RAG

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to all those who have supported and guided us throughout the completion of this graduation project.

First and foremost, We extend our deepest thanks to our supervisors, Dr Wael Badawy and T.A. Mohamed Tawfik, for their invaluable guidance, encouragement, and constructive feedback, which were instrumental in shaping this project. Their expertise and support have been a source of inspiration throughout our journey.

We are also profoundly grateful to ERU for providing the necessary resources, facilities, and a conducive environment for the successful execution of this project.

Table of Contents

List of Figures

List of Abbreviations

- AI: Artificial Intelligence
- RAG: Retrieval-Augmented Generation
- CRM : Customer Relationship Management
- LLMs: Large Language Models

Abstract

The banking industry is increasingly under pressure to meet the growing demands of customers who expect fast, efficient, and accessible service across digital platforms. Traditional customer service models, such as in-branch visits, call centers, and basic AI chatbots, often fail to address the complex and diverse needs of modern customers. These methods are slow, resource-intensive, and often inefficient when handling tasks like loan applications, account openings, and the verification of documents. Existing AI solutions, although capable of providing basic assistance, lack real-time data retrieval and struggle with maintaining context over multi-turn conversations. As a result, customers experience delays, frustration, and poor service quality, which negatively impacts customer satisfaction and retention, and increases operational costs.

This project aims to explore the potential of Retrieval-Augmented Generation (RAG) technology to revolutionize customer service in the banking industry. RAG combines the strengths of real-time information retrieval with advanced natural language generation, enabling systems to provide accurate, context-aware, and dynamic responses to customer queries. Unlike traditional chatbots, RAG systems can pull in relevant, up-to-date information from various sources, allowing them to respond intelligently to complex, multi-step queries. By integrating this technology, banks can offer personalized, efficient, and scalable customer support, thereby improving service quality, reducing response times, and lowering operational costs.

This paper examines the limitations of current customer service solutions, the potential of RAG technology, and its applications within the banking sector. It highlights how RAG can address key issues such as information retrieval, multi-turn dialogue, and real-time accuracy, providing a solution that better meets customer needs. This study outlines the broader impact of implementing RAG in banking, including enhanced customer satisfaction, increased operational efficiency, and improved competitive positioning. Ultimately, this research proposes RAG as a transformative technology that can reshape the future of banking customer service, enabling financial institutions to deliver a superior and more responsive customer experience in an increasingly digital world.

1.1 overview

The world is rapidly shifting towards a more digital focus, especially in the banking industry. Traditional banks are undergoing major digital transformations to meet the needs of both new and existing customers who demand a more tailored and individualized banking experience through digital channels [4]. To achieve this, banks and financial institutions must adopt a digital transformation strategy centered around customer experience by analyzing, interacting, and understanding customer needs[38].

In the following pages we'll explain the problems the current bank sector faces, the evolution of customer services, what is RAG, how it works and how it's implementation in the system is more helpful than your normal AI models, the main objectives of the project, related work,…

1.2 Problem Description

In today's digital-first world, banking customers expect seamless, quick, and reliable access to services, whether they are applying for loans, opening new accounts, or verifying required documents. Traditional customer service methods, such as call centers or in-branch visits, are resource-intensive, time-consuming, and often inconvenient for users. Although banks have implemented AI-powered chatbots and virtual assistants to automate support, these systems face significant limitations: they cannot retrieve real-time data, often fail to provide accurate answers for complex or multi-turn queries, and struggle to adapt to dynamic user needs [3].

These limitations result in delayed responses, incomplete or irrelevant information, and customer dissatisfaction. For example, a customer looking for detailed guidance on loan application or documentation requirements may not receive the correct, up-to-date answers through existing chatbot systems, forcing them to seek human assistance, thereby increasing service time and operational costs. As customer expectations for digital convenience and efficiency continue to grow, these inefficiencies pose a significant challenge for financial institutions striving to retain customers and improve service quality [23].

To address this problem, there is a need for an advanced, intelligent system that can provide factually accurate, context-aware, and personalized responses to customer queries in real time. Retrieval-Augmented Generation (RAG) technology offers a promising solution by combining the strengths of real-time information retrieval and natural language generation. By deploying RAG-based systems, banks can enhance their customer service capabilities, improve user satisfaction, and optimize operational efficiency, positioning themselves to meet the evolving demands of modern customers.

RAG is an AI framework that combines the strengths of traditional information retrieval systems with the capabilities of generative large language models (LLMs) by integrating information retrieval techniques [3]. It enables companies to retrieve and use their data from various internal sources for better generative AI results. Because the source material is derived from trusted data, RAG reduces or even eliminates errors and incorrect outputs, generating relevant and accurate information[45].

1.3 Project Motivation

This project is motivated by the increasing need to improve challenges in customer service within the banking sector. Many customers struggle with understanding the documentation and procedural requirements for services like loans and account setup. Current systems often fail to deliver timely, accurate, and personalized responses, leading to frustration and inefficiencies [2]. As banking becomes increasingly digital, customer expectations for support are at an all-time high. Banks face challenges in sustaining these interactions while managing costs and operational efficiency [5]. This project employs RAG, a state-of-the-art AI technology, to revolutionize the way banks handle customer inquiries by combining generative AI with real-time information retrieval [3]. The goal is to automate routine queries, ensure accuracy, and provide 24/7 personalized support, thereby enhancing customer satisfaction and easy to use information.

1.4 Main Objectives

1. Enhance customer experience.
2. Reduce staff workload.
3. Save time for customers.
4. Decrease paperwork and errors.
5. Simplify information retrieval.
6. Develop hybrid models where AI handles simple inquiries and escalates complex cases to human agents.
7. Ensure accuracy and up-to-date information.
8. Boost efficiency with automation.

1.5. Applications of the Project

1. Document Requirements Guide for Loan Applications.

2.Account Setup Documentation Assistance.

3. Credit Card Application Guidance.

4.Documentation for International Services.

5. Insurance Services and Claims Documentation.

6. Compliance and Regulatory Information.

7. Savings and Investment Account Documentation.

1.6 Background

The banking industry has undergone significant transformation over the past two decades, driven by technological advancements and changing customer preferences [4]. Historically, banks relied on in-person interactions to serve customers.

Early History: In the 17th century, with the establishment of major banks like the Bank of England (1694), customer service was personalized and conducted face-to-face, using handwritten ledgers for record-keeping.

19th Century: The industrial revolution brought network expansion, formalized customer service, and the emergence of branch managers as key relationship builders.

20th Century:

- Early 1900s: The telephone enabled remote customer service for basic inquiries.
- 1980s: Early Customer Relationship Management (CRM) systems emerged, allowing banks to track interactions and preferences for personalized service [5].

2000s:The rise of the internet, mobile banking apps, and social media expanded customer service to multiple channels, enhancing accessibility and convenience.

2010s:Digital tools like chatbots, virtual assistants, and AI-powered solutions surged, addressing some service challenges but leaving gaps in personalization and handling complex queries [4].

2020s: The release of GPT-3 marked a pivotal shift in AI capabilities, combining generative models with information retrieval to deliver accurate, context-aware responses [1]. While GPT-3 was transformative, persistent challenges such as compliance, data accuracy, and personalization remained.

By 2023, the integration of Retrieval-Augmented Generation (RAG) into banking processes began to address these gaps. RAG's ability to combine real-time data retrieval with generative AI allows banks to streamline customer service, enhance efficiency, and reduce operational costs [3]. This innovation enables banks to offer precise, personalized assistance while meeting the growing demand for transparent and accurate information in real time.

Linking Historical Trends to RAG:

This progression highlights the evolving priorities of banking—from face-to-face interactions to digital automation. Challenges like complex document management and rising customer expectations pushed banks to adopt solutions like RAG. Unlike earlier tools, RAG combines trusted data retrieval with generative AI, setting a new benchmark for accuracy, speed, and personalization in customer service.

Related Work

2.1 Introduction

Recent advancements in AI have significantly impacted customer service systems. Retrieval-Augmented Generation (RAG) is a notable approach for improving AI-driven customer service by enhancing the contextual relevance and accuracy of responses. OpenAI's GPT-3 demonstrated the potential for few-shot learning, enabling AI models to handle a wide variety of queries with minimal training data.

In the banking domain, McKinsey & Company (2021) emphasized the importance of AI in reducing operational costs and meeting rising customer expectations . Additionally, Accenture (2022) highlighted how banks have increasingly adopted AI technologies, including RAG, to deliver personalized, real-time support while addressing compliance challenges.

Rajpurkar et al. (2018) explored question-answering systems and highlighted the limitations of existing models in handling unanswerable or ambiguous queries, which RAG addresses by integrating trusted data sources. Meanwhile, conversational AI systems, such as chatbots, have been widely studied for their efficiency and scalability. However, their inability to manage complex inquiries has prompted the integration of hybrid models where AI and human agents collaborate effectively.

2.2 Related Work

Alternative Methods for the Same Problem

Chatbot Enhancements Huang et al. (2020) proposed an enhanced chatbot system that utilizes reinforcement learning to improve customer satisfaction in banking services. Their approach focuses on learning from interactions to better handle customer inquiries, which contrasts with RAG's method of leveraging retrieval-based information for response generation.

End-to-End Memory Networks: Sukhbaatar et al. (2015) introduced end-to-end memory networks for question-answering tasks. This method emphasizes the use of memory components to store and retrieve relevant information, differing from RAG's approach but aiming to solve similar problems in customer service.

Same Method for Different Problems

Healthcare Customer Service: Kwon et al. (2021) applied RAG to healthcare customer service, aiming to provide accurate and timely responses to patient inquiries. This study demonstrated how RAG could be used to integrate medical databases and patient records to enhance service delivery in the healthcare sector, highlighting the versatility of the method beyond banking.

Legal Document Review: Wu et al. (2021) utilized RAG for legal document review and summarization, showcasing its application in extracting relevant information from vast legal texts and generating concise summaries to assist legal professionals.

Similar Methods for Similar Problems

Hybrid AI Models: Zhu et al. (2019) explored the use of hybrid AI models combining rule-based systems with machine learning to handle complex customer service tasks in telecommunications. Their approach is similar to RAG in integrating multiple sources of information to improve response accuracy and relevance, addressing the similar issue of handling diverse and complex customer queries.

Knowledge-Enhanced Neural Networks: Chen et al. (2019) proposed knowledge-enhanced neural networks for customer support, which integrate structured knowledge bases with neural models to provide more accurate and contextually relevant responses.

Related Problems in the Domain

Fraud Detection and Prevention: Lin et al. (2018) discussed the application of AI in fraud detection within the banking sector, which is a related problem domain. While focusing on security rather than customer service, their work highlights the importance of accurate data retrieval and real-time processing, principles that align with the goals of RAG in enhancing customer service.

Sentiment Analysis for Customer Feedback: Zhang et al. (2020) examined sentiment analysis techniques for analyzing customer feedback in the banking sector. Understanding customer sentiment can help improve service quality, and integrating such analysis with RAG could further enhance customer interactions.

Knowledge Graphs in Customer Service: Sun et al. (2020) explored the use of knowledge graphs to enhance customer service by providing a structured way to store and retrieve information. This method shares similarities with RAG's integration of retrieval-based techniques but focuses more on the organization and accessibility of knowledge.

Specific Use Cases

- **Document Requirements Guidance**: Providing accurate, real-time information on documentation required for various banking services.
- **Account Setup Assistance**: Helping customers through the process of setting up new accounts with personalized, step-by-step guidance.
- **Loan Application Support**: Offering detailed information and answering queries related to loan applications and approvals.
- **Compliance and Regulatory Information**: Ensuring customers have access to the latest regulatory requirements and compliance information.

3.1 Problem Statement

The current methods for delivering customer support in the banking industry are unable to meet the growing expectations of modern users who demand instant, accurate, and context-aware solutions. Traditional approaches, such as in-branch visits, call centers, and basic AI chatbots, are inefficient in handling complex tasks like loan applications, account openings, and documentation verification. These systems lack the ability to retrieve real-time information and generate precise, tailored responses, leading to delayed service, customer frustration, and operational inefficiencies. To overcome these challenges, an advanced solution leveraging Retrieval-Augmented Generation (RAG) technology is required to improve accuracy, enhance user satisfaction, and streamline customer service delivery in the banking sector.

3.2 Problem Analysis

1. Multi-channel Customer Interaction

   - Customers expect support across multiple channels (such as live chat, phone, email, and social media). Managing these channels effectively is complex.

2. Artificial Intelligence with Limited Understanding

   - Despite advancements in AI, understanding natural language can still be limited in some cases (especially with dialects or ambiguous phrases).

3. Managing High Customer Expectations

   - Customers expect instant responses and quick resolutions to all inquiries, which can put pressure on support teams.

4. Employee Retention and Training

   - Some companies struggle with continuously training customer service staff to keep up with technological advancements and customer expectations.

5. Handling Complex Inquiries

   - AI systems may not be able to handle complex or special cases that require human intervention.

3.3 Project Visibility Study

1. Project Objectives

- **Enhance Customer Experience:** Improve the customer experience in banks by providing fast and accurate customer support services [36].

- **Reduce Staff Workload:** Minimize the workload on customer service staff by automating routine tasks [37].

- **Ensure Accurate and Relevant Information:** Deliver precise and up-to-date information using Retrieval-Augmented Generation (RAG) technology [38].

- **Boost Efficiency with Automation:** Automate information retrieval and reduce the time needed for customer interactions [39].

2. Target Audience

- **Banks and Financial Institutions:** Institutions aiming to enhance customer service and operational efficiency [40].

- **Bank Customers:** Customers relying on digital banking services who require quick and accurate responses [41].

- **Customer Service Teams:** Teams seeking to reduce effort and increase efficiency in handling customer queries [42].

3. Project Impact

- **Improved Customer Satisfaction:** Increase customer satisfaction by meeting high expectations and providing instant responses [43].

- **Operational Efficiency:** Enhance operational efficiency for banks by reducing errors and manual workloads [44].

- **Increased Financial Literacy:** Help customers better understand financial services by offering accurate and easy-to-understand information [45].

- **Enhanced Financial Inclusion:** Encourage less digitally-savvy customers to adopt and trust digital banking services [46].

4. Competitor Analysis

- **Existing Solutions:**
  - Traditional systems relying entirely on human support teams [47].
  - Chatbots with limited capacity to handle complex queries effectively [39].
- **Strengths of Proposed Solution:**
  - Integrates advanced AI with real-time information retrieval to ensure accuracy [38].
  - Offers 24/7 customer support with tailored responses for individual needs [37].
- **Weaknesses of Current Market:**
  - Difficulty understanding dialects or ambiguous language in current systems [39].
  - Ineffectiveness in handling complex cases requiring human intervention [47].

5. Feasibility Analysis

- **Technical Feasibility:**
  - Utilizing RAG technology, which relies on retrieving trusted data sources, makes the project technically viable [38].
- **Economic Feasibility:**
  - Reduces operational costs by automating processes and lowering dependency on human labor [37].
- **Time Feasibility:**
  - The project can be executed within a defined timeline due to the availability and maturity of the required technologies [48].

3.4 Project Time Scheduling

Duration	Activities	Status
septemper-october	searching for a project idea	Done
October 20th- October 28th	background, main objectives	Done
October 20th- October 28th	project applications	Done
November 1st- November 8th	problem statement and definition	Done
November 9th- November 12th	problem analysis	Done
November 23rd -November 27th	project visibility study	Done
December 1st - December 6th	project background and survey	Done
December 9th - December 14th	project analysis	Done
December 16th - December 23rd	project design	Done
December 25th - December 27th	abstract and conclusion	Done
December 28th	preparing for the Graduation project report	Done
_____	project design	_____
_____	project implementation	_____
_____	project testing	_____
_____		_____
_____		_____

3.5 Functional Requirements

1. Information Retrieval

- Retrieves data from various sources, including databases, PDFs, and APIs. Supports advanced retrieval techniques like vector-based search and keyword matching [29].

2. Dynamic Knowledge Retrieval

- The system must fetch relevant external data (e.g., documents, articles) in real-time to enhance the generative model‘s output [3][30].

3. Hybrid Search Integration

- Combine keyword and semantic search techniques for accurate and context-aware information retrieval [29][31].

4. Contextual Data Fusion

- Efficiently merge retrieved content with the input query to generate coherent and contextually relevant responses [3][32].

5. Text Generation

- Generates coherent and contextually relevant text based on retrieved information. Supports formats such as summaries, direct answers, or explanations [33].

6. Support for Modular RAG Frameworks

- Enable naive, advanced, and modular configurations for flexibility in RAG implementations [29][33].

7. Multimodal Capabilities

- integrate text, images, and other data types into retrieval and augmentation processes[30][31].

8. Explainability Features

- Provide clear references for retrieved content used in generated responses to enhance transparency and trust [32][34].

9. Continuous Learning from Feedback

- Incorporate user feedback to refine retrieval models and improve generation accuracy dynamically [30][33].

10. Fine-Tuning Options

- Support domain-specific fine-tuning of both retrieval and generation models for specialized use cases [34][35].

11. User Interaction

- Offers an intuitive interface for users to input complex queries and receive structured outputs. Allows interactive feedback to refine results, ensuring relevance and accuracy [29].

12. Technical Support

- Includes a feature to report errors or inaccurate results, enabling continuous system improvement [29].

3.5 Non-Functional Requirements

1. **Performance**

- Responds to simple queries within 1.5-2 seconds and handles large datasets efficiently [29][30].

2. **Security**

- Implements encryption protocols (e.g.,HTTPS) to secure sensitive data and interactions. Includes safeguards against common attacks like SQL Injection and Cross-Site Scripting (XSS) [3][30].

3. **Data Privacy**

- Ensures compliance with GDPR and other privacy regulations, including encrypted storage and secure API calls [3][33].

4. **Accuracy**

- Enhances response accuracy using human-in-the-loop feedback. Maintains a minimum accuracy of 97% in retrieval and generation with less than 2% hallucination. Ensures answers are grounded in clear, verifiable sources [3][29][34].

5. **Availability**

- Ensures 99.9% uptime, with recovery time for failures not exceeding 5 minutes [3].

6. **Low Latency**

- Response generation time must not exceed 1.5 seconds for typical queries under normal server loads [29][30].

7. **Scalability**

- Support scaling to millions of queries daily across multiple regions while maintaining consistent performance [31][32].

8. **High Accuracy Rate**

- Maintain a minimum accuracy of 97% in retrieval and generation, with less than 2% hallucination [3][34].

9. **Resource Efficiency**

- Optimize memory and CPU/GPU usage through techniques like vector compression and on-demand indexing [32][35].

10. **Cross-Platform Deployment**

- Support cloud-based (AWS, Azure) and on-premises deployment options [33][31].

11. **Usability**

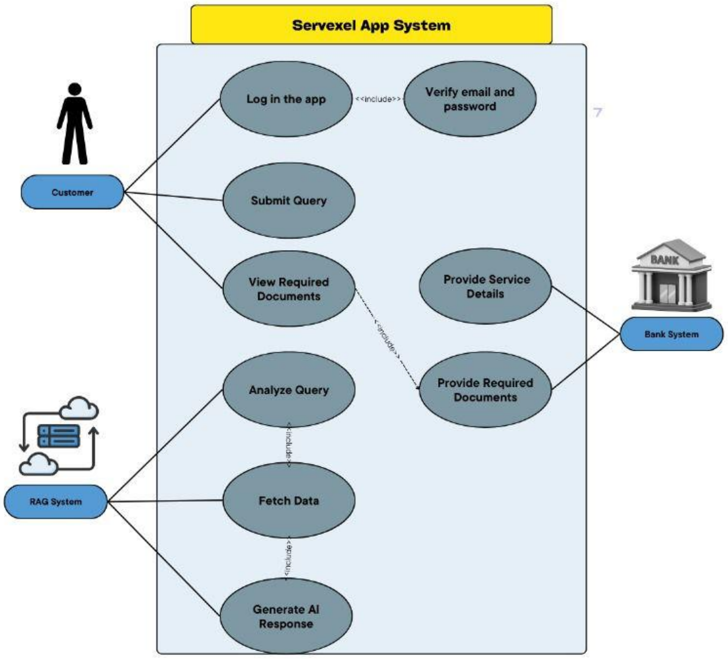
- Provide a clean and intuitive interface for developers to configure and query the system [30][31].

12. **Traceability**

- Ensure all generated outputs are traceable back to their source documents for validation [29][34].

3.7 Use case

Figure 1



3.8 Use case scenario

Retrieve Information

Actor	Customer (Primary Actor)	RAG System (Supporting Actor)
Scenario	**1**-Customer logs into the system. **2**-Customer selects the option to retrieve specific information. **3**-Customer views the information and confirms its accuracy.	**3**-RAG System validates the customer's request. **4**-RAG System retrieves the relevant data from the Database. **5-** RAG System displays the retrieved information to the Customer.
Alternative	If the customer is not registered or lacks proper authorization. If the request format is incorrect or missing required fields.	The system notifies the Customer of the error and prompts for correction**.** The system denies access and prompts the customer to register or contact support."
Pre-condition	The customer must already be registered in the system and have valid login credentials	The database must contain the relevant information.
Post-condition	If successful, the requested information is displayed to the Customer	If unsuccessful, appropriate error messages or guidance are provided

Display Results

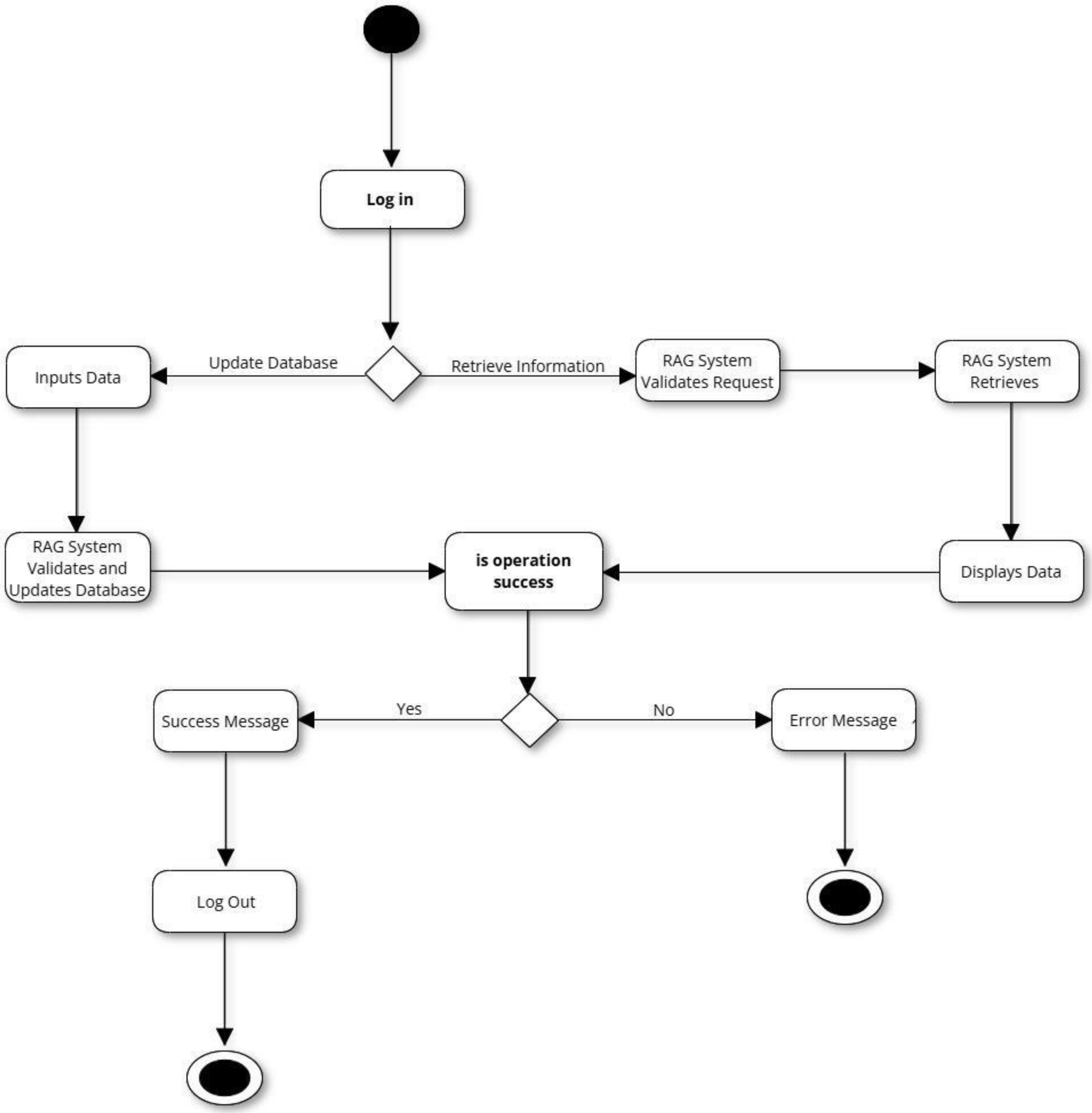
Actor	Customer (Primary Actor)	RAG System (Supporting Actor)
Scenario	**1-**Customer requests to display the search results. **3-** Customer reviews the displayed results.	**2-** RAG System processes the request and fetches the relevant results. **4-** RAG System ensures that the data displayed matches the query parameters.
Alternative	If nothing is found, the customer is informed by the system.	The RAG System recommends probable corrigendum or alternative keywords**.**
Pre-condition	The customer must have submitted a valid query.	The RAG System should be connected to the database and running.
Post-condition	The results are displayed to the customer.	In case no results are found, an alert or guidance is sent to the customer.

Update Database

Actor	**Customer (Primary Actor)**	**RAG System** (Supporting Actor)
Scenario	**1-**Customer provides new or updated information. **2-**Customer submits the request to update the database.	**3-** RAG System validates the provided information**.** **4-** RAG System updates the database with the new information**.** **5-**RAG System confirms the update and notifies the Customer.
Alternative	If the provided information is incomplete or incorrect.	The system notifies the Customer of errors and prompts for corrections.
Pre-condition	The Customer has the proper authorization and valid information to update the database.	The database is accessible for updates.
Post-condition	The database is successfully updated with the new information.	The system ensures the updated data is stored securely and accurately.

3.9 Activity Diagram

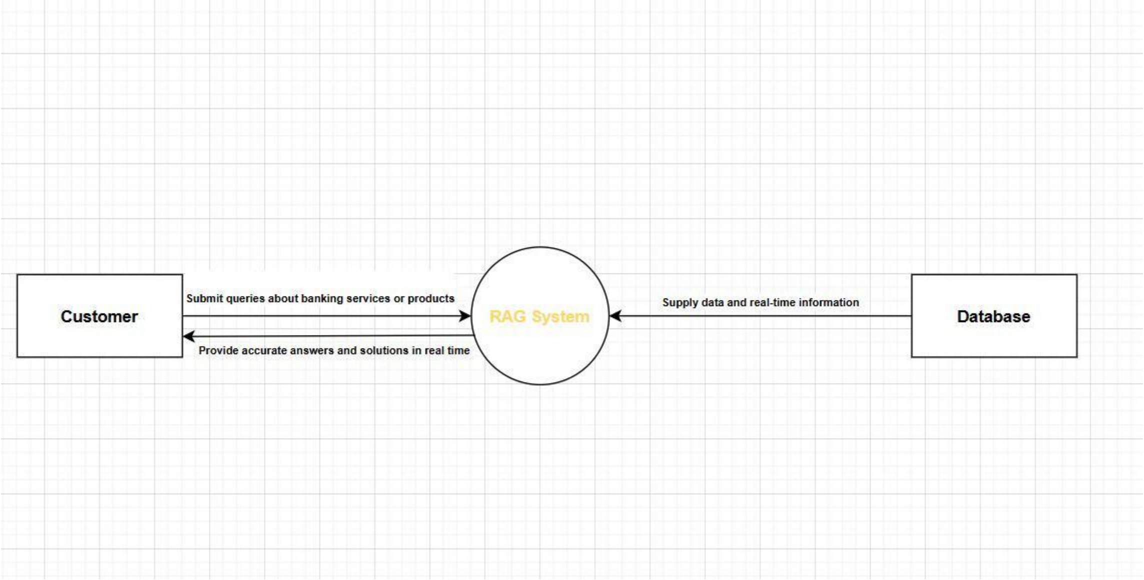
Figure 2



3.10 Data Flow Diagram

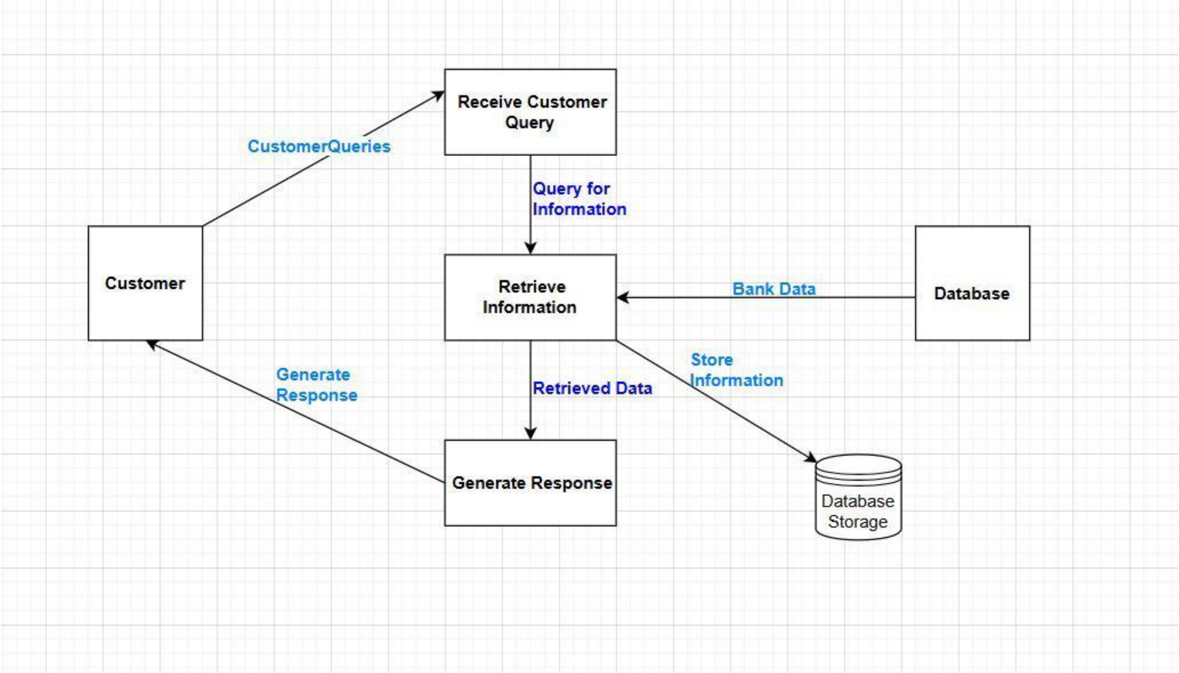
3.10.1 Context diagram

Figure 3-1


Customer
Submit queries about banking services or products
Provide accurate answers and solutions in real time
RAG System
Supply data and real-time information
Database

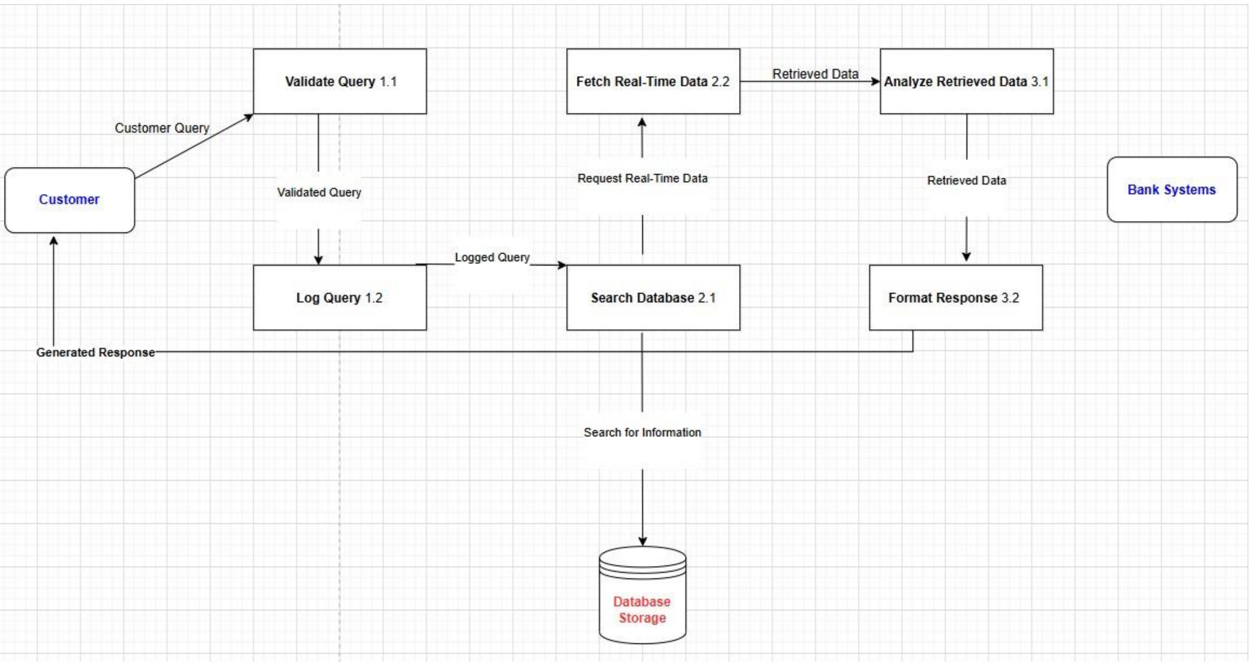
3.10.2 Level 0 diagram

Figure 3-2



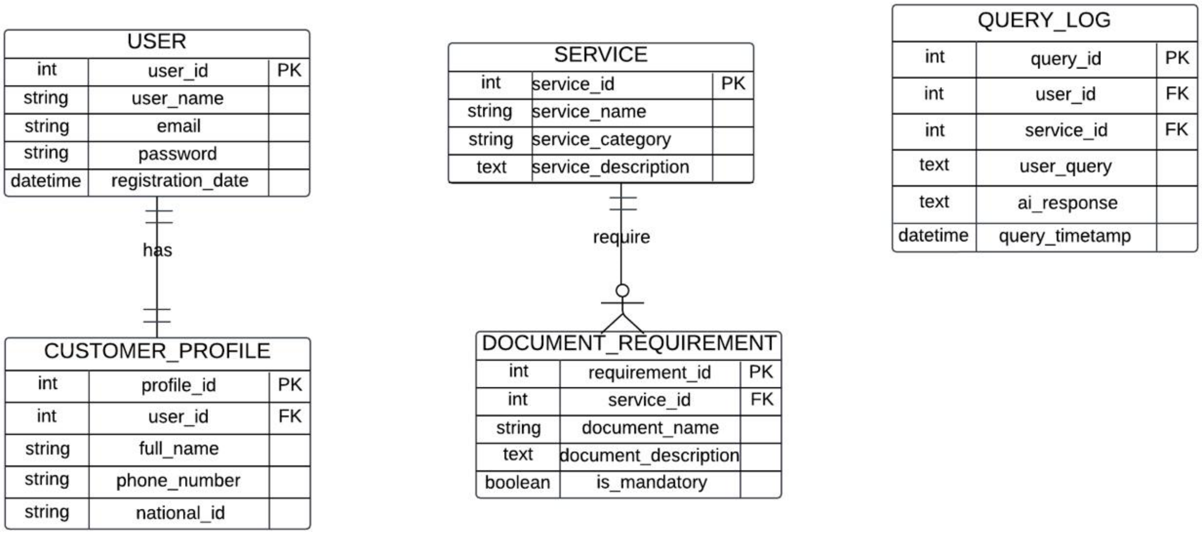
3.10.3 Level 1 Diagram

Figure 3-3



3.11 Entity Relationship Diagram

Figure 4



USER		
int	user_id	PK
string	user_name	
string	email	
string	password	
datetime	registration_date	

CUSTOMER_PROFILE		
int	profile_id	PK
int	user_id	FK
string	full_name	
string	phone_number	
string	national_id	

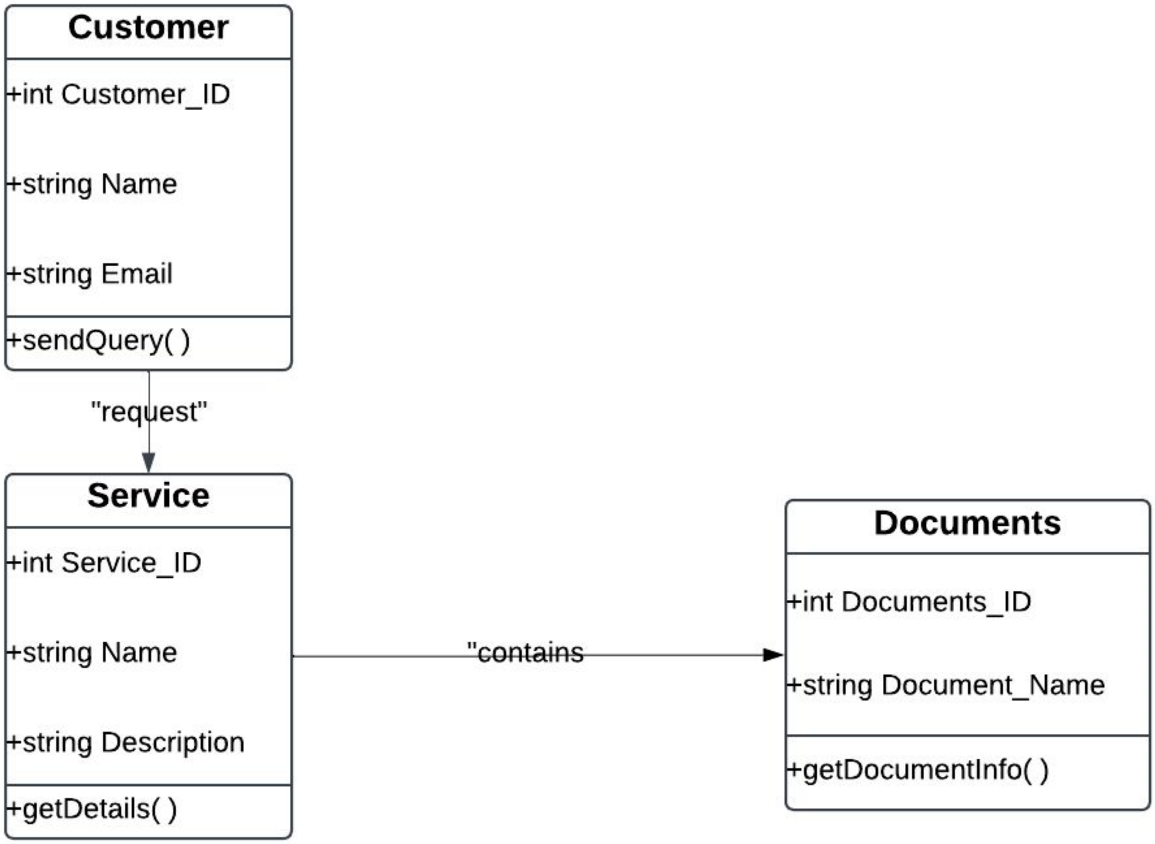
SERVICE		
int	service_id	PK
string	service_name	
string	service_category	
text	service_description	

DOCUMENT_REQUIREMENT		
int	requirement_id	PK
int	service_id	FK
string	document_name	
text	document_description	
boolean	is_mandatory	

QUERY_LOG		
int	query_id	PK
int	user_id	FK
int	service_id	FK
text	user_query	
text	ai_response	
datetime	query_timetamp	

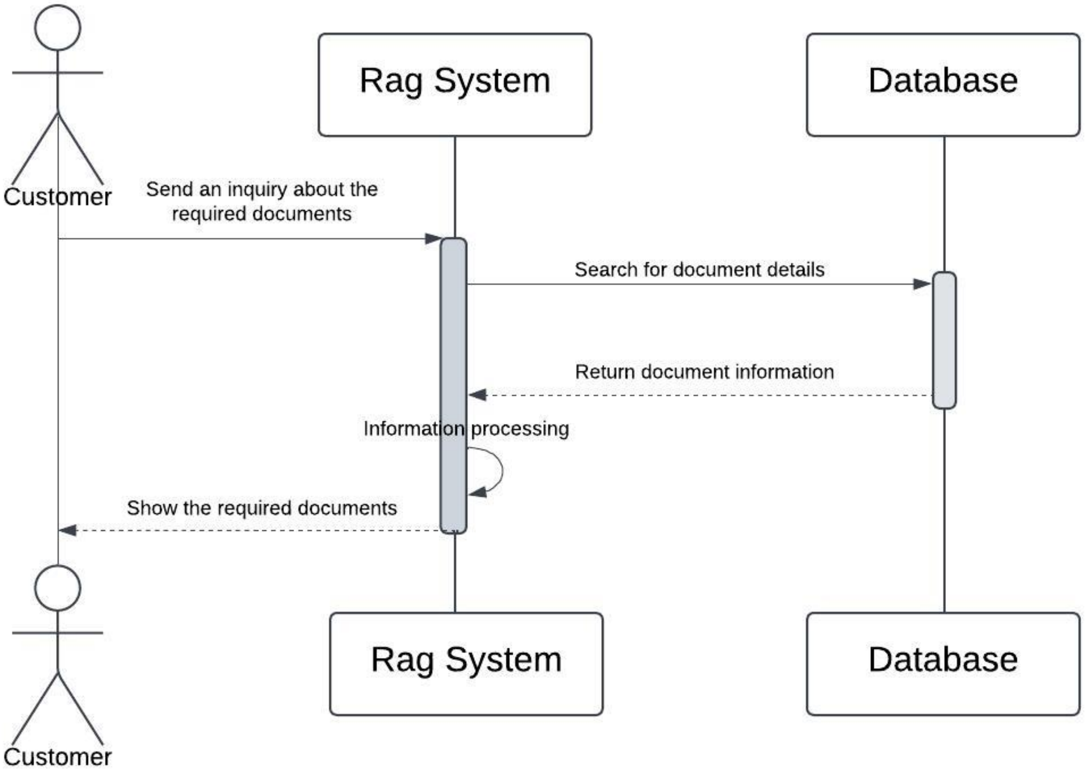
3.12 Class diagram

Figure 5



3.13 Sequence diagram

Figure 6

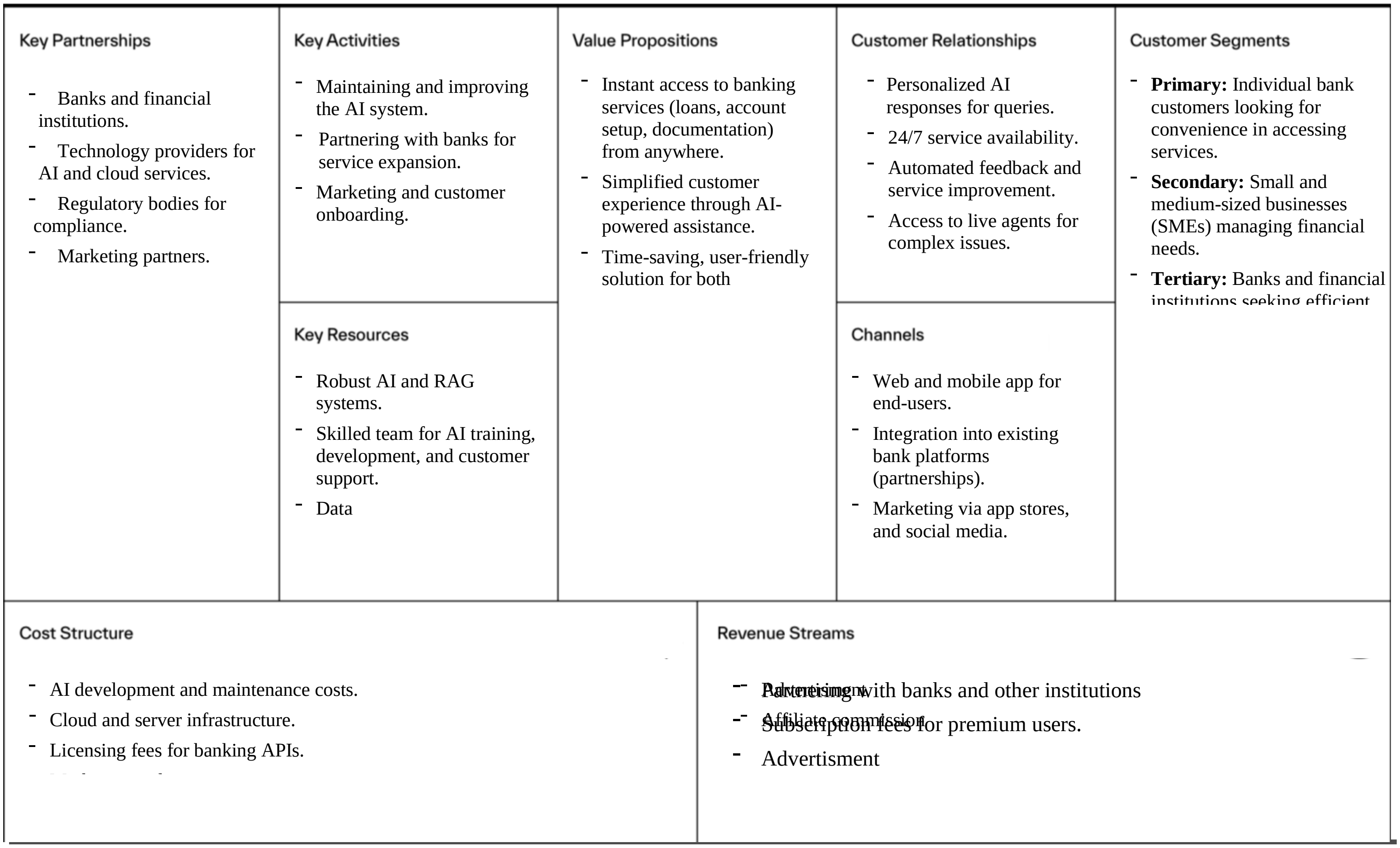

Customer
Rag System
Database
Send an inquiry about the required documents
Search for document details
Return document information
Information processing
Show the required documents
Customer
Rag System
Database

Background and Survey of Related Work

1. **Natural Language Processing in Banking**
   Natural language processing (NLP) has revolutionized customer service in banking by enabling automated chatbots and virtual assistants. Examples include Erica by Bank of America and Eva by HDFC Bank, which provide users with instant assistance for common queries [17]. However, these systems often struggle with complex requests requiring dynamic information retrieval, which is where RAG technology can offer a significant advantage.
2. **Retrieval-Augmented Generation (RAG)**
   RAG systems have emerged as a promising solution for enhancing conversational AI. By combining language generation with information retrieval, RAG models can provide factually accurate and context-aware responses [3]. Microsoft and Google have successfully employed RAG in applications such as document summarization, legal support, and technical customer service.
3. **Transformers and Knowledge Retrieval**
   The advent of transformers, like BERT and GPT, has significantly improved natural language understanding and generation [18]. Combining these models with retrieval mechanisms enables real-time adaptation to specific queries, making them ideal for dynamic domains like banking [19].
4. **AI-Powered Banking Solutions**
   AI-driven platforms are transforming banking operations, enabling faster loan processing, fraud detection, and customer support. Examples include Zest AI's underwriting platform and Kasisto's conversational AI solutions. These systems demonstrate the potential for AI to improve efficiency but highlight the need for adaptable and knowledge-driven tools like RAG to address diverse customer needs [20].
5. **Open-Domain Question Answering (ODQA)**
   Open-domain question answering systems like Dense Passage Retrieval (DPR) and DrQA serve as precursors to RAG systems [21]. These frameworks leverage knowledge retrieval from large corpora to generate precise answers, which is crucial for banking scenarios requiring compliance with ever-changing regulations [22].
6. **Customer Expectations in Digital Banking**
   A 2022 survey by Deloitte revealed that 78% of banking customers prefer self-service options for routine tasks but demand human-like interactions for complex queries [23]. The study underscores the necessity of AI systems that combine automation with intelligence and contextual understanding [24].
7. **Limitations of Current Chatbot Systems**
   Research has highlighted the limitations of rule-based and end-to-end NLP systems in handling ambiguous or multi-turn conversations [25]. For example, chatbots often fail to provide accurate answers when faced with evolving customer requirements [26]. RAG, by integrating real-time retrieval, offers a solution to these challenges.
8. **Real-World Applications of RAG in Banking**
   RAG frameworks have been explored in applications such as automated customer support in

healthcare and law [27]. The applicability of these systems to banking, which shares the need for accurate, context-sensitive responses, is increasingly evident.

Business Model



Key Partnerships

- Banks and financial institutions.
- Technology providers for AI and cloud services.
- Regulatory bodies for compliance.
- Marketing partners.

Key Activities

- Maintaining and improving the AI system.
- Partnering with banks for service expansion.
- Marketing and customer onboarding.

Key Resources

- Robust AI and RAG systems.
- Skilled team for AI training, development, and customer support.
- Data

Value Propositions

- Instant access to banking services (loans, account setup, documentation) from anywhere.
- Simplified customer experience through AI-powered assistance.
- Time-saving, user-friendly solution for both

Customer Relationships

- Personalized AI responses for queries.
- 24/7 service availability.
- Automated feedback and service improvement.
- Access to live agents for complex issues.

Channels

- Web and mobile app for end-users.
- Integration into existing bank platforms (partnerships).
- Marketing via app stores, and social media.

Customer Segments

- **Primary:** Individual bank customers looking for convenience in accessing services.
- **Secondary:** Small and medium-sized businesses (SMEs) managing financial needs.
- **Tertiary:** Banks and financial institutions seeking efficient

Cost Structure

- AI development and maintenance costs.
- Cloud and server infrastructure.
- Licensing fees for banking APIs.

Revenue Streams

- Partnering with banks and other institutions
- Subscription fees for premium users.
- Advertisment
- Advertisement
- Affiliate commission

References

1. Brown, T., et al. (2020). "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems, 33, 1877-1901.
2. Rajpurkar, P., et al. (2018). "Know What You Don't Know: Unanswerable Questions for SQuAD." arXiv preprint arXiv:1806.03822.
3. Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Advances in Neural Information Processing Systems, 33, 9459-9474.
4. McKinsey & Company. (2021). "The Future of AI in Banking: Improving Customer Experience While Managing Costs."
5. Accenture. (2022). "AI in Banking: Enhancing Customer Experience and Compliance."
6. OpenAI. (2020). "Introducing GPT-3: Language Models are Few-Shot Learners."
7. Accenture Banking Technology Vision 2022. "From AI to RAG: Transforming Customer Service in Banking."
8. Huang, J., et al. (2020). "Enhancing Chatbot Responses with Reinforcement Learning." Journal of Artificial Intelligence Research, 67, 235-250.
9. Kwon, Y., et al. (2021). "Application of Retrieval-Augmented Generation in Healthcare Customer Service." Healthcare Informatics Research, 27(4), 295-304.
10. Zhu, X., et al. (2019). "Hybrid AI Models for Enhanced Customer Service in Telecommunications." International Journal of Intelligent Systems, 34(7), 1234-1251.
11. Lin, C., et al. (2018). "AI-Powered Fraud Detection in Banking: Challenges and Solutions." Journal of Financial Crime, 25(3), 654-670.
12. Sukhbaatar, S., et al. (2015). "End-to-End Memory Networks." Advances in Neural Information Processing Systems, 28, 2440-2448.
13. Wu, P., et al. (2021). "Using Retrieval-Augmented Generation for Legal Document Review." Journal of Information Technology & Politics, 18(1), 22-34.
14. Chen, T., et al. (2019). "Knowledge-Enhanced Neural Networks for Customer Support." IEEE Transactions on Knowledge and Data Engineering, 31(7), 1306-1319.
15. Zhang, X., et al. (2020). "Sentiment Analysis Techniques for Customer Feedback in Banking." Expert Systems with Applications, 156, 113456.
16. Sun, Z., et al. (2020). "Knowledge Graphs in Customer Service: Enhancing Information Retrieval and Interaction." Journal of Knowledge Management, 24(6), 1289-1302.
17. Zamani, H., et al. (2020). "Neural Retrieval in Question Answering." *Journal of Artificial Intelligence Research*. https://jair.org/index.php/jair
18. Devlin, J., et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL-HLT Proceedings*. https://arxiv.org/abs/1810.04805
19. Radford, A., et al. (2019). "Language Models Are Few-Shot Learners." *OpenAI Research Paper*. https://arxiv.org/abs/2005.14165
20. McKinsey & Company (2021). "The AI Imperative in Financial Services." https://www.mckinsey.com
21. Karpukhin, V., et al. (2020). "Dense Passage Retrieval for Open-Domain Question Answering." *EMNLP Findings*. https://arxiv.org/abs/2004.04906
22. Chen, D., et al. (2017). "Reading Wikipedia to Answer Open-Domain Questions." *ACL Proceedings*. https://arxiv.org/abs/1704.00051
23. Deloitte Insights. (2022). "Digital Banking Maturity." https://www2.deloitte.com
24. PwC (2023). "The Future of Banking Customer Experience." https://www.pwc.com

25. Jurafsky, D., & Martin, J.H. (2021). *Speech and Language Processing*. (3rd ed.).
26. Shah, P., et al. (2018). "Building User-Centric Task-Oriented Dialog Systems." *NAACL HLT Proceedings*. https://arxiv.org/abs/1801.03277
27. Yang, L., et al. (2021). "RAG for Contextual Legal Question Answering." *Proceedings of AAAI*. https://ojs.aaai.org
28. EY Global (2022). "How Artificial Intelligence is Driving Banking Transformation."https://www.ey.com
29. AI Explorer Series. (2023). Evolving RAG Systems for LLMs: A Guide to Naive, Advanced, and Modular RAG.\
30. Rothman, D. (2024). RAG-Driven Generative AI: Build Custom Retrieval-Augmented Generation.
31. LangChain Team. (2023). RAG with Langchain: Building Powerful LLMs with RAG & Langchain.
32. Practical AI Solutions. (2023). Hybrid Search With RAG: Hands-on Guide to Building Real-Life Production-Grade Applications with RAG.
33. Data Science Experts. (2023). Unlocking Data with Generative AI and RAG: Enhance Generative AI Systems by Integrating Internal Data.
34. Mallahyari, H. (2024). A Practical Approach to Retrieval-Augmented Generation Systems. Available: https://mallahyari.github.io/rag-ebook/
35. Research Paper. (2023). "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape, and Future Directions."
36. J. W. Manyika, Five Ways to Drive Experience-Led Growth in Banking, McKinsey & Company, 2022.
37. D. Smith, How Retrieval Augmented Generation is Redefining Business Operations and Efficiency, SunTec Group, 2023.
38. IBM Research, What is RAG (Retrieval Augmented Generation)?, IBM, 2023.
39. A. Kumar, RAG in Customer Service Chatbots: Generative AI, Kommunicate, 2023.
40. P. Gupta, What Is Digital Transformation in the Banking Industry?, Zebra Technologies, New York, 2021.
41. T. Hansen, 7 Digital Banking Customer Experience Trends in 2024, EPAM Startups & SMBs, 2024.
42. R. Lewis, The 4 Phases of Digital Transformation in Banking, LivePerson, 2023.
43. S. White, Focus on Customer Experience for Digital Transformation Success in Banking, Finextra, 2023.
44. L. Anderson, Digital Transformation in Banking and Financial Services, Mad Devs, 2023.
45. J. Chen, Top Customer Experience Trends in Banking Industry Transforming 2024, Quixy, Cambridge University Press, 2024.
46. P. Roberts, Digital Transformation in Banking: Enhancing CX and Scaling DeFi, FinTech Magazine, Oxford, 2024.
47. A. Kumar, "RAG in Customer Service Chatbots - Generative AI," IEEE Transactions on AI Systems, 39(2), March 2023, 456-462.
48. R. Taylor, RAG in Financial Services: Use-Cases, Impact, & Solutions, HatchWorks, Academic Press, 2023.
49. Finn, T., & Downie, A. (2024, August 15). Digital transformation banking. IBM. https://www.ibm.com/think/topics/digital-transformation-banking