MSc Project Proposal: FIN-RAG – Financial Document Parsing using Retrieval-Augmented Generation

# 1. Introduction

The rise in unstructured financial content on the internet, especially in the form of IPO filings, annual reports, and investor presentations, has created challenges for investors and analysts who seek quick, accurate access to essential financial data. FIN-RAG, short for Financial Document Parsing using Retrieval-Augmented Generation, is proposed as a modern AI-based solution to this problem. The project integrates web scraping, semantic retrieval, and large language model (LLM) capabilities to deliver real-time question answering over financial documents. With the increasing dependence on digital platforms for regulatory filings and financial disclosures, the need for a context-aware and explainable AI pipeline has become imperative. FIN-RAG meets this demand by combining best-in-class open-source tools such as Selenium, BeautifulSoup, FAISS, HuggingFace MiniLM, LangChain, and Groq LLaMA-3 to offer a scalable, user-friendly application. This proposal outlines the problem domain, objectives, methodology, and the critical literature that justifies this innovative approach to financial data parsing.

# 2. Problem Statement

Financial data such as IPO listings, regulatory disclosures, and corporate financial reports are typically hosted in semi-structured or unstructured formats. These documents are often dynamically rendered through JavaScript-heavy web interfaces or published as PDFs without metadata, posing significant difficulties for automated information retrieval systems. Stakeholders such as retail investors, compliance officers, analysts, and researchers must manually extract key financial indicators, resulting in inefficiency, delays, and potential errors in decision-making. Furthermore, traditional scraping techniques and rule-based QA systems struggle with domain-specific semantics and structural variability in financial documents. The lack of standardised APIs exacerbates the problem, especially in emerging markets where financial portals like Chittorgarh.com serve as primary sources. A key gap lies in the inability of existing tools to provide precise, explainable, and contextually relevant answers to natural language queries over these documents. FIN-RAG directly addresses this issue by applying Retrieval-Augmented Generation (RAG) to bridge unstructured content with structured query responses. The system uses semantic embeddings and transformer-based language models to understand user intent and retrieve the most relevant document segments. It then generates focused answers while minimising hallucination risks. This approach not only streamlines information access but also aligns with the increasing demand for AI tools that ensure transparency, compliance, and efficiency in financial workflows.

# 3. Aims and Objectives

The primary aim of this project is to develop an AI-based system that enables accurate and explainable question answering from unstructured financial content using a Retrieval-Augmented Generation (RAG) pipeline. This system FIN-RAG targets content in dynamic HTML webpages and static PDF documents related to IPOs and regulatory filings.

The project is structured around the following objectives:

1. To design a robust scraping module using Selenium and BeautifulSoup for capturing dynamic web content.

2. To implement PDF parsing functionality using PyPDF2 for extracting textual information from IPO documents.

3. To clean, preprocess, and segment the extracted data using recursive chunking.  
4. To embed document segments using HuggingFace MiniLM and index them using FAISS for semantic search.

5. To integrate Groq's LLaMA-3 model for generating context-aware answers based on retrieved chunks.

6. To develop a user-friendly Streamlit frontend with secure login, session tracking, and query logging via MongoDB.

7. To evaluate the system’s performance using real-world IPO cases and benchmark it against traditional QA tools.

Research Strategy: A design science approach will be employed, with iterative prototyping, prompt engineering, and empirical testing forming the core methodology. LangChain will orchestrate the retrieval and inference logic. Evaluation will focus on answer accuracy, retrieval precision, system responsiveness, and user feedback.

# 4. Legal, Social, Ethical and Professional Considerations

The FIN-RAG system adheres to legal and ethical standards throughout its development and deployment lifecycle. All financial data scraped or processed is sourced from publicly available platforms that do not require authentication or breach terms of service. No personally identifiable information (PII) is collected, stored, or shared, ensuring full compliance with GDPR and other data protection laws. User credentials are securely stored using SHA-256 hashing in a local SQLite database, and query logs are maintained locally through MongoDB. From an ethical perspective, care has been taken to minimize risks associated with LLM hallucinations by enforcing strict prompt formatting, validating retrieval relevance, and displaying source context alongside generated answers. The system’s outputs are designed to be explainable, non-biased, and traceable. Socially, FIN-RAG promotes transparency and democratizes access to financial information for non-technical users, thereby enhancing financial literacy. Professionally, the development aligns with best practices in AI engineering, including modular coding, open-source libraries, and documentation. Ethical software engineering principles such as transparency, fairness, and accountability have guided the system architecture, particularly in the design of user interfaces, data storage, and output validation mechanisms.

# 5. Background

The proposed FIN-RAG system builds upon the convergence of three mature areas: web scraping, semantic vector search, and large language models (LLMs) for question answering. Traditional financial data extraction techniques rely heavily on static scrapers or rule-based systems, which are fragile when dealing with inconsistent formats or evolving web structures. Recent literature identifies web scraping as a reliable technique when structured APIs are unavailable. Bharathi et al. (2024) emphasize the utility of Selenium and BeautifulSoup in automating data extraction from JavaScript-heavy web pages. Similarly, Fawei (2024) and Gašpar et al. (2023) have explored rule-based QA systems for structured legal documents, but these approaches lack generalizability and adaptability in finance, where terminology and structure vary significantly.

Advances in NLP have shifted the paradigm from rule-based systems to deep learning models. Transformers, particularly models like BERT, RoBERTa, and GPT, have enabled machines to interpret human language with contextual understanding. Basha et al. (2023) and Chen et al. (2024) document how contextual embeddings improve named entity recognition, sentiment analysis, and domain-specific information extraction. This development paved the way for Retrieval-Augmented Generation (RAG), a hybrid approach that combines semantic search with generation-based reasoning. Joshi et al. (2023) and Rathod & Anurag (2024) demonstrate the efficiency of RAG pipelines in enterprise document parsing, highlighting advantages like reduced fine-tuning costs and enhanced modularity.

Setty et al. (2024) argue that document chunking and retrieval re-ranking significantly impact RAG output accuracy, a finding corroborated in this project’s implementation. Zeng et al. (2024) discuss privacy considerations in RAG architectures and propose mitigation strategies such as source-aware responses an approach embedded in FIN-RAG’s design. In the finance domain, RAG's relevance is underscored by works like Wang et al. (2025), who built an assistant for retail investors using LLM-based parsing of IPO documents. Malali (2025) also showcases the use of RAG in automating compliance checks, further validating FIN-RAG’s practical potential.

From a technical perspective, FIN-RAG leverages a range of proven technologies. Selenium handles dynamic scraping, while PyPDF2 extracts PDF content. Text chunks are embedded via HuggingFace’s MiniLM model, indexed in FAISS, and retrieved using cosine similarity. LangChain serves as the orchestration layer for retrieval and prompt structuring. Groq’s LLaMA-3 model generates answers with minimal latency and high contextual precision. The entire system is presented via Streamlit, with secure login (SQLite) and chat logging (MongoDB). The modularity of each component ensures scalability, debuggability, and maintainability.

This project thus contributes a novel, hybrid pipeline to the growing body of financial NLP systems. It is aligned with current research trends, employs well-established libraries and practices, and introduces innovations in prompt engineering and dual-mode document ingestion (HTML + PDF). Its ability to offer accurate, traceable answers from complex financial content makes it highly relevant to both academic and industrial FinTech audiences.

# 6. References

[1] A. S., B. R., and M. M., “Web Scraping Techniques and Applications: A Literature Review,” ResearchGate, 2023.

[2] V. V. Bharathi et al., “Web Scraping: Techniques and Applications,” IJETT, vol. 72, no. 1, pp. 81–84, 2024.

[3] B. Fawei, “NLP-Based Rule Learning from Legal Text for QA,” AJRCS, vol. 17, no. 7, pp. 31–40, 2024.

[4] A. Gašpar et al., “Evaluation of Rule-Based QA,” Language Resources and Evaluation, 2023.

[5] Y. Chen et al., “AI Methods in NLP: A Review,” Highlights in Science, Engineering and Technology, vol. 85, 2024.

[6] J. Carrasco Ramírez, “NLP Advancements,” JAIGS, vol. 3, no. 1, 2024.

[7] M. Basha et al., “Advancements in NLP,” E3S Web of Conferences, vol. 399, 2023.

[8] H. Liu and T. Chen, “Neural Architectures in NLP,” Procedia Computer Science, vol. 176, pp. 327–336, 2020.

[9] P. Joshi et al., “Multi-Model RAG Pipeline,” Avasant Labs, 2023.

[10] P. Rathod and Anurag, “Efficient RAG Systems,” Intel Corp., 2024.

[11] S. Setty et al., “Improving RAG Retrieval,” Stanford University, 2024.

[12] S. Zeng et al., “Privacy Issues in RAG,” 2024.

[13] N. Malali, “RAG in Financial Document Processing,” Int. J. of Management Tech., vol. 12, no. 3, Apr. 2025.

[14] J. Wang et al., “LLM-RAG for Financial QA,” Conf. on Modern Management & Innovation, Apr. 2025.

[15] I. Iaroshev et al., “Evaluating RAG Models for Finance,” Applied Sciences, vol. 14, no. 20, Oct. 2024.