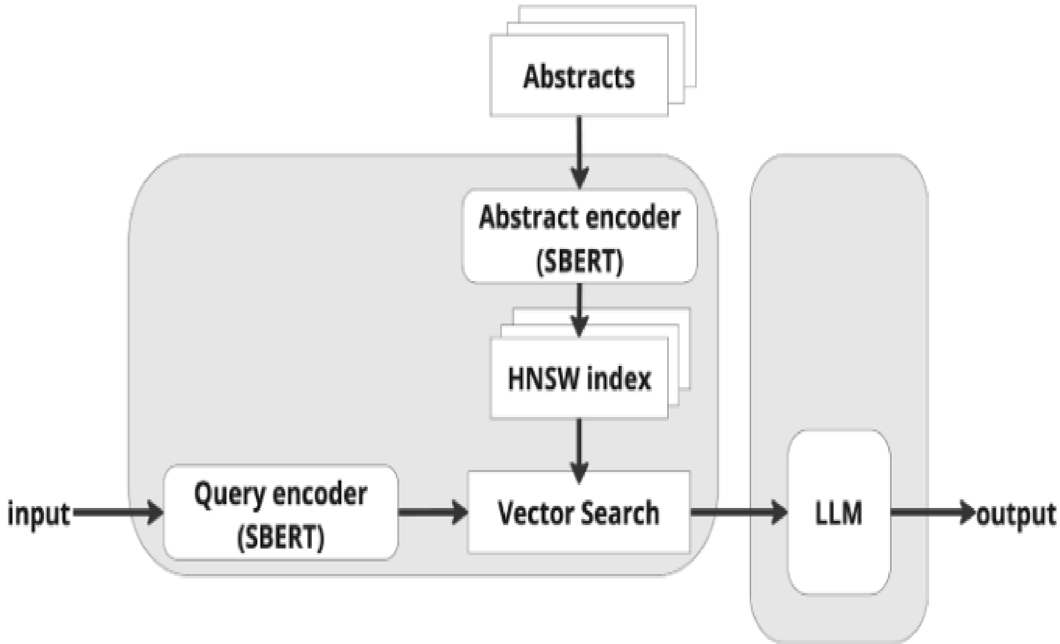
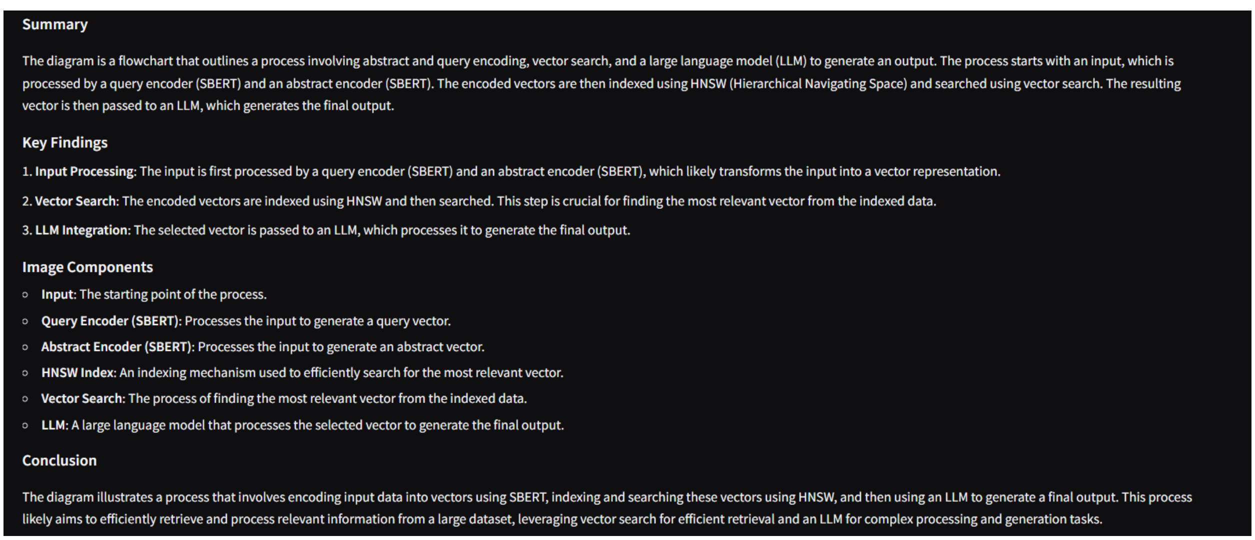


Figure 6. Shows sample of diagram extracted from the document.



Summary

The diagram is a flowchart that outlines a process involving abstract and query encoding, vector search, and a large language model (LLM) to generate an output. The process starts with an input, which is processed by a query encoder (SBERT) and an abstract encoder (SBERT). The encoded vectors are then indexed using HNSW (Hierarchical Navigating Space) and searched using vector search. The resulting vector is then passed to an LLM, which generates the final output.

Key Findings

1. **Input Processing**: The input is first processed by a query encoder (SBERT) and an abstract encoder (SBERT), which likely transforms the input into a vector representation.
2. **Vector Search**: The encoded vectors are indexed using HNSW and then searched. This step is crucial for finding the most relevant vector from the indexed data.
3. **LLM Integration**: The selected vector is passed to an LLM, which processes it to generate the final output.

Image Components

- **Input**: The starting point of the process.
- **Query Encoder (SBERT)**: Processes the input to generate a query vector.
- **Abstract Encoder (SBERT)**: Processes the input to generate an abstract vector.
- **HNSW Index**: An indexing mechanism used to efficiently search for the most relevant vector.
- **Vector Search**: The process of finding the most relevant vector from the indexed data.
- **LLM**: A large language model that processes the selected vector to generate the final output.

Conclusion

The diagram illustrates a process that involves encoding input data into vectors using SBERT, indexing and searching these vectors using HNSW, and then using an LLM to generate a final output. This process likely aims to efficiently retrieve and process relevant information from a large dataset, leveraging vector search for efficient retrieval and an LLM for complex processing and generation tasks.

Figure 7. illustrate output produced from the system of the diagram in **Figure 6**.

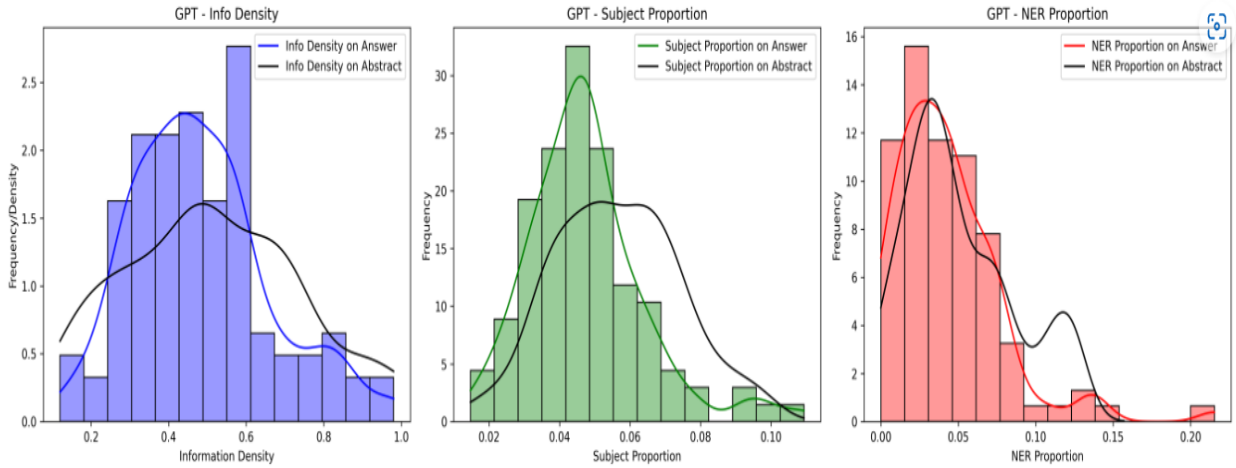
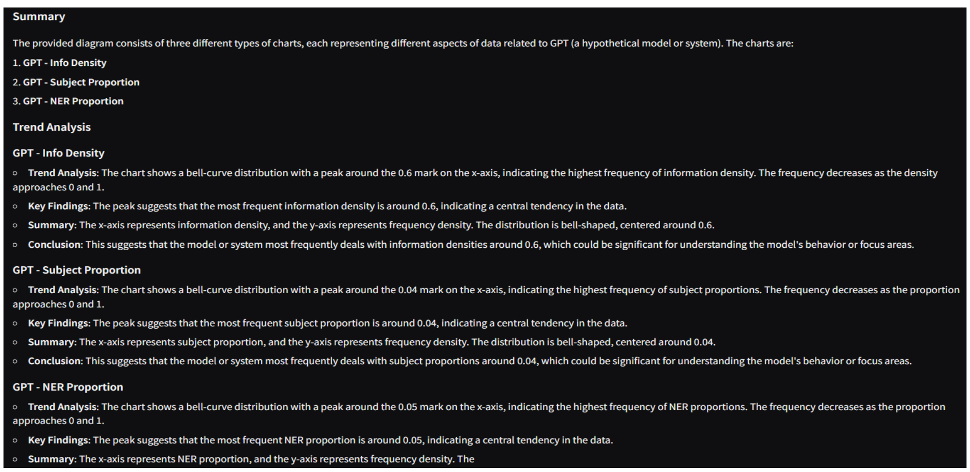


Figure 8. Shows sample of diagram extracted from the document.



Summary

The provided diagram consists of three different types of charts, each representing different aspects of data related to GPT (a hypothetical model or system). The charts are:

1. **GPT - Info Density**
2. **GPT - Subject Proportion**
3. **GPT - NER Proportion**

Trend Analysis

GPT - Info Density

- **Trend Analysis:** The chart shows a bell-curve distribution with a peak around the 0.6 mark on the x-axis, indicating the highest frequency of information density. The frequency decreases as the density approaches 0 and 1.
- **Key Findings:** The peak suggests that the most frequent information density is around 0.6, indicating a central tendency in the data.
- **Summary:** The x-axis represents information density, and the y-axis represents frequency density. The distribution is bell-shaped, centered around 0.6.
- **Conclusion:** This suggests that the model or system most frequently deals with information densities around 0.6, which could be significant for understanding the model's behavior or focus areas.

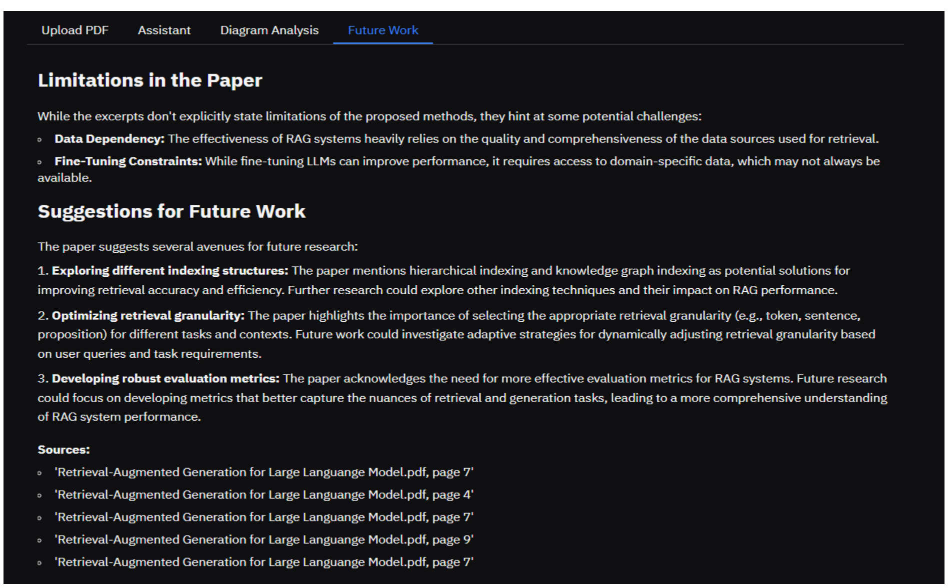
GPT - Subject Proportion

- **Trend Analysis:** The chart shows a bell-curve distribution with a peak around the 0.04 mark on the x-axis, indicating the highest frequency of subject proportions. The frequency decreases as the proportion approaches 0 and 1.
- **Key Findings:** The peak suggests that the most frequent subject proportion is around 0.04, indicating a central tendency in the data.
- **Summary:** The x-axis represents subject proportion, and the y-axis represents frequency density. The distribution is bell-shaped, centered around 0.04.
- **Conclusion:** This suggests that the model or system most frequently deals with subject proportions around 0.04, which could be significant for understanding the model's behavior or focus areas.

GPT - NER Proportion

- **Trend Analysis:** The chart shows a bell-curve distribution with a peak around the 0.05 mark on the x-axis, indicating the highest frequency of NER proportions. The frequency decreases as the proportion approaches 0 and 1.
- **Key Findings:** The peak suggests that the most frequent NER proportion is around 0.05, indicating a central tendency in the data.
- **Summary:** The x-axis represents NER proportion, and the y-axis represents frequency density. The

Figure 9. illustrate output produced from the system of the diagram in **Figure 8**.



Upload PDF | Assistant | Diagram Analysis | Future Work

Limitations in the Paper

While the excerpts don't explicitly state limitations of the proposed methods, they hint at some potential challenges:

- **Data Dependency:** The effectiveness of RAG systems heavily relies on the quality and comprehensiveness of the data sources used for retrieval.
- **Fine-Tuning Constraints:** While fine-tuning LLMs can improve performance, it requires access to domain-specific data, which may not always be available.

Suggestions for Future Work

The paper suggests several avenues for future research:

1. **Exploring different indexing structures:** The paper mentions hierarchical indexing and knowledge graph indexing as potential solutions for improving retrieval accuracy and efficiency. Further research could explore other indexing techniques and their impact on RAG performance.

2. **Optimizing retrieval granularity:** The paper highlights the importance of selecting the appropriate retrieval granularity (e.g., token, sentence, proposition) for different tasks and contexts. Future work could investigate adaptive strategies for dynamically adjusting retrieval granularity based on user queries and task requirements.

3. **Developing robust evaluation metrics:** The paper acknowledges the need for more effective evaluation metrics for RAG systems. Future research could focus on developing metrics that better capture the nuances of retrieval and generation tasks, leading to a more comprehensive understanding of RAG system performance.

Sources:

- 'Retrieval-Augmented Generation for Large Languange Model.pdf, page 7'
- 'Retrieval-Augmented Generation for Large Languange Model.pdf, page 4'
- 'Retrieval-Augmented Generation for Large Languange Model.pdf, page 7'
- 'Retrieval-Augmented Generation for Large Languange Model.pdf, page 9'
- 'Retrieval-Augmented Generation for Large Languange Model.pdf, page 7'

Figure 10. illustrates the "Future Work" feature that automatically extracts limitations and suggested research directions from uploaded documents.

Table 4. Summary of the system performance

	Number of Document per review	Time spent to embed textual content	Time spent analyzing and extracting information from the diagram	Time spent for AI Assistant to answer user query
AI Research Assistant	Up to 10 documents	62 seconds for 10 documents (155 pages)	5 minutes for 50 Diagram	within 3 seconds

4. **Benefit to mankind**

Research Paper AI Assistant revolutionizes how researchers, students, and professionals access and comprehend academic information. Designed to simplify the often tedious and time-consuming process of reviewing and synthesizing research papers, it empowers users by providing conversational interaction with document content. Its unique features, such as diagram analysis and automatic extraction of key problems and "Future Work," enable quicker insights into complex topics and foster innovation. By enhancing access to technical knowledge and reducing barriers to understanding, the system promotes inclusive education and global collaboration. It has a profound socio-economic impact by accelerating research and innovation, driving advancements across industries, and democratizing access to high-quality academic insights. Its scalability ensures accessibility for individuals, academic institutions, and corporate R&D teams worldwide, making it a transformative tool for fostering knowledge-driven growth.

5. **Innovation and Entrepreneurial Impact**

Research Paper AI Assistant fosters a culture of innovation and entrepreneurship by bridging the gap between academic research and practical application. By streamlining the understanding of complex research papers, the system empowers researchers and entrepreneurs to identify opportunities for innovation more effectively. Its ability to extract future work highlights and generate insights from diagrams accelerates the ideation process, reducing time-to-market for new inventions and discoveries. The system democratizes access to academic knowledge, enabling a broader audience to engage with research and pursue entrepreneurial ventures. By integrating cutting-edge AI technology into academic workflows, the Research Paper AI Assistant exemplifies how innovation can drive efficiency and inspire new business models. Its scalability across academic, industrial, and corporate domains positions it as a catalyst for fostering creativity, collaboration, and entrepreneurial growth globally.

6. Potential commercialization

Research Paper AI Assistant offers immense commercialization potential as a versatile solution for academia, research institutions, and industries. Its scalability and robust architecture make it ideal for integration into digital libraries, publishing platforms, and corporate R&D systems. Academic institutions can leverage it to enhance research efficiency and the student learning process. Industries can adopt it to streamline innovation processes. The chatbot assistant, diagram analysis, and future work extraction features make it attractive to a wide range of users, from individual researchers to global organizations. Additionally, the system's ability to save time, reduce cognitive load and improve access to critical insights positions it as a valuable subscription-based service. Partnerships with universities, publishers, and tech companies can drive adoption and revenue, creating a sustainable commercialization model. The Research Paper AI Assistant not only addresses market needs but also aligns with the growing demand for AI-driven solutions in education and research.

7. Acknowledgment

Deepest gratitude is extended to Ammar Helmey Iskandar Mohd Rayhan for his invaluable guidance, continuous support, and insightful feedback throughout this research. His expertise in system architecture development and model selection has been crucial to the success of this work. Special thanks to Dr. Ezza for providing guidance throughout the entire journey.

8. Authors' Biography

Ammar Helmey Iskandar Mohd Rayhan is a final-year student pursuing a Bachelor's degree in Intelligent System Engineering at BIS. He is highly motivated to expand his knowledge in machine learning and artificial intelligence, with a current focus on generative AI. Ammar actively participates in hackathons and innovation competitions, where he has honed his problem-solving skills and developed a keen interest in real-world applications of AI.

Dr. Ezzatul Akmal Kamaru-Zaman is a lecturer at Universiti Teknologi MARA (UiTM), Malaysia, specializing in artificial intelligence and data analytics. With a Ph.D. in Computer Science from UiTM and a Master's degree in Advanced Computer Science (Artificial Intelligence) from The University of Manchester, her expertise encompasses data science, machine learning, and big data technologies. She has contributed to impactful research, including the development of models for handling noisy clinical trial data, and her work has been published in international journals and conferences. Passionate about teaching and innovation, Dr. Ezzatul actively mentors students and engages in interdisciplinary projects to advance AI-driven solutions in education and healthcare.

Muhammad Fareed ‘Aidil Rozaidi is a final-year student at UiTM, working toward a Bachelor’s degree in Intelligent System Engineering. His passion for artificial intelligence and its real-world applications has driven him to explore innovative solutions that address critical challenges such as in education and healthcare. Aidil aspires to contribute to advancements in intelligent systems as he is deeply committed to bridging the gap between theoretical and practical applications.

Dr. Azlin is a Senior Lecturer at Universiti Teknologi MARA (UiTM), Malaysia, with extensive expertise in artificial intelligence, machine learning, and data analytics. She holds a Ph.D. from Universiti Teknologi Malaysia (2016), a Master of Science in Intelligent Knowledge-Based Systems from Universiti Utara Malaysia (2003), and a Bachelor of Information Technology with a major in Artificial Intelligence from Universiti Utara Malaysia (2001).