Introduction

Background and Context

The financial technology (fintech) sector has experienced unprecedented growth over the past decade, revolutionizing how individuals and organizations manage their finances. Traditional methods of financial planning, which often involved manual bookkeeping, in-person consultations with financial advisors, or static budgeting tools, are being rapidly replaced by innovative digital solutions. These advancements are driven by the increasing adoption of artificial intelligence (AI), natural language processing (NLP), and big data technologies.

Today, personal finance assistants, chatbots, and robo-advisors dominate the landscape. Companies such as Mint, YNAB (You Need a Budget), and Wealthfront provide tailored recommendations and automated insights, enabling users to track expenses, set financial goals, and make informed investment decisions. Similarly, conversational AI systems like Cleo and Erica (developed by Bank of America) empower users to interact with their financial data seamlessly through chat-based interfaces. These systems rely on sophisticated machine learning models and extensive data training to deliver personalized advice.

This transition to digital approaches is fueled by the demand for convenience, accessibility, and real-time assistance. Unlike traditional financial planning, which required significant time and expertise, modern systems democratize financial management by making tools intuitive and readily available. Users can now access insights from anywhere, receive instant responses, and benefit from AI-driven strategies that were previously limited to expert consultations. This is particularly beneficial because it empowers all users, regardless of their financial literacy, in past times these were services tailored and solely available to a wealthy minority of the population.

Despite these advancements, most existing solutions rely on extensive training of AI models, requiring large datasets and computational resources. This dependency introduces challenges in scalability, adaptability to niche user needs, and maintaining up-to-date knowledge bases. The next frontier in fintech innovation lies in leveraging more dynamic, efficient, and adaptable systems that can generate accurate and contextually relevant responses without traditional model retraining.

Problem Statement and Research Objectives

The adoption of Retrieval-Augmented Generation (RAG) systems in financial planning has significantly improved the efficiency and accuracy of conversational AI systems. Most existing implementations rely on vector store RAG systems, where embeddings of text are indexed and searched for semantic similarity. While effective, this approach inherently has limitations in capturing the intricate, structured relationships that are critical for certain fields like financial planning, where interdependencies among concepts such as expenses, investments, goals, and risks are highly relational.

Graph-based RAG systems, which focus on semantic relationship searches through graph structures, offer an alternative by encoding relationships explicitly and dynamically. This capability can potentially enhance the contextual understanding of financial data, enabling a more nuanced and interconnected representation of user goals, spending patterns, and investment strategies. However, graph-based RAG systems have yet to be explored in the domain of financial planning.

Moreover, current AI solutions in this domain often require substantial time, resources, and expertise to train and fine-tune large language models (LLMs). This dependency creates barriers for individuals and organizations that lack access to vast datasets and computational power, limiting the democratization of these advanced tools. By investigating the feasibility of a “zero-shot” approach—

relying solely on prompt tuning without model retraining—this research aims to reduce these barriers and make financial planner chatbots more accessible and affordable.

This thesis addresses the gap by introducing a graph-based RAG system for financial planning and exploring how a prompt-tuning approach can perform in this context. The findings are expected to contribute to the development of practical, cost-effective AI tools that are truly available to a broader audience while exploring the unique advantages that graph structures may bring to this domain.

The primary objective of this thesis is to design and evaluate a financial planner chatbot based on a graph RAG system, using only prompt tuning to generate meaningful outputs. Secondary objectives include:

1. Assessing the performance of the graph RAG system compared to vector-based RAG systems in providing relevant and accurate financial recommendations.
2. Exploring the usability and practicality of a zero-shot approach in real-world applications.
3. Identifying the advantages and constraints of graph-based semantic search in capturing financial relationships.

By addressing these objectives, this research seeks to contribute to the broader field of AI in fintech, offering insights into the applicability of graph-based RAG systems and the potential of zero-shot methods to make AI solutions more accessible and impactful.

Scope and Delimitations

This research focuses on designing and evaluating a financial planner chatbot using a graph-based Retrieval-Augmented Generation (RAG) system, relying exclusively on a "zero-shot" prompt-tuning approach. While the techniques explored in this study may eventually prove useful for developing generic domain-specific chatbots, the primary focus is limited to the financial planning domain. The scope does not extend to creating a generalized framework for graph-based chatbots, nor will this research address the implementation of RAG systems using alternative architectures such as vector-based approaches, as these are extensively covered in existing literature.

Instead, this work embraces an exploratory approach, experimenting with various techniques and configurations necessary for the system's development. This includes: • Investigating different methods for graph search to improve semantic relationship retrieval. • Exploring prompt configurations and tuning strategies to enhance the chatbot's response quality. • Experimenting with answer-generation techniques tailored to the financial planning context.

The research adopts a "free-for-all" methodology, allowing for flexibility in testing and integrating different approaches, tools, and algorithms as required to achieve the project's objectives. However, the study will remain strictly within the boundaries of graph-based RAG systems and will not delve into unrelated technologies or frameworks.

By maintaining this focus, the research aims to deliver targeted insights into the application of graph RAG systems for financial planning, without attempting to generalize findings beyond the specified domain or architecture.

Research Questions

This research seeks to explore the potential of a graph-based Retrieval-Augmented Generation (RAG) system for financial planning by addressing the following key questions:

1. Graph RAG System Effectiveness
   - How effectively can a graph RAG system capture the semantic relationships necessary for delivering personalized and contextually accurate financial advice?
2. Zero-Shot Prompt-Tuning Feasibility

- Can a zero-shot prompt-tuning approach generate meaningful and accurate financial recommendations without requiring extensive model training?

3. Comparison with Vector-Based RAG Systems
   - How does the performance of a graph-based RAG system compare to that of vector-based systems in terms of relevance, contextual understanding, and adaptability within the financial planning domain?
4. System Practicality and Scalability
   - What are the practical advantages, limitations, and potential challenges of using a graph RAG system for real-world financial planning applications?

These questions guide the exploration and evaluation of the proposed chatbot, helping to identify both its technical potential and its practical implications.

Methodology

This study was carried out in two main stages: research and system design and development. Each stage played a crucial role in the realization of a financial planner chatbot based on a graph-based Retrieval-Augmented Generation (RAG) system, ensuring both theoretical grounding and practical implementation.

Research

The research phase comprised two primary objectives:

1. Survey of Graph-Based RAG Systems

To understand the underlying principles and methodologies for implementing a graph-based RAG system, an extensive review of relevant academic literature was conducted. Papers discussing techniques for retrieval, indexing, and answer generation within graph-based architectures were critically analyzed. This was essential to gain a comprehensive understanding of existing approaches and their applicability to constructing the proposed system.

2. Review of Fintech Chatbot Systems

A secondary review focused on exploring existing works in the domain of fintech chatbots, particularly those employing RAG systems. This inquiry provided valuable insights into how domain-specific challenges—such as contextual relevance, user intent understanding, and financial data representation—were tackled. The findings from this review informed design decisions tailored to the financial planning domain.

These two research streams ensured that the system design was grounded in established techniques while addressing the unique requirements of the target application.

System Design and Development

The system was implemented in several phases, each addressing distinct components required for the chatbot's functionality.

1. Knowledge Extraction

Tools and libraries such as LangChain, Beautiful Soup, and APIs from Gemini and OpenAI were used to extract financial planning knowledge from various web sources. Information was processed into a graph structure with nodes and edges, where both contained textual notations. For example:

- Node: "CAR LEASE"
- Edge: "IS A SOURCE OF"
- Target Node: "DEBT"

Additionally, the graph captured multiple relationships per node, as illustrated by:

- "CAR LEASE" -> "REQUIRES" -> "MONTHLY PAYMENTS".

The graph structure was stored in Neo4j, while the embedded representations of textual content were stored in a vector database using MongoDB. This dual representation allowed for flexibility in retrieval strategies.

2. Knowledge Retrieval

The primary focus was on retrieving knowledge from the graph embeddings rather than the graph structure itself. Techniques such as k-means clustering and cosine similarity were employed to identify relevant nodes and edges. These methods enabled efficient retrieval of semantically similar data points based on user queries.

3. Subgraph Extraction

Retrieved nodes and edges formed the basis for constructing a subgraph, representing the subset of information most relevant to a given query. Several techniques for subgraph extraction were explored, which will be elaborated upon in subsequent sections.

4. Answer Generation

The extracted subgraph was processed to generate natural language responses. This was accomplished using the Gemini API, leveraging various models to transform structured graph data into coherent and contextually appropriate outputs.

5. Evaluation and Testing

To assess the performance of the chatbot, a set of handcrafted queries covering diverse financial scenarios was created. The system's responses were compared against those from other systems and general-purpose models. Evaluation was based on the following metrics:

- Accuracy: The correctness of the information provided relative to the query.
- Factual Correctness: Whether the response aligned with verified financial knowledge.
- Efficiency: The speed and computational cost of generating a response.
- Robustness: The system's ability to handle ambiguous or incomplete queries.

This methodological framework ensured a systematic exploration and implementation of the proposed graph-based RAG system, providing a basis for both technical and practical contributions to the field of financial planning.

Structure of the Thesis

This thesis is organized into several chapters, each addressing a specific aspect of the research process and contributing to the overall goal of developing a financial planner chatbot based on a graph-based Retrieval-Augmented Generation (RAG) system. The structure is as follows:

- Chapter 1: Introduction

The introduction provides an overview of the research, including the background, problem statement, research objectives, and motivation for the study. It also outlines the scope and delimitations of the work, and presents the research questions and hypotheses guiding the study.

- Chapter 2: Literature Review

This chapter reviews relevant literature in the areas of fintech chatbots, RAG systems, and graph-based knowledge representations. It includes a discussion of existing approaches to chatbot development in the financial planning domain and surveys different techniques for graph-based RAG systems, providing the foundation for the research.

- Chapter 3: Methodology

The methodology chapter describes the research design and methods employed to build the proposed financial planner chatbot. It includes a detailed explanation of the data collection and preprocessing steps, the design and development of the system, the retrieval techniques used, and

the natural language generation process. The evaluation strategy and performance metrics are also outlined in this chapter.

- Chapter 4: System Architecture and Implementation

This chapter provides a detailed description of the system architecture, including the components involved in knowledge extraction, graph construction, retrieval, and natural language generation. The implementation process is also discussed, with specific attention to the tools, frameworks, and databases used to develop the system.

- Chapter 5: Evaluation and Results

In this chapter, the results of the system's evaluation are presented. The performance of the chatbot is assessed based on various metrics such as accuracy, factual correctness, efficiency, and robustness. The findings from the evaluation experiments are discussed in detail, and the implications of the results are analyzed.

- Chapter 6: Conclusion and Future Work

The conclusion chapter summarizes the key findings of the research, reflects on the contribution of the work, and discusses its limitations.-It also suggests potential avenues for future research, including improvements to the system and further applications of graph-based RAG approaches in other domains.