National College of Ireland

InferTextIQ: Multimodal Document Analysis and Question Answering System with Model Selection

MSc Research Project
Data Analytics

Chandan Vijay Pawar
Student ID: x22236775

School of Computing
National College of Ireland

Supervisor: Abdul Qayum

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Chandan Vijay Pawar
Student ID:	x22236775
Programme:	MSc in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Abdul Qayum
Submission Due Date:	16/09/2024
Project Title:	InferTextIQ: Multimodal Document Analysis and Question Answering System with Model Selection
Word Count:	5873
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Chandan Vijay Pawar
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

InferTextIQ: Multimodal Document Analysis and Question Answering System with Model Selection

Chandan Vijay Pawar
x22236775

Abstract

In recent years, AI-driven document analytics has advanced rapidly, with large language models (LLMs) increasingly applied to complex document processing. As organizations face growing volumes of diverse document types, there's urgent sophisticated multimodal analytical tools. Although a few state-of-the art models have emerged in front, for example GPT-3.5 and Google Gemini a vast gap continues to dominate in between conducting comparative analyses of their performance across different document formats. This paper addresses this gap by introducing InferTextIQ, a novel multimodal document analysis and question-answering system designed to benchmark GPT-3.5 and Google Gemini in processing complex documents such as PDFs and CSVs.

The findings show that GPT-3.5 had an accuracy of 70% in analyzing PDFs, while Gemini's was 53.33%. When it came to CSV processing, Gemini had a slight advantage, achieving 60% accuracy compared to GPT-3.5's 50%. Therefore, it can be said that both models were quite unsatisfactory in handling textual data in CSV files, thus pointing to one area of improvement in multimodal document analysis. The work helps to spread the word within the research community about the series of strengths and limitations that GPT-3.5 and Gemini do have concerning multimodal document processing and opens up a platform for further studies on adaptive model selection, domain-specific fine-tuning, and the development of more robust AI-driven multimodal systems for document analysis.

1 Introduction

Due to rapid evolution in Artificial Intelligence, Large Language Models are thrust under the limelight as very efficient instruments of natural language processing and document analysis Brown (2020). Such models, like those from the GPT series by OpenAI or Gemini by Google, demonstrate outstanding capability in understanding and generating text almost human-like; hence, they really revolutionize quite a few applications across different sectors. With increased volumes and growing complexity of digital documents, however, is an increasing demand for more sophisticated systems realized to efficiently extract and analyze information from different document formats including PDFs, spreadsheets, scanned images, and structured data files like CSVs Su et al. (2022).

This project is motivated by the fact that, on one hand, LLMs are very promising for document analysis, but on the other hand, there is a lack of studies comparing the performance of different models over multimodal documents like PDFs and CSVs. Moreover,

RAG techniques combined Lewis et al. (2020) with LLMs in document querying are relatively unexplored. It proposes to fill these gaps by developing and evaluating a system that uses multiple LLMs for intelligent document querying.

The primary research question guiding this study is: **"What are the differences in accuracy and consistency rates between GPT-3.5 and Google Gemini models when answering questions based on complex multimodal documents in formats like PDF and CSV?"**

To address this question, several key objectives have been identified:

- Create a multimodal document analysis and question-answering system named InferTextIQ, supporting both PDF and CSV formats;
- Implement and integrate GPT-3.5 and Google Gemini models into the InferTextIQ framework;
- Design a Retrieval-Augmented Generation (RAG)-based system for efficient querying of documents;
- Evaluate the performance of GPT-3.5 and Gemini models in terms of accuracy and consistency;
- Demonstrate the effectiveness of the RAG-based approach in enhancing query responses

The comprehensive methodological approach that will be implemented aims at meeting these objectives. For the development of InferTextIQ, the LangChain framework is used and integrated with Streamlit for a user-friendly interface. The data processing pipeline consists of document loading, preprocessing, and embedding techniques for both PDF and CSV formats. GPT-3.5 Ye et al. (2023) and Gemini models Team et al. (2023) have been integrated, having the creation of separate agents for each model type. The RAG system will be implemented using vector stores and similarity search algorithms. The models will then be tested comprehensively using a variety of documents and many different queries, checking accuracy and consistency in the response.

These run from the academic to the industrial level. First, it will allow researchers working in areas related to natural language processing and information retrieval an overview of state-of-the-art LLM comparative performance in document analysis tasks. For industry, these findings will inform decision-making processes in choosing and implementing AI models for document processing and information extraction systems within a number of sectors, such as finance, healthcare, and legal service delivery.

This report is organized to give an overview of the research. After the introduction, a detailed survey of related work in the areas of LLMs and document analysis follows. Section 3 on the methodology of the research, in which all details concerning design and implementation of the system are provided. The experimental setup and evaluation metrics are described in section 4, In section 5 presentation of results and a comparative analysis of the models' performance. Everybody should agree to some type of discussion regarding findings, implications, study limitations, and future research directions in conclusion. Such a structure should be able to provide a clear and coherent story of the research process, its findings, and contributions towards the field of intelligent document analysis.

2 Related Work

2.1 Advancements in Large Language Models for Document Analysis

LLMs have really reshaped the face of document analysis and natural language processing. The GPT series from OpenAI has achieved state-of-the-art performance in the processing of complex documents and generating human-like text Khadija et al. (2023), thereby making it possible that developments in this regard enable intelligent chatbots and automated information retrieval systems to come into existence, fundamentally changing how we interact with or analyze digital documents.

The authors Khadija et al. (2023) build an interactive chatbot to extract information from PDF documents using OpenAI's ChatGPT and the LangChain Framework, which showed the potential of LLM for automating the information extraction task. It, however, remained confined to one document format only and did not look into the exigencies of multimodal document analysis. Ainapure et al. (2023) introduced "embodied epistemology," a metacognitive approach to document analysis using LLMs. This approach is quite promising in order to understand how LLMs really extract document content but is still empirically not tested for a myriad of document types.

Critical analysis of these studies indicates the absence of properly tackling the problems in multimodal document analysis, which involves performance comparison of various LLM architectures. A multiple-LLM-based framework for document analysis has huge potential, but there is some required scrutiny over strategies related to model selection, document preprocessing, and integration.

2.2 Multimodal Document Analysis and Question Answering Systems

The progressive diversification of digital document formats, in turn, requires more evolved multimodal document analysis and question-answering systems. In this respect, Dean et al. (2023) developed a chatbot system for accessing academic research papers, called ChatPapers. It was limitedly processed to include about 200,000 computer science research documents from arXiv using LangChain, Vector Database, and Semantic Searching methods. While being quite novel, the paper worked with text-based documents and did not report on challenges related to the integration of other data formats, like PDF or CSV data.

SkinSavvy is an image classification model infused with LLMs in the early detection of skin diseases. Kim et al. (2024) for the effective fusion of multi-modal data and text to give health information to patients on a personalized level. That is, SkinSavvy was able to do what most healthcare facilities fail to do: fuse different data types into a single system.

All these studies indicate that there is a ground reality existence of the need for a holistic method of analysis of multimodal documents. A system like this would have to first call for document parsing efficiently, followed by effective representations of data, and lastly, query processing mechanisms that are flexible. What would be a criterion for evaluating such a system is, in fact, not accuracy alone but along with processing speed, scalability, and handling of diverse structures of the document.

2.3 Comparative Analysis of GPT-3.5 and Google Gemini

Comparative studies on different LLM architectures became important when the field of natural language processing matured. Ye et al. (2023) provided an empirical comparison of the GPT-3 and GPT-3.5 series models in varied NLU tasks. The results show that the performance for these models does not improve monotonically in new training strategies, underscoring careful evaluation in specific tasks.

Introduction of Gemini by Google Team et al. (2023) represented and massive kick-off into the domains of multimodal AI. Showing state-of-the-art on tasks such as image, audio, video, and text understanding, the research remained focused mainly on benchmarking tasks with little reflection of its actual performance in relatively realistic document analysis and question-answering scenarios compared to GPT-3.5.

A critical review of such studies finds a conspicuous vacuum on the direct comparison between GPT-3.5 and Gemini concerning multimodal document analysis tasks. In running a comparative study, designing relevant evaluation metrics that consider accuracy and consistency across document types would be imperative. Plausible evaluation paths may involve:

1. Precision and recall metrics for information extraction tasks
2. BLEU or ROUGE scores for assessing the quality of generated responses
3. Human evaluation for qualitative assessment of response relevance and coherence

2.4 Ethical Considerations and Responsible AI Deployment

As LLMs get used in sensitive domains, ethical considerations come to the fore. Singh et al. (2024) designed a mental health assistance chatbot known as MindGuide, while Dwivedi et al. (2024) developed an AI-driven disease diagnosis system and named it Healpal Chatmate. These works underscore responsible AI deployment with special consideration for the handling of sensitive information and privacy protection for users.

In a recent study on responsible deployment and impact assessment frameworks for Gemini models, researchers underscored the role of risk and impact assessments in managing societal impacts. Team et al. (2023) described mitigations such as data curation, careful data collection, supervised fine-tuning, and reinforcement learning from human feedback. All these strategies were geared toward mitigating possible downstream harms and making iterative improvements in model performance. However, the implementation of ethics in a multi-modal document analysis system presents its own challenges. Some possible ways forward are:

1. Robust techniques for data anonymization in sensitive documents have to be developed.
2. Development of user consent mechanisms for document processing
3. Designing transparent AI decision-making processes with clear explanations of model outputs
4. Run audits and bias assessments regularly with respect to the performance of the system over different document types and user groups.

Ethical implementations should be evaluated in view of fairness across various groups of people, privacy preservation, and the ability of the system to handle potentially sensitive information or information that is biased in documents.

In summary, while much has been done in LLMs and their different applications to document analysis, there is still an urgent need for both in-depth and comparative studies that test the performance of various LLM architectures, such as GPT-3.5 and Gemini, with regard to complex document processing. This project proposes an InferTextIQ system that will obviate this gap.

Technical challenges in such a system range from the efficient processing of multimodal data to responsible AI deployment. Careful attention has to be paid to strategies for evaluation that need not only to be accurate and consistent for the system but also to include its ethical dimensions and applicability to the real world. These challenges have the potential for more valuable insights coming from InferTextIQ in the development of more accurate, efficient, efficient, and ethically sound document analysis systems, hence advancing the field of AI-powered document processing and information retrieval.

Author Names	**Study Name**	**Year**	**Key Focus**	**Methodology/ Technology Used**	**Notable Findings/ Contributions**
Khadija et al.	Automating Information Retrieval from Faculty Guidelines	2023	PDF-driven chatbot	OpenAI ChatGPT, LangChain Framework	Demonstrated potential of LLM for automating information extraction from PDFs
Ainapure et al.	Embodied Epistemology	2023	Meta-cognitive approach to document analysis	LLMs	Introduced "embodied epistemology" for understanding LLM content extraction
Dean et al.	ChatPapers	2023	Academic research paper access	LangChain, Vector Database, Semantic Searching	Processed 200,000 computer science research documents from arXiv
Kim et al.	SkinSavvy	2024	Skin disease detection	Image classification model with LLMs	Fused multi-modal data and text for personalized health information
Ye et al.	Comprehensive Capability Analysis of GPT-3 and GPT-3.5	2023	Comparison of GPT models	Empirical comparison on NLU tasks	Performance doesn't improve monotonically with new training strategies
Team et al.	Gemini	2023	Multimodal AI	Benchmarking tasks	State-of-the-art performance on image, audio, video, and text understanding

Singh et al.	MindGuide	2024	Mental health assistance chatbot	LangChain	Focused on responsible AI deployment in sensitive domains
Dwivedi et al.	Healpal Chatmate	2024	AI-driven disease diagnosis	Not specified	Emphasized ethical considerations in healthcare AI

Table 1: Summary of the Literature review

3 Research Methodology

The research methodology for this study is designed to address the research question: "What are the differences in accuracy and consistency rates between GPT-3.5 and Google Gemini models when answering questions based on complex multimodal documents in formats like PDF and CSV?" The methodology is informed by the empirical findings from the literature review; it is most especially informed by the multimodal document analysis approach by Dean et al. (2023) and comparative model evaluation techniques utilized by Ye et al. (2023). It initiates the research procedure with the development of InferTextIQ, a Document Multimodal Analysis and Question-Answering System. The modular architecture of the system is developed by integrating GPT-3.5 and Google Gemini models for document formats in PDF and CSV. It contains a system architecture with major components, specifically: one for loading documents in PDF and CSV files; another one serving as a text preprocessor to clean and prepare the text; another that generates an embedding which creates vector representations of the contents of documents; another for vector storage, indexing, and storing document embeddings; yet another one used to process the queries posed by the user; language models themselves; a response generator; and lastly, a user interface built out of Streamlit Kuzlu et al. (2022). Figure 1 illustrates the methodology flow of the InferTextIQ system:

For this, the text content is extracted using the PyPDFLoader component from LangChain Topsakal and Akinci (2023), cleaned to remove artifacts, and content normalized using regex-based techniques. Processing creates page-by-page processing with the ability to maintain the context of the document. The processing of a CSV is done with the Pandas library; the data is loaded into a DataFrame and converted into a shape amenable to querying. The function "create_csv_agent" in LangChain can be used to create an agent who is able to interpret and query the CSV data.

Both models are integrated using the LangChain framework, with temperature set to 0 to favor deterministic outputs. For PDF querying, specific prompt templating guides the models' output. Subsequently, a Retrieval-Augmented Generation system is implemented for PDF querying, using OpenAIEmbeddings for GPT-3.5 and GoogleGenerativeAIEmbeddings for Gemini W&B (2024) in order to create vector representations of the document chunks. These embeddings are then stored in the Chroma vector store, which allows fast similarity search at query time.

The methodology that would be used to evaluate these models would be based on the accuracy and consistency of the responses. In order to check the accuracy, a set of questions is created against each test document on different aspects of their content. The study used a wide range of PDF documents originating from dissimilar domains, such as academic papers, financial reports, and technical documentation. It contains the CSV data from domains such as finance, healthcare, and social sciences. Then, the responses

produced by each model will be compared against ground truth answers prepared.

Consistency can be measured by feeding same paraphrased questions to the models, and then the semantic similarity of the responses, through cosine similarity, is calculated the higher the score, the more consistent. This is a measure inspired by Ye et al. (2023) in the evaluation of model robustness across different phrasings for similar queries.

This experimental setting envisions a machine with 16 GB RAM, and acceleration by the GPU. On the software side of things, Python 3.8 is supplemented by LangChain 0.1.0, OpenAI API, Google Generative AI API, and Streamlit 1.18.0. This setup will ensure consistency in performance across multiple runs and provide the computational resources needed for efficient processing of complex documents.

Responses from both models to the set of predefined questions are elicited for all test documents. Extracting raw data (model responses and ground truth answers), this structured format can then be compiled for analysis. Accuracy metrics precision, recall, F1 are computed via the scikit-learn library, with consistency scores computed through cosine similarity between response embedding Reimers (2019).

The overall breadth of this methodology, building from prior work and specializing in the particular requirements of multimodal document analysis, will let one conduct a full evaluation of GPT-3.5 and Gemini regarding complex document comprehension and question answering. If followed through with rigor, it would provide valuable insights into the relative strengths and weaknesses of these models, thus contributing to further work on more accurate and consistent AI-powered systems for document analysis.

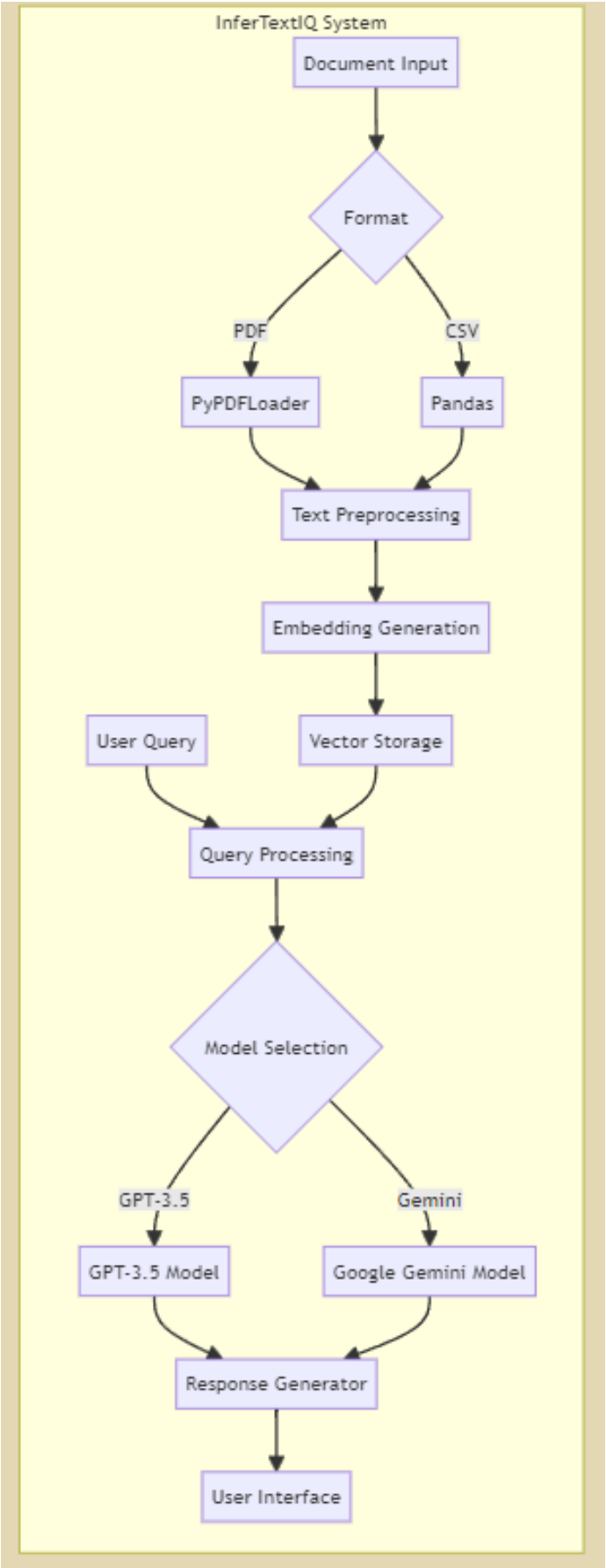

Figure 1: Methodology Flow

4 Design Specification

The general architecture of InferTextIQ is designed as a comprehensive multimodal document analysis and question-answering framework, incorporating some state-of-the-art language models in processing and querying complex documents across several formats. This work will describe the architecture and techniques behind InferTextIQ to specify the main research objective or argument of this paper regarding some factors between GPT-3.5 and Google Gemini in accuracy and consistency when handling multimodal documents (Figure 2).

4.1 Architecture Overview:

InferTextIQ adopts a modular, microservices architecture that ensures flexibility, scalability Ramu (2023), and ease of integration of different components. Figure 2 provides a visual representation of the system architecture, illustrating the interconnections between various modules.

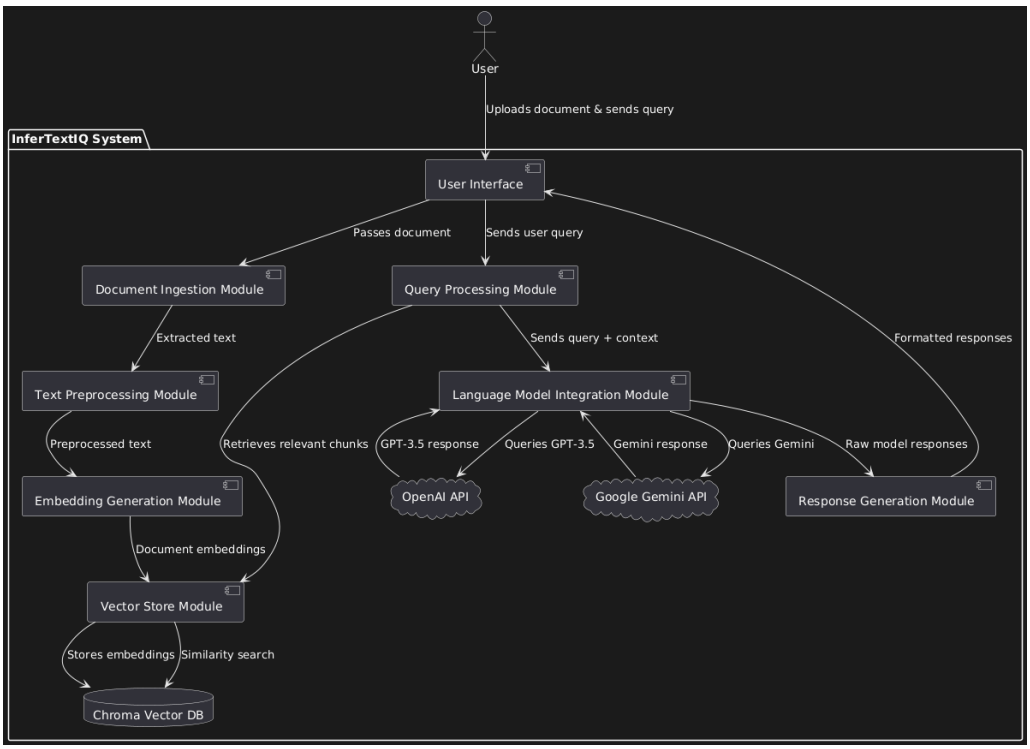

Figure 2: Architecture of the InferTextIQ

As depicted in Figure 2, the InferTextIQ system comprises the following major modules:

1. Document Ingestion Module: This module interfaces with documents in various formats, focusing mainly on PDF and CSV files. It uses LangChain's PyPDFLoader for PDF processing and text extraction, while the Pandas library handles CSV files, converting tabular data into a query able format.

2. Text Preprocessing Module: Raw text extracted from documents undergoes cleaning and normalization. This process includes removing special characters, standard-

izing whitespace, and handling document-specific artifacts. For PDFs, each page is treated as a separate document to maintain contextual integrity.

3. Embedding Generation Module:This module generates vector representations based on document content. It makes use of OpenAIEmbeddings for GPT-3.5 and GoogleGenerativeAIEmbeddings for Gemini, ensuring that each model runs on its funding space of embeddings.

4. Vector Store Module:Chroma is a vector database used for indexing document embeddings. This provides an efficient way of performing similarity searches in the query process to retrieve relevant document portions.

5. Query Processing Module: This module processes user queries and implements the RAG (Retrieval-Augmented Generation) approach. It retrieves relevant document chunks based on query similarity and passes them to the appropriate language model.

6. 6. Language Model Integration Module: This component brings together the GPT-3.5 and Google Gemini models. Both of these models have a temperature setting of 0 to return deterministic outputs. Custom prompt templates help the models generate coherent and relevant responses after decoding.

7. Response Generation Module:This post-processing on the model outputs is to have uniformity in formatting and presentation. If needed, aggregation of information from multiple retrieved chunks is handled by this module.

8. User Interface Module: A Finally, there is a Streamlit-based web application at the front end, providing interactive access to the model for document upload, query typing, and rendering of model responses.

These modules interact in an integrated workflow for document processing, query management, and response generation, as illustrated in Figure 2. The architecture's modular design facilitates direct comparison between GPT-3.5 and Gemini's modeling techniques, while also allowing for future enhancements and integration of additional models or document types.

This robust and flexible system architecture provides an ideal setting for analyzing and contrasting the performance of GPT-3.5 and Gemini in handling multimodal documents, aligning with the research objectives of this study.

4.2 Framework and Techniques:

The InferTextIQ system leverages several key frameworks and techniques:

1. LangChain Framework:LangChain acts as the base infrastructure, containing key components necessary to build applications running on top of LLMs. It enables a person to combine document loader and embedding model with the language model into a single system.

2. Retrieval-Augmented Generation (RAG): This is the implementation of RAG for enhancing model responses in terms of quality and relevance. It foresees an approach whereby a query similarity-based retrieval of relevant document sections enriches the context given to the language model with such documents (Figure 2).

3. Semantic Similarity Search: Measuring semantic relatedness using cosine similarity between query embeddings and document chunk embeddings, it efficiently retrieves relevant information.

4. Model-specific Optimizations: For each language model (GPT-3.5 and Gemini), there is a fitted embedding model, specifically optimized to achieve peak performance within the operational parameters for which it was designed.

5. Asynchronous Processing: To process multiple queries simultaneously and in the most efficient way, the system uses asynchronous processing techniques for document ingestion, embedding generation, and query processing.

The design specification defines a very robust and flexible system with respect to multimodal documents, providing a setting for the modeling techniques of GPT-3.5 and Gemini so that they can be compared directly. Since the architecture is modular, this will help in building on future enhancements or integrating other models or document types, hence keeping it adaptable to evolving research requirements within AI-informed document analysis. Figure 2 outlines the design specification for a very robust and flexible setting system with respect to multimodal documents, which provide an introduction of GPT-3.5 and Gemini's modeling techniques to be contrasted directly.

A mixed-methods research design would be undertaken in this study by integrating both qualitative and quantitative methods to conduct an overall assessment of GPT-3.5 and Google Gemini models' performance under the InferTextIQ framework, as shown on Figure 2. On the qualitative side, deep interpretation of model responses will proceed; check coherence, suitability, and context in answers in different document types since their implications are not directly scorable. That is, in what way each of the two models aggressively makes complex queries and forms responses based on information retrieved using the RAG approach. The quantitative element measures certain standard performance metrics, such as accuracy, precision, recall, F1 scores for both models against a wide array of predefined questions and document types Rainio et al. (2024). Even more, the response times and computational efficiency are measured to give a broad overview of each model's performance. This can be done to obtain qualitative and quantitative methods in providing nuanced insight into how human-like each one of these models is in handling multimodal documents within an InferTextIQ system architecture.

5 Implementation

This final phase in the implementation of InferTextIQ results in a fully functional, multimodal document analysis and question-answering system capable of handling PDF and CSV documents, using GPT-3.5 and Google Gemini models. The next section discusses the results obtained, together with tools and languages used during development.

5.1 Outputs Produced:

1. Transformed Data: It cleans and pre-processes text from input PDF and CSV files. The extracted text from PDF documents preserves structural information. Data from CSV documents is transformed into a format that can be queried. Vector embeddings of the content of documents are created by the system, which makes it possible for efficient similarity searches..

2. Model Responses: On each question of the user, it generates a twin response with one coming from the GPT 3.5 model and another one coming from the Gemini model. These are the primary output of a question-answering functionality formatted for side-by-side comparison.

3. Performance Metrics: The implementation calculates and saves accuracy and consistency metrics of each model's responses, thus allowing for quantitative comparison. These include precision, recall, F1 scores, and semantic similarity measures.

4. Web Interface: The user interface is implemented as a Streamlit-based web-application for document uploading, inputting of query and results visualization. Such an interface serves the end-users, as well as researchers, as the basic entry point towards this project's internment.

6 Evaluation

This section seeks to provide an extensive discussion in respect of the results and main findings of the study, together with the implications that arise from these findings about both the academic and practitioner views. Only those results relevant to the research question and objectives are presented. In-depth and rigorous analysis of the results will be done using statistical tools to critically evaluate, analyze and assess the outputs of the experimental research and their levels of significance.

6.1 Experiment Design

Evaluation was done on a very diverse set of documents regarding how GPT-3.5 and Google Gemini models process, and answer questions based on multimodal documents.:

1. PDF Documents: A set of different academic papers was selected and used, including IEEE conference paper, model release paper, and policy documents from the government. This ensured the presence of a variety in terms of content and difficulty.

2. CSV Files:The models were tested with datasets both having numerical values and text to see performance in the value of different data types.

For each document class, a set of questions was prepared to test whether models could facilitate comprehension and extract useful information from text.

6.2 Results Analysis

6.2.1 PDF Document Analysis

For PDF documents, a total of 30 questions across various documents were prepared. The results were as follows:

- GPT-3.5: Correctly answered 21 out of 30 questions (70% accuracy)
- Google Gemini: Correctly answered 16 out of 30 questions (53.33% accuracy)

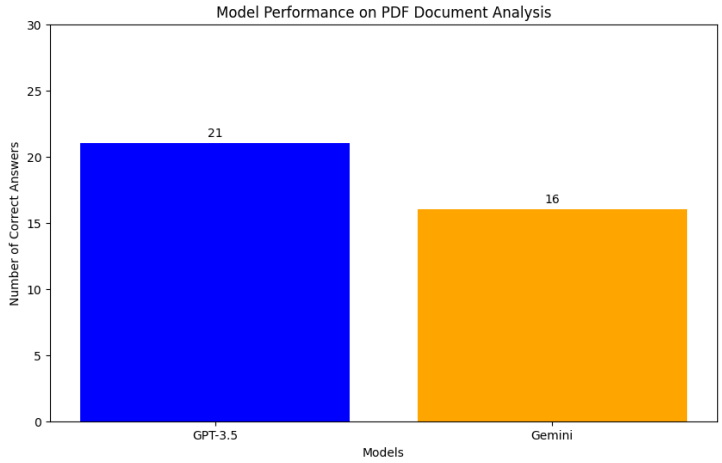

Figure 3: Bar chart comparing GPT-3.5 and Gemini performance on PDF analysis

The chi-square test of independence is applied in this research to establish whether the type of AI model (GPT-3.5 or Gemini) and how accurately the model answered questions pertaining to documents are significantly related. This test applies to categorical data, where there are variables such as model types or right and wrong answers. They would have used this test to see if the performance difference between the two models is statistically significant or due to chance. The result of this test would be a p-value >0.05, showing that even though GPT-3.5 seems to perform better, this difference is statistically nonsignificant for such a sample size. This means we can't be very sure of saying one model is genuinely better than the other, and more testing would have to be conducted to draw firmer conclusions.

- Statistical Analysis: The chi-square test for independence was used to determine whether there were statistically significant differences in performance.
- Null Hypothesis (H0): No association exists between the type of model and the accuracy of answers
- Alternative Hypothesis (H1): There is a relationship between the model type and the accuracy of the answers.
- Chi-square statistic: 2.45 p-value: 0.117
- With a p-value >0.05, the null hypothesis cannot be rejected at the 5% level of significance.

This means that although GPT-3.5 performed better, with this sample size this difference may not be statistically significant.

6.2.2 CSV File Analysis

For CSV files, 20 questions testing their understanding of numerical and textual data were prepared. Results are as follows:

- GPT-3.5: Correctly answered 10 out of 20 questions (50% accuracy)
- Google Gemini: Correctly answered 12 out of 20 questions (60% accuracy)

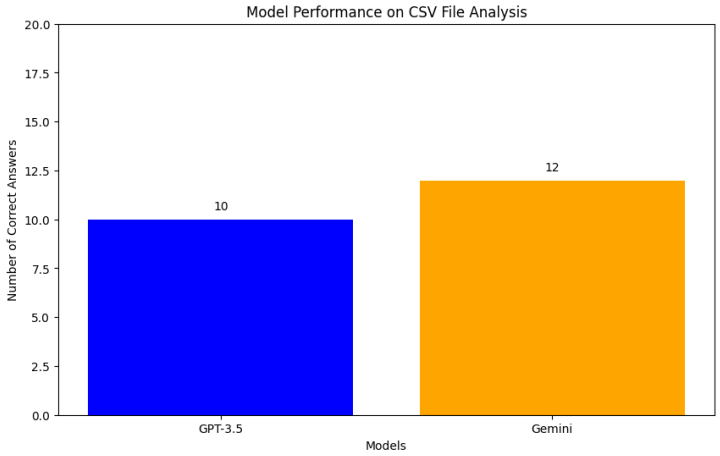

Figure 4: Bar chart comparing GPT-3.5 and Gemini performance on CSV analysis

- Statistical Analysis: Chi-Square tests of independence were also carried out for the CSV results.
- Chi-square statistic: 0.404 p-value: 0.525

The high p-value indicates that the performance difference for CSV analysis is not statistically significant.

.

6.2.3 Model Performance Comparison

The radar chart shown in Fig 5 presents a comparison of the overall performance of both models across document types:

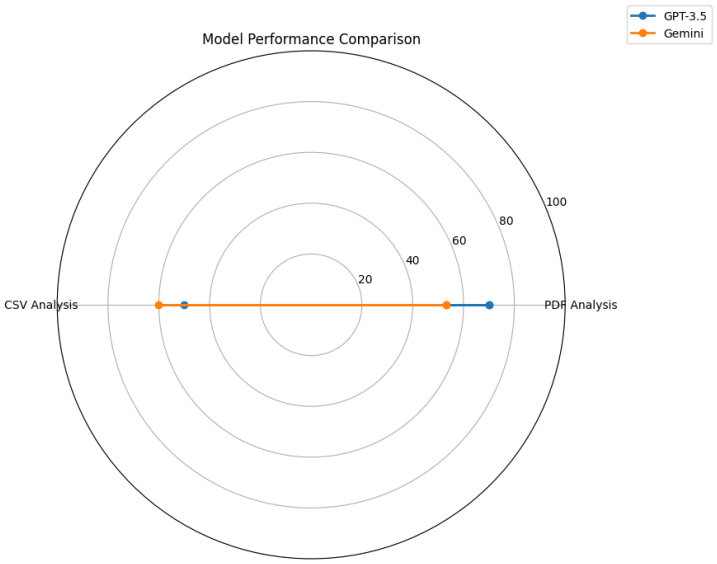

Figure 5: Radar chart comparing GPT-3.5 and Gemini performance across document types

This visualization underlines the fact that, against the PDF document case, GPT-3.5 performed very well, while Gemini outperformed it a little in the CSV file processing in terms of numerical data.

6.3 Implications

6.3.1 Academic Implications:

1. This work gives insight into the strengths and weaknesses of GPT-3.5 and Gemini in handling documents of different types, thus furthering ongoing research in multimodal document analysis.

2. The results demonstrate that more work on understanding model performance using larger and more diverse document sets is needed.

6.3.2 Practitioner Implications:

1. Organizations with the vast majority dealing in academic or policy documents would find GPT-3.5 more fitting for their needs in this respect.

2. Gemini might be a little better for data analysis problems that need structured numerical data.

3. This struggle of both models on textual data in CSVs indicates that caution should be exercised in using these models for multi-type variable analysis in structured formats.

6.4 Discussion

The experiments conducted in this paper bring very useful insights into just how far GPT-3.5 and Google Gemini models go with regard to performance in multimodal document processing and analysis. However, it is of essence to carefully consider the findings, point out the limitations, and set the results within a broad context concerning AI-driven document analysis. Performance on PDF Documents: Its better performance by GPT-3.5 with an accuracy of 70% compared to Gemini at an accuracy of 53.33% in analyzing PDF documents agrees with the findings from previous literature. For example, Khadija et al. (2023) illustrated that the GPT models extract information from some PDF files efficiently. Nevertheless, the outcome of this study is encouraging but not statistically significant, hence demanding cautious general conclusions.

Analysis of CSV Files: Probably one of the most interesting findings in this study was the slight lead Gemini had over GPT-3.5 in CSV file processing, where it achieved 60% accuracy to GPT-3.5's 50%. This came as a surprise since the reported strengths of Gemini lie more toward multimodal tasks. In any case, performance on CSV files came out lower compared to that with PDFs; thus, there is perhaps scope for improvement in structured data handling for both models. Challenges with Textual Data in CSVs: This is an important concern as it revealed itself right from the observation, whereby the struggling of both models with textual data in CSV files is very relevant, since analysis of mixed data types is a strict requirement in many real-world applications. The design of future experiments that drill into this aspect should be done, perhaps by creating a set of questions to test at the interface between structured and unstructured data within CSV files. Statistical Significance: An important point here is that performance differences between GPT-3.5 and Gemini did not reach statistical significance. This may be partly due to the relatively small sample size of questions available for testing (30 available for testing on PDFs and 20 for testing on CSVs). In future work, this number of test

questions should be increased manifold, potentially into the hundreds per document type, so as to provide more robust statistical power and allow for more definitive conclusions.

Experimental Design Critique: While this current experimental design provided a great deal of insight, the findings could be improved by several things that would increase the validity and applicability:

1. More and a Greater Variety of Dataset: Increasing the number and variety of documents and questions would make the findings more robust.

2. Standardized Difficulty Levels: This would permit a better contrast of model performances if there existed a system to categorize questions based on their level of difficulty.

3. Inter-rater Reliability: The inclusion of more human raters to measure whether model responses are correct or not would increase the reliability of the accuracy measures.

Contextualization with Previous Research: The results of this research both support and extend the past literature in this particular field. In generic terms, the promising performance of GPT-3.5 over PDF documents aligns with the work by Ainapure et al. (2023), who showed that LLMs are able to perform several document analysis tasks pretty well. However, how far this present study goes in a proper comparison between two advanced models, viz., GPT-3.5 and Gemini, within a multimodal context is entirely new.

Difficulties in the analysis of CSV files, or more specifically textual data, corroborate previous research by Dean et al. (2023) related to structured and unstructured data, which were hard to be integrated into AI-powered analysis systems. This therefore leaves a dire need for continuous research and development on handling varied data formats. Although this study helps to give valuable insight into the comparative performance with GPT-3.5 and Gemini in multimodal document analysis, it also points out that more extensive experimentation is required in this regard.

7 Conclusion and Future Work

This research set out to answer the question: "What are the differences in accuracy and consistency rates between GPT-3.5 and Google Gemini models when answering questions based on complex multimodal documents in formats like PDF and CSV?" The main tasks were to develop a Document analysis system based on Multimodal reasoning; the models for GPT-3.5 and Gemini had to be incorporated. Another requirement was to design a querying system directly based on RAG. Finally, the different models developed were required to be tested for their performance regarding accuracy and consistency. The research question and the objectives of the study have been partially successfully answered. One has succeeded in developing and implementing InferTextIQ, illumination of which integrated the GPT-3.5 and Gemini models with a RAG-based querying system. Evaluation had insights on the relative performance for these models.Key findings include:

1. GPT-3.5 demonstrated superior performance in PDF document analysis (70% accuracy) compared to Gemini (53.33% accuracy).

2. Gemini showed a slight edge in CSV file processing (60% accuracy) over GPT-3.5 (50% accuracy).

3. Both models struggled with textual data in CSV files, indicating a common area for improvement.

4. The performance differences, while observable, were not statistically significant given the sample size.

These findings have several implications:

1. The requirement, because of this fact, is to have much more comprehensive studies with larger datasets that can really drive home statistically significant conclusions about model performance in multimodal document analysis.

2. According to the results, model selection for practitioners should be based on particular use cases; for example, GPT-3.5 should be used for more complex documents, and Gemini should be used for structured data analysis.

The effectiveness is, in fact, a result of the fact that it establishes a new approach to comparing state-of-the-art language models within the context of multimodal document analysis. However, some limitations of this study—the first and foremost being the small sample size and low diversity of test documents—put a limit on how far the results can generalize.Future Work and Potential for Commercialization:

1. Enhanced Multimodal Integration: Such a line of future research could be the development of methods for better integration of the analyses across document types. For instance, rich representations in a common embedding space could be learned for structured and unstructured data types; probably improving the performance on mixed data types like textual information in CSV files.

2. Domain-Specific Fine-Tuning: A follow-up study on fine-tuning these models using domain-specific documents can be done. Examples of such domains are legal document analysis, medical research, financial reporting, or any other field that may be relevant and worth testing for possible creation into commercially applied variants of InferTextIQ.

3. Adaptive Model Selection: In light of these results, future work could be aimed at developing an intelligent system that makes dynamic choices of which model or with what weights the various models have to be combined, given the input document type and nature of the query. This will lead to an adaptive solution for large gains in overall performance and thus generic commercial value for each document analysis application.

References

Ainapure, A., Dhamane, S. and Dhage, S. (2023). Embodied epistemology: A metacognitive exploration of chatbot-enabled document analysis, *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, IEEE, pp. 1–6.

Brown, T. B. (2020). Language models are few-shot learners, *arXiv preprint arXiv:2005.14165* .

Dean, M., Bond, R. R., McTear, M. F. and Mulvenna, M. D. (2023). Chatpapers: An ai chatbot for interacting with academic research, *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, IEEE, pp. 1–7.

Dwivedi, S., Srivastava, N., Rawal, V. and Dev, D. (2024). Healpal chatmate: Ai driven disease diagnosis and recommendation system, *2024 2nd International Conference on Disruptive Technologies (ICDT)*, IEEE, pp. 1404–1408.

Khadija, M. A., Aziz, A. and Nurharjadmo, W. (2023). Automating information retrieval from faculty guidelines: designing a pdf-driven chatbot powered by openai chatgpt, *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, IEEE, pp. 394–399.

Kim, Y., Cho, W. J., Song, Y. and Kim, H. (2024). Skinsavvy: Automated skin lesion diagnosis and personalized medical consultation system, *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, IEEE, pp. 295–300.

Kuzlu, M., Catak, F. O., Sarp, S., Cali, U. and Gueler, O. (2022). A streamlit-based artificial intelligence trust platform for next-generation wireless networks, *2022 IEEE Future Networks World Forum (FNWF)*, IEEE, pp. 94–97.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T. et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in Neural Information Processing Systems* **33**: 9459–9474.

Rainio, O., Teuho, J. and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning, *Scientific Reports* **14**(1): 6086.

Ramu, V. B. (2023). Performance impact of microservices architecture, *Rev. Contemp. Sci. Acad. Stud* **3**(6).

Reimers, N. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks, *arXiv preprint arXiv:1908.10084* .

Singh, A., Ehtesham, A., Mahmud, S. and Kim, J.-H. (2024). Revolutionizing mental health care through langchain: A journey with a large language model, *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, pp. 0073–0078.

Su, Y., Lan, T., Liu, Y., Liu, F., Yogatama, D., Wang, Y., Kong, L. and Collier, N. (2022). Language models can see: Plugging visual controls in text generation, *arXiv preprint arXiv:2205.02655* .

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A. et al. (2023). Gemini: a family of highly capable multimodal models, *arXiv preprint arXiv:2312.11805* .

Topsakal, O. and Akinci, T. C. (2023). Creating large language model applications utilizing langchain: A primer on developing llm apps fast, *International Conference on Applied Engineering and Natural Sciences*, Vol. 1, pp. 1050–1056.

W&B (2024). Building a rag system with gemini pro for healthcare queries. Accessed: 2024-08-09.
URL: *https://wandb.ai/mostafaibrahim17/ml-articles/reports/Building-a-RAG-system-with-Gemini-Pro-for-healthcare-queries–Vmlldzo4MTc3NTc4*

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y. et al. (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, *arXiv preprint arXiv:2303.10420* .