Literature Review

This literature review will begin by providing an overview of the state of the art technologies, frameworks and methodologies used in RAG applications using a graph kmnowledge base. This will include preprocessing of documents, knowledge graph creation, embedding, retrieval of information from storage and generation using LLMs. Later we will discuss about some relevant studies about RAG applications in non-finance fields, particularly medical applications, because they may be helpful when targetting domain specific usage of the RAG architecture. And at the end we will summarize some studies about using RAG, data quality, etc, for FinTech applications, specially financial planning, which will be the focus of the proposal of this document.

Preprocessing and Data Storage

Graph databases have emerged as specialized systems designed to store and manage data in graph formats, emphasizing the relationships between data points. Technologies such as Neo4j, Amazon Neptune, and ArangoDB offer scalable solutions for handling complex graph structures. These databases enable efficient querying and traversal of relationships, making them ideal for applications requiring high-performance retrieval and manipulation of interconnected data, which is fundamental for Graph RAG systems.

Effective storage and retrieval in Graph RAGs rely on optimized data structures and indexing methods that facilitate rapid access to relevant nodes and edges. Techniques such as adjacency lists, path indexing, and graph partitioning are employed to manage large graphs efficiently. Query languages like Cypher and Gremlin allow for expressive querying capabilities, enabling RAG systems to retrieve pertinent subgraphs or relationships necessary for informed content generation.

Building upon these foundational technologies, **Peng et al. (2024)** discussed the construction and indexing of graph databases in the context of GraphRAG, emphasizing both open-source and self-constructed graph databases. They explored various indexing methods, including graph, text, and vector indexing, to improve retrieval efficiency [1]. Their work highlights the importance of tailored indexing strategies in enhancing the performance of RAG systems.

Furthermore, the integration of knowledge graphs into RAG systems has shown significant benefits. **Hu et al. (2024)** demonstrate how knowledge graphs serve as rich information sources, providing relevant entities and their relationships, which lead to more coherent and accurate responses from language models [2]. By leveraging the structured information in knowledge graphs, RAG systems can generate content that is both contextually relevant and semantically accurate.

In specific domains like medicine, the structure of the graph data becomes even more critical. **Cui et al. (2024)** explore the application of Directed Acyclic Graph (DAG) task decomposition in the medical domain. They propose that hierarchical graph structures, such as DAGs, provide a systematic way to process complex medical tasks, enabling large language models to organize and represent inter-task dependencies more efficiently. This allows for more structured task execution and the use of metadata for task nodes, thereby enhancing the management and understanding of medical knowledge. Their research shows that DAG structures, especially in retrieval-augmented generation, can optimize medical task processing by enabling a clear breakdown of subtasks [3].

When considering the choice between different database systems, it's essential to analyze the trade-offs to ensure optimal performance. Compared to traditional relational databases, graph databases offer superior performance in handling complex, interconnected data due to their inherent design focused on relationships. While relational databases excel in structured, tabular data management, they are less efficient for operations requiring multiple joins or traversals. This emphasizes the advantages of graph databases in supporting the dynamic retrieval needs of RAG systems.

Adding to this perspective, **Jin et al. (2024)** discuss the performance benefits of graph-based vector databases, such as HNSW, used in Retrieval-Augmented Generation (RAG) systems. They highlight that vector databases structured as graphs offer significant advantages in terms of efficient search and retrieval compared to other types of databases. Specifically, graph databases are shown to be more efficient in managing large-scale vector searches through hierarchical structures, reducing computational complexity and enhancing query performance. This contrasts with traditional databases that may not be optimized for high-dimensional vector searches. Their work demonstrates that incorporating graph databases into RAG systems can improve knowledge retrieval efficiency, particularly for applications requiring high-accuracy vector similarity searches [4].

Moreover, the organization of data within graphs can further enhance retrieval processes. **Chang et al. (2024)** introduce CommunityKG-RAG, a novel framework that leverages community structures in knowledge graphs (KGs) to enhance retrieval-augmented generation systems. They emphasize that traditional RAG systems suffer from poor integration of structured knowledge, limiting their effectiveness in fact-checking tasks. By focusing on community detection within KGs, CommunityKG-RAG provides a more contextually rich retrieval process compared to standard databases or flat knowledge structures. This approach significantly enhances the retrieval of relevant information by utilizing the multi-hop relationships in KGs, showing the advantages of graph-based data structures over other types of databases in complex, fact-checking applications [5].

An essential aspect of improving retrieval in RAG systems is the use of semantic embeddings. Semantic embeddings represent words, phrases, or entities in continuous vector spaces, capturing their meanings and relationships. In RAG systems, these embeddings are crucial for mapping between the language model and the knowledge base. They enable the system to retrieve semantically relevant information based on the context of the input, facilitating more coherent and accurate generation.

To generate these embeddings, various techniques are employed. Traditional methods like Word2Vec and GloVe capture semantic relationships based on word co-occurrence, while contextual models like BERT and GPT account for word meaning based on surrounding context. For graph data, methods like Node2Vec and GraphSAGE generate embeddings that capture structural information within the graph. Combining these approaches allows RAG systems to create rich representations that reflect both linguistic and relational properties.

In the context of textual graph knowledge bases—where both nodes and edges have textual notations—pre-trained language models prove particularly beneficial. **He et al. (2024)** stated that using models such as **SentenceBERT** can enhance the system's ability to understand and retrieve relevant information [6]. This approach leverages the strengths of language models in capturing semantic nuances, thereby improving the retrieval process in RAG systems.

Semantic embeddings not only improve retrieval relevance but also enable better generalization across different contexts. Applications extend to semantic search, recommendation systems, and personalization. The benefits include more accurate responses, improved user satisfaction, and the ability to handle ambiguous or novel queries effectively. By integrating semantic embeddings, RAG systems become more adept at handling the complexities of natural language and the intricacies of user intent.

Concluding this discussion, **Peng et al. (2024)** further elaborated on how GraphRAG enhances retrieval performance by incorporating semantic embeddings derived from graph structures. This integration allows for a more nuanced understanding of relationships within the data, improving the quality of information retrieval for downstream tasks [1]. Their insights reinforce the critical role of semantic embeddings in advancing the capabilities of RAG systems.

retrieval

Retrieval is a critical component in Graph Retrieval-Augmented Generation (Graph RAG) systems, determining the relevance and accuracy of the information drawn from the knowledge base for the generation process. Unlike traditional RAG systems that retrieve unstructured text, Graph RAGs leverage the structural properties of graphs, retrieving not just textual information but also relational data embedded within nodes, edges, and subgraphs. This allows for more nuanced responses, which are essential for applications requiring deep contextual understanding and interconnected knowledge. Efficient retrieval methods are essential for ensuring that the system can effectively process large, complex graph structures while maintaining response accuracy and speed. In this section, we explore the key retrieval techniques employed in Graph RAGs, including query processing, optimization strategies, and evaluation metrics.

The query processing stage in Graph RAG systems is designed to map a user's natural language input to relevant graph structures. This involves transforming the input into a query that can be understood by the underlying graph database. Techniques such as graph traversal algorithms, entity linking, and semantic matching are often employed to extract nodes, edges, or subgraphs relevant to the query. The unique structure of graphs allows for retrieval methods that can account for both direct and indirect relationships between entities, thus providing richer context for the generation model.

He et al. (2024) propose a k-nearest neighbors (k-NN) retrieval approach that identifies relevant nodes and edges by measuring the cosine similarity between the query and the embeddings of graph elements. This method retrieves the top-k most relevant nodes and edges, ensuring that the query is mapped to meaningful graph structures and relationships, providing the system with precise data for subsequent generation tasks [6].

Hu et al (2024). demonstrate how WeKnow leverages structured knowledge from graphs to improve the accuracy and relevance of information retrieved by language models, especially for complex queries requiring relational understanding, but they do it in a more insteresting way. They proposed a domain specific RAG model based on a scrapping mechanism, making the graph knowledge base the domains to which to scrap the relevant information, making it easy to assemble a bigger knowledge base with less resources such as memory or compute to store and process data. [7]

Chang et al. (2024) contribute to improving retrieval processes in fact-checking by integrating knowledge graphs (KGs) with RAG systems. They introduce a community-aware retrieval mechanism that leverages multi-hop pathways within KGs to improve the precision of retrieved information. This enhances the relevance and contextual accuracy of the data retrieved for fact-checking tasks, surpassing traditional retrieval methods. Their system also performs well in zero-shot learning settings, adapting to new queries without additional fine-tuning, making it highly scalable. [5]

Due to the potentially large size of graph databases, especially in real-world applications, optimization strategies are necessary to enhance the efficiency of retrieval. Techniques such as graph indexing, approximate nearest neighbor search, and multi-stage retrieval methods are utilized to narrow down the search space and reduce computational overhead. In Graph RAGs, the trade-off between retrieval accuracy and speed is crucial, particularly in applications requiring real-time responses. Hybrid retrieval methods, which combine both parametric (e.g., neural models) and non-parametric techniques, are increasingly popular for balancing performance with scalability.

To optimize retrieval efficiency and ensure only relevant information is passed through, He et al. (2024) introduce a method that constructs subgraphs based on the Prize-Collecting Steiner Tree (PCST) algorithm. This method assigns relevance scores (or "prizes") to nodes and edges, constructing a subgraph that maximizes relevance while minimizing retrieval cost. The adaptation of PCST for both nodes and edges allows for efficient subgraph construction that maintains manageable size without sacrificing important relational data [6].

One of the limitations of graph RAG systems is what is called multi-hop reasoning, which means arriving to answers that require several steps of reasoning in the knowledge graph to get there. Fang et al. (2024) [8] proposes a solution to this problem, via the creation of ground based reasoning chains, which are chains of nodes and edges that are used to construct a response to a query, basically, constructing a chain of thought where given any point in the chain you can move forward and the last node is the answer to the query.

Cui et al. (2024) contribute to improving retrieval in Retrieval-Augmented Generation (RAG) systems by proposing the Bailicai framework. They highlight the importance of refining retrieval processes to reduce noise from irrelevant documents and optimizing the timing of retrieval in large language models (LLMs). By utilizing adaptive strategies, such as self-knowledge boundary identification, Bailicai ensures that retrieval is only invoked when necessary, improving efficiency. This approach addresses the challenge of document noise and improves accuracy in medical retrieval tasks, contributing significantly to the integration of external knowledge in LLMs. [3]

Jin et al. (2024) make significant contributions to retrieval in RAG systems by introducing the RAGCache framework. They focus on optimizing retrieval processes by caching intermediate knowledge states, effectively reducing redundant computations in RAG systems. By leveraging multilevel dynamic caching strategies, RAGCache enhances retrieval efficiency, especially for high-frequency document requests, which are cached to minimize retrieval latency. The study also discusses the performance bottlenecks of retrieval steps in RAG and how caching intermediate results can drastically cut down on retrieval and computational costs. [4]

Wu et al. (2024) make significant contributions to the retrieval process by introducing a U-retrieve method, which enhances the retrieval efficiency of RAG systems in medical applications. U-retrieve balances global awareness and indexing efficiency by generating summaries at different graph levels and performing top-down matching. This ensures that the system retrieves highly relevant entities and their relationships across multiple layers of medical knowledge. Their method enables precise, evidence-based responses to medical queries, dramatically improving the reliability of retrieval-augmented language models in critical medical scenarios. [9]

Graph Neural Networks (GNNs) are a class of neural networks tailored to process graph-structured data by capturing the dependencies between nodes via message-passing mechanisms. GNNs extend traditional neural networks by incorporating the topology of graphs into the learning process, enabling models to learn representations that reflect both the features of individual nodes and their structural context within the graph.

Several GNN architectures, such as Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and Graph Recurrent Neural Networks, are particularly relevant to RAG systems. These architectures differ in how they aggregate and propagate information across the graph. For instance, GATs leverage attention mechanisms to weigh the importance of neighboring nodes, which can enhance the retrieval process in RAGs by focusing on more relevant relationships.

Peng et al. (2024) highlighted the role of GNNs in GraphRAG systems, explaining that GNNs are used to represent graph structures and enhance the retrieval and generation processes. GNN-based retrievers are particularly suited for tasks involving complex relationships in graph-structured data [1]

In Graph RAG systems, GNNs are utilized to encode graph knowledge bases into embeddings that can be integrated with language models. Applications include improving entity recognition, relation extraction, and providing context-aware responses. GNNs enable the system to understand complex interactions within the graph, leading to more accurate retrieval and generation of information in response to user queries.

He et al. (2024) uses a GAT to encode the subgraph response from the retrieval phase using a Pooling operation, followed by a multilayer perceptron and finally using a text embedder to turn the output in textual form, yielding is a smaller amount of tokens required to perform a question-answer operation. [6]

Fang et al. (2024) proposes de use of a GNN in the retrieval phase of the RAG system to retrieve the top k relevant triplets (head, edge, tail) in the knowledge graph. [8]

Efficient query processing is critical for the performance of Graph RAG systems. Techniques such as graph traversal algorithms, indexing strategies, and approximate nearest neighbor search are employed to retrieve relevant information swiftly. Optimizing these processes ensures that the retrieval component does not become a bottleneck in the generation pipeline.

Hu et al. (2024) highlight the importance of efficient graph querying and representation learning, embedding entities and relations into a shared space with the language model to facilitate seamless interaction between structured and unstructured data. [2]

Optimization strategies in Graph RAGs focus on reducing latency and computational overhead. Methods include precomputing embeddings, using caching mechanisms, and employing parallel processing. Additionally, pruning irrelevant parts of the graph and leveraging hierarchical structures can improve retrieval efficiency.

Peng et al. (2024) explored how GraphRAG enhances query processing through techniques like query expansion and decomposition. These methods ensure more precise retrievals by refining the original query with additional context or breaking it down into smaller, more manageable parts [1]

Cui et al. (2024) contribute to query optimization within the context of RAG systems by introducing mechanisms that determine whether a medical query requires external retrieval or can be resolved using internal model knowledge. Their self-knowledge boundary identification module optimizes the retrieval process by avoiding unnecessary external data retrieval, reducing computational time and resource consumption. The combination of this method with task decomposition ensures that each query is handled efficiently, either through internal knowledge or by accessing the relevant external documents. This dynamic and targeted retrieval process offers significant improvements in query processing for medical applications. [3]

Jin et al. (2024) introduce several key optimizations for query processing in RAG systems. One of the primary contributions is the development of dynamic speculative pipelining, which overlaps the retrieval and inference stages to minimize query processing time. This technique ensures that as retrieval results are generated, they are immediately used in inference, reducing latency. Additionally, their cache-aware scheduling and a prefix-aware replacement policy ensure that frequently retrieved documents are quickly accessible, improving the overall query performance. This approach is shown to significantly reduce end-to-end query time, offering a scalable solution for systems handling large volumes of complex queries. [4]

Huang et al. (2023) introduce optimization strategies such as RAC and DKS, which focus on retrieving passages that are knowledge-relevant. This optimization ensures that the system retrieves the most meaningful passages for a given query, improving both retrieval and generation performance. Their work significantly enhances query efficiency, particularly by focusing on dense knowledge matching rather than relying solely on token-level matching. [10]

Chang et al. (2024) optimize the query process by introducing a hierarchical, community-driven retrieval system within their knowledge graph framework. This approach dynamically selects the most relevant communities and sentences based on their relationship to the query, improving both the speed and accuracy of fact-checking. The zero-shot capability of their system means that query optimization occurs without the need for re-training, making the solution efficient and scalable for various fact-checking and retrieval tasks. [5]

Wu et al. (2024) introduce several optimizations to the query process through their U-retrieve mechanism, which involves top-down matching of queries with graph-based structures. This approach minimizes the search space by dynamically indexing and matching queries with relevant graph nodes, enabling faster and more accurate retrieval. Additionally, the use of hierarchical graph structures allows the system to retrieve information at varying levels of detail, ensuring that even complex medical queries are handled efficiently and effectively. [9]

Huang et al. (2023) contribute to the field of retrieval by introducing the concept of Dense Knowledge Similarity (DKS) and Retriever as Answer Classifier (RAC). These components enhance the retrieval process by ensuring that the retrieved passages are not only label-relevant but also knowledge-relevant, which addresses the issue of missing key knowledge-relevant passages in the retrieval process. Their experiments show significant improvements in retrieval metrics such as Recall@1 on the MSMARCO dataset and R-Precision on the KILT-WoW dataset. [10]

Evaluating the performance of retrieval techniques in Graph RAGs involves assessing both the accuracy of the retrieved subgraphs and the efficiency of the retrieval process. Metrics such as precision, recall, and F1-score are commonly used to measure retrieval relevance. Additionally, specific metrics related to graph-based retrieval, such as subgraph quality and path relevance, are considered. Speed metrics, including query latency and retrieval time, are also critical for applications where real-time generation is required. Benchmarking on standard datasets helps ensure that retrieval methods can generalize to various domains while maintaining performance.

generation

The integration of RAGs into NLP tasks enhances language models by providing access to external knowledge during text generation. This approach improves the factual accuracy and relevance of outputs in tasks such as question answering, dialogue systems, and summarization. By retrieving pertinent information from graph knowledge bases, RAGs can fill knowledge gaps inherent in standalone language models.

Semantic parsing involves converting natural language into structured, machine-understandable representations. Graph RAGs contribute to this process by retrieving and incorporating relevant graph structures that reflect the semantic meaning of the input. This integration allows for more precise interpretation of queries and the generation of responses that are aligned with the underlying knowledge base.

RAG-based representations enhance NLP models by embedding retrieved knowledge directly into the generation process. This fusion of retrieval and generation enables models to produce outputs that are contextually enriched and factually consistent with the external knowledge base. Techniques such as joint training and attention mechanisms over retrieved content are explored to maximize the synergy between retrieval and generation components.

Peng et al. (2024) addressed limitations in traditional NLP models by leveraging graph structures to retrieve relational knowledge. They demonstrated that GraphRAG outperforms conventional RAG systems by incorporating both textual and structural information, enabling more context-aware responses [1]

non-finance rag studies

In healthcare, Graph RAGs assist in clinical decision support by retrieving relevant medical knowledge in response to patient data. They enable personalized treatment plans by considering complex relationships between symptoms, diseases, and treatments. This application requires handling sensitive data and ensuring compliance with privacy regulations.

Cui et al. (2024) demonstrate that the Bailicai framework enhances the performance of LLMs in medical tasks. By integrating medical knowledge injection and self-knowledge boundary identification, the framework effectively mitigates hallucinations, a common issue in medical text generation. Bailicai surpasses proprietary models like GPT-3.5 and achieves state-of-the-art performance across several medical benchmarks. The researchers focus on curating datasets from the UltraMedical dataset and ensuring that the model learns from high-quality medical data. The framework's ability to handle medical question-answering tasks with precision makes it a significant advancement in medical AI applications. [3]

Wu et al. (2024) propose a graph-based Retrieval-Augmented Generation (RAG) framework specifically designed for medical applications, called MedGraphRAG. This framework constructs a hierarchical graph from medical documents, medical textbooks, and authoritative medical dictionaries, significantly enhancing the capability of Large Language Models (LLMs) to process and retrieve medical information. By employing a graph-based structure, the system is able to preserve complex relationships between medical entities, providing superior retrieval accuracy compared to flat databases or simple key-value retrieval systems. The hierarchical nature of the graph allows for more precise linking and retrieval, ensuring that the generated responses are evidence-based and contextually accurate. [9]

The primary focus of Wu et al. (2024) is on enhancing LLM performance in the medical domain. Their MedGraphRAG system addresses the specific needs of medical professionals by ensuring that generated responses are backed by evidence from credible medical sources, such as textbooks and peer-reviewed papers. The hierarchical graph structure they employ links user-provided documents with established medical knowledge, ensuring transparency and interpretability in medical diagnoses and recommendations. MedGraphRAG's performance on multiple medical Q&A benchmarks, including PubMedQA, MedMCQA, and USMLE, demonstrates its effectiveness, surpassing even fine-tuned models in generating accurate, evidence-based medical information. [9]

Graph RAGs enhance social network analysis by generating insights based on the structure and dynamics of social interactions. They can identify influential nodes, detect communities, and predict relationship evolution. This information is valuable for marketing strategies, trend analysis, and understanding social phenomena.

In GIS, Graph RAGs facilitate the retrieval of spatial and relational data for applications like route planning, resource management, and environmental monitoring. By integrating various geospatial datasets, RAG systems can generate comprehensive analyses and support decision-making processes that consider multiple geographic factors.

The work of He et al. (2024) represents a significant advance in the world of domain specific applications using the RAG architecture. By creating a framework of questions and answers for a graph RAG system, it paves the path for more structured and less hallucination-proned domain specific chatbots. [10]

finance rag studies

Summary of “Domain Specific Data Quality Framework” by Komal Azram Raja (2024)

Overview

The thesis develops a **Retrieval-Augmented Generation (RAG)** framework to improve **data quality management** in the financial sector, focusing on **contextual data validation** [11]. Financial data poses unique challenges due to regulatory, security, and consistency requirements. The study emphasizes the need for **domain-specific models** to address these issues, as traditional tools often lack the necessary **context awareness** . [11]

Methods and Implementation

The framework leverages **Large Language Models (LLMs)** with **knowledge graphs** to improve the accuracy of data quality processes [11]. Key technologies include **Hugging Face embeddings** and **Llama 2** models for query processing. Several advanced techniques were employed for data collection, preprocessing, and building the architecture:

1. **Web Scraping for Data Collection**
   - **Scrapy** and **BeautifulSoup** were used to extract relevant financial content, focusing on **publicly available financial blogs and articles** .
   - The collected data was cleaned, removing unnecessary HTML tags, and structured into **PDF documents** for further analysis.

2. **Embeddings and Model Architecture**
   - **Hugging Face’s all-mpnet-base-v2** was used to generate **sentence embeddings**, mapping text into a **768-dimensional space** for better semantic understanding.
   - The RAG framework employs a **vector index store** for faster query retrieval and more accurate contextualization.

3. **System Architecture and Configuration**
   - The system integrates **Llama 2** with pre-trained models consisting of **7 billion parameters** for improved response quality.
   - A **query engine** leverages the indexed vector store to provide **context-aware answers** to financial data-related queries.

Evaluation Using Financial Data

The framework was tested with **three use cases** relevant to financial data quality:

1. **Timeliness Check**
   - Query: “How to check if stock data is stale?”
   - Result: The framework provided detailed recommendations for evaluating stale stock data, going beyond traditional **timestamp checks** (Atlan, n.d.).

2. **Completeness Check**
   - Query: “What strategies should we use if customer income data is missing?”
   - Result: The framework suggested context-aware strategies such as **regression imputation** and explained the consequences of various imputation techniques (Finance Train, n.d.).

3. **Consistency Check**
   - Query: “Are these symbols ($, €, &&, ?) valid currencies?”
   - Result: It recommended **ISO-based currency validation**, outperforming regex-based methods used by **SQL Server DQS** (Microsoft, n.d.).

These results demonstrate that the **RAG-based approach provides actionable insights** not achievable with standard data quality tools [11].

Results and Analysis
The responses were evaluated using the **RAGAS framework** (Doe, Smith, & Turing, 2023). While the RAG model showed high **answer relevancy and recall**, it could benefit from improving **context precision** and **faithfulness**. The absence of **harmful content** confirms the framework's safety and reliability [11].

Comparison with existing tools, such as **Talend**, **Informatica**, and **SQL Server**, shows the RAG framework's ability to provide **context-aware recommendations** [11]. For instance, the **RAG model** not only detected missing values but also suggested **root-cause analysis** for financial anomalies.

Limitations and Future Enhancements
The thesis [11] highlights two key challenges :

1. **Scalability and Performance Constraints**:
   - The use of **Google Colab** limited scalability due to **memory constraints** and lack of GPU resources .
2. **Domain-Specific Knowledge Limitations**:
   - Future enhancements include **training the model with real-world datasets** from **Jira logs** and **Notion entries** to improve performance .

Conclusion
This research shows that **RAG-based frameworks**, integrated with **knowledge graph databases**, can significantly improve **data validation processes** for the financial industry. By using domain-specific insights, organizations can achieve **better compliance**, **reduced false positives**, and **more accurate data governance** [11]. Further developments will focus on deploying **local RAG models** for **real-world testing** to overcome scalability issues.

Summary of "Scaling AI Adoption in Finance: Modelling Framework and Implementation Study" by Sepanosian, Milosevic, and Blair (2024)

Overview
The financial services industry is experiencing a transformation driven by **Artificial Intelligence (AI)**, with applications spanning **fraud detection, algorithmic trading**, and **personalized financial services** [12]. This paper presents two key contributions:
1. **An industry-oriented framework** to guide scalable AI adoption in financial institutions.
2. A **proof-of-concept implementation** for retirement planning using **multi-agent systems and RAG (Retrieval-Augmented Generation)** technologies.

The framework and proof-of-concept aim to help financial organizations integrate **advanced AI techniques**, from **basic task automation** to **complex multi-agent systems**. These solutions address both immediate business challenges and future developments, ensuring that financial institutions can remain competitive in an evolving landscape.

Methods and Implementation
1. **Framework for Scalable AI Adoption** The framework categorizes AI solutions into problem areas such as **information retrieval, data analysis, content generation, decision support, and task automation** [12]. The goal is to enable financial institutions to **progress from basic AI capabilities**, such as **customer query handling**, to more advanced applications, such as **personalized financial advice**. Central to the framework is the **RAG technology**, which ensures the use of up-to-date, accurate information without needing constant model retraining.

2. **Multi-Agent System for Retirement Planning** The proof-of-concept implementation was developed using **crewAI**, a **multi-agent orchestration framework** (João, 2024). Four specialized agents were deployed to address various aspects of retirement planning:
   - **Retirement Policy Expert**: Provides information on **contribution limits** and **regulations**.
   - **Industry Expert**: Delivers insights on **investment trends** and **average savings**.
   - **Advisory Expert**: Aggregates the information and delivers **personalized advice**.
   - **Quality Assurance (QA) Agent**: Ensures **accuracy** by cross-referencing information and flagging inconsistencies.

   These agents interact sequentially, using **RAG tools and LangChain functionalities** to dynamically retrieve relevant data. Each agent works autonomously within the framework, delegating tasks and providing **contextual insights** for customer inquiries [12].

Results and Findings

In testing the system with a **fictional client** (John Doe), the agents generated **personalized reports** comparing his retirement savings with industry averages (Figure 5, [12]). The agents were able to:

1. Identify that Doe's investments were performing **competitively**.
2. Advise him on potential **additional contributions** to reach his goals.
3. Provide **regulatory insights** based on recent updates from the Australian Taxation Office.

The agents' collaboration yielded valuable results, but challenges emerged related to **explainability** and **non-deterministic outputs** inherent to LLMs. The **QA agent** often required manual review due to **inconsistent or ambiguous outputs** from the advisory agent.

Key Observations

- **Transparency and Explainability**: LLM-based agents produced **stochastic outputs**, making it difficult to **trace decision paths** or ensure predictable results [12]. Solutions may involve **auditing agent behaviors** and **tracking tool usage**.
- **Security Risks**: Using external tools like **OpenAI's API** introduced concerns about **data security**. While **local deployment** of agents is feasible, it comes with **high computational costs**.
- **Scalability Challenges**: Handling complex data types (e.g., tables, images) added **preprocessing overhead**. Redundant API calls reduced performance, especially when agents were not **effectively sharing memory** [12].

Limitations and Future Directions

- **Consistency in Outputs**: Current agent definitions were relatively **generic**, resulting in **inconsistent advice** across different test runs. Future work will refine the agents to **align with specific retirement planning goals** [12].
- **Integration with Knowledge Graphs**: A promising direction is **integrating knowledge graphs** to enhance the agents' contextual awareness.
- **Distributed Multi-Agent Systems**: Future systems may distribute agents across multiple environments, improving **resilience and performance** [12].

Conclusion

The study demonstrates that **multi-agent systems** powered by **LLMs and RAG** tools offer significant potential for financial services, particularly in **personalized advisory applications**. However, the success of these technologies depends on balancing **performance, transparency, and security**. Financial institutions can leverage these findings to **scale AI adoption** responsibly and **continuously adapt** to new challenges [12].

Summary of "Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations" by de Zarzà, de Curtò, Roig, and Calafate (2024)

Overview

[13] proposes an innovative financial planning framework that combines traditional budgeting models with AI-powered large language models (LLMs) to enhance financial management at both individual and cooperative (household) levels. The core objective is to democratize financial planning by providing personalized recommendations that optimize savings and budgeting through accessible AI tools.

Key Methods and Framework

The study focuses on two interconnected financial planning models:

1. Individual Budget Optimization Model
   - The model aims to maximize savings by distributing monthly income efficiently across expense categories (e.g., rent, groceries, and utilities).
   - A utility function prioritizes saving while meeting necessary expenses, defined as:

$$U(I, E) = \alpha \log(S) - \beta \sum_{o=1}^{n} w_o \log(E_o)$$

where I is income, S is savings, E_o are categorized expenses, and w_o are personalized preference weights.

- The LLM-generated recommendations serve as a guiding baseline to suggest feasible allocations, balancing real-world financial needs with optimization.

2. Cooperative Budgeting Model for Households
   - Extends the individual model to address shared financial responsibilities among household members, such as rent and grocery costs.
   - The cooperative model ensures financial equity among members by considering individual priorities while optimizing the total savings:

$$\mathrm{TS} = \sum_{o} = 1^n \left(I_o - \sum_{z} E_{\mathrm{zo}} \right)$$

where TS is the total household savings, I_o is each member's income, and $E_{\{\mathrm{zo}\}}$ are their respective expenses per category z.

- LLM-informed recommendations offer advice on efficient shared budget planning and propose strategies to minimize redundant expenses (e.g., shared subscriptions).

Financial Planning with LLMs

The integration of LLMs offers several advantages (([13])):

- Personalization: Tailors recommendations based on individual and household-specific data.
- Dynamic Adaptability: Adjusts recommendations to reflect economic trends and evolving financial goals.
- Enhanced Savings Strategies: Includes advice on building emergency funds, retirement savings, and managing short-term vs. long-term priorities.
- Consumption Smoothing: Uses cooperative game theory principles to maintain consistent living standards over time, ensuring long-term financial stability.

Financial Planning Case Studies

[13] evaluates scenarios using both individual and cooperative models, guided by LLM recommendations. Key insights include:

1. Short-term Budgeting: The LLM suggests emergency fund allocations and prioritizes debt repayments.
2. Retirement Planning: In a household scenario, LLMs guide members to optimize savings, balancing retirement contributions with daily expenses.
3. Long-term Financial Goals: A child fund is recommended to accommodate future expenses (e.g., education, healthcare).

Cooperative Financial Planning within EC Framework

The extended coevolutionary (EC) theory is employed to model the interaction of financial agents (individuals or households) as an adaptive system. Agents iteratively adjust their spending based on LLM recommendations and their financial environment. For example, given an agent o and a time t:

$$E_o(t+1) = (1-\beta)\Big(E_o(t) + \alpha \nabla U_o\Big(E_{o(t)}, E_{-o}(t)\Big)\Big) + \beta R_{o,t}$$

where $R_{o,t}$ is the LLM recommendation and α and β control learning and LLM influence. ([13])

This approach ensures resilience in dynamic financial landscapes by enabling households to collaboratively optimize spending strategies.

Results and Evaluation

[13] demonstrates that LLM-generated financial recommendations yield optimized budget plans that are economically sound and align with best practices in personal finance. However, human oversight is still essential to validate LLM suggestions and avoid potential errors. The combination of LLM insights with traditional optimization models shows significant improvements in:

- Savings rates
- Cooperative financial management
- Adaptability to changing income and expenses

Bibliography

[1] BOCI PENG *et al.*, “Graph Retrieval-Augmented Generation: A Survey,” vol. 37, no. 4.

[2] Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao, “GRAG: Graph Retrieval-Augmented Generation.”

[3] Cui Long, Yongbin Liu, Chunping Ouyang, and Ying Yu, “Bailicai: A Domain-Optimized Retrieval-Augmented Generation Framework for Medical Applications.”

[4] Chao Jin *et al.*, “RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation.”

[5] Rong-Ching Chang and Jiawei Zhang, “CommunityKG-RAG: Leveraging Community Structures in Knowledge Graphs for Advanced Retrieval-Augmented Generation in Fact-Checking.”

[6] Xiaoxin He *et al.*, “G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering.”

[7] Zetian Hu, Weijian Xie, Xuefeng Liang, Yuhui Liu, Kaihua Ni, and Hong Cheng, "WeKnow-RAG: An Adaptive Approach for Retrieval-Augmented Generation Integrating Web Search and Knowledge Graphs."

[8] Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald, "TRACE the Evidence: Constructing Knowledge-Grounded Reasoning Chains for Retrieval-Augmented Generation."

[9] Junde Wu, Jiayuan Zhu, and Yunli Qi, "Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation."

[10] Wenyu Huang, Mirella Lapata, Pavlos Vougiouklis, Nikos Papasarantopoulos, and Jeff Z. Pan, "Retrieval Augmented Generation with Rich Answer Encoding."

[11] Komal Azram Raja, "Domain Specific Data Quality Framework."

[12] Thomas Sepanosian, Zoran Milosevic, and Andrew Blair, "Scaling AI adoption in finance: modellingframework and implementation study."

[13] I. de Zarzà, J. de Curtò, Gemma Roig, and Carlos T. Calafate, "Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations."

[14] Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald, "REANO: Optimising Retrieval-Augmented Reader Models through Knowledge Graph Generation," vol. 1.