Generative AI's Role in Auditing: An Exploration of Automated IFRS Compliance

Gagan Sardana (20200440), Pranoti Dhaiwal (21201254), Shivani Bhardwaj (22200976)

A Capstone submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business, University College Dublin

18th August 2023

Sponsor & Mentor (EY): Eda Mae Badilla, Nabanita Roy

Supervisor: Dr. Michael MacDonnell

Head of School: Professor Anthony Brabazon

Dedication

To our family, friends, and mentors, whose unwavering support and belief in us have been the pillars of this journey. This capstone project, a culmination of countless hours and persistent endeavours, is a testament to the collective faith, guidance, and encouragement we have received throughout our master's journey. May this work not only showcase the knowledge and skills we have acquired but also stand as a tribute to all who have walked this path with us.

Table of Contents

List of Figures ... vi

List of Tables ... vii

Preface ... viii

Acknowledgements ... ix

Executive Summary ... x

Chapter 1 - Introduction ... 1

1.1 IFRS Standards: A Pillar of Financial Reporting ... 1

1.2 Potential of Generative Artificial Intelligence ... 1

1.3 Research Objectives ... 3

1.4 Significance of Research ... 3

1.4.1 Advancing Auditing Efficiency ... 3

1.4.2 Enhancing Accuracy and Consistency ... 3

1.4.3 Enabling Informed Decision-Making ... 3

1.4.4 Expanding Accessibility ... 4

1.4.5 Contributing to Research and Innovation ... 4

1.4.6 Shaping the Future of Financial Auditing ... 4

Chapter 2 - Literature Review ... 5

2.1 Artificial Intelligence in Auditing: Opportunities and Challenges ... 5

2.2 Machine Learning in Auditing: Opportunities and Challenges ... 6

2.3 Natural Language Processing in Auditing ... 7

2.4 Large Language Models in Auditing ... 8

2.5 Prompt Engineering ... 9

2.6 Potential of Knowledge Graph in Auditing ... 11

2.7 Ethical Considerations in Auditing ... 12

2.8 BloombergGPT: A Large Language Model for Finance ... 13

2.9 IFRS Standards: Previous and Current Development ... 14

Chapter 3 - Methodology ... 16

3.1 Comprehensive Workflow Overview ... 16

3.1.1 Data Acquisition ... 16

3.1.2 Data Processing ... 17

3.1.3 Semantic Analysis with GPT-3.5 model ... 17

3.1.4 UI Development ... 18

3.1.5 Evaluation ... 18

3.2 Business Process Model 18
3.3 Data Used 19
3.3.1 Tesco PLC Annual Report 19
3.3.2 Marks & Spencer PLC Annual Report 20
3.3.3 Content Overview 20
3.4 User Interface 20
3.4.1 Overview of User Interface 20
3.4.2 Components of User Interface 22
3.5 Software Development 26
3.5.1 High-Level Architecture 26
3.5.2 Low-Level Architecture 27
3.6 Implementation 30
3.6.1 Software and Tools Used 30
3.6.2 Other Essential Libraries 32
3.6.3 Chroma Vector Database 32
3.6.4 Code Implementation 32
3.6.5 Prompt Engineering 35
3.7 Software Testing 38
3.7.1 Functional Testing 38
3.8 Performance Evaluation Metric 43
3.8.1 Evaluation Strategy 43
3.8.2 Success Metrics 43
3.8.3 Comparative Analysis 44
3.8.4 Thresholds and Benchmarks 44
3.8.5 Feedback and Iteration 44
Chapter 4 - Process Flow 45
4.1 Initialisation 45
4.2 User Interface Display 45
4.3 Document Upload 45
4.4 Document Processing 45
4.5 Conversational Retrieval Chain Setup 46
4.6 Question-Answer Interface 46
4.7 Error Handling 46
4.8 End Interaction 47
Chapter 5 - Results and Findings 48

5.1 Prompts (Questions Asked) 49
5.1.1 Standard Prompts 49
5.1.2 Detailed Prompts 49
5.2 Results Interpretation 49
5.2.1 Accuracy and Prompt Styles 50
5.2.2 Scalability and Replication 50
5.2.3 Prompts and Manipulation of Accuracy 50
5.2.4 Text Processing Improvement 50
5.3 Overall Implications 50
Chapter 6 - Challenges 52
6.1 Rate Limit 52
6.2 Token Limitations 52
6.3 Text Extraction Challenges 52
Chapter 7 - Future Enhancement 53
7.1 Fine-Tuning w/wo Knowledge Graph 53
7.1.1 With Knowledge Graph 53
7.1.2 Without Knowledge Graph 53
7.2 Tabular Extraction from Annual Reports 53
7.3 One-Shot Prompting 53
7.4 Few-Shot Prompting 53
7.5 Multiple File Processing 54
7.6 Support for Multiple Document Types 54
7.7 Report Generation 54
7.8 Integration with Other Financial Standards 54
Chapter 8 - Conclusion 56
References 57

List of Figures

Figure 1: AI Landscape 2
Figure 2: Comprehensive Workflow.. 17
Figure 3: User Interface - 1.. 21
Figure 4: User Interface - 2.. 21
Figure 5: User Interface - Successful Uploading of the Annual Report.. 22
Figure 6: User Interface - Unsuccessful Uploading of the Annual Report 22
Figure 7: User Interface - Processing of Annual Report 23
Figure 8: User Interface - Successful Processing of Annual Report.. 23
Figure 9: User Interface - Selecting Standard Questions 24
Figure 10: User Interface - Selecting Custom Questions.. 24
Figure 11: User Interface - Chat History of Previous Questions.. 25
Figure 12: User Interface - Different Prompt Styles 25
Figure 13: Software Architecture - High-Level Architecture 26
Figure 14: Software Architecture - Low-Level Architecture 28
Figure 15: Implementation - Creation & Storing of Embeddings.. 33
Figure 16: Implementation - Example of Single Embedding.. 33
Figure 17: Implementation - Embeddings in Chorma DB.. 34
Figure 18: Implementation - Code for Conversation Retrieval 34
Figure 19: Implementation - Code for Question Answer Mechanism 35
Figure 20: Testing Result - No Database before Processing the Annual Report.. 39
Figure 21: Testing Result - Database after Processing the Annual Report 39
Figure 22: Testing Result - Processing of Annual Report at the Backend.. 40
Figure 23: Testing Result - Post Processing Message 40
Figure 24: Testing Result - Message Indicating that the Document Type is not Permitted ... 40
Figure 25: Testing Result - Standard Question Response.. 41
Figure 26: Testing Result - Custom Question Response 41
Figure 27: Testing Result - No Data Extraction Error Message.. 42
Figure 28: Testing Result - Chat History 42
Figure 29: Testing Result - Chat History Results after clicking Clear Chat History Button...... 43
Figure 30: Process Flow Diagram 46
Figure 31: Metric Results of Tesco PLC Annual Report 2022 48
Figure 32: Metric Results of Marks and Spencer Annual Report 2023 48

List of Tables

Table 1: Tesco PLC Test Results 48
Table 2: M&S PLC Test Results 48

Preface

This capstone project is the summary of our learnings and knowledge acquired from the modules that have been taught throughout our master's programme. It signifies the conclusion of our master's in business analytics. All subsequent work was conducted in collaboration with EY. It is co-authored by Gagan Sardana, Pranoti Dahiwal, and Shivani Bhardwaj.

The utilisation of Generative AI in the auditing domain holds significant importance, especially given its relatively uncharted territory within the current market landscape. As a result, the primary objective of this research-oriented capstone project is to develop an auditing tool through the application of artificial intelligence. The core purpose is to harness AI's potential to enhance auditing practices and leverage its capabilities in this field.

Dublin,	Gagan Sardana
18th August 2023	Pranoti Dhaiwal
	Shivani Bhardwaj

Acknowledgements

We extend our heartfelt gratitude to Dr. Michael MacDonnell, our programme director and capstone project supervisor, for his unwavering support and guidance throughout this journey. His insights have been invaluable to the progression and success of our project.

Our appreciation also goes to Nabanita Roy (EY), who, alongside Dr. Michael MacDonnell, diligently mentored us. Their combined efforts in conducting weekly meetings, reviewing our progress, and offering insightful feedback have been instrumental in shaping our work.

Lastly, our sincere thanks to the Micheal Smurfit Graduate School of Business, UCD, for granting us this enriching opportunity.

Executive Summary

The automation of the interpretation of International Financial Reporting Standards (IFRS) is becoming crucial in today's financial landscape. This necessity arises from the potential benefits of increased efficiency, improved accuracy, and rigorous adherence to these intricate standards. Central to this transformation is Artificial Intelligence (AI), particularly its Natural Language Processing (NLP) capabilities, which allow for rapid analysis of complex IFRS documentation.

AI-driven technologies offer streamlined processes, ensuring consistent compliance, swift adaptation to IFRS updates, and a marked reduction in human errors. This is paramount when considering the vast array of annual reports, which emanate from a plethora of industries, each with its distinct financial transactions and reporting practices. Manually examining these data-rich reports and aligning them with IFRS standards is not only time-intensive but also susceptible to errors.

Given the depth and breadth of IFRS standards, a profound understanding of accounting principles is essential, which might be challenging for those outside the financial domain. AI-powered tools can be game-changers in this context, processing vast amounts of data and providing accurate insights. Such tools bridge the knowledge gap, enabling professionals across various industries to accurately interpret and apply IFRS standards.

Our developed tool is one such platform that will help in finding accurate results with quick information retrieval. Individuals can use this tool to quickly retrieve relevant information with explanations, saving time and effort.

Chapter 1 - Introduction

The integration of automation into financial tools has reached a point of undeniable indispensability. Its profound advantages manifest through a remarkable capacity to significantly amplify efficiency, precision, and cost-effectiveness across a diverse spectrum of financial processes. This transformative influence ripples outward, yielding multifaceted benefits that encompass diminished human errors, expedited task completion, elevated sophistication in data analysis, unwavering adherence to compliance standards, and the seamless facilitation of strategic decision-making by seasoned experts. The crowning achievement of this paradigm shift? A surge in productivity leads to the realisation of unparalleled financial outcomes.

1.1 IFRS Standards: A Pillar of Financial Reporting

International Financial Reporting Standards (IFRS) play a pivotal role in shaping the global financial landscape. Developed and maintained by the International Accounting Standards Board (IASB), these standards serve as the bedrock upon which financial reporting is built.

Currently, the IFRS framework encompasses a comprehensive set of standards, totalling over 50, which address various aspects of financial reporting, including revenue recognition, lease accounting, financial instruments, and more. This vast framework provides a unified language for financial reporting, ensuring that financial information is presented consistently, transparently, and comparably across different organisations and jurisdictions.

The significance of IFRS standards reverberates through the global financial ecosystem. They not only cultivate unwavering consistency in reporting practices but also illuminate the path toward enhanced transparency, bolstered investor assurance, streamlined cross-border transactions, and the seamless exchange of financial insights. As financial operations continue to span international horizons, the relevance and impact of IFRS standards remain paramount, shaping the contours of modern financial reporting.

1.2 Potential of Generative Artificial Intelligence

The advent of Generative Artificial Intelligence, epitomised by pioneering models such as GPT-3.5 and GPT-4, marks a profound shift in the landscape of natural language comprehension and generation. These models stand as a testament to the remarkable strides achieved, showcasing an unparalleled prowess in not only comprehending but also seamlessly generating text that emulates human expression. Their versatile capabilities have ignited a revolution across diverse domains, spanning from the realm of linguistic translation to dynamic content creation and even intricate problem-solving.

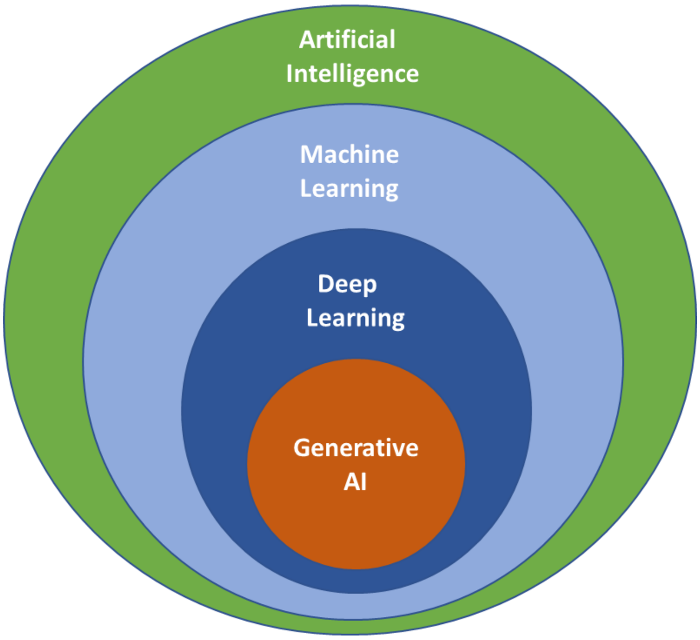

Figure 1: AI Landscape

Source: *https://medium.com/@amol-wagh/whats-generative-ai-explore-underlying-layers-of-machine-learning-and-deep-learning-8f99272e0b0d*

Intriguingly, the integration of these advanced models within the intricate tapestry of financial auditing beckons a new era of possibilities. The discerning application of these AI powerhouses has the potential to unleash a wave of transformative effects. At the heart of this potential lies the profound impact of automation – a force that has become increasingly indispensable in today's fast-paced financial landscape. By harnessing the innate abilities of GPT-3.5 and GPT-4, financial auditing processes stand poised to transcend the traditional confines, surging forward with amplified levels of efficiency, accuracy, and agility.

This innovative convergence delves into the very essence of automation. The intricate complexities that define financial auditing – from dissecting intricate regulations to discerning nuanced anomalies within vast datasets – can be seamlessly navigated through the capabilities of these models. The symbiotic partnership between human expertise and AI augmentation can yield insights that are not only profound but also timely, forming the cornerstone of informed decision-making.

The implications extend beyond the confines of mere process enhancement. They reach into the realm of redefining how financial auditing interfaces with data, standards, and interpretations. By channelling the prowess of GPT-3.5 and GPT-4, auditors can potentially engage in a dynamic dialogue with financial documents, extracting nuanced insights and unearthing hidden correlations. This transformation isn't just about expediting operations; it's about empowering auditors with an expanded cognitive toolkit that enables them to traverse uncharted territories of financial scrutiny.

1.3 Research Objectives

- Develop an IFRS compliance assessment software solution powered by the OpenAI's GPT-3.5 model.
- Develop the software to accurately interpret and analyse financial documents, extracting relevant information related to IFRS standards.
- Enable the software to provide concise and well-substantiated responses to queries about IFRS compliance based on evidence extracted from annual reports.
- Enhance the software's usability and user-friendliness, catering to both novice and expert users in the field of financial auditing.
- Evaluate the accuracy, efficiency, and effectiveness of the software in comparison to traditional manual methods of IFRS compliance assessment.

1.4 Significance of Research

1.4.1 Advancing Auditing Efficiency

The research aims to revolutionise the auditing process by automating the interpretation and analysis of IFRS standards. This will lead to faster and more accurate assessments, reducing the time and effort required for compliance.

1.4.2 Enhancing Accuracy and Consistency

By leveraging GPT-3.5, the software solution ensures consistent and well-founded responses, minimising the potential for human errors in interpreting complex financial regulations.

1.4.3 Enabling Informed Decision-Making

The software empowers auditors and financial professionals with real-time insights and accurate information, facilitating strategic decision-making based on reliable data.

1.4.4 Expanding Accessibility

The user-friendly nature of the software makes IFRS compliance assessments accessible to a broader range of users, democratising access to advanced auditing tools.

1.4.5 Contributing to Research and Innovation

This research project bridges the realms of finance, technology, and artificial intelligence, contributing to the advancement of knowledge and innovation in both fields.

1.4.6 Shaping the Future of Financial Auditing

The successful implementation of the software solution has the potential to reshape how financial audits are conducted, setting a precedent for a more efficient, accurate, and technology-driven approach.

In the ensuing chapters, we present a detailed exploration of our approach to automating financial auditing. We will discuss the methodology, emphasising the role of advanced AI technologies, and describe the process of data extraction and analysis from financial documents. The results section will highlight the performance and capabilities of our software solution in the context of IFRS compliance assessment. Additionally, we'll address challenges encountered during the project's development and consider future research directions. This document aims to provide a thorough understanding of our efforts in enhancing the efficiency and accuracy of financial auditing using AI.

Chapter 2 - Literature Review

2.1 Artificial Intelligence in Auditing: Opportunities and Challenges

Artificial Intelligence (Al) has significantly reshaped various industries, including auditing. Research conducted by Omoteso (2012) and Baldwin, Brown, and Trinkle (2006) presents a comprehensive view of the potential applications of AI in the auditing domain.

Omoteso (2012) provides a detailed review of the role of AI in auditing, particularly focusing on expert systems and neural networks. Expert systems, as described by Omoteso, are intelligent agents that mimic human expert decision-making, aiding in simplifying complex problems and improving the quality of auditing. Arnold et al. (2004), show that a well- executed combination of user and intelligent aid can enhance decision-making in auditing.

Eining and Dorr's (1991) research highlighted the impact of expert systems on knowledge acquisition. The study shows that those who were assisted by an expert system performed better than those without such aid, hinting at the educational implications of expert systems in training novice auditors. Furthermore, Omoteso discusses the use of expert systems m different audit types, such as risk assessment, highlighting the versatility of these systems.

In contrast, neural networks, which mimic the human brain's functioning, are beneficial in making predictions based on past trends, thus enhancing audit judgments. For instance, Green and Choi (1997), developed a neural network fraud classification model, aiming to alert auditors to perform substantive testing as soon as any financial statement is classified as fraudulent.

Baldwin, Brown, and Trinkle (2006), in their seminal paper, delve into the potential of AI in auditing, focusing on how cross-disciplinary collaboration involving AI researchers could improve auditing practices. They identify more than 400 individual audit tasks, including analytical review procedures, classification, materiality assessments, internal control evaluation, risk assessment, going-concern decisions, and bankruptcy prediction, which could benefit from AI implementation.

The authors also discuss the potential of more complex AI approaches, such as genetic algorithms, neural networks, fuzzy systems, and hybrid systems for various audit tasks. Baldwin et al. argue that these complex AI applications are crucial for tackling complex audit tasks, and their potential for improvement should be fully investigated.

Both Omoteso (2012) and Baldwin et al. (2006) underscore the necessity for further research in the field of AI in auditing. Omoteso points out the gaps in assessing the financial costs and benefits of AI systems in auditing, practical litigation implications, implications for small and medium audit firms, and auditor independence. On the other hand, Baldwin et al. call for accounting researchers to bridge the gap between the business and accounting domains and the computer science and AI domains.

In conclusion, the advent of AI in auditing, through expert systems and neural networks, promises improved efficiency and decision-making quality. Yet, there remain gaps in our understanding of the full implications and potential applications of these technologies. Future research can help fill these gaps and further enhance the integration of AI in the audit field.

2.2 Machine Learning in Auditing: Opportunities and Challenges

Machine learning presents considerable opportunities for the audit profession, revolutionising the way auditors conduct their work. By utilising large datasets, machine learning allows auditors to identify patterns, trends, and financial anomalies that may otherwise remain unnoticed (Dickey, Blanke, & Seaton, 2019). Moreover, pioneering research by Sifa et al. (2019) demonstrates the application of machine learning techniques in auditing, particularly in the extraction and classification of statements of information (Sols) from financial documents.

However, there are substantial challenges to implementing machine learning in auditing. Obtaining relevant and high-quality data, especially non-financial data, can be difficult due to legal and ethical restrictions. Additionally, there is the concern of human bias influencing the audit process when employing machine learning tools (Dickey, Blanke, & Seaton, 2019).

The implementation of machine learning also has implications for audit documentation standards, necessitating a shift in practice from justifying procedures to detailing the evaluation and application of data analysis. Nevertheless, advances in AI and machine learning technologies offer novel ways to automate and enhance audit processes, even offering flexibility through unsupervised models that can adapt to changing requirements (Sifa et al., 2019).

However, the quality of training data and appropriate model selection are crucial for successful implementation. Therefore, further research and development are needed, especially with the advent of more advanced LLM, such as the OpenAI series of GPT 3.5 models, that could potentially improve these results.

Ultimately, as the field continues to evolve, auditors will need to adapt, not only to use these technologies effectively but also to understand the broader implications they present. The literature indicates that machine learning will significantly alter the audit profession in the future, and while the exact nature of this impact is still uncertain, top audit firms are already using machine learning tools (Dickey, Blanke, & Seaton, 2019). Meanwhile, the work by Sifet al. (2019) provides a roadmap for future research, contributing significantly to the literature on AI applications in auditing.

2.3 Natural Language Processing in Auditing

Natural Language Processing is the field of Artificial intelligence that focuses on the interaction between computers and human language by enabling computers to understand, interpret, and generate human language in a meaningful and useful manner (Lutkevich,2021).

NLP focuses on developing algorithms and models that can process and analyse text by extracting relevant information and driving insights from it. It uses various techniques and approaches to tackle these, including sentiment analysis, named entity recognition, text classification, and part-of-speech tagging (Lutkevich,2021).

Auditing involves the examination of financial annual reports and other financial information to ensure that they are accurate and comply with accounting standards and regulations. Natural Language Processing (NLP) can be used in auditing to analyse large volumes of financial data, including text-based data such as emails, memos, and other documents.

(Fisher, Garnsey, and Hughes, 2016) address the fact that NLP in accounting, auditing, and finance can help to improve the accuracy and efficiency of financial reports as it extracts data from unstructured text documents such as reports, and articles, and identifies patterns that are difficult for humans to categorise. This process helps in reducing errors increases speed and improves accuracy (Fisher, Garnsey, and Hughes, 2016).

Further, it helps in getting insights into auditing and finance by creating methodologies using sentiment analysis, data toning, and pattern recognition methods that summarise the financial reports for decision-making purposes (Fisher, Garnsey, and Hughes, 2016).

(Fisher, Garnsey, and Hughes, 2016) discuss that some of the most used algorithms such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision trees, Latent Sentiment Indexing (LSI), Support Vector Regression (SVM), and Word Embedding can help in identifying anomalies and patterns by extracting data from unstructured text, automating categorisation of transactions and by generating summaries.

Despite these advances, there are a few challenges in using NLP**.** These challenges are accurately interpreting and analysing financial data because of the lack of standardisation, and data quality issues because of errors and inconsistencies.

Overall, Natural Language Processing (NLP) has the potential to revolutionise the field of accounting, auditing, and finance by enabling more efficient and accurate analysis of financial data however some challenges need to be addressed appropriately to ensure the accuracy and ethical use of this technology.

2.4 Large Language Models in Auditing

Large Language Models are advanced machine learning models that are capable of processing and generating human-like text, they show significant highlights in the Natural Processing tasks such as text generation, translation, and question-answering (Hore, 2023).

The Transformer architecture plays a crucial role in the development and effectiveness of Large Language Models (LLMs) in natural language processing. By utilising input embeddings and positional encoding, transformers enable LLMs to convert text inputs into numerical representations and capture the order and meaning of words in a sentence. The encoder-decoder structure, along with the attention mechanism, allows LLMs to process the entire input sequence simultaneously, resulting in faster and more accurate language understanding and generation.

The ability of Transformers to handle complex connections between words and capture long- term dependencies makes them superior to traditional recurrent neural networks (RNNs) and LSTMs. Overall, the Transformer architecture is a fundamental component that empowers LLMs to achieve state-of-the-art performance in various NLP tasks, making it a crucial advancement (Menon,2023) in different fields such as auditing.

There are various components and techniques of LLM employed in the automated auditing system such as the text pre-processing step which not just involves character and word-based pre-processing but also domain-specific pre-processing steps such as the detecting and normalising legal references, accounting terms, and identifying section titles which help in enriching the pre-processing task semantically (Araci, 2019).

Different approaches are followed such as the character n-grams, or a Bag-of-Words representation that considers the joint vector-space representation of all words in the block. Additionally, neural language models, such as doc2vec paragraph embeddings, provide context-aware vector embeddings. Further, Unsupervised ranking involves computing the similarity between requirements and Statements of Interest (Sols) based on text representations and ranking the Sols accordingly. All of this is achieved by using binary classifiers or multi-layer classifiers.

Despite this, all the traditional sentiment analysis models do not perform well in the financial domain due to the unique vocabulary and language used in the financial text. To overcome this algorithms such as FinBert, and word embedding are used for large amounts of auditing data, by doing this they capture the financial language and make it more effective for analysis. (Araci, 2019). While there are commonly used dictionaries for financial terms such as the General Inquirer (GI), Harvard IV-4 (HIV4), Diction, and Loughran and McDonald's (LM) word lists (Araci, 2019).

Overall, LLMs and the associated techniques have significantly improved automated auditing processes by enabling better language understanding, document representation, and sentiment analysis in the financial domain. These advancements contribute to more efficient and accurate auditing practices, ultimately benefiting the field of finance.

2.5 Prompt Engineering

The rapid progression of artificial intelligence (AI) and language models has opened new opportunities for transforming traditional audit processes. Two recent papers - Lo (2023) and Gu et al. (2023) - present ground-breaking approaches to harness the potential of AI in auditing, highlighting the importance of prompt engineering and co-piloted auditing.

Lo (2023) introduces the CLEAR Framework, a user-centric approach to formulating prompts that result in more effective interactions with AI language models. The framework emphasises the importance of Clarity, Logic, Explicitness, Adaptability, and Reflective evaluation. Clear and concise prompts, along with logically structured instructions, allow AI language models to generate more relevant and precise responses, improving their overall efficiency in various tasks. Users' adaptability, based on AI models' performance, is underscored to ensure the generation of content that addresses specific challenges. This iterative process, supported by constant evaluation and improvement, optimises the model's performance, paving the way for enhanced information literacy.

Meanwhile, Gu et al. (2023) introduce the concept of co-piloted auditing, a collaborative approach where AI models and human auditors work in partnership to deliver comprehensive audits. The authors demonstrate this by fine-tuning GPT-4, a foundation model, for tasks such as financial ratio analysis, text mining, and journal entry testing. Key to their approach is the use of chain-of-thought (CoT) prompting, demonstrating the model's ability to comprehend and respond to complex auditing tasks.

The co-piloted auditing process involves iterative interactions between auditors and foundation models, where the AI serves as an intelligent assistant analysing audit evidence, providing insights, and facilitating decision-making. The integration of foundation models into the audit process offers several benefits such as improved risk management, increased efficiency, enhanced decision-making, and robust continuous learning. To maximise the potential of co-piloted auditing, the authors highlight the need for intuitive user interfaces, dynamic task allocation, effective feedback mechanisms, and thorough documentation systems.

The fusion of the CLEAR Framework's principles with co-piloted auditing opens new possibilities for auditing. By fine-tuning prompts based on the CLEAR Framework and incorporating AI models into the auditing process, we can improve Al's understanding of complex tasks, thus optimising audit outcomes. This combined approach also promises the potential to create a robust learning environment where AI models and auditors can learn from each other, improving the efficiency and effectiveness of audit processes. The future of auditing appears brighter with these developments, heralding a shift towards increased digitisation and automation.

In conclusion, both studies provide significant insights into the transformative potential of AI and prompt engineering in the field of auditing. As the integration of AI into traditional processes continues, these studies contribute to a better understanding of how to utilise AI effectively in auditing, providing a roadmap for future exploration in the field.

2.6 Potential of Knowledge Graph in Auditing

The application of knowledge graphs (KGs) in auditing represents a significant development in the finance field, bridging the structured representation of data with complex relationships between entities. Ahmadi et al. (2022) and Liu et al. (2020) lay the foundation for auditing- specific KGs, discussing the key challenges such as non-standard named entities and variations of noun phrases. They suggest methods for KG construction and potential improvements via human-in-the-loop solutions (Ahmadi et al., 2022).

Further, Liu et al. (2020) propose an innovative KG-based enterprise audit platform to standardise and aggregate data, enhancing audit decision-making. They emphasise the importance of both structured and unstructured data, employing named entity recognition models and web crawler technology for data acquisition.

Beyond the realm of auditing, knowledge graphs have been proven effective across domains, as Lin et al. (2021) highlight. They have improved semantic understanding, search and recommendation systems, data integration, and the discovery of hidden patterns in various sectors.

Specifically in finance and auditing, Mackie and Dalton introduce the idea of query-specific KGs, presenting a more focused overview of a domain, as opposed to traditional KGs derived from ontologies or text corpora. The adoption of this approach, especially when incorporating auditing standards like IFRS, can result in a specialised representation of domain-specific information, addressing the challenges of knowledge representation in finance.

In summary, knowledge graphs hold substantial promise for auditing and finance, offering enhanced data understanding and improved decision-making.

Nevertheless, their implementation presents challenges requiring innovative solutions, including the potential incorporation of LLM. As the complexity of financial systems continues to grow, it's clear that the future of audit information and enterprise auditing will greatly benefit from further exploration and refinement of KGs.

2.7 Ethical Considerations in Auditing

As the field of AI continues to advance, incorporating ethical considerations is becoming increasingly vital, particularly when employing large language models (LLMs) for auditing and other financial applications. In the literature, scholars like (O'Leary, 2022) and (Svetlova, 2022) discuss the challenges and risks associated with the use of LLMs and conversational AI in finance, emphasising the importance of ethical considerations and transparency in AI systems. These factors encompass the need for addressing biases, ensuring fairness, mitigating systemic risks, and integrating compliance with regulations.

Key ethical areas for auditors using LLMs include Bias and Fairness Auditing (Pineiro- Martin et al., 2023), which involves scrutinising the fairness of AI recommendations while considering diversity, representation, and equal treatment of users. Privacy and Data Security are also vital, with the need for adherence to privacy regulations and the implementation of robust measures against unauthorised access or misuse of data. The importance of transparency is paramount, ensuring clear insights into the data sources, decision-making processes, and system behaviour.

Identifying potential vulnerabilities in AI systems that could be manipulated for malicious purposes is essential. It's crucial to implement preventative measures, such as content filtering and verification mechanisms, to protect against such risks.

Furthermore, interdisciplinary collaboration can enable a comprehensive understanding and addressing of ethical concerns. Collaborative efforts between researchers, developers, policymakers, and ethicists ensure that diverse perspectives are considered in the design and development of LLMs. Coupling this with continuous feedback from users promotes transparency, accountability, and the responsible use of LLMs, leading to more ethically sound, user-centric applications.

In conclusion, integrating ethical considerations into the design and development of LLMs for auditing is crucial. It not only fosters the creation of reliable, trustworthy tools for assessment, evaluation, and decision-making but also ensures the responsible use of AI, thereby reinforcing public trust and acceptance.

2.8 BloombergGPT: A Large Language Model for Finance

In the continuously evolving landscape of financial technology, Natural Language Processing (NLP) has emerged as a game-changer. "BloombergGPT: A Large Language Model for Finance" offers a groundbreaking perspective on the role of LLMs tailored specifically for the financial sector (Wu et al., 2023).

The paper underscores the multifaceted applications of NLP in finance, such as sentiment analysis and question answering. While LLMs have demonstrated their prowess across a myriad of tasks, the authors identify a notable void: the absence of a specialised LLM for finance. Hence, they introduce BloombergGPT, a behemoth model equipped with 50 billion parameters, fine-tuned on financial data (Wu et al., 2023).

The foundation of any potent model lies in its dataset. The authors meticulously curated a dataset boasting 363 billion tokens from Bloomberg's extensive databases. In a bid to harmonise domain specificity with adaptability, this dataset was enriched with an additional 345 billion tokens from general-purpose collections (Wu et al., 2023).

The intricacies of training a model as vast as BloombergGPT are detailed in the paper. It delves into the hurdles faced and the ingenious strategies adopted. A comprehensive reading would illuminate the optimisation methodologies, hardware setups, and other large-scale training nuances. BloombergGPT's performance claims are anchored in rigorous testing. It was put to the test against standard LLM benchmarks and a range of finance-centric tasks. The model's training on the mixed dataset proved beneficial, as it outshone its peers in financial tasks without sacrificing performance on generic benchmarks. The authors also hint at proprietary benchmarks, reflecting the model's envisioned applications (Wu et al., 2023).

While numbers paint a picture, qualitative samples offer a glimpse into BloombergGPT's real-world utility. These examples, although briefly mentioned, demonstrate its potential in practical financial scenarios. The authors candidly discuss the model's constraints, emphasising its ethical utilisation. They touch upon the broader ramifications of BloombergGPT's integration into the financial realm and urge for its responsible deployment (Wu et al., 2023).

In conclusion, BloombergGPT marks a significant stride in the quest for a financial domain-specific language model. Through meticulous training and assessment, the model has showcased its potential, heralding a transformative era in financial technology. While this review encapsulates the paper's essence, a detailed exploration of the document would offer profound insights into its contributions and nuances (Wu et al., 2023).

2.9 IFRS Standards: Previous and Current Development

The automation landscape for IFRS standards has seen the rise of several software platforms, each aiming to bolster efficiency and compliance in financial reporting. Notable among these are Workiva, SAP S/4HANA, and Oracle Financial Services Analytical Applications (OFSAA). Each tool brings to the table unique capabilities to facilitate financial reporting, compliance tasks, and data management, all in sync with IFRS guidelines. Yet, a consistent gap emerges in their ability to provide definitive answers to specific queries within individual standards.

Identifying this void, we've ventured into pioneering territory. We've crafted a specialised software solution that automates the auditing of IFRS standards, harnessing the capabilities of OpenAI’s GPT 3.5 models’ and advanced embedding techniques. Unlike conventional tools, our application delves deep into the annual reports, interpreting them with unparalleled precision. It's engineered to grasp textual nuances, empowering it to respond to IFRS-related queries with a clear "Yes" or "No/NA", each answer enhanced by precise, context-driven reasoning.

Our solution addresses a pressing need: a tool that not only extracts specific details from annual reports but also interprets them through the IFRS prism. By tapping into advanced embedding techniques, the application ensures its responses are accurate and backed by relevant justifications.

In conclusion, while many platforms have made commendable advancements in IFRS automation, our software solution carves a distinct niche. It directly addresses the specific challenge of answering IFRS queries using annual reports. This innovation promises to redefine the way financial professionals engage with and interpret IFRS standards, setting the stage for more accurate compliance, informed decisions, and streamlined financial processes.

Chapter 3 - Methodology

In our pursuit to automate the complex task of auditing IFRS standards using annual reports, we embarked on an innovative journey harnessing cutting-edge technological advancements. Central to our methodology is the utilisation of word embedding, a technique rooted in Natural Language Processing (NLP). In the world of NLP, word embeddings translate words or phrases into numerical vectors in a continuous space. This transformation captures the semantic essence of words, allowing for the preservation of intricate linguistic nuances.

With the textual content of annual reports now translated into structured numerical forms using word embeddings, we integrated this data with OpenAI's artificial intelligence platform. OpenAI, celebrated for its proficiency in natural language understanding, holds the capability to dissect complex textual data with pinpoint accuracy. This integration empowers our methodology to pinpoint relevant passages within annual reports and align them meticulously with the demanding requirements of IFRS standards.

To manage the vast and high-dimensional data generated through word embeddings, we turned to Chroma DB. This database is specifically crafted to handle high-dimensional vectors, making it a seamless fit for our needs. It ensures that the numerical representations of text are stored efficiently and can be retrieved without hitches, enhancing the robustness and scalability of our system.

Collectively, the synergy between word embeddings, OpenAI’s GPT 3.5 models’, and Chroma DB marks a transformative step in automating IFRS compliance assessments. This blend of technologies not only accelerates the auditing process but also elevates its precision and consistency. Our approach embodies a paradigm shift, showcasing how technological convergence can revolutionise the interpretation and application of intricate financial standards. The results? Enhanced efficiency, unparalleled accuracy, and a gold standard in compliance assessments.

3.1 Comprehensive Workflow Overview

3.1.1 Data Acquisition

Obtained annual reports directly from the official websites of Tesco PLC and M&S. These reports, spanning hundreds of pages, are essential for the IFRS standards compliance assessment. The data was carefully selected to ensure a representative and comprehensive dataset for the software solution's operations.

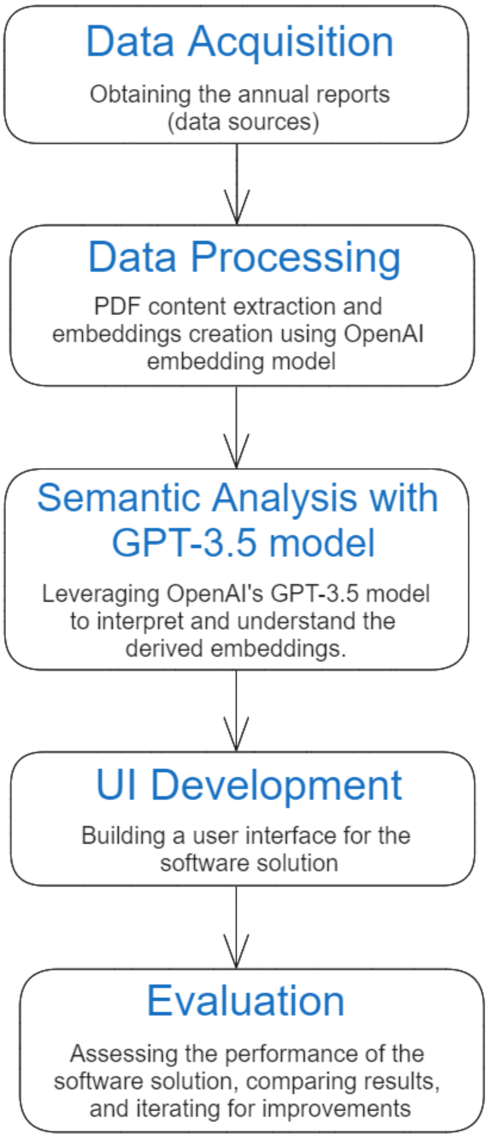

Figure 2: Comprehensive Workflow

3.1.2 Data Processing

Processed the textual content of these annual reports, transforming it into structured numerical forms. The data, primarily contained in tables and financial sections, was converted into embeddings using OpenAI GPT 3.5 “text-embedding-ada-002” model. This numerical representation captures the semantic relationships between words, which is crucial for the software to understand and assess compliance based on IFRS standards.

3.1.3 Semantic Analysis with GPT-3.5 model

Employed OpenAI's GPT 3.5 “text-davinci-003” model to interpret and understand the embeddings derived from the annual reports. This advanced language model leverages its natural language capabilities to provide deep insights into the financial data, facilitating accurate IFRS standards compliance assessments. The model's proficiency in analysing complex textual representations ensures that the software can discern intricate IFRS requirements and correlate them with relevant passages in the reports.

3.1.4 UI Development

A user-friendly interface was developed to provide both novice and expert users easy access to the tool. The interface allows users to upload annual reports, input queries related to IFRS standards, and receive clear, evidence-based responses. It bridges the gap between the financial and non-financial domains, ensuring that users from various backgrounds can effortlessly understand the intricacies of the auditing process.

3.1.5 Evaluation

The software's performance was meticulously evaluated using the annual reports of Tesco PLC and M&S. The evaluation strategy focused on comparing the software's outputs against answers reviewed by experts. Precision, recall, and F1 score were the primary metrics employed to gauge the software's accuracy and effectiveness. Feedback loops were implemented, which facilitated iterative refinement of the software based on real-world results.

3.2 Business Process Model

The business process model serves as a strategic blueprint, outlining the sequential execution of vital steps within our software. As our software operates, it offers a robust auditing solution that fundamentally transforms the compliance landscape. By seamlessly harnessing the power of software, it adeptly addresses IFRS queries using the contents of annual reports.

In this endeavour, our primary stakeholders encompass accountants and financial professionals, essential. Simultaneously, employees engaged in financial operations find value in this tool as it streamlines their activities. Recognising the global significance of IFRS standards, our software helps as a beacon as IFRS helps organisations in positioning themselves optimally in international capital markets. The tool's ability to generate IFRS-compliant financial statements enhances credibility and familiarity, drawing global investments.

The essence of our software resides in its accessibility and inclusivity. It bridges the gap between the financial and non-financial domains, offering a platform that demystifies the complex world of auditing. Individuals from various backgrounds can effortlessly grasp the intricacies of the auditing process and discern suitable answers aligned with IFRS standards.

The true value proposition of our software crystallises when considering its unparalleled capabilities. It possesses an unmatched ability to swiftly process massive volumes of data, thereby drastically reducing the time required for compliance assessments. This accelerated pace is achieved without compromising the critical facet of accuracy – a remarkable fusion that propels compliance workflows into a new era. Moreover, its scalability empowers organisations to efficiently manage diverse sizes of annual reports, thus adroitly addressing varying demands.

Beyond speed and efficiency, cost reduction becomes a tangible benefit. By automating the compliance assessment process, our software minimises dependency on labour-intensive tasks, resulting in significant financial savings.

The final testament to its value is the mitigation of dependency on the intricate language of accounting. Our software acts as a translator, decoding complex accounting jargon into a language that stakeholders from diverse backgrounds can easily comprehend, fostering collaboration and informed decision-making.

In essence, our software doesn't just automate; it revolutionises the compliance assessment landscape. It empowers professionals, enhances credibility, accelerates processes, and optimises resources. As a beacon of innovation, it reshapes how compliance with IFRS standards is achieved, guiding organisations toward a future of efficiency, accuracy, and global recognition.

3.3 Data Used

For our research, the primary sources of data were the annual reports of two leading retail companies: Tesco PLC and Marks & Spencer (M&S). These reports were directly procured from the official websites of the respective companies, ensuring authenticity and relevance.

3.3.1 Tesco PLC Annual Report

- Source: Official Tesco PLC website.

- Length: The report spans 220 pages with a file size of 9.52 MB.
- Year: The report covers the financial year ending 2022.
- Link: https://www.tescoplc.com/media/an0cp1co/tesco-annual-report-2022.pdf

3.3.2 Marks & Spencer PLC Annual Report

- Source: Official Marks & Spencer website.
- Length: The document is 236 pages long and has a file size of 8.49 MB.
- Year: This report pertains to the financial year ending 2023.
- Link:https://corporate.marksandspencer.com/sites/marksandspencer/files/2023-06/M%26S_2023_Annual_Report.pdf

3.3.3 Content Overview

While both reports are comprehensive, encompassing various facets of the companies' operations and performance, our primary focus was on the financial sections. These sections, vital for our research, contain detailed financial statements, accompanying notes, and related disclosures. It's worth noting that while these sections are predominantly tabular, they also encompass textual content that provides context and elaboration on the presented figures.

3.4 User Interface

3.4.1 Overview of User Interface

The user interface of our software solution simplifies the auditing process by providing an intuitive and efficient way to interact with OpenAI's GPT 3.5 models' capabilities for information extraction and conversation. It was built using Streamlit, a powerful and user-friendly Python library specifically designed for creating interactive web applications. Streamlit was chosen as the framework for the user interface due to its simplicity, ease of use and flexibility.

The application allows users to seamlessly upload an annual report and initiate the auditing process with a single click. Once the report is processed and the data is stored in a vector database, users can engage in a conversational interface to query the extracted data. The application offers two options for asking questions: "Standard Questions" and "Custom Questions." In the "Standard Questions" option, users can choose from a list of predefined questions specifically related to the IFRS 16 accounting standard. On the other hand, the "Custom Question" option empowers users to input their inquiries, offering more flexibility. When a question is submitted, the application promptly retrieves the relevant information from the vector database and provides the AI-generated response. Furthermore, the processing time for each query is displayed, providing transparency and insights into the application's efficiency.

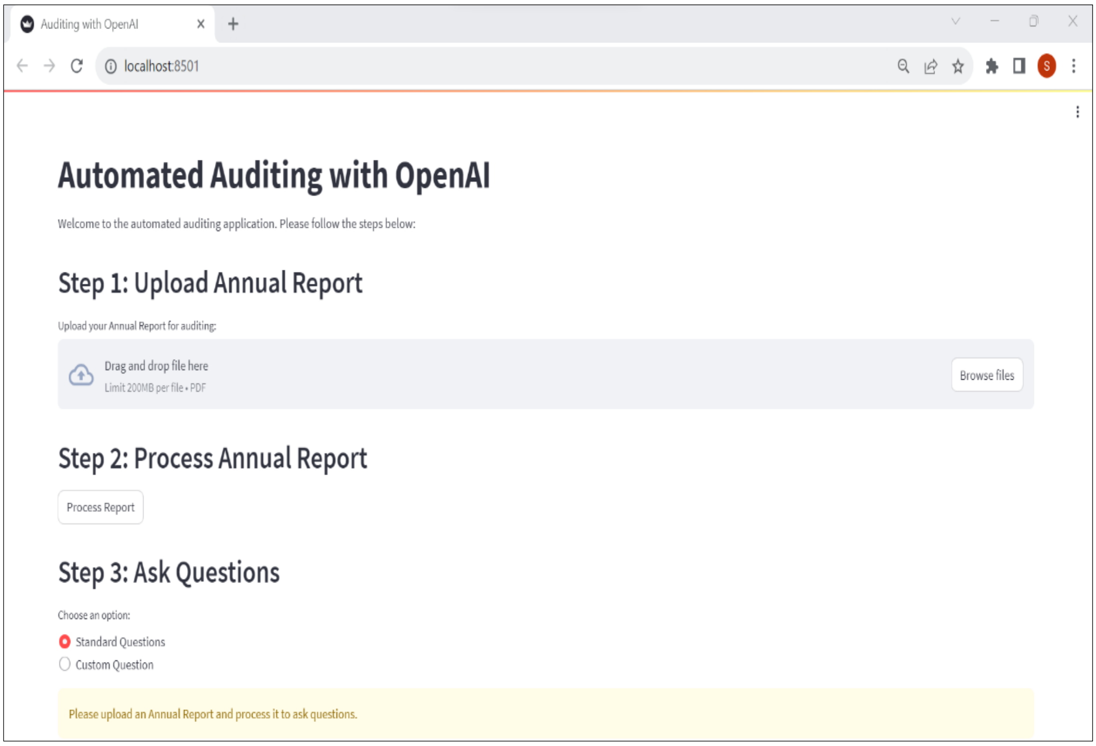

Figure 3: User Interface - 1

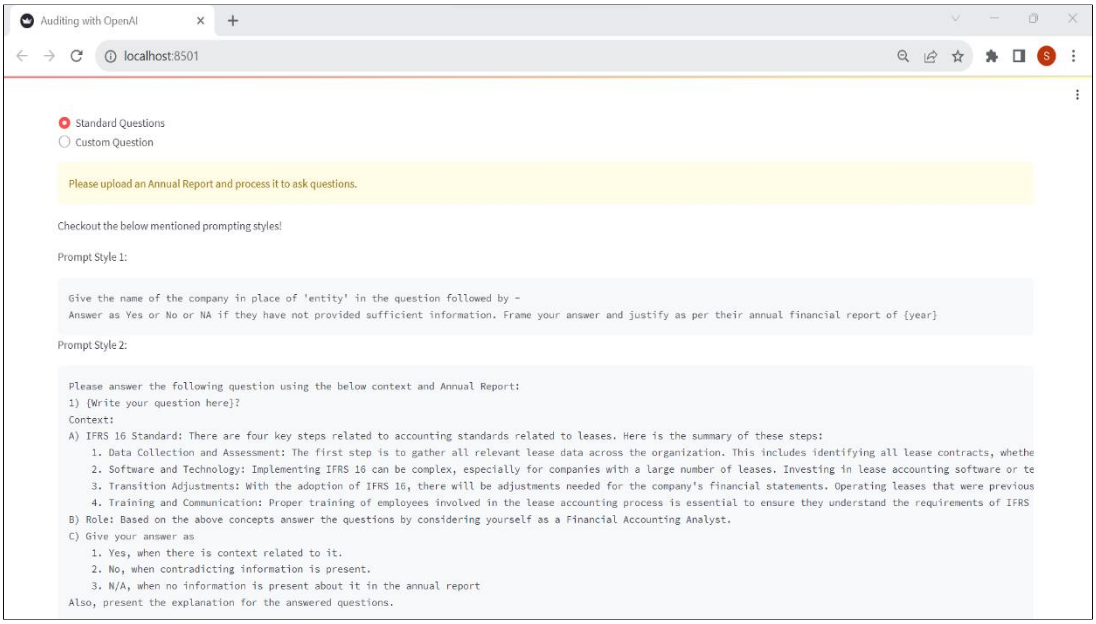

Figure 4: User Interface - 2

To enhance user experience and continuity, the application maintains a chat history that records user queries and the corresponding AI responses. Users can easily review previous interactions, making it convenient to track the conversation flow and past queries. Additionally, the user interface includes detailed information about prompting.

3.4.2 Components of User Interface

The application provides steps to be followed for AI auditing, each setup is discussed in detail below.

3.4.2.1 *Step 1: Upload Annual Report*

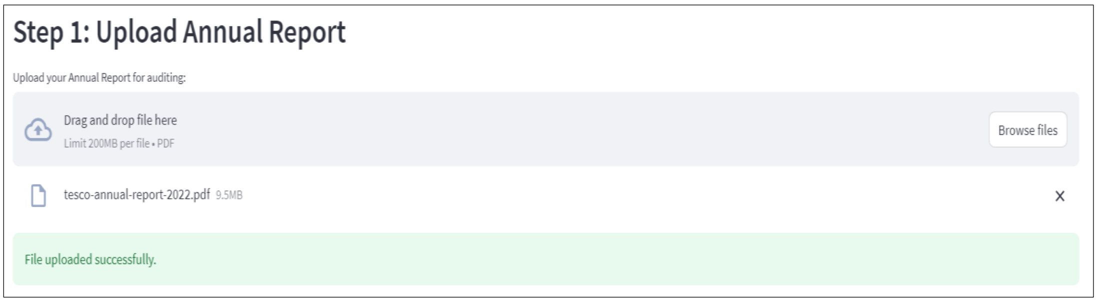

Figure 5: User Interface - Successful Uploading of the Annual Report

The first step of the user interface provides users with a convenient and straightforward way to initiate the auditing process by uploading their annual reports. The application presents a user-friendly interface where users are prompted to browse their local file system and select the annual report they want to audit. This is facilitated using the

"file_uploader" component provided by Streamlit, which allows users to select a PDF file from their computer. The interface allows users to upload only one PDF file at a time.

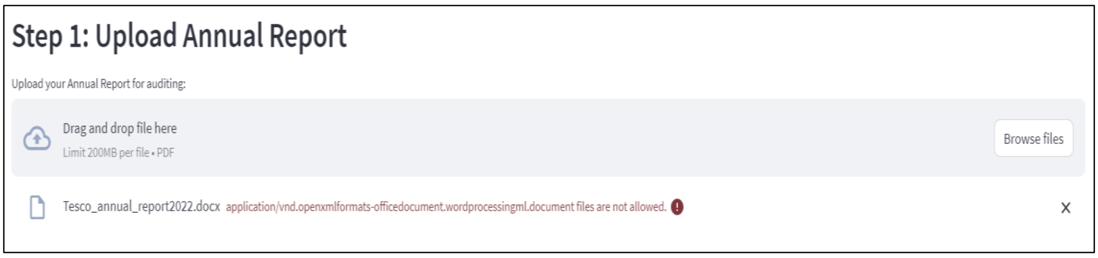

Figure 6: User Interface - Unsuccessful Uploading of the Annual Report

Upon selecting the file, the application processes the upload, and a success message is displayed to confirm that the file was successfully uploaded. This feedback ensures that users are aware of the successful file upload and are ready to proceed to the next steps of the auditing process.

3.4.2.2 Step 2: Process Annual Report

n this step of the user interface, users can initiate the processing of the uploaded annual report. The processing involves converting the textual content of the PDF document into vector embeddings, which will later be used for information retrieval and answering user queries.

Figure 7: User Interface - Processing of Annual Report

Figure 8: User Interface - Successful Processing of Annual Report

relevant information is stored in the vector database. The processed data is now ready for retrieval and answering user queries in Step 3 of the user interface.

3.4.2.3 Step 3: Ask Questions

This step is designed to facilitate users in querying the processed data to gain insights and information from the annual report in a conversational manner.

In the "Standard Questions" option, users are presented with a curated list of predefined questions specific to the International Financial Reporting Standard (IFRS) 16 accounting standard. These questions are provided to extract crucial information from the annual report regarding its compliance with the IFRS 16 standard. Users can choose from the list of standard questions, which cover various aspects of the accounting standard related to leases, modifications, practical expedients, lessor, lessee, sale, and leaseback transactions, and more.

When users select a question from the list and click "Submit Standard Question," the application uses the vector database to retrieve the relevant information and provide the AI-generated answer to the selected question. The AI response is displayed on the interface, along with the time taken to process the query. This feature enables users to quickly access specific information from the annual report, streamlining the auditing process and assisting in compliance assessment with the IFRS 16 standard.

Figure 9: User Interface - Selecting Standard Questions

In the "Custom Question" option, users have the flexibility to formulate their questions related to the annual report. This option empowers users to seek information tailored to their specific auditing requirements, allowing them to dig deeper into specific aspects or address unique concerns within the report.

The custom question option is provided to try various prompting styles (prompt engineering) to set the context and optimise the query to obtain more accurate responses. Similar to the standard questions, the AI response and the time taken for processing are displayed to the user.

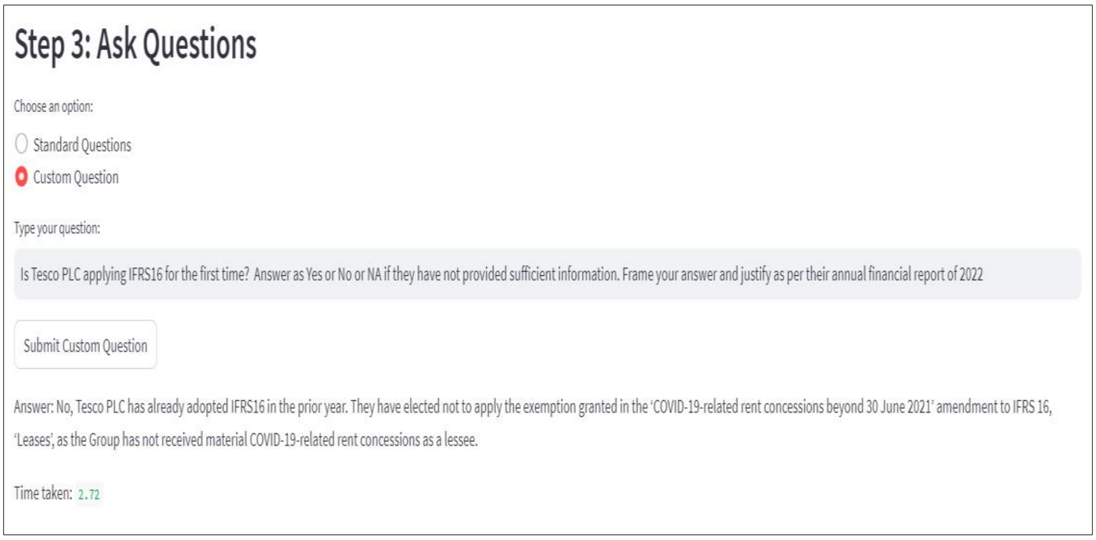

Figure 10: User Interface - Selecting Custom Questions

The user interface maintains a chat history to preserve a record of all user queries and the corresponding AI responses during the entire interaction. This chat history feature allows users to review their previous conversations, providing continuity and context to their auditing process. Users can easily track the conversation flow, and access past queries and responses.

In addition to the question options, the user interface includes detailed information about prompting styles that users can follow when framing their questions. These styles offer guidelines on how to structure queries effectively, considering the context of the IFRS 16 (Auditing Standard).

Figure 11: User Interface - Chat History of Previous Questions

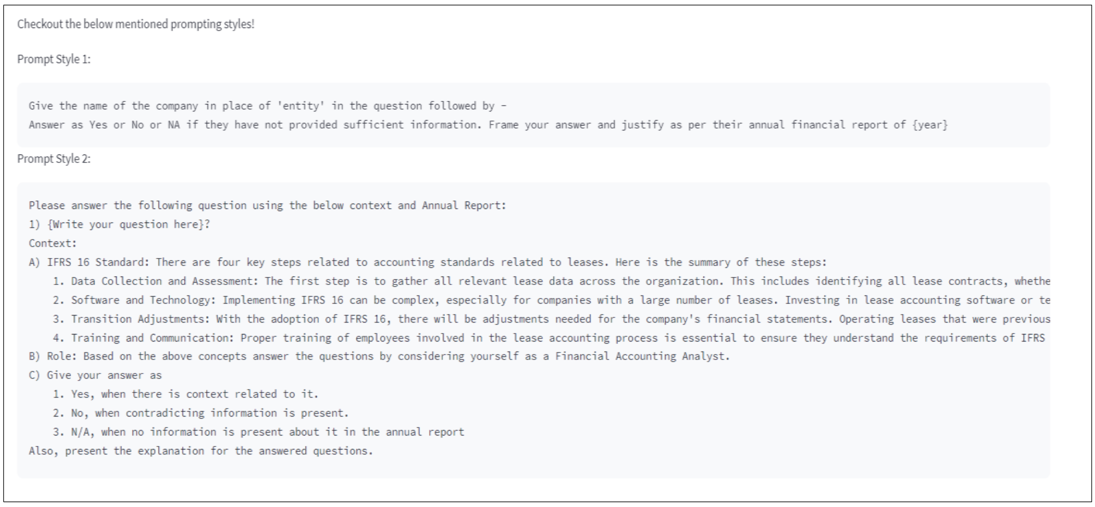

Figure 12: User Interface - Different Prompt Styles

Users can copy the prompting styles with one click. Following the prompting styles enhances the accuracy and relevance of the AI-generated answers, ensuring that users obtain valuable and contextually accurate insights from the annual report.

3.5 Software Development

3.5.1 High-Level Architecture

3.5.1.1 Front End

3.5.1.1.1 User Interface (Streamlit Web Framework)

Allows users to interact with the system, by uploading the pdf and getting the response of the software.

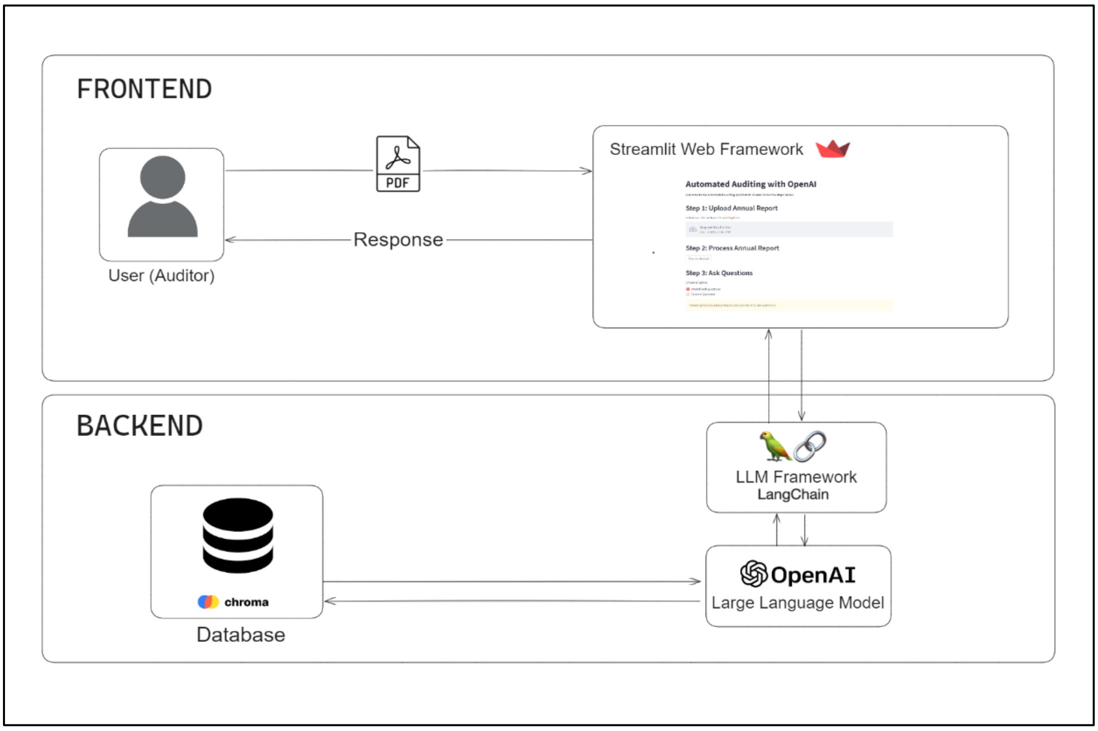

Figure 13: Software Architecture - High-Level Architecture

3.5.1.1.2 PDF Uploading and Processing Initiation

Users can use the Streamlit interface to upload PDF files locally. The application utilises libraries for PDF processing such as the "PyPDFLoader" to extract text and relevant information from the PDFs.

3.5.1.1.3 Query Input and Answer Display

After entering the query, users can trigger a process that analyses the query and extracts relevant information from the uploaded PDFs. The software's response to user queries is generated based on the processed Pdf content and displayed within the Streamlit app.

3.5.1.2 Backend

3.5.1.2.1 Embedding Engine using OpenAI's GPT 3.5 model's API and LangChain

The system converts textual data into embeddings, which are numerical representations that capture the semantic meaning of the text. This process is adept at transforming chunks of PDF content into these embeddings using the OpenAI's GPT 3.5 "text-embedding-ada-002" model's API (default). Additionally, it can convert user queries into embeddings, facilitating a semantic search. To provide the most accurate results, it performs a vectorised search to identify the most similar answers to user questions. Once these potential answers are identified, the system ranks and selects the best-matched records from its vector database to present to the user.

3.5.1.2.2 Vector Database (ChromaDB)

The primary purpose of the system is to efficiently store and retrieve embeddings. It achieves this by storing embeddings of document content, ensuring that the vast amounts of textual data are represented in a compact and meaningful manner. Beyond mere storage, the system also supports semantic search. This allows users to find similar embeddings based on their query embeddings, ensuring that results are not just exact matches but also contextually relevant and semantically similar to the user's intent.

3.5.1.3 Data Flow

3.5.1.3.1 Document Uploading

When using the Web App, the process begins with the "Document Upload" step, where the user uploads a PDF document. Once uploaded, the "Content Extraction" phase is initiated, during which the Document Processor meticulously extracts the content and divides it into manageable chunks. These chunks then undergo the "Embedding Transformation" stage, where the Embedding Engine transforms each chunk into its corresponding embedding. After this transformation, the embeddings are securely stored in a specialised database named "ChromaDB".

3.5.1.3.2 Query Processing

When a user poses a question through the Web App, it is swiftly converted into an embedding using the advanced Large Language Framework. Utilising this embedding, a search is conducted within "ChromaDB" to identify similar content embeddings that might hold the answer. This leads to the "Answer Generation" phase, where a sophisticated Conversational Agent takes the helm. It ranks the search results, meticulously selects the best matches, and crafts a comprehensive answer. Finally, in the "Displaying Answer" stage, this well-formed response is elegantly presented to the user via the Graphical User Interface, ensuring clarity and ease of understanding.

3.5.2 Low-Level Architecture

3.5.2.1 Initialisation and Configuration

To initiate the "Environment Setup", it's crucial to utilise the load_dotenv function. This function is specifically designed to load environment variables from a .env file. The significance of the .env file is that it typically houses sensitive information, ensuring that such details are not hardcoded directly into scripts, thereby enhancing security. Once this setup is complete, the next step involves extracting the OPENAI_API_KEY from the loaded environment variables. This particular API key is a gateway to OpenAI's vast array of services, in our case we used text-davinci-003 which is the default model of GPT 3.5 provided. Given its importance and the access it grants, it's imperative that this key remains confidential and is treated with the utmost care.

3.5.2.2 Session State Setup

To kick off the process with the Streamlit application, it's vital to initialise the session state variables. Session state acts as a powerful tool within the Streamlit framework, serving a pivotal role in ensuring a seamless user experience. It allows developers to store and manage user-specific data, ensuring that as users interact with the application, their preferences, inputs, and other relevant data are consistently maintained. This capability not only enhances user engagement but also ensures continuity between different interactions with the application.

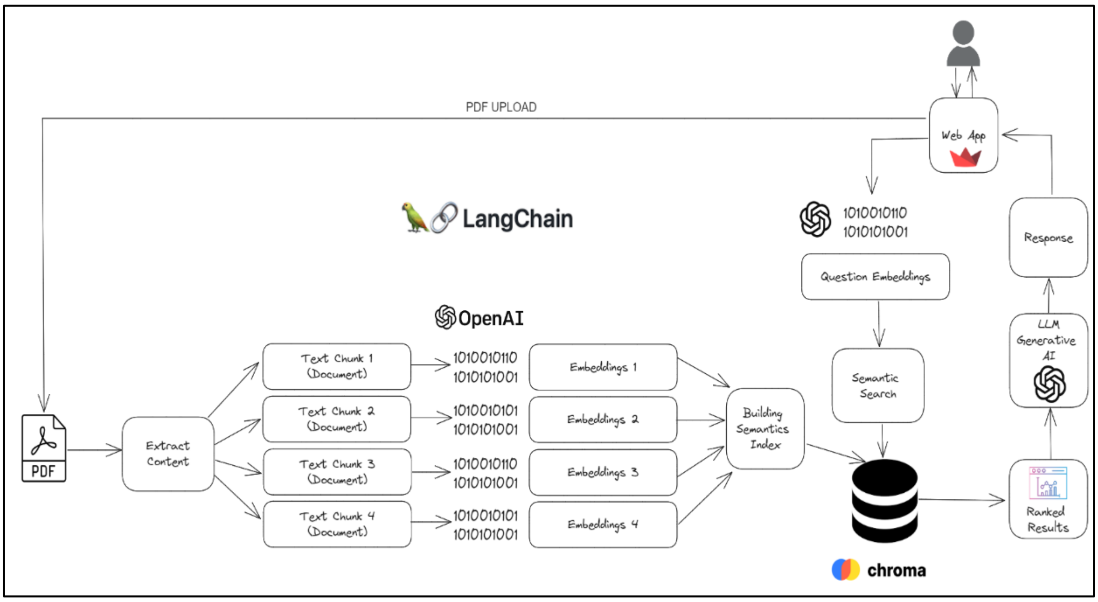

Figure 14: Software Architecture - Low-Level Architecture

3.5.2.3 Core Functionalities

The PyPDFLoader is a specialised utility designed to efficiently handle PDF documents. By loading the contents of a provided PDF and segmenting it into manageable chunks, it simplifies the process of content analysis and processing. On the other hand, the OpenAI GPT 3.5 “text-embedding-ada-002” embeddings model plays a critical role in the transformation of raw text content. It converts this content into embeddings - numerical representations of text - which streamline search, comparison, and analytical tasks. However, while these utilities enhance content processing, it's essential to be mindful of potential roadblocks like API rate limits. Recognising this, the Rate Limit Handling section has been incorporated. APIs, including OpenAI’s GPT 3.5 models’, often implement rate limits as a safeguard against misuse. Our system ensures that the number of requests made to the OpenAI’s GPT 3.5 “text-davinci-003” model’s API (default) remains within permissible bounds. It cleverly divides the document into chunks suitable for processing without surpassing the rate limit. Additionally, it intelligently calculates the necessary delays between consecutive requests, ensuring the application operates smoothly without encountering rate limit constraints.

3.5.2.4 Vector Database Management

Chroma stands as an indispensable utility tailored for the meticulous storage and management of document embeddings. By housing these numerical representations of the document's content, Chroma ensures that searching and retrieving pertinent information becomes a streamlined and efficient process. Its design focuses on maximising the potential of embeddings, allowing users to delve deep into the document and access information with unparalleled precision and speed.

3.5.2.5 Conversational Handling

The Conversational Retrieval Chain serves as a dynamic sequence of operations, meticulously designed to fetch the most pertinent information aligned with user queries. By interfacing directly with the vector database, it ensures the delivery of accurate and contextually relevant responses. Complementing this chain is the Conversation Buffer Memory, a pivotal utility crafted for preserving the history of interactions. By storing past conversations, it provides the system with the capability to reference previous exchanges. This not only aids in delivering more coherent replies but also empowers the system to offer context-aware responses, enhancing the overall user experience and making interactions feel more intuitive and personalised.

3.5.2.6 Error Handling and Feedback

At the heart of the application lies a robust system of checks and feedback mechanisms. The "File Upload Checks" step ensures that before any processing takes place, the application diligently verifies the uploaded file's validity. It flags any incompatible file types or formats, guaranteeing that only the right files move forward in the processing chain. As the file undergoes processing, the "Processing Feedback" mechanism plays a pivotal role. Should any hitches arise - be it due to API rate limits or parsing challenges - the system swiftly captures these errors and presents them to the user. These feedback messages are more than mere notifications; they provide valuable guidance, directing users on potential next steps or corrective measures. Lastly, the "Query Handling" component ensures a seamless user experience. Before allowing queries, users are prompted to process a document, ensuring the system has a data repository to scour. And if any wait times emerge, whether due to rate limits or inherent processing durations, users are kept in the loop, ensuring transparency and managing expectations effectively.

3.5.2.7 Cleanup and Resource Management

Ensuring efficient use of resources and maintaining user data privacy is paramount in the system's operations. The "Temporary Files" protocol ensures that any uploaded files are saved only temporarily, strictly for processing purposes. Once their role in the processing is fulfilled, these files are promptly deleted. This not only aids in freeing up system resources but also underscores a commitment to user data privacy, ensuring that no residual data remains on the system. Complementing this approach is the "Database Management" strategy. Before embarking on a new processing session, any pre-existing databases are purged. This ensures that the system always operates with the freshest data, guaranteeing accuracy and preventing any potential data overlap or redundancy. Together, these measures reflect a dedication to efficiency, accuracy, and user trust.

3.6 Implementation

This part explains the design and requirements for a functional system that can process PDF documents, interact with OpenAI's GPT 3.5 models', and provide conversational capabilities through a web interface.

3.6.1 Software and Tools Used

We have used different software tools and libraries to achieve the project objectives, the tools used are as follows:

3.6.1.1 Python as Primary Programming Language

Python is chosen as the primary programming language for development due to the ease of use, readability, and wide range of libraries that facilitates various tasks including data processing, and natural language processing.

3.6.1.2 Streamlit for Web Application Interface

Streamlit, a Python library, is used to create the web application's user interface. With Streamlit, you can easily convert data scripts into interactive web applications. It provides widgets and tools for creating dynamic visualisations, input fields and interactive components.

3.6.1.3 OpenAI's GPT 3.5 models' for Natural Language Processing

OpenAI's advanced language models offer a range of capabilities that are pivotal for extracting meaningful insights from textual data. These capabilities can be grouped into three main categories:

3.6.1.3.1 Embeddings

Embeddings refer to the representation of words or phrases in a numerical vector space. OpenAI's GPT 3.5 models are trained on vast amounts of text, enabling them to capture the semantic essence of words and phrases. By transforming textual data into these embeddings, we can capture the contextual and semantic relationships between different pieces of information. This numerical representation is vital for tasks like similarity measurements and clustering (OpenAI Platform, 2022).

3.6.1.3.2 Semantic Search

Semantic search is the ability to understand the intent and contextual meaning of search queries. Instead of merely matching keywords, semantic search delves into the deeper layers of language, ensuring that the results are contextually and semantically relevant to the query. OpenAI's GPT 3.5 models excel in this domain, allowing for more nuanced and precise searches, especially when navigating complex documents like annual reports (Kyriacou, K.,2023).

3.6.1.3.3 Natural Language Generation

One of the flagship capabilities of OpenAI's GPT 3.5 model's is Natural Language Generation (NLG). NLG involves generating coherent and contextually relevant text based on certain input parameters. This feature can be harnessed to provide explanations, generate summaries, or even answer queries in a human-like manner. Given a specific prompt or question, the model can produce detailed and accurate responses, making it invaluable for tasks like interpreting IFRS standards (OpenAI Platform, 2022).

In essence, OpenAI's GPT 3.5 model's capabilities in embeddings, semantic search, and natural language generation provide a comprehensive toolkit for navigating and understanding the complexities of financial documents, ensuring accuracy, efficiency, and depth in our auditing processes.

3.6.1.4 "dotenv" for Managing Environment Variables

The "dotenv" library is used to manage environment variables. It ensures that sensitive information, such as API keys or other configuration details, remains secure and separate from the source code.

3.6.2 Other Essential Libraries

Libraries like **os, shutil,** and **tempfile** are utilised for file operations. **os** provides a way to interact with the operating system, **shutil** assists in file operations, and **tempfile** is used for creating temporary files and directories.

3.6.3 Chroma Vector Database

The chosen database for this project is Chroma, optimised for storing and retrieving vector embeddings. Vector embeddings are numerical representations of data points, often used in machine learning and data analysis. The decision to use Chroma could be due to its compatibility with embeddings and its efficient mechanisms for storing and retrieving such data.

3.6.4 Code Implementation

The code is organised in a modular structure, emphasising clarity and ease of maintenance. Key components included are:

3.6.4.1 Document Embedding and Database Creation (process_documents)

This function is pivotal, in processing uploaded PDFs, dividing them into chunks, and storing embeddings into Chroma. It also starts by initialising embeddings and loading the document. It then calculates the chunk size, ensuring efficient processing without exceeding API limits.

```python
def process_documents(pdf_file_name, vectordb_directory):
    """
    Process and load documents into a vector database.

    Args:
    - pdf_file_name (str): Path to the PDF file to process.
    - vectordb_directory (str): Directory for storing vector data.

    Returns:
    - Chroma: Vector database with the processed documents.
    """

    # Check and clean up any previous database instances
    if persist_directory and os.path.exists(persist_directory):
        shutil.rmtree(persist_directory)

    # Initialize embeddings and document loader
    embeddings = OpenAIEmbeddings(openai_api_key=OPENAI_API_KEY)
    loader = PyPDFLoader(pdf_file_name)
    pages = loader.load_and_split()

    if not pages:
        return "not_parsed"
    else:
        total_pages = len(pages)
        chunk_size = calculate_chunk_size(total_pages)

    # Initialize Chroma for vector storage and set up embedding function
    vectordb = Chroma(persist_directory=vectordb_directory, embedding_function=embeddings)

    # Split pages into manageable chunks
    chunks = [pages[i:i + chunk_size] for i in range(0, len(pages), chunk_size)]

    # Delay time to respect API rate limits
    delay_time = 30  # Increase this if you're still hitting rate limit

    # Process chunks and handle rate limit exceptions
    for i, chunk in enumerate(chunks):
        print(f"Processing chunk {i+1} of {len(chunks)}")
        while True:
            try:
                vectordb.add_documents(chunk)
                vectordb.persist()
                break
            except Exception as e:
                if 'Rate limit reached' in str(e):
                    print(f"Rate limit reached. Error: {e}. Sleeping for {delay_time} seconds.")
                    time.sleep(delay_time)
                else:
                    raise # If it's a different error, raise it
        # Sleep for 30 seconds after processing each chunk
        print(f"Sleeping for 30 seconds after processing chunk {i+1}")
        time.sleep(30)

    return vectordb
```

Figure 15: Implementation - Creation & Storing of Embeddings

```
In [16]:  df['document'][0]
executed in 16ms, finished 16:38:31 2023-08-08
Out[16]: 'Marks and Spencer Group plc   Annual Report & Financial Statements 2023\nMarks and Spencer Group plc
\nAnnual Report & Financial Statements 2023Reshaping'

In [19]:  for i in df['embedding'][0]:
              print(i)
executed in 249ms, finished 16:39:15 2023-08-08
-0.01577959612238528
-0.003240916008337816
-0.005855709425781834
-0.0040454679007300215
0.005148482141272919
0.01985426098638512
-0.009563784297017514
-0.001805375131676195
-0.01019315161582302
-0.0227610286931403
0.031092031519227507
0.01581852540674504
-0.02013974675119546
0.007779496404307701
-0.011322119488707918
0.0166360534168163
0.018945895495947506
-0.0048500193274609305
0.008694349022262252
```

Figure 16: Implementation - Example of Single Embedding

3.6.4.2 Database using Chroma DB

The generated Database is stored in Chroma DB for storing and retrieving vector embeddings. It stores the data locally in the local file system when creating applications around Chroma. It performs tasks like vector embedding, retrieving them, and performing a semantic search for vector embedding.

	collection_uuid	embedding	document	id
0	304c0e8b-	1be09cb4-	[0.01333631 -0.02464686 -0.01606331 ... 0.00211505 -0.00655129	Serving	a11f948d-
1	304c0e8b-	d089eea9-	[-0.00154864 -0.02789945 -0.02207751 ... 0.01946592 -0.00394605	Serving our	a11f948e-
2	304c0e8b-	afeadcf1-6	[0.00224226 -0.03061258 -0.01470399 ... 0.02820338 0.00051883	1 Tesco PLC Annual Report and Financial Statements 2022	a11f948f-3
3	304c0e8b-	6a42e437-	[0.00666085 -0.0112467 -0.01011487 ... 0.020581 -0.03491745	Tesco was built to be a champion for customers, serving	a11f9490-
4	304c0e8b-	591f1a01-	[-0.02496547 -0.01453802 -0.0266596 ... 0.0216166 -0.01162253	Î" Alternative performance measures (APMs)	a11f9491-
5	304c0e8b-	b2bde166-	[-0.00065617 -0.02806642 0.00130825 ... 0.01311686 -0.01202379	Tesco at a glance	a11f9492-
6	304c0e8b-	a1c3353b-	[0.00505381 -0.01087383 0.00560612 ... 0.01961503 -0.01572012	Our purpose	a11f9493-
7	304c0e8b-	dd100c53-	[-0.00488173 -0.02242958 0.00099201 ... 0.02378855 -0.02278582	Our purpose in action	a11f9494-
8	304c0e8b-	ae2460d4-	[0.00405088 -0.03167528 -0.00193724 ... 0.01266227 -0.02290708	Serving our	a11f9495-
9	304c0e8b-	496ce792-	[0.00517599 -0.03014809 -0.00627976 ... 0.04031066 -0.02396956	Our purpose in actionOur purpose in action continued	a11f9496-
10	304c0e8b-	eb7ef004-	[-0.01838822 -0.02457047 -0.01417426 ... 0.0251385 -0.00879781	Chairmanâ€™s statement	a11f9497-
11	304c0e8b-	002f9074-	[-0.00706164 -0.02196804 -0.03272927 ... -0.00442871 -0.0052861	and strong partnerships we have built with suppliers. Many	a11f9498-
12	304c0e8b-	07263a3b-	[-0.02123347 -0.0246969 -0.00286565 ... 0.00945782 -0.01185558	The momentum in our business is strong	a11f9499-
13	304c0e8b-	d37da07b-	[-0.01972832 -0.01793238 -0.00507049 ... 0.02584531 -0.00849357	on health and sustainability. This strategic priority is all	a11f949a-
14	304c0e8b-	a0d210f5-	[-0.01222228 -0.02276105 -0.00420204 ... 0.01959605 -0.01861288	Taken together, these priorities will ensure we do the	a11f949b-
15	304c0e8b-	e4542e02-	[-0.00096095 -0.01998383 -0.01339633 ... 0.01137191 -0.00948418	Our sustainability efforts are grounded in doing the	a11f949c-

Figure 17: Implementation - Embeddings in Chorma DB

3.6.4.3 Conversational Retrieval (get_ conversation_ chain)

It is responsible for implementing a retrieval mechanism that obtains pertinent answers in response to user queries. This mechanism leverages the vector embeddings that have been stored in the Chroma database. Furthermore, it interfaces with the OpenAI's GPT 3.5 Language Model (LLM) to construct a sequence of conversation interactions, often referred to as a "conversation chain."

```python
def get_conversation_chain(vectorstore):
    """
    Create a retrieval conversation chain.

    Args:
    - vectorstore (Chroma): Vector database with processed documents.

    Returns:
    - ConversationalRetrievalChain: Initialized conversational retrieval chain.
    """
    memory = ConversationBufferMemory(memory_key='chat_history', return_messages=True)
    conversation_chain = ConversationalRetrievalChain.from_llm(OpenAI(temperature=0),
                                                              retriever=vectorstore.as_retriever(),
                                                              memory=memory)
    return conversation_chain
```

Figure 18: Implementation - Code for Conversation Retrieval

3.6.4.4 Question Answer Mechanism (question_answer)

Given a user query, this function retrieves the most relevant answer from the database. The function converts the user's question into embeddings, performs a semantic search on the database, and then formulates a response using OpenAI's GPT 3.5 LLM.

3.6.4.5 Database and Code Integration

The seamless integration between the codebase and the Chroma database is nothing short of exemplary. Central to this flawless synergy is the "process_documents" function, which stands out as the linchpin in the entire integration mechanism.

Entrusted with a critical responsibility, this function ensures that embeddings, once extracted from the OpenAI's GPT 3.5 "text-embedding-ada-002" model's API (default), are accurately channelled into the Chroma database. But the function's prowess doesn't end there. The retrieval process, an equally vital component, has been fine-tuned to perfection. Through meticulous optimisation, it guarantees swift returns on user queries, ensuring that each response is not only prompt but also contextually apt. In a nutshell, the "process_documents" function sets the rhythm, orchestrating an impeccable flow of information that amplifies both the efficiency and effectiveness of the system.

```python
def question_answer(query, vectordb, chat_history):
    """
    Retrieve the answer for a given query from the vector database.

    Args:
    - query (str): User query.
    - vectordb (Chroma): Vector database.
    - chat_history (list): History of previous conversations.

    Returns:
    - tuple: Time taken and the AI-generated answer.
    """
    qa = get_conversation_chain(vectordb)
    time.sleep(20)  # Sleep for 20 seconds before processing the query
    start = time.time()
    res = qa(query)
    answer = res['answer']
    end = time.time()

    chat_history.append((query, answer))  # Append the user query and AI response to the chat history
    return (round(end - start, 2),answer)
```

Figure 19: Implementation - Code for Question Answer Mechanism

3.6.5 Prompt Engineering

Prompt engineering emerges as both an art and a science, playing a pivotal role in the realm of query formulation. Its essence lies in crafting queries that optimise the accuracy, relevance, and contextual awareness of a language model's responses. Given the profound reliance of our system on OpenAI's GPT-3.5 for generating document embeddings and fielding user queries, the importance of adept prompt engineering cannot be overstated. However, it's essential to recognise that while GPT-3.5 boasts immense power and capability, its generality can be a double-edged sword. Absent well-engineered prompts, the model might veer towards responses that lack precision, stray from contextual relevance, or are outright erroneous. This underscores the significance of guiding the model with meticulously structured prompts. By doing so, we can effectively navigate its vast knowledge base, channelling it to yield the most desired and fitting responses.

3.6.5.1 Contextual Information

In our approach to harnessing the full potential of the model, we've incorporated strategies that feed it with enriched contextual information. This is particularly crucial when dealing with specific types of documents or zeroing in on particular sections of a report that pique a user's interest. Take, for example, a query about a specific financial standard like "IFRS 16". In such a scenario, the prompt isn't merely framed around the standard itself. Instead, it's meticulously structured to elucidate that the context in question pertains to lease contracts. By doing so, we guide the model, narrowing its focus and directing it to extract and provide information that is directly relevant to the user's inquiry. This level of precision in prompt crafting ensures that users receive answers that are not just accurate but also contextually apt, thereby elevating the overall user experience.

3.6.5.2 Explicit Instructions

A key strategy in optimising the model's response lies in providing it with clear and direct instructions concerning the format or nature of the expected answer. This ensures that the model's output aligns closely with the user's intent. For instance, if a query is crafted in a manner that anticipates a binary 'Yes' or 'No' answer, we make it a point to explicitly instruct the model to respond within that specific format. This principle shines particularly in the custom questions section. Here, users have the liberty to frame their questions, and to guarantee the relevance and clarity of the model's answers, it's equipped with unambiguous contextual cues. By doing so, we ensure that the model remains attuned to the user's needs, delivering responses that are not only accurate but also presented in the desired format, enhancing user satisfaction and system reliability.

3.6.5.3 Iterative Refinement

The journey to perfecting our query prompts was an iterative one, characterised by continuous refinement based on the model's performance. The initial prompts we used leaned towards being broad and general. However, as we observed the model's responses and gauged the precision of the retrieved information, it became evident that tweaks and refinements were necessary. For instance, when confronted with an overgeneralised answer concerning lease contracts, we didn't settle for the broad strokes. Instead, we honed the prompt, steering it towards more specific territories like "operating leases" or "financial leases." This approach of iterative refinement ensured that the model's outputs progressively improved in accuracy and relevance, aligning more closely with the user's intended query, and providing answers with the desired depth and specificity

3.6.5.4 Prompt Styles

To optimise the interface between the user's queries and the model's responses, we embarked on an explorative journey testing various prompting styles. Our goal was to discern which style consistently drew out the most accurate and comprehensive answers from the model. As a testament to this endeavour, the system now presents users with two distinct prompt styles. Each of these styles is structured in its unique way, designed to shape and guide the model's response in a particular direction. By offering this variety, we empower users to select the approach that best aligns with their query needs, ensuring that the model's output is both precise and rich in detail. This dual-style mechanism underscores our commitment to delivering an adaptive and user-centric experience.

3.6.5.5. Prompt Examples

3.6.5.5.1 Contextual Information

- Original Question: "Tell me about the leases."
- Enhanced Prompt: "Based on the Annual Financial Report for 2022, provide details about the entity's engagement with lease contracts, specifically referring to the IFRS 16 standards."
- Rationale: By specifying the context (Annual Financial Report for 2022) and the standard (IFRS 16), the model will provide a more tailored and accurate response about lease contracts.

3.6.5.5.2 Explicit Instructions

- Original Question: "What can you say about lease contracts?"
- Enhanced Prompt: "Provide a 'Yes' or 'No' answer: Does the entity in the Annual Report follow the IFRS 16 standards for lease contracts?"
- Rationale: The model is explicitly directed to provide a binary response, making the information more straightforward for the user to interpret.

3.6.5.5.3 Iterative Refinement

- Original Prompt: "Tell me about the entity's leases."
- Enhanced Prompt: "Elaborate on the entity's operating leases as mentioned in their Annual Financial Report for 2022, especially in context to the IFRS 16 standards."

- Rationale: The refined prompt narrows down the focus to "operating leases" and relates it to a specific financial standard, ensuring a more detailed response.

3.6.5.5.4 Prompt Styles

- Style 1: "Considering the Annual Financial Report of 2022, does the entity apply IFRS 16 standards for their lease contracts? Justify your answer based on the report."
- Style 2: "Please answer the following question using the Annual Report of 2022 as a reference: Does the entity comply with IFRS 16 standards when accounting for lease contracts? Provide a 'Yes', 'No', or 'N/A' response and justify your answer."
- Rationale: Different prompt styles guide the model's response in varied ways, with one providing a more open-ended approach and the other a structured binary response.

3.7 Software Testing

3.7.1 Functional Testing

Functional testing focuses on evaluating the overall functionality of the application to ensure that it meets its specified requirements and performs its intended functions correctly. The testing scenarios covered various use cases, including successful file processing, error handling, AI-generated answers, chat history management, and no data extraction.

3.7.1.1 Functional Testing Scope

The functional testing of the Automated Auditing Application covered the following scenarios:

- Uploading a valid pdf file and verifying successful processing and data extraction.
- Uploading an invalid file format (e.g., non-pdf file) and confirming appropriate error handling.
- Asking standard questions and custom questions to validate ai-generated answers.
- Testing the behaviour when no data is extracted from the report.
- Verifying the updating and clearing of the chat history.

3.7.1.2 Test Environment

The test environment was set up with the necessary dependencies and libraries, including `streamlit`, `langchain`, and other modules used in the application.

3.7.1.3 Functional Test Cases

The functional test cases were designed to validate different aspects of the application's functionality and user interactions. Each test case covers specific scenarios related to the functional requirements.

- **Test Case 1: Uploading a Valid PDF File and Successful Processing**

Description: This test case validates the application's ability to process a valid PDF file successfully and extract information from it.

Steps:

a) Upload a valid PDF file to the application.

b) Click on the "Process Report" button to initiate the processing.

c) Observe the application's response.

Expected Result:

The application should display a success message indicating that the file was processed successfully. The vector database should be updated with the extracted information.

Figure 20: Testing Result - No Database before Processing the Annual Report

Figure 21: Testing Result - Database after Processing the Annual Report

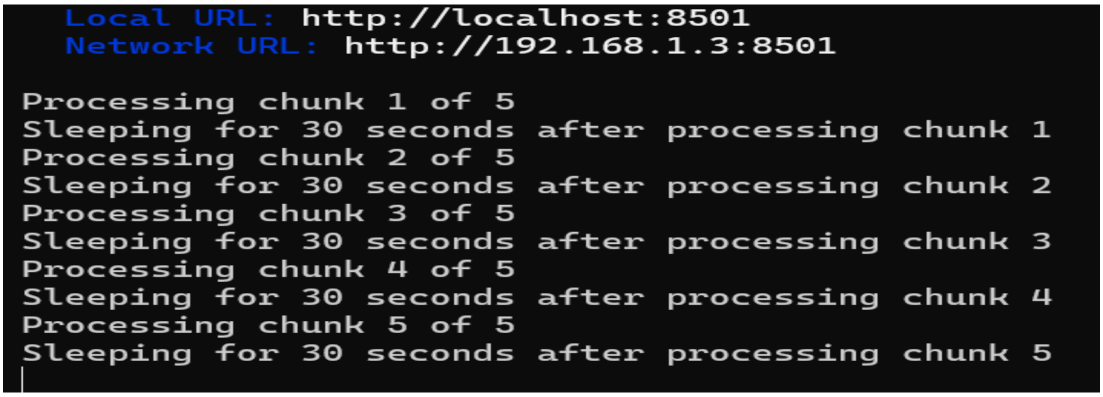

Figure 22: Testing Result - Processing of Annual Report at the Backend

Figure 23: Testing Result - Post Processing Message

- **Test Case 2: Uploading an Invalid File Format and Error Handling**

Description: This test case assesses the application's response when attempting to upload an invalid file format (e.g., a non-PDF file).

Steps:

a) Attempt to upload a file in a format other than PDF (e.g., a Word document or an image file).

b) Observe the application's response.

Expected Result: The application should display an appropriate error message, informing the user that only PDF files are accepted for processing.

Result: Message Displaying that document type is not allowed.

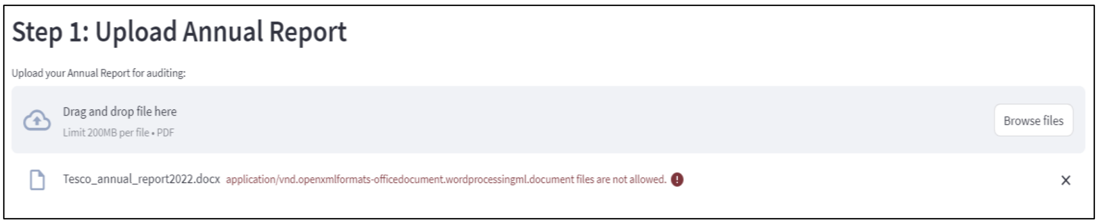

Figure 24: Testing Result - Message Indicating that the Document Type is not Permitted

- **Test Case 3: Asking Standard Questions and Custom Questions**

Description: This test case verifies the AI-generated answers for both standard and custom questions.

Steps:

a) Upload a valid PDF file and process it.

b) Ask a standard question from the provided list.

c) Ask a custom question related to the processed data.

Expected Result: The application should display relevant AI-generated answers for both standard and custom questions, along with the time taken to generate the response.

4. Results:

Standard Question Response:

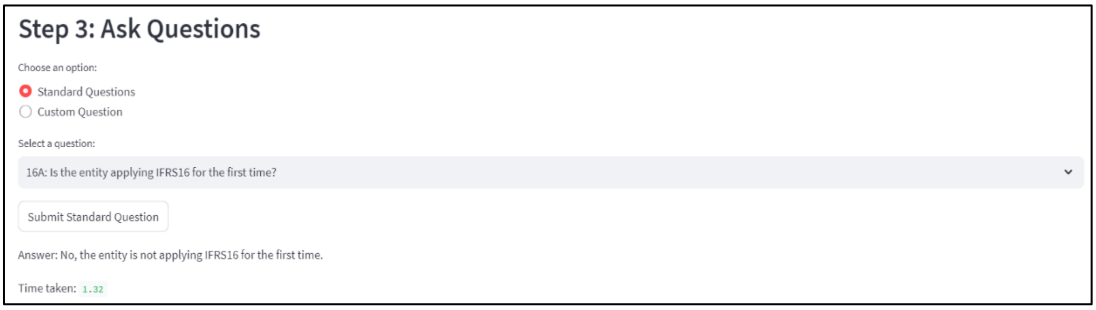

Figure 25: Testing Result - Standard Question Response

Custom Question Response:

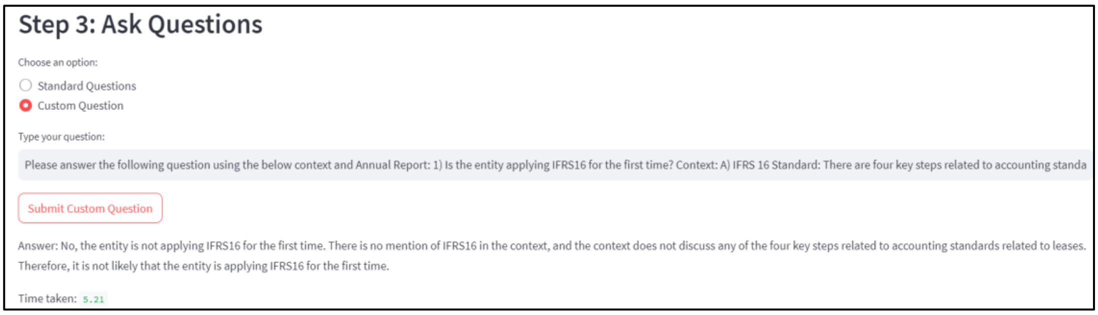

Figure 26: Testing Result - Custom Question Response

- **Test Case 4: No Data Extraction from the Report**

Description: This test case evaluates the application's behaviour when no data is extracted from the uploaded report.

Steps:

a) Upload an empty PDF file.

b) Process the report.

Expected Result: The application should handle the situation gracefully and inform the user that no data was extracted.

Result:

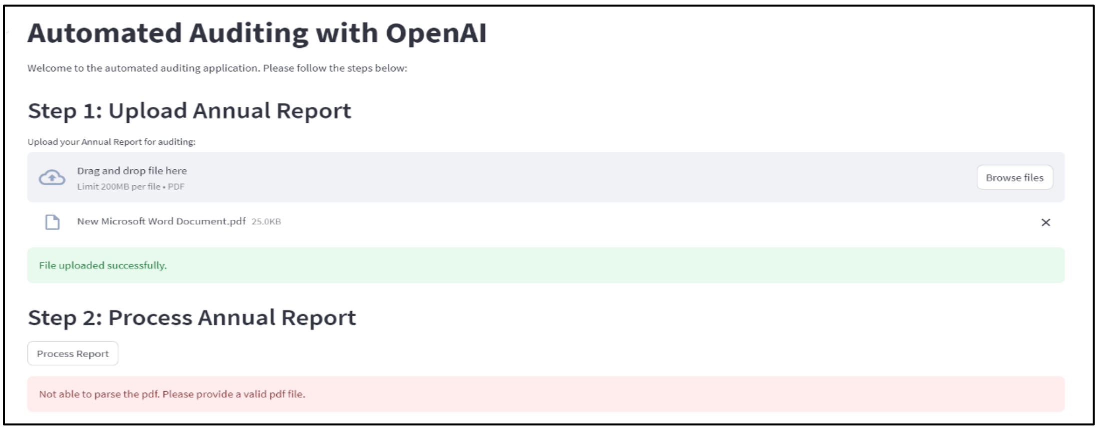

Figure 27: Testing Result - No Data Extraction Error Message

- **Test Case 5: Verifying Chat History Updates and Clearing**

Description: This test case ensures that the chat history is updated with user queries and AI responses. It also checks the functionality to clear the chat history.

Steps:

a) Upload a valid PDF file and process it.

b) Ask a few questions using the application's interface.

c) Verify that the chat history is correctly updated with user queries and AI-generated responses.

d) Click on the "Clear Chat History" button.

Expected Result:

The chat history should be updated with the user's queries and corresponding AI responses. After clicking the "Clear Chat History" button, the chat history should be emptied.

Results:

Chat History Results:

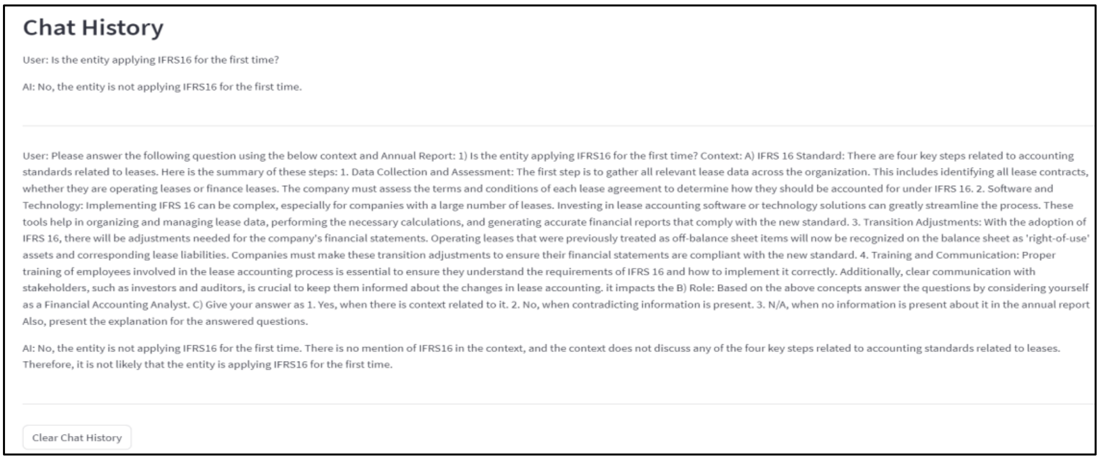

Figure 28: Testing Result - Chat History

After pressing clear chat history: The chart history disappeared successfully.

Figure 29: Testing Result - Chat History Results after clicking Clear Chat History Button

All functional test cases were executed, and the application performed as expected, meeting the specified requirements. The functional testing process verified the correct behaviour of the application under various scenarios.

3.8 Performance Evaluation Metric

3.8.1 Evaluation Strategy

Our primary strategy for evaluating the performance of our software solution was a direct comparison approach. We juxtaposed the outputs generated by the software against the actual answers provided by domain experts. This comparative method offered a clear and direct way to gauge the accuracy and reliability of the tool's responses. While we didn't employ traditional evaluation techniques, this direct comparison approach served as a robust method, given the unique nature of our tool and its application in the auditing domain.

3.8.2 Success Metrics

The primary metrics chosen to assess the tool's performance were accuracy and the F1 score. The accuracy provided a straightforward measure of how often the tool's responses matched those of the experts. Precision ensures that our tool's interpretations are trustworthy, and recall ensures thoroughness, making certain that no relevant compliance is overlooked. The F1 score, being the harmonic mean of precision and recall, offered a balanced measure of the tool's ability to correctly identify relevant responses while avoiding false positives.

The Rationale for Choosing Metrics: Choosing these metrics was particularly pertinent as our solution is aimed at assisting auditors in answering IFRS standards questions. Given the critical nature of financial auditing, the emphasis was on ensuring that the tool's outputs were both accurate and reliable, mirroring the precision expected in the auditing profession.

3.8.3 Comparative Analysis

While there aren't existing tools in the market that perform the exact function as ours, we validated our tool's outputs by comparing them with manual audits. This comparison served as the gold standard, given that manual audits by experts are currently the most trusted method for interpreting and answering IFRS standards questions.

3.8.4 Thresholds and Benchmarks

While we didn't have a fixed benchmark for our tool's performance, our aspirational target was to achieve an accuracy rate of 60% or higher. This threshold was set to ensure that our tool would provide a significant value proposition, offering reliable insights and responses for a majority of queries.

3.8.5 Feedback and Iteration

The iterative nature of our development process was evident in our approach to refining the tool's performance. Based on the evaluation results, we revisited our prompting style and provided additional context, optimising the tool to generate more accurate and contextually relevant responses.

Chapter 4 - Process Flow

This part will describe the sequence of steps or stages involved in a particular process or system. For your Automated Auditing with OpenAI's GPT 3.5 model's application, the process flow can be described as follows:

4.1 Initialisation

- Begin by loading the environmental variables that store configuration settings. This is followed by initialising session states by providing a starting point for user interactions
- Customise the Streamlit application by configuring page settings.

4.2 User Interface Display

- Display Welcome Message: Present users with a warm welcome message along with concise instructions on how to navigate and utilise the application effectively. This introduction sets a friendly tone and guides users through the process.
- Provide Users with the Option to Upload an Annual Report: Enable users to easily upload their annual reports through a user-friendly interface. This functionality empowers users to initiate the data processing and analysis workflow seamlessly.

4.3 Document Upload

- Users submit annual reports as PDFs: During this phase, users provide their annual reports by uploading files in PDF format. This step ensures that the necessary data is accessible for processing.
- Verify file type (PDF exclusively): The system validates the uploaded files to ensure they are exclusively in PDF format.
- Display appropriate success or error messages: Upon successful upload, users receive a confirmation message indicating the completion of the upload process. Conversely, if any issues arise during the upload, the system displays relevant error messages to guide users through resolving the problem.

4.4 Document Processing

- Splitting into Individual Pages: The uploaded document is segmented into individual pages to facilitate focused processing and analysis.
- Optimal Chunk Size Calculation: Calculating the optimal chunk size is crucial to ensure efficient document processing.

- Text-to-Vector Conversion: Utilising OpenAI's GPT 3.5 "text-embedding-ada-002" model's API (default), the text data extracted from the document is converted into vector embeddings.
- Embedding Storage in Chroma: The generated embeddings are stored within a vector database called Chroma.

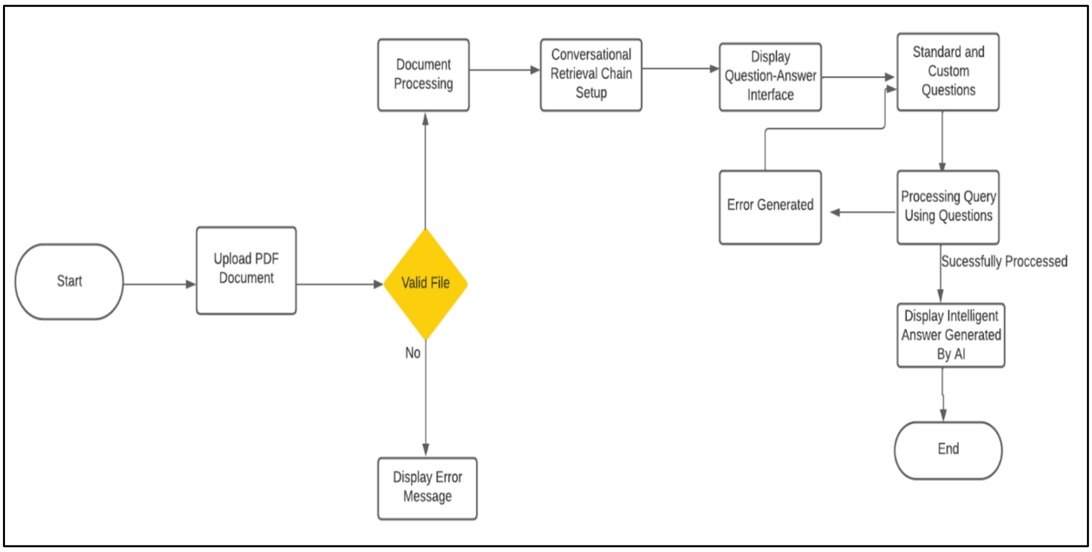

Figure 30: Process Flow Diagram

4.5 Conversational Retrieval Chain Setup

- Initialising Retrieval Chain: At this stage, the conversational retrieval chain is established by integrating the vector database and OpenAI's GPT 3.5 "text-davinci-003" language model. This fusion forms the foundation for generating responses to user queries.
- AI-Generated Answer Provision: The retrieval chain's purpose is to leverage the vector database's embeddings and the language model's capabilities to generate AI-driven answers in response to user queries. This dynamic interaction ensures that users receive contextually relevant and informative responses.

4.6 Question-Answer Interface

- Present users with choices to pick standard questions or provide custom queries.
- Channel user queries through the conversational retrieval chain.
- Showcase AI-generated answers on the interface.
- Retain a chat history for user queries and AI response monitoring.

4.7 Error Handling

- Check for rate limits from the OpenAI's GPT 3.5 models' API and handle them by delaying requests.

- Display appropriate error messages for any issues during processing or querying.

4.8 End Interaction

- Provide users with the option to start over or end their session.
- Clear session states and data for privacy.

Chapter 5 - Results and Findings

The findings derived from the application are showcased through the utilisation of standard inquiries aligned with IFRS 16, employing two distinct prompt formats. To assess the application's performance, examinations were conducted employing the 2022 annual report of Tesco PLC and the 2023 annual report of M&S. The testing encompassed all three prompt variants, leading to the following discernible findings:

Table 1: Tesco PLC Test Results

Question Type	Precision	Recall	F1 Score	Accuracy
Standard Question	1	0.6	0.75	0.63
Prompt Style 1	1	0.5	0.67	0.56
Prompt Style 2	1	0.8	0.89	0.86

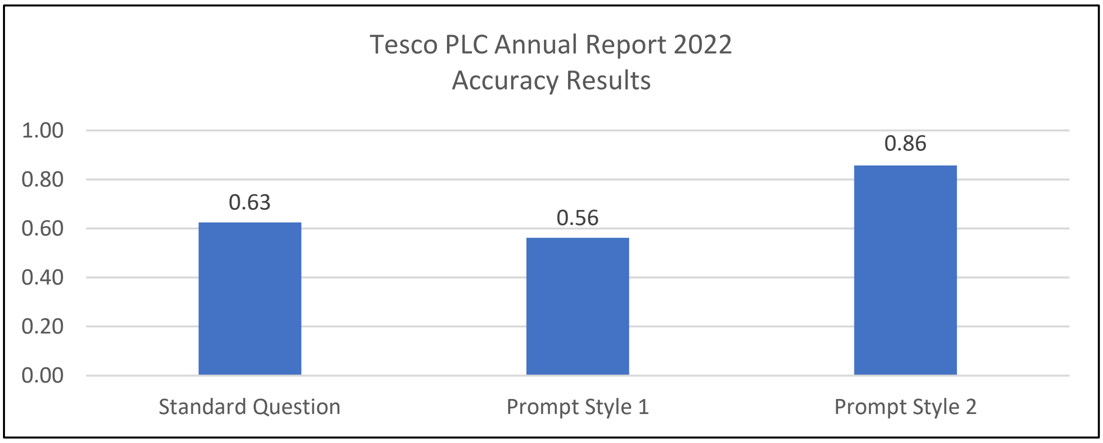

Figure 31: Metric Results of Tesco PLC Annual Report 2022

Table 2: M&S PLC Test Results

Question Type	Precision	Recall	F1 Score	Accuracy
Standard Question	1	0.6	0.75	0.63
Prompt Style 1	0	0.0	0.00	0.29
Prompt Style 2	1	0.4	0.57	0.57

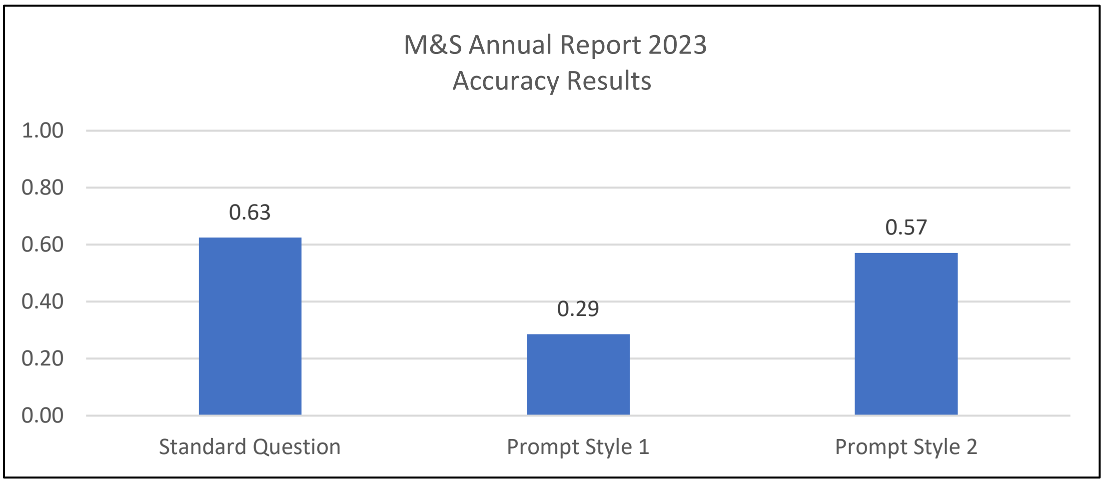

Figure 32: Metric Results of Marks and Spencer Annual Report 2023

5.1 Prompts (Questions Asked)

5.1.1 Standard Prompts

For both Tesco and M&S, the standard prompt style shows consistent results across both precision, recall, F1 score, and accuracy. The precision is perfect (1.0), meaning that the identified instances were all relevant. The recall is 0.6, indicating that 60% of the actual relevant instances were successfully identified. The F1 score is 0.749, which is the harmonic mean of precision and recall. This balanced score suggests reasonably effective performance in capturing relevant information. The accuracy, at 0.625, indicates that 62.5% of all instances were correctly classified.

5.1.2 Detailed Prompts

- **Prompt Style 1:** In this prompt style, the results vary depending on the type of information being evaluated. For Tesco, precision remains perfect (1.0), while recall drops slightly to 0.5, indicating that only half of the relevant instances were captured. This discrepancy leads to an F1 score of 0.667, which is slightly lower than the standard prompt's F1 score. The accuracy also decreases to 0.5625.

  For M&S, precision becomes undefined due to the absence of true positives (TP) and false positives (FP). As a result, the recall is also 0, the F1 score becomes undefined, and the accuracy drops significantly to 0.2857, indicating poor accuracy.
- **Prompt Style 2:** In this prompt style, the precision remains perfect (1.0) for both Tesco and M&S. However, there are variations in the recall values. For Tesco, recall remains at 0.8, indicating that 80% of the relevant instances were identified. This results in an F1 score of 0.889, which is higher than the F1 score from the standard prompt, suggesting better overall performance. The accuracy increases to 0.8571, implying good performance in identifying the responses to the standard's questions.

  For M&S, recall drops to 0.4, indicating that only 40% of the relevant instances were captured. This results in an F1 score of 0.571, which is lower than the standard prompt's F1 score. The accuracy remains the same as the F1 score at 0.571, reflecting the balance between correct and incorrect answers.

5.2 Results Interpretation

In interpreting the results of our study, we turn our attention to the efficacy of our application, a focus meticulously honed through the comprehensive analysis of outcomes yielded by the meticulous auditing of annual reports from renowned entities such as Tesco and M&S. These observations have a dual purpose: they not only shed light on our accomplishments to date, but they also serve as a crucial base upon which future improvements can be built.

5.2.1 Accuracy and Prompt Styles

The analysis reveals that Tesco's report achieved the highest accuracy when utilising "Prompt Style 2," while M&S's report yielded the best results with standard questions of IFRS 16. Notably, both reports achieved the same accuracy (63%) when using standard questions. Additionally, our application achieved an impressive average accuracy of 71.5% with "Prompt Style 2 for both reports." These results showcase the application's capability to adapt to different prompt styles and deliver competitive accuracy levels.

5.2.2 Scalability and Replication

The level of precision attained in this comprehensive analysis establishes a robust groundwork, indicating a high likelihood of replicating similar outcomes across diverse IFRS standards.

5.2.3 Prompts and Manipulation of Accuracy

The analysis additionally underscores the notable impact of prompts on the outcomes' accuracy. This recognition serves as a catalyst for our exploration into potential enhancements, specifically the conception of generalised prompts. The intention behind these prompts is to establish a uniform framework that consistently elicits comparable levels of accuracy across a spectrum of diverse annual reports.

5.2.4 Text Processing Improvement

The notable convergence of a 63% accuracy rate in response to standard questions across Tesco and M&S reports underscores the significance of refining our text processing capabilities. By systematically addressing the nuances inherent in text processing, we unlock a promising avenue for substantial and impactful accuracy improvements.

5.3 Overall Implications

The core objective of our research was to pave the way for an advanced IFRS compliance assessment software solution, underpinned by Generative AI using OpenAI's GPT-3.5 model. Through rigorous evaluations employing the annual reports of Tesco PLC and M&S, our solution has manifested its potential, adeptly interpreting, and analysing financial documents to extract pertinent IFRS-related insights. Performance metrics, spanning accuracy to the F1 score, have substantiated our solution's capabilities. Distinct prompt styles exhibited varying degrees of efficacy, with certain styles notably outperforming others. Although our results are encouraging, they also unveil avenues for enhancements, particularly in refining text processing and optimising prompt efficiency. In essence, our findings not only validate our efforts in achieving the set research objectives but also underscore the transformative potential of our software solution in the realm of IFRS compliance assessments.

Chapter 6 - Challenges

6.1 Rate Limit

Navigating OpenAI's GPT 3.5 model's API's rate limits was a key challenge. It allowed only 150,000 tokens per request and limited three queries per minute. This posed efficiency issues for processing large documents and user queries.

Solution: A system was devised to intelligently divide document content into chunks to fit token limits. Delays between requests were introduced strategically, ensuring adherence to rate limits without affecting user experience.

6.2 Token Limitations

Text sent to OpenAI's GPT 3.5 model is tokenised, with a cap on tokens per request. Going beyond this limit leads to errors.

Solution: The system was made token-limit-aware. Before sending a text for embedding or queries, the token count is checked to remain within limits. Large documents are segmented to ensure each part adheres to token restrictions.

6.3 Text Extraction Challenges

Extracting text from tables in PDF financial reports, while retaining context, is tough due to structured formats.

PDFs aren't optimised for text extraction, potentially leading to unordered text or formatting anomalies.

Solution: Optimal extraction tools are vital. While PyPDFLoader is used, integrating advanced tools, and testing continuously can enhance accuracy. Specialised libraries or services might be needed for table-specific extractions.

Chapter 7 - Future Enhancement

7.1 Fine-Tuning w/wo Knowledge Graph

The current system leverages the general capabilities of OpenAI's GPT 3.5 model. However, for specific domains like financial auditing, there's potential for model improvement through fine-tuning.

7.1.1 With Knowledge Graph

A knowledge graph can serve as a structured representation of domain-specific information. Fine-tuning the model with data from a knowledge graph can enhance its understanding and make its outputs more contextually relevant.

7.1.2 Without Knowledge Graph

Even without a structured knowledge graph, the model can be fine-tuned using annotated financial datasets to improve its performance in the auditing domain.

7.2 Tabular Extraction from Annual Reports

Financial reports often contain vital data in tabular formats. While the current text extraction provides a general overview, targeted table extraction can yield more structured and actionable insights.

Solution: Integrate specialised libraries or tools like Tabula or Camelot for precise table extraction. Once extracted, these tables can be processed to create embeddings or directly queried for specific data points.

7.3 One-Shot Prompting

One-shot prompting is the idea of providing the model with a single, well-crafted prompt and receiving a detailed and accurate response, reducing back-and-forth interactions (Kristian, 2023).

Solution: Research and experiment with various prompt formulations to optimise the quality of responses. Utilise feedback loops to continually refine prompting strategies based on real-world results.

7.4 Few-Shot Prompting

It operates by providing the model with a set of examples to guide its output generation. Instead of relying on a vast knowledge graph or a singular prompt, few-shot prompting leverages the power of examples to induce context and direction. This method can be especially beneficial when dealing with nuanced topics where a singular prompt might be insufficient and a vast knowledge graph impractical. Although, this method could be resource intensive as compared to other prompting techniques (Tam, 2023).

Solution: Devote research efforts towards identifying the optimal number and nature of examples for various tasks within the financial auditing domain. This can be achieved by iteratively experimenting with different sets of examples and gauging the accuracy and relevance of the model's output. The feedback loops established during the one-shot prompting phase can further refine the examples used, ensuring the model remains contextually aligned with the auditing domain's requirements.

7.5 Multiple File Processing

The current system processes a single file at a time. In real-world scenarios, auditors might need to process multiple reports simultaneously.

Solution: Implement a batch processing mechanism, allowing users to upload multiple files. The system can then process these files concurrently or in a queue, providing aggregated results or insights.

7.6 Support for Multiple Document Types

Expanding the system's capability to process various file formats increases its utility.

Solution: Integrate libraries that can handle diverse file types. For instance, python-docx for DOCX files. Ensure that the extraction logic is adaptable to different document structures.

7.7 Report Generation

To provide users with a more comprehensive overview, interactive visual representations can be invaluable.

Solution: Integrate data visualisation libraries like Plotly or Bokeh in the Streamlit app. Visualise key metrics, trends, or even embeddings to provide users with a more intuitive understanding of the data.

7.8 Integration with Other Financial Standards

Broadening the system's scope can make it a one-stop solution for multiple auditing needs.

Solution: Gradually incorporate modules or functionalities tailored to other financial standards beyond IFRS 16. This involves both data processing adaptations and UI modifications to accommodate new query types.

These enhancements, while ambitious, can propel the system to new heights, making it more versatile, user-friendly, and insightful. They also open avenues for academic and industrial research, ensuring the platform remains at the forefront of AI-driven financial auditing.

Chapter 8 - Conclusion

In conclusion, the application presents a well-structured process flow that efficiently guides users through the intricacies of analysing annual reports and obtaining AI-generated insights for the IFRS standards. By adhering to a systematic sequence of steps, the application seamlessly integrates cutting-edge technology to streamline the auditing process.

The initialisation phase ensures a smooth start, while the user interface display provides users with clear instructions and options to upload their annual reports. Robust document processing techniques are employed to break down uploaded PDFs into manageable segments and convert textual content into vector embeddings using OpenAI's GPT 3.5 “text-embedding-ada-002” model’s API (default). These embeddings are stored in the Chroma vector database for future reference.

The setup of a conversational retrieval chain brings together Chroma's database and OpenAI's GPT 3.5 “text-davinci-003” language models, enabling the application to generate accurate and contextually relevant answers to user queries. The user-friendly question-answer interface empowers users to explore standard questions or input their custom inquiries, enhancing the interaction with the AI-powered system.

To ensure robustness, the application incorporates error-handling mechanisms, such as managing potential API rate limits and providing appropriate error messages when needed. Lastly, users have the flexibility to choose between starting a new session or concluding their current one, with session data cleared to maintain privacy.

In essence, the application orchestrates a sophisticated synergy between Generative AI technology, document processing techniques, and user interface design. By following this well-defined sequence, the application offers a comprehensive and user-centric solution for automating auditing processes while maintaining efficiency, accuracy, and a high standard of user experience.

References

Ahmadi, N., Sand, H. and Papotti, P., 2022. Building A Knowledge Graph for Audit Information. In EDBT/ICDT Workshops.

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. [online] Available at: https://arxiv.org/pdf/1908.10063.pdf

Arnold, V., Collier, P.A., Leech, S.A. and Sutton, S.G., 2004. Impact of intelligent decision aids on expert and novice decision-makers' judgments. Accounting & Finance, 44(1), pp.1-26.

Baldwin, A.A., Brown, C.E. and Trinkle, B.S., 2006. Opportunities for artificial intelligence development in the accounting domain: the case for auditing. Intelligent Systems in Accounting, Finance & Management: International Journal, 14(3), pp.77-86.

Caruana, V. (2023) Semantic search: next big thing in search tech. Available at: https://www.algolia.com/blog/product/semantic-search-the-next-big-thing-in-search-engine-technology/.

Chroma (2023). Getting started. Available at: https://docs.trychroma.com/getting-started.

Chroma (2023). Embeddings. Available at: https://docs.trychroma.com/embeddings.

Dickey, G., Blanke, S. and Seaton, L., 2019. Machine learning in auditing. The CPA Journal, 89(6), pp.16-21.

Eining, M.M. and Dorr, P.B., 1991. The impact of expert system usage on experiential learning in an auditing setting. Journal oflnformation Systems, 5(1), pp.1-16.

Fisher, I.E., Garnsey, M.R. and Hughes, M.E., 2016. NLP in Accounting, Auditing and Finance: Intelligent Systems. Accounting, Finance and Management, 23(3), pp.157-214.

Green, B.P. and Choi, J.H., 1997. Assessing the risk of management fraud through neural network technology. Auditing, 16, pp.14-28.

Gu, H., Schreyer, M., Moffitt, K. and Vasarhelyi, M.A., 2023. Artificial Intelligence Co- Piloted Auditing. Available at SSRN 4444763.

Hore, S. (2023). An Introduction to Large Language Models (LLMs). [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2023/03/an-introduction-to- large-language-models-llms/.

Kristian (2023) "ChatGPT Prompt engineering tips: Zero, one and few shot prompting," All About AI, 5 February. Available at: https://www.allabtai.com/prompt-engineering-tips-zero-one-and-few-shot-prompting/.

Kyriacou, K. (2023) Semantic search with OpenAI and Chroma. Available at: https://kkyr.io/blog/semantic-search-with-openai-and-chroma/

Langchain-Ai (2023) GitHub - langchain-ai/langchain:Building applications with LLMs through composability. Available at: https://github.com/langchain-ai/langchain.

Langchain-Ai (2023) Error while loading saved index in chroma db · Issue #2491 · langchain-ai/langchain. Available at: https://github.com/langchain-ai/langchain/issues/2491.

Langchain (2023). QuickStart. Available at: https://python.langchain.com/docs/get_started/quickstart.

Lin, J., Zhao, Y., Huang, W., Liu, C. and Pu, H., 2021. Domain knowledge graph-based research progress of knowledge representation. Neural Computing and Applications, 33, pp.681-6.

Liu, F., Wang, R., Yang, Y. and Zhang, J., 2020, November. A Preliminary Approach of Constructing a Knowledge Graph-based Enterprise Informationized Audit Platform. In 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) (pp. 126-131). IEEE.

Lo, L.S., 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. The Journal of Academic Librarianship, 49(4), p.102720.

Lutkevich, B. (2021). What is Natural Language Processing? An Introduction to NLP. [online] SearchEnterpriseAI. Available at: https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP.

Mackie, I. and Dalton, J. (2022). Query-Specific Knowledge Graphs for Complex Finance Topics. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2211.04142.

Menon, P. (2023). Introduction to Large Language Models and the Transformer Architecture. [online] Medium. Available at: https://rpradeepmenon.medium.com/introduction-to-large-language-models-and-the-transformer-architecture-534408ed7e61

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T. and Trajanov, D. (2020). Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. IEEE Access, [online] 8, pp.131662–131682. doi:https://doi.org/10.1109/ACCESS.2020.3009626.

O'Leary, D.E., 2022. Massive data language models and conversational artificial intelligence: Emerging issues. Intelligent Systems in Accounting, Finance and Management, 29(3), pp.182-198.

Omoteso, K., 2012. The application of artificial intelligence in auditing: Looking back to the future. Expert Systems with Applications, 39(9), pp.8490-8495.

OpenAI Platform (2022). Available at: https://platform.openai.com/docs/guides/embeddings/what-are-embeddings.

OpenAI Platform (2022). Available at: https://platform.openai.com/docs/introduction/key-concepts.

OpenAI Platform (2022). Available at: https://platform.openai.com/docs/guides/gpt/function-calling.

OpenAI Platform (2022). Available at: https://platform.openai.com/docs/models/how-we-use-your-data.

Piñeiro-Martín, A., Garci-a-Mateo, C., Docío-Fernández, L. and Pérez, M. del C.L. (2023). Ethical Challenges in the Development of Virtual Assistants Powered by Large Language Models. [online] Preprints.org. doi:https://doi.org/10.20944/preprints202306.0196.v1.

Sifa, R., Ladi, A., Pielka, M., Ramamurthy, R., Hillebrand, L., Kirsch, B., Biesner, D., Stenzel, R., Bell, T., Lübbering, M. and Nütten, U., 2019, September. Towards automated auditing with machine learning. In Proceedings of the ACM Symposium on Document Engineering 2019 (pp. 1-4).

Streamlit Docs (2023) Get started. Available at: https://docs.streamlit.io/library/get-started.

Svetlova, E., 2022. AI ethics and systemic risks in finance. AI and Ethics, 2(4), pp.713-725.

Tam, A. (2023) "What are Zero-Shot prompting and Few-Shot prompting," MachineLearningMastery.com [Preprint]. Available at: https://machinelearningmastery.com/what-are-zero-shot-prompting-and-few-shot-prompting/.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D. and Mann, G., 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.