Enhancing Bank Website Accessibility and User Experience through an AI-Driven Information Assistant

Artur Avagyan
BS in Data Science American University of Armenia Yerevan, Armenia
artur_avagyan2.@edu.aua.am

Davit Davtyan
BS in Data Science American University of Armenia Yerevan, Armenia
davit_davtyan2.@edu.aua.am

Supervisor: Aram Butavyan
American University of Armenia Yerevan, Armenia

***Abstract*—The Bank Information Retrieval Assistant (BIRA) is an innovative chatbot developed on the Retrieval Augmented Generation (RAG) framework, leveraging LangChain and Streamlit technologies to offer real-time, intelligent customer assistance. Utilizing the OpenAI API, BIRA uses data (English & Armenian) scraped from the ConverseBank and InecoBank websites to provide precise, on-demand responses to customer inquiries. Designed to operate 24/7, BIRA significantly enhances customer experience by delivering immediate and accurate responses concerning bank services and products, thus eliminating the need for customers to navigate through complex website structures. By automating interactions with a chatbot that understands and responds to user needs effectively, BIRA offers personalized, swift, and effortless assistance. Initially launched in English, the successful implementation led to the subsequent integration of Armenian language support, replicating all developmental stages to fit multilingual interactions and broaden user accessibility. This paper outlines the main objectives of BIRA, addressing common issues such as inaccurate search functionalities and poor user guidance on bank websites, and reflecting on the system's ability to substantially reduce customer service overheads. Looking forward, continuous enhancements are planned through updates to ChromaDB and the potential integration of a user feedback mechanism to dynamically improve response accuracy. These advancements will ensure BIRA remains at the cutting edge of technology, continually evolving to meet user expectations and expanding its impact within the banking industry.**

I. Introduction

The Bank Information Retrieval Assistant (BIRA) is an advanced chatbot built on the Retrieval Augmented Generation (RAG) framework and developed using LangChain and Streamlit technologies. It aims to enhance customer satisfaction by providing immediate responses to the questions given by the customers. It mainly utilizes the API of OpenAI and uses the scraped data from the ConverseBank and InecoBank websites. Designed to operate 24/7, BIRA offers accurate responses to inquiries about the bank's services, products, and eliminating the need to navigate through complex website structures. Machine learning (ML) and Artificial Intelligence (AI) play a crucial role in enhancing customer service within the banking sector. By integrating technologies like this chatbot, banks can offer more personalized, quick, and effortless assistance to users' inquiries. This helps users with the irritating process of endless browsing through the pages of the websites. The opportunity to integrate a chatbot with their website helps to solve this problem. Furthermore, it can analyze the behavior of the user and preferences to give more enhanced and specific answers by encouraging further interaction with asking follow-up questions if needed to clarify the user's needs or to deepen the dialogue.

A. Main Objectives of the BIRA

The development of BIRA has been designed and motivated by the need to resolve the following issues that revolutionize customer interaction and service in banking.

- Inaccurate Search Functionality: Many users experience difficulties with the ineffective search features of bank websites, which fail to return relevant or accurate results. The search often returns results that contain the keyword but do not address the user's actual intent or query. This can frustrate users who need specific information but find themselves shifting through pages of irrelevant content. This not only leads to customer dissatisfaction but also limits the ability to quickly access necessary information.
- Lack of Contextual Understanding: Traditional search engines on these websites fail to understand the context behind a user's query. For example, a query for "loan application process" might return all documents containing the word "loan," including those related to loan payment

or loan rates, rather than a step-by-step guide to applying for a loan.
- Poor User Guidance: Even when the correct information is found, it is often presented in a format that lacks clear guidance on what steps the user should take next. This can leave users confused about how to proceed, especially if the information is embedded in a lengthy document or dispersed across multiple pages.
- Insufficient Customer Support: Customer support channels in banks, such as call centers or email helpdesks, face significant challenges, particularly during high-demand periods. During peak times, such as financial year-ends or public holidays, the volume of inquiries can overwhelm traditional support channels. Banks often struggle to scale their operations quickly enough to handle these spikes, leading to long waiting times. Thus, the quality of assistance provided can vary significantly depending on the agent's expertise and availability. This inconsistency can lead to unresolved issues or incorrect information being provided to customers, affecting their trust and satisfaction.

B. Potential Benefits of the BIRA

The Bank Information Retrieval Assistant (BIRA) offers significant benefits that can transform the landscape of customer service within the banking sector. Each of these advantages not only improves the efficiency of service delivery but also enhances the user experience in a meaningful way. One of the primary benefits of BIRA is its ability to reduce the workload on customer support representatives. By automating responses to users' inquiries that are typically found on the bank's website, BIRA allows human agents to focus their expertise on more complex, nuanced customer needs. This shift can lead to a higher quality of service for complicated issues, as representatives can spend more time addressing individual concerns without being overwhelmed by high volumes of routine inquiries. Consequently, this leads to increased job satisfaction among staff and reduces burnout, which contributes to better overall customer service. BIRA significantly enhances accessibility to banking services. For individuals who have limited physical access to bank branches — whether due to geographic constraints, physical disabilities, or a preference for digital interactions — BIRA provides a key service. By ensuring that these users can obtain immediate assistance and get information they need, BIRA not only extends the bank's reach but also improves engagement. Adaptability is another critical advantage of BIRA. Unlike static FAQ systems, BIRA is designed to learn from each interaction. This learning capability allows the assistant to adjust its responses based on individual user behaviors and preferences over time. Such personalized interactions can significantly improve user satisfaction as the system becomes better at predicting and meeting the specific needs of each user. In conclusion, the potential benefits of BIRA range from operational efficiencies to enhanced user satisfaction and accessibility.

II. LITERATURE REVIEW

A. GPT (Generative Pre-trained Transformer)

Yenduri et al. (2023) provide an in-depth analysis of Generative Pre-trained Transformers (GPT), discuss the evolution of GPT models from their inception, focusing on their significant impact on natural language processing tasks and their ability to confer effectively, which has made them popular in both academic and industrial contexts. There are various applications and the emerging challenges that GPT models face, such as issues related to training data size, model bias, and the need for computational efficiency. Furthermore, Yenduri et al. (2023) delve into potential solutions to these challenges, proposing future research directions to enhance the functionality and accessibility of GPT models across different fields. The detailed exploration of GPT technologies and their implications serves as a valuable resource for understanding the current landscape and potential future developments in the area of advanced language models.

B. Large Language Models (LLM)

The GPT (Generative Pre-trained Transformer) series has been at the front line of Large Language Models (LLM) development. Beginning with GPT, which demonstrated the power of transformers trained on diverse internet text, to GPT-4, these models have progressively increased in size and capability. Each release has brought improvements in understanding and generating natural language, making these models highly effective for a wide range of tasks. Authors Yenduri et al. (2023) delve into the evolution of GPT models. Key to the discussion is the identification of enabling technologies such as deep learning, cloud computing, and edge computing that have facilitated the rise of GPT models. The authors critically assess the challenges facing current GPT models, such as data bias and computational demands, and propose potential solutions that may guide future developments in the technology. Large language models (LLMs) like OpenAI's GPT series have revolutionized natural language processing by enabling a wide range of applications from conversational AI to content generation. However, they are not without their challenges, and we identified two common issues. One notable issue is the lack of source citation in their responses, which can make it difficult for users to verify the accuracy and reliability of the information provided. This is particularly problematic in academic and professional settings where source credibility is crucial. Additionally, the data used to train these models can become outdated, as the models do not automatically update with new information. Even when their pre-training data does not address the topic at hand, generative models always respond with confidence. Thus, users may be misled by the models' occasional creation of believable text. This results in responses that may be based on outdated facts or norms, which limits their effectiveness in the dynamically evolving world.

C. Retrieval Augmented Generation (RAG) Overview

In the evolving field of Retrieval-Augmented Generation (RAG) systems, the study by Gao et al. (2023) provides

a thorough review of various RAG models, focusing on their developmental paradigms—Naive RAG, Advanced RAG, and Modular RAG. These models incorporate state-of-the-art technologies to enhance the capabilities of language models by integrating external data sources for improved information accuracy and relevance (Gao et al., 2023). Naive RAG is highlighted as the foundational RAG model. It employs basic retrieval and generation processes to augment the scope of language models. However, it faces limitations such as misaligned retrieval accuracy and the generation of contextually irrelevant content. These challenges lead to the exploration of more advanced RAG models to better meet complex needs (Gao et al., 2023). Advanced RAG builds on the Naive model by incorporating sophisticated retrieval mechanisms that optimize the accuracy and relevancy of the information. This model enhances both the indexing and retrieval phases, leading to more accurate generation outcomes (Gao et al., 2023). Modular RAG represents the most advanced paradigm, offering a flexible architecture that allows for the modification and reconfiguration of components to meet specific user requirements and adapt to different data types. This adaptability is a significant advancement in the development of RAG systems, which makes it suitable for a broader range of applications (Gao et al., 2023). For BIRA, which operates on scraped data from bank websites, encodes them into vector representations using an embedding model and stores in vector database, and retrieves answers based on user queries, the Advanced RAG model appears to be the most aligned with its operational framework. This model's focus on refining both the retrieval of relevant information and its subsequent generation makes it particularly suitable for BIRA. Advanced RAG's capabilities to optimize information accuracy through indexing and retrieval techniques ensure that the responses generated are not only precise but also highly relevant to the user's specific inquiries. This alignment supports the objective of providing efficient and reliable customer service through an AI-driven chatbot in the banking sector (Gao et al., 2023).

D. AI Applications for Banking

In the study conducted by Petersson et al. (2023), the authors explore the dimensions of customer experiences with real-life text-based banking chatbots through a qualitative approach. Their research identifies critical factors influencing customer satisfaction and highlights the difference between user expectations and chatbot performance. The study pays importance to human-like interactions, including personality traits and even the use of emojis, which notably improve the customer experience with banking chatbots. These human-like features significantly mitigate the negative impacts of miscommunication errors, particularly in simpler tasks where users expect swift and accurate assistance (Petersson et al., 2023). Authors suggest that informing customers about what chatbots can and cannot do increases user satisfaction by aligning expectations with the chatbot's capabilities. This approach is important for complex tasks, where the study finds that a well-informed user is more receptive to the assistance provided by the chatbot, despite potential errors in understanding complex queries. This customer-centric approach in the development and implementation of chatbots, shows us that the success of these AI-driven tools heavily relies on their ability to adapt to and address user needs effectively. It gives solid evidence about what features improve user experience and gives valuable insights into our methodology and approach, which connects the gap between technological advancements and user satisfaction.

E. Website Accessibility and User Interaction

In their study, Shukla et al. (2020) delve into the transformative impacts of Artificial Machine Intelligence (AMI) on the banking sector, specifically through the deployment of chatbots. Their research highlights the significant shift from conventional banking interactions to automated digital conversations, facilitated by advancements in artificial intelligence and machine learning technologies. The paper discusses the evolution of chatbots from simple rule-based systems to sophisticated AI-driven agents capable of complex and nuanced interactions. Shukla et al. (2020) address the critical challenges and limitations faced by current chatbot technologies, such as the need for more advanced natural language processing abilities to fully understand and respond to user requests in a contextually relevant manner. So, it offers valuable insights into both the potential benefits and the areas requiring further development.

F. Gap Analysis in Current Technology

Given that our proposed Bank Information Retrieval Assistant (BIRA) currently operates without requiring personal customer data, the study by Lappeman et al. (2023) marks the significance of privacy and trust even when planning future enhancements that might involve personal data handling. The study delves into the complex relationship between trust, privacy concerns, and user willingness to disclose personal information in the context of digital banking chatbot services. The authors identify three key dimensions of trust — i) Brand: the trust associated with the bank's brand itself. ii) Cognitive: trust based on the logical understanding and beliefs about the service's reliability and competence. iii) Emotional: trust driven by feelings and the relational history with the service. These dimensions influence how much personal information customers are willing to disclose, which directly impacts the functionality and personalization capabilities of chatbots. They found that privacy concerns significantly decrease user willingness to disclose personal information, creating necessity for banks to build trust and address privacy concerns proactively. Interestingly, the study found that customers are likely to disclose less information when interacting with their preferred banking brand's chatbot as opposed to a fictitious brand. This suggests that existing customer relationships might influence expectations and interactions with automated services. Authors recommend that for chatbots to be effective in enhancing digital banking services, institutions must not only ensure robust privacy protections but also actively work to build all dimensions of trust. There has to be a balance between personalization and privacy in the context of banking chatbots. For

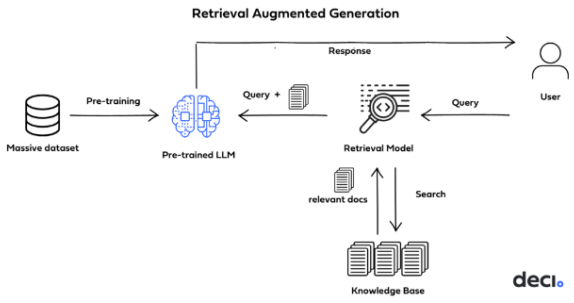

Fig. 1. The diagram describes the workflow of a Retrieval Augmented Generation (RAG) system. It starts with a pre-trained language model (LLM) that has been trained on a large dataset. When a user submits a query, the system activates a retrieval model that searches a knowledge base for relevant documents. These documents are combined with the user's query to enhance the LLM's response. Finally, the LLM generates a response that is delivered back to the user, leveraging both its pre-trained knowledge and the specific information retrieved to ensure relevance and accuracy.

BIRA, which currently operates without accessing personal customer data, the findings suggest an opportunity to build trust through reliable information delivery and potentially create a way for future capabilities that may involve personal data, ensuring that privacy and trust considerations are at the front line of development and communication strategies.

III. METHODOLOGY

To lay the groundwork for our Capstone project, it is essential to explore the underlying technologies that drive our AI-driven information assistant. Here, we delve into the backbone of our implementation.

A. Framework

As stated in the Literature Review section, Retrieval Augmented Generation (RAG) addresses some of the key limitations of LLMs by combining the generative capabilities of models like GPT with up-to-date data retrieval. So, RAG systems enhance the response quality of LLMs by dynamically retrieving relevant documents or data as context during the generation process. This approach is important when the model needs to provide factual, up-to-date information or when handling domain-specific queries that require external knowledge. Another vital component of our AI-driven assistant is the utilization of embeddings. In the context of AI models like those developed by OpenAI, embeddings play a critical role in handling and interpreting the vast amounts of text data these models engage with. Embeddings convert high-dimensional textual data into lower-dimensional vectors while preserving semantic relationships, making them essential for tasks like semantic search, clustering, and classification. OpenAI embeddings, in particular, are designed to capture deep semantic meanings. These embeddings use the advanced capabilities of GPT models, allowing for a more nuanced understanding of the language and its generation. This property is crucial for our project as it enhances the AI assistant's ability to accurately match user queries with relevant information from bank websites. The OpenAI API allows the efficient use of OpenAI embeddings to be smoothly integrated into our project's architecture. It acts as a gateway to advanced AI models like GPT-4, enabling our project to utilize cutting-edge language processing technologies. The API makes it easier to include sophisticated machine learning models into our application without requiring a large amount of computing infrastructure or specific expertise. We got concentrated on improving the user experience instead of getting distracted by the underlying technological complexity thanks to its straightforward implementation. Additionally, the economic aspect of using the OpenAI API is crucial for project scalability and sustainability. The API's pricing model, which varies depending on the model used and the volume of data processed, requires careful consideration. Our hands-on experience with various models offered through the API has underscored the importance of balancing cost with computational power. More powerful models like GPT-4, while providing enhanced capabilities, come at a higher cost, prompting us to carefully manage resources to maintain both efficiency and effectiveness in our service delivery. This strategic approach ensures that our AI-driven assistant remains both economically viable and technologically advanced. Building on the foundational use of OpenAI embeddings, it brings to the essential role of vector databases in our AI-driven assistant's architecture. Vector databases are specialized storage systems designed to efficiently store and manage vector embeddings, which are the output from models like GPT when processing text into numerical formats that capture semantic meanings. These databases are essential because they allow for the quick retrieval of vectors through similarity searches by finding the closest matches to a user's query vector in real time. By efficiently indexing and retrieving vector embeddings, these databases help in significantly reducing the response time of the AI assistant, which enhances the user experience. The ability to handle high-dimensional data with minimal latency is what makes vector databases absolutely essential in deploying advanced AI models into user-facing applications. With this context on the importance of vector databases, our choice of ChromaDB for the project was driven by specific needs and advantages that it presents, particularly in terms of scalability, speed, and compatibility with our existing technological framework. The use of ChromaDB extends into another crucial area of our project — prompt engineering. Efficient retrieval of relevant vectors from ChromaDB is harmonized by the skillful crafting of a prompt template, which is designed to carefully get the most accurate and contextually appropriate responses from the AI model. This process of prompt engineering is not just about understanding the model's language capabilities but also about crafting the template in a way to maximally generalize any query structure that can make the most of the stored embeddings. This relationship between the stored embeddings in ChromaDB and the crafted prompt template significantly boosts the efficiency and effectiveness of our AI-assistant's responses. Building on this foundation provided by our selection of ChromaDB for efficient data management and the crafting process of prompt engineering, we further enhance our project by incorporating Langchain and Streamlit. These frameworks are important for bringing all the technological

components together into a functional chatbot application that is both robust and user-friendly. The integration of LangChain in our project enables the uninterrupted connection of ChromaDB and OpenAI APIs. LangChain specializes in complex query handling, which is essential given the sophisticated nature of user interactions expected with our chatbot. With it, we ensure our application can access and utilize the data stored in ChromaDB, effectively use the vector embeddings and the optimized prompt to deliver precise responses to user queries. Streamlit, on the other hand, plays a role in the user interface design of our chatbot. Known for its effectiveness, Streamlit allows us to build and deploy interactive web applications rapidly. It provides an intuitive platform for users to interact with our AI-driven assistant, which makes the technology accessible to a broader audience without compromising on functionality.

B. Data Collection and Selection Process

The data collection process involved scraping ConverseBank and Inecobank's websites, storing each page as a separate markdown file, processing them to get embeddings and storing them in ChromaDB. Our approach to scraping the bank websites involved a process designed to handle the often messy and inconsistent structure of these web pages. Given the diversity and complexity of the website layouts, we developed a scraping strategy that was both flexible and robust, capable of extracting a wide array of information types while maintaining the integrity of the data. We implemented specific strategies for each bank. The website of ConverseBank presented unique challenges due to having two operating websites — the old and a newer version. The new website operates only on several web pages and lacks complete information, frequently redirecting users back to the old website for more detailed content. Given this situation, we opted to completely scrape the old website, as it contains the complete and necessary information essential for our data collection. We developed custom scripts to navigate its specific layout and extract data efficiently, particularly focusing on areas where dynamic content changes and updates are frequent.

In contrast, Inecobank's website had a more static nature but featured complex nested HTML structures that required precise parsing strategies. Our approach here involved deeper analysis and customization in the script to accurately extract information from these nested elements without losing context. We used Selenium for navigating and interacting with the websites dynamically to manage and respond to JavaScript-rendered content. We get access to up-to-date information that only appears as a result of user interactions (e.g., filling forms or clicking buttons). For parsing the HTML content and extracting data, BeautifulSoup was employed due to its ease of use and ability to navigate through the HTML tree structure effectively. The key challenge was the inconsistent HTML structures across both banks' web pages. To address this, we crafted custom scraping scripts to each website page but built on a common framework that allowed for reusable code and techniques. This framework was designed to be adaptable, with parameters that could be adjusted for different page layouts and content types. One specific challenge we faced was extracting data from HTML tables and converting it into a more manageable format. Our solution was to write a script to transform any html table to markdown. This approach involves identifying HTML table elements within the scraped web pages and then converting these elements into markdown format. Markdown was chosen because it offers a clear and concise way to represent tables and other formatted data in plain text, which is easier to handle and integrate into our database system. Additionally, markdown files are lightweight and versatile, suitable for version control systems and compatible with a wide range of software platforms and tools, which facilitates further processing and analysis. It also allows for easy conversion to HTML or other formats if needed, which gives great flexibility in how the data can be used and displayed. Overall, our scraping methodology is driven by the need to manage complex and variable data in a way that maximizes flexibility, efficiency, and consistency.

C. Ethical Considerations While Scraping

In the development of our Bank Information Retrieval Assistant (BIRA), ethical considerations are preserved. To ensure our practices adhere to both legal standards and ethical guidelines concerning data collection and privacy, we have implemented several key strategies. Respecting robots.txt: Our scraping tools are configured to comply with the directives specified in the 'robots.txt' file of each website. This file is intended to communicate with web crawlers and inform them of which areas of the website should not be accessed or indexed. By following these restrictions, we respect the website administrators' guidelines and avoid scraping data that they have designated as off-limits.

Server Load Management: We take great care to ensure that our scraping activities do not overload the servers of ConverseBank and Inecobank. Overloading servers can lead to slower website performance or even cause downtime, which can affect not only the bank but also its customers. To prevent this, we carefully manage the frequency and volume of our requests to ensure they are within acceptable limits. Reasonable Throttling: To further minimize the impact on the banks' web servers and mitigate any potential misinterpretation of our actions as hostile (e.g., a Distributed Denial of Service, or DDoS, attack), we implement reasonable throttling in our scraping scripts. This involves setting deliberate pauses between our requests to ensure that our data collection is more in line with typical human browsing speeds rather than automated scripts, which can aggressively pull data at high speeds.

D. Data Exploration

For ConverseBank, we scraped and stored each webpage as a separate markdown file, resulting in a total of 153 markdown files. This includes 7 PDF files that were converted into markdown to maintain consistency in data format. The markdown format was chosen for its readability and ease of integration with our database systems, allowing for straightforward data manipulation and analysis. Our data collection

from Inecobank was more focused. We specifically targeted the Individual section of the bank's website, scraping all relevant sub links within this section. This targeted approach resulted in approximately 70 markdown files, and captures a detailed snapshot of the services offered to individual customers. An important aspect of our data collection was the inclusion of both English and Armenian language data. This dual-language support enables BIRA to serve a broader user base and deliver more personalized and accessible service.

E. Topics of BIRA (ConverseBank)

BIRA is well-equipped to handle a wide range of user queries.

General Customer Support

- General Information: BIRA can provide essential details about the bank's operations and values.
- Branches: Users can get information on branch locations, operational hours, and contact details.
- Historical Background: BIRA can share insights into the bank's history and evolution over the years.
- Shareholders and Investors: Provides information on shareholder composition and investor opportunities.
- Management: Details about the management team and board members.
- Accounting & Policies: Information on the bank's accounting practices and operational policies.
- FAQ: Answers to frequently asked questions can be provided instantly.
- Ratings: BIRA can share the ratings and financial stability reports.

Individual Banking Services

- Loans: Detailed explanations of different loan types, eligibility criteria, and application processes.
- Cash & ApplePay: Guidance on cash handling services and how to set up and use ApplePay.
- Deposits: Information on various deposit schemes, interest rates, and terms.
- Cards: Details about credit and debit card options and benefits.
- Property Sale: Information on properties for sale through the bank.
- Transactions: Assistance with conducting and managing transactions.
- State Agencies: Information regarding interactions with state agencies.
- Account Opening & Maintaining: Help with starting new accounts and maintaining existing ones.
- Flexible Rate: Information on interest rates and financial products with flexible rates.
- Insurance: Details on available insurance products.
- Custodian Services: Information about custodian services offered by the bank.
- Corporate Bonds: Guidance on investing in corporate bonds issued by the bank.

Business Banking Services

- Loans: Insights into business loan products, including terms and application guidelines.
- Deposits: Details on business deposit accounts and benefits.
- Trade Finance: Support with trade finance services.
- Payroll Projects: Information on managing company payroll through the bank.
- Leasing: Options and details for leasing agreements.
- Transfers: Assistance with business transfer procedures.
- Factoring: Explaining the benefits and process of invoice factoring.
- Financial Institutions: Information related to services provided to other financial institutions.
- Account Opening & Maintaining: Help with opening and managing business accounts.
- eCommerce: Support for setting up and managing eCommerce banking needs.
- Applications: Help with filling out and submitting various banking applications.
- Bonds: Information on bond options for businesses.
- Reverse Repo Transactions: Details about reverse repo options.
- Forward Contracts: Assistance with forward contract services.

BIRA's ability to access and retrieve information from these categorized data points ensures that it can provide precise, contextually relevant responses to user queries.

F. Topics of BIRA (InecoBank)

For InecoBank, BIRA is equipped to handle topics related to individual banking needs. **Account Information**

- Account Packages Terms & Conditions: BIRA can provide details on the terms and conditions of various account packages offered by InecoBank
- Special Accounts Terms & Conditions: Specifics on special account offerings.

Savings and Investments

- Deposits: Information on fixed and variable rate deposit accounts, including terms, interest rates.
- Savings Accounts: Details on different savings accounts available, their features, and benefits.
- Investment Services: Guidance on investment products such as funds, bonds, and other securities offered by InecoBank.

Loans

- Consumer Loans: Information about eligibility, interest rates, repayment terms, and application procedures for personal loans.
- Credit Lines: Details on credit lines available for individuals, their revolving credit options.
- Mortgage Loans: Details on mortgage products, including requirements, rates, and application tips.

Cards

- Standard Cards: Features and benefits of standard debit and credit cards, application processes, and usage tips.
- Premium Cards: Features of premium card offerings, including reward programs, travel benefits, and higher credit limits.

- Special Offers Cards: Promotional cards that come with time-limited offers.

Digital Banking, Transfers and Payments, Insurance

- Digital Banking: Instructions and support for using online and mobile banking platforms.
- Other Information about Money Transferring: Details on international money transfer services, including fees, transfer limits, and required documentation.
- Insurance: Information on insurance products available through the bank, such as life insurance, health insurance, and property insurance.

IV. RESULTS

After successfully verifying that the Bank Information Retrieval Assistant (BIRA) functioned as intended on a local server, we progressed to deploying the assistant on Streamlit. This shift makes BIRA accessible to a broader audience. The initial tests of BIRA demonstrated its capability to address a wide range of banking queries. For ConverseBank, the assistant properly handled inquiries related to loans, mortgages, deposits, cards, property sales, transactions, and account maintenance. Similarly, for InecoBank, BIRA effectively provided information on accounts, savings, loans, cards, mortgages, digital banking, investments, and insurance money transfers. These results confirmed that the assistant's functionality aligns well with the intended design, and the scraped data utilized by the model effectively supports this broad query scope. To ensure the reliability of BIRA's responses, we implemented a manual testing procedure. This involved comparing the assistant's generated responses directly with the actual data from the websites of ConverseBank and InecoBank. By doing so iteratively for a large number of queries, we recorded the accuracy of the information provided by BIRA and made necessary adjustments to improve its performance. During subsequent testing phases, different prompt templates were tested to analyze and optimize performance. This process involved refining the prompt template to enhance the clarity and specificity of information retrieval, thereby maximizing the accuracy of responses provided by BIRA. Through these iterative refinements, BIRA achieved a high degree of response accuracy, and gives precise and contextually relevant answers to user queries. Despite the high accuracy in common query responses, challenges were encountered when addressing less common or more complex questions. These instances showed the potential areas for improvement in the NLP components of BIRA. To optimize BIRA's performance, several large language models were tested, starting with GPT-3.5-turbo-0125. After experimenting with various models, we finalized the use of GPT-4-turbo-2024-04-09, which demonstrated impressive performance with contextually relevant answers. Initially, the development and testing of BIRA were conducted solely using the English version of both the scraped data and the query/response generation. This focused approach allowed us to establish a foundation of the development process by dealing with a single language dataset. After successfully finalizing the English version and confirming its functionality and accuracy, we then expanded the system's capabilities by integrating Armenian language support. This extension involved passing through all the stages of development that were initially conducted for the English version, including scraping Armenian data, adapting the NLP models for Armenian language processing, and iterating through testing phases to ensure the system's accuracy and effectiveness in handling queries in Armenian. This progression allowed us to offer a multilingual service, making BIRA accessible to a broader audience. Here are several examples of user queries and BIRA's generated responses:

- User: Who are the founding shareholders of Inecobank?
- BIRA: The founding shareholders of Inecobank CJSC are Mr. Avetis Baloyan, who owns 39.97% of the shares, and Mr. Karen Safaryan, who owns 34.58% of the shares.
- User: What is the maximum number of visa digital cards for each customer?
- BIRA: The maximum number of Visa Digital cards that each customer can have in each currency (AMD, USD, EUR, RUB) is 2 cards.

V. DISCUSSION

Issues Encountered and Resolutions Throughout the development of our project, we faced several technical challenges, particularly related to managing large datasets with ChromaDB, and deployment issues with Streamlit. As the project progressed, the volume of data stored in ChromaDB grew significantly, which posed challenges in terms of storage capacity and efficiency. The large size of ChromaDB made it difficult to manage locally due to resource constraints. We explored several solutions. **AWS S3 Bucket:** We initially explored using AWS S3 for its high availability and data redundancy, which makes it an attractive option for large-scale data storage. However, we encountered challenges in integrating S3 directly with ChromaDB due to compatibility and latency issues in real-time data retrieval.

Azure Blob Storage: Azure Blob Storage was another option considered for its robustness and integration capabilities with various databases and applications. Similar to AWS, the integration complexity put us off this solution.

Successful Implementation with Git Large File Storage (LFS): Eventually, we resolved it using Git Large File Storage (LFS) which provided a rationalized solution for handling large files and datasets associated with our project. Git LFS integrates well with our existing development workflows, allowing us to version large files as efficiently as code. This integration made managing large datasets without compromising on operational efficiency and made collaboration and version control more manageable.

Deployment Challenges with Streamlit

Deploying the Streamlit application proved more complex than anticipated, primarily due to dependency management and compatibility issues. Streamlit, while user-friendly and efficient for creating web apps, presented challenges when integrating with the complex dependencies required by our AI models and data handling systems.

We documented and managed all dependencies in a 'requirements.txt' file, ensuring that all necessary libraries and their

correct versions were clearly defined. This approach helped to mitigate issues related to missing or conflicting dependencies.

We did an iterative approach to testing and deployment, where smaller changes were tested individually before full-scale deployment. This method helped identify specific issues related to Streamlit compatibility and allowed us to address them systematically without disrupting the entire application.

Streamlit community and support forums proved to be invaluable. Many of the issues we faced had been encountered and documented by other users, and community-provided solutions in some sense helped resolve our deployment challenges.

VI. CONCLUSION

The development of the Bank Information Retrieval Assistant (BIRA) represents a significant achievement in the integration of advanced AI technologies within the banking sector. This project successfully merged the capabilities of Retrieval-Augmented Generation (RAG) systems with the practical needs of modern banking.

BIRA has effectively demonstrated its ability to handle a wide range of banking-related queries with high accuracy and relevancy. Using data from ConverseBank and InecoBank and utilizing advanced NLP techniques, the assistant has managed to improve the accessibility of information and reduce the workload on bank staff. The project has contributed to the field by showcasing how RAG-based models can be practically implemented to improve customer service and operational efficiency in financial institutions. Additionally, the incorporation of multilingual support has broadened the usability of BIRA, for a diverse customer base.

The project has successfully met its objectives, with BIRA providing reliable, accurate, and timely responses to user inquiries. The iterative testing and development phases have ensured that the assistant is not only functional but also aligns closely with user needs and expectations. The use of Streamlit for deployment enables wider accessibility and real-time feedback from users.

Here is the deployed version: bira-ai.streamlit.app

Potential Future Developments and Improvements

Looking ahead, BIRA needs continuous development and improvement. Given its foundation in the RAG model, the system will benefit from periodic updates to ChromaDB and enhancements to its NLP capabilities. To ensure these updates are seamless and do not disrupt service, implementing a robust continuous integration and continuous delivery (CI/CD) framework using Azure DevOps is planned. This approach will facilitate smooth and efficient updates and maintenance of the system.

Moreover, to enhance the accuracy and user satisfaction further, we plan to integrate a human-based feedback mechanism. This feature will allow users to rate BIRA's responses, which can be used to fine-tune the assistant's performance. Collecting user feedback in this manner will help in making necessary updates more targeted and effective, ensuring that BIRA continues to meet the evolving needs of its users.

VII. REFERENCES

[1]Yenduri, G., Ramalingam, M., Selvi, G. C., Y, S., Srivastava, G., Maddikunta, P. K. R., ... Vasilakos, A. V. (2023). GPT (Generative Pre-trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *arXiv preprint arXiv:2305.10435*. Retrieved from https://arxiv.org/abs/2305.10435

[2]Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*. Retrieved from https://arxiv.org/abs/2312.10997

[3]Petersson, A. H., Pawar, S., & Fagerstrøm, A. (2023). Investigating the factors of customer experiences using real-life text-based banking chatbot: A qualitative study in Norway. *Procedia Computer Science, 219*, 697-704. https://doi.org/10.1016/j.procs.2023.01.341

[4]Shukla, V. K., Vyas, S., Mishra, V. P., Suhel, S. F., & Sharma, S. K. (2020). Conversation to Automation in Banking Through Chatbot Using Artificial Machine Intelligence Language. *Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization*. https://doi.org/10.1109/ICRITO48877.2020.9197825

[5] Lappeman, J., Marlie, S., Johnson, T., & Poggenpoel, S. (2023). Trust and digital privacy: Willingness to disclose personal information to banking chatbot services. *Journal of Financial Services Marketing, 28*(1), 337-357. https://doi.org/10.1057/s41264-022-00154-z