Orclever
Vol. 5 No. 1 (2025)
DOI: 10.56038/ejrnd2026089703
www.orclever.com
The European Journal of Research and Development
e-ISSN: 2822-2296

An AI-Based Question–Answering System for Corporate Documents: VK ArtiFin

Zeynep Örpek[1*], Büşra Tural[1], Zeynep Destan[1]

[1] KFT Bilişim Sistemleri A.Ş.

0000-0001-7130-9118

0000-0003-3645-8761

0009-0006-1448-4159

* Corresponding author: kft.zeynep.orpek@vakifkatilim.com.tr

DOI: 10.56038/ejrnd2026089703

Received: 2025-09-16 · Revised: 2025-11-20 · Accepted: 2025-12-25 · Published: 2025-12-30

Abstract

The banking sector, due to its intense regulatory landscape and constantly changing legislation, has become a sector where rapid access to accurate information is critical. The increasing variety of financial products, the proliferation of digital banking, and the tightening of regulatory processes are increasing the need for instant access to current legislation and internal procedures among corporate employees. However, current information access methods largely rely on manual communication channels like phone calls or text messaging, resulting in wasted time, human errors, process delays, and operational losses. Artificial intelligence (AI) technologies are playing a transformative role in information management and process optimization in the banking sector. Developed in this context, VK ArtiFin, an AI-based question-answer system, stands out as an innovative solution capable of generating fast, accurate, and context-appropriate answers to questions about internal documents such as regulations and procedures. The VK ArtiFin question -answer system aggregates different document formats into a data repository and analyzes content at the semantic level using large language models (LLMs). This model, capable of understanding users' sequential and contextual questions, facilitates direct access to information, reduces manual workload, and increases operational efficiency. This question-answer system allows users to answer questions on relevant documents or their own documents. Findings obtained from the use of the question-answer system reveal that artificial intelligence-supported information access systems accelerate digital transformation in the banking sector, improve employee experience, and increase reliability in corporate decision-making processes. This study demonstrates that AI is shaping the future of banking not only as a technical tool but also as a strategic value creator.

Keywords: Artificial Intelligence (AI), Large Language Models (LLM), Natural Language Processing (NLP), Question-Answer Systems, In-House Artificial Intelligence Applications, Financial Technologies

1. Introduction

With the advancement of technology, the banking and finance sector is one of the sectors rapidly becoming digital. Due to its large customer base, the banking and finance sector is one of the most data-intensive areas. Furthermore, this sector's regulatory structure stands out as one of the most complex. The constantly changing and evolving regulatory landscape, the increasing diversity of financial products, and the proliferation of digital banking applications have made rapid access to accurate and up-to-date information critical to corporate success.Due to the nature of banking operations, employees must access timely and accurate information regarding regulations, internal procedures, and operational processes. However, current information access methods largely rely on manual searches, email correspondence, or direct employee communication, leading to issues such as lost time, process delays, and operational errors.In recent years, AI technologies have offered revolutionary opportunities in information management and process optimization in industries with large amounts of data, such as banking. Natural language processing (NLP) techniques, in particular, are being used to understand human language, analyze text, and extract contextual meaning from it. This allows for automatic analysis and classification of texts contained in documents with complex structures and languages. In this context, LLMs are AI models trained with billions of parameters. LLMs learn not only words but also contextual relationships within sentence and paragraph structures, and have the potential to produce meaningful, accurate, and contextual answers to users' natural language questions. LLM-based systems can semantically analyze unstructured

data from various document formats. This allows LLMs to accelerate information access and minimize human errors. By using these technologies, bank employees not only access accurate information quickly but also manage informed decision-making processes more securely, efficiently, and transparently.In this context, the VK ArtiFin question-answer system, designed and developed by Vakıf Katılım and KFT Bilişim Sistemleri, is an innovative, AI -based corporate information management system. The VK ArtiFin question-answer system combines all documents used within the organization into a single data repository and performs semantic analysis using LLMs. It provides instant, accurate, and contextual answers to users' questions about legislation, procedures, or internal policy documents. This automates manual information access processes, increases operational efficiency, and transforms decision-making processes into more reliable ones.This study proposes the VK ArtiFin question-answer system as an exemplary model for the applicability of AI-based information access and management systems in the banking sector. The study examines the system's architecture, data processing methods, and operational outputs. The results reveal the impact of AI-based information access solutions on organizational efficiency, employee experience, and digital transformation in banking.

Today, conversational agents powered by AI and especially transformer-based LLMs have radically transformed the way we access information and interact with digital systems[1]. Compared to traditional enterprise systems, chatbots offer two key advantages. First, they allow users to intuitively ask questions using natural language and receive conversational responses. Second, they are becoming increasingly proficient at executing complex search tasks, significantly simplifying problem-solving and decision-making across a variety of domains. [2] In their study, Bhattacharya et al. examine how banks are using AI applications to enhance the customer experience during the digitalization process. The research examines banking processes, and the opinions and findings obtained from bankers reveal the AI features used in banking platforms and demonstrate that chatbot use cases are ranked by customer importance [3] .In their study, Fazlija et al. emphasize that tightening regulations aimed at strengthening financial stability have created increased costs and complex software development processes for banks. They examine the potential of GenAI, and particularly LLMs, to reduce software development burdens during these challenging times. They also propose workflows that combine LLMs with human expertise, demonstrating that the presented methods can significantly support the development of regulatory compliance software [4].In their study, Cao et al. examine the transformative impact of LLMs in the financial sector. They emphasize the multifaceted benefits LLMs offer, including improving customer support processes, strengthening fraud detection, and enhancing market analysis and forecasting capabilities. However, they also highlight critical challenges such as user trust and ethical risks. The research highlights the importance of collaboration between industry and academia for the responsible and reliable integration of LLMs into financial applications [5].In their study, Veturi et al. examine the RAG (Retrieval-Augmented Generation) method, which enables LLMs to be supplemented with a knowledge base to provide accurate and reliable answers to customer questions. An end-to-end RAG system was developed for a real retail call center, combining customer queries with relevant documents and providing agents with response suggestions. Automated and human evaluations demonstrate that this system produces more accurate and relevant responses than existing BERT-based methods. Consequently, it has been demonstrated that RAG-enhanced LLMs can significantly reduce agent workload [6].In their study, Lithgow et al. present a new benchmark, FinDoc-RAG, developed to evaluate the performance of RAG-based document processing systems in the financial sector. Four current RAG architectures were tested with over 600 question-answer pairs generated from 46 documents from a banking institution. The documents contained dense numerical information and complex layouts, such as product descriptions, investment guides, legal policies, and marketing materials. The evaluation results show that while the models achieve high accuracy in basic information extraction (0.91), performance degrades significantly (0.44) on challenging questions requiring the merging of multiple documents. The study demonstrates that different RAG approaches have varying strengths depending on the complexity of financial documents, demonstrating that FinDoc-RAG is an important standard for measuring progress in this field [7].The study by Lajčinová et al. examines how effective LLM is in accurately detecting user intent in fixed-response chatbots used on banking websites. The study compares the fine-tuned SlovakBERT model with both raw and fine-tuned versions of multilingual generative models such as Llama 8B instruct and Gemma 7B instruct. The findings show that SlovakBERT achieves higher accuracy in in-scope classification and lower rates of out-of-scope mismatches. These results suggest that SlovakBERT stands out as the most successful model for this use case. [8].Chua et al. present an LLM-based agent architecture called Ryt AI that allows customers to perform basic banking transactions using natural language. This solution stands out as the world's first regulatory-approved application that directly handles banking transactions through a chat interface, unlike previous chatbots designed for advisory or support purposes. Built on the closed-source ILMU model, developed entirely in-house by the bank, this architecture integrates four separate LLM agents: Guardrails, Intent, Payment, and FAQ, transforming multi-step screen processes into a single, integrated dialogue flow. The system ensures security and compliance with deterministic

security layers, human verification, and a stateless audit architecture that facilitates record tracking. The findings demonstrate that natural language-based interfaces can reliably perform basic financial transactions with appropriate security measures and regulatory compliance [9].AI, particularly LLM- and RAG-based systems, have the potential to enhance many functions in the banking sector, such as operational efficiency, decision support processes, and customer service. However, the integration of these technologies into banking applications poses significant risks in terms of ethics, privacy, and regulatory compliance. Literature indicates that LLM-based systems can create fairness issues due to biased training data, leading to unequal service delivery among different demographic groups [10]. Furthermore, the "black box" nature of AI makes it difficult to explain how banking decisions are made and makes it difficult to track accountability [11]. While the RAG architecture provides real-time information access, it creates additional privacy risks due to the processing of high volumes of customer data, making compliance with GDPR and U.S. banking privacy laws critical [12].In their study, Ferdaus et al. examine the trust, ethics, and transparency issues arising from the rapid development of AI. Challenges such as bias, risk of attack, misuse, and the inability to explain decisions undermine user trust. They emphasize that the need for regulation, ethical oversight, and accountability increases as usage in critical areas such as finance and healthcare. The study offers fundamental principles and recommendations for developing trustworthy and responsible AI. On the regulatory front, frameworks such as GDPR, PCIDSS, and the AI Act mandate transparency, fairness, and explainability in the use of AI in banking [13].

2. Methodology

This section describes the data sources, software components, and methods used for the AI-based question-answer system. The question-answer system is built on a modular and scalable architecture that allows users to make context-sensitive queries on both corporate documents and files they upload.

2.1. Dataset Description and Document Flows

The question-answer system works through two different data processing flows, File-Based and Document Management System, so that users can make context-sensitive queries on documents.As part of the File-Based Question-Answering Flow, users can upload their own documents to the system. Uploaded files are converted to text format using the RapidOCR and Docling libraries, parsed, segmented, and indexed. This structure ensures high accuracy by allowing contextual queries to be made only on the relevant document. This allows users to answer questions on their own documents.Document Management System Flow allows for Q&A on a total of 2,344 documents currently residing in the corporate document management system. These documents consist of 1 PowerPoint, 273 Excel, 310 PDF, and 1,760 Word files. Documents are categorized by content, such as procedures and regulations. This classification simplifies the data indexing process, provides users with quick contextual access to documents, and increases accuracy.

2.2. Dialog Analysis

A dialogue analysis mechanism was designed to correctly classify user-received questions. Expressions such as greetings, thanks, help, and exit commands are responded to with rule-based responses. Question-based messages are routed to the RAG pipeline. A LangGraph-based multi-agent intent detection approach was tested experimentally; however, it was not used in the final question-answer system because it required additional LLM calls for each query and increased hardware costs.

2.3. Security, Prompt Engineering, and Compliance

A comprehensive prompt engineering process was implemented to prevent the model developed for the question-answer system from generating out-of-scope or undesirable answers not covered in the documentation. This process employed structural prompting, role prompting, and chain-of-thought suppression approaches. Queries containing personal data or requiring external information were restricted, thus ensuring compliance with KVKK and data privacy requirements.

2.4. System Architecture and Software Components

The developed question-answer system is built on a modular architecture to ensure high performance and scalability. The architectural layers and their definitions are shown in Table 1.

Table 1: Application Layers

Layer	Description
Information Indexing Layer - LlamaIndex	It was used for document parsing and semantic search. It formed the basis of the RAG pipeline, enabling queries to be matched with relevant document fragments.
Model Inference Layer - vLLM	vLLM was preferred for its low-latency response generation and high parallel query capacity. The PagedAttention mechanism enabled efficient use of GPU memory.
Vector Database - ChromaDB	It is used for storing embeddings and performing fast similarity searches. Semantic accuracy is optimized with the BGE-M3 and Multilingual-e5 embedding models.
Service Layer - FastAPI	It is structured as an API layer that acts as a bridge between the user interface and the model. Its modular REST architecture simplifies integration with enterprise systems.

2.5. Testing and Integration

Extensive testing and integration studies were conducted on the VK ArtiFin question-answer system, which emerged as a result of the development efforts. These studies aimed to evaluate the system in terms of accuracy, performance, security, and integration. The main test titles and study details of the test studies carried out are asFunctional Tests: OCR, file uploading, indexing, and query response accuracy were tested. In these tests, the consistency of the responses with the document was checked. Tests were conducted on 2,837 test scenarios, and the average accuracy was measured as 80%.Performance Tests: Sub-second response times were observed for concurrent user queries with the vLLM Paged Attention mechanism.Security Tests: Prompt injection, XSS and data leakage tests within the scope of KVKK were implemented and the effectiveness of the filter mechanisms was verified.Regression Tests: Previous development environment versions were compared and no critical errors or loss of functionality were detected.

2.6. LLM and Embedding Model Analysis

Meta-llama/Llama-3.1-70B-Instruct, NVIDIA Llama-3.1-Nemotron-70B-Instruct-HF, and meta-llama/Llama-3.3-70B-Instruct models were studied as LLM models. The Meta-llama/Llama-3.1-70B-Instruct model provided higher accuracy in text understanding and answer generation than previous models and was used in the final question-answer system.When BGE-M3 and Multilingual-e5 were used together as embedding models, higher accuracy was achieved than when used alone, and they were used together. Other LLM embedding models (BAAI/bge-multilingual-gemma2, gte-Qwen 2-7B Instruct) did not demonstrate sufficient success and were therefore not used.

2.7. Application Features and User Interaction

The question-answer system provides an interface where users can upload documents or query existing corporate documents within the document management system. Users enter their questions in natural language, and the system generates responses and directs the user to the relevant document. Security filters and KVKK-compliant operation support a secure user experience.A comprehensive user guide has been developed to ensure users can use the VK ArtiFin question-answer system effectively and efficiently. This guide covers the system's basic functions, document uploading and querying steps, security warnings, and possible error handling procedures. The user guide was shared with both internal users and the pilot test group, and contributed significantly to system adoption and improved user experience.

2.8. Challenges Encountered and Solutions Implemented

Various technical and operational problems were encountered throughout the question-answer system development and integration process, and systematic approaches were implemented to address these problems. Table 2 lists the major problems identified during the project and the solutions developed to address them.

Table 2: Problems and Implemented Solutions

Problem	Implemented Solutions
English Answers	Translate models & language detection, LLM model selection
Responses Containing Repetition	Post-processing transactions, LLM model optimization
Non-Document Responses	Parameter optimization, Prompt engineering
Low Accuracy Rate	Post/preprocess transactions, embedding ve retrieval optimization, LLM model improvements, index-based indexing, adding abbreviations to the index, directing the user to the relevant document
Banned Questions	Pre-process banned word blocking, Llama Guard

3. Results

The implemented AI-based question-answer system successfully processed the institution's large document pool consisting of 13 different directories, was tested on a total of 2,837 question-answer scenarios and achieved an accuracy rate of 80%. This performance demonstrates that the system both adapts to the institution's document structure and significantly improves information access processes. The fact that the developed solution is not limited to specific units but is designed for use by all employees has enabled it to generate direct value in a wide range of areas, from operational processes and compliance and auditing to human resources and legal matters. Rapidly resolving uncertainties frequently encountered by employees in their daily workflow has eliminated the need for manual information searching, thus reducing process errors, ensuring standardization and increasing operational efficiency. As a result, the project makes a strong contribution to the bank's digital transformation goals, providing an effective AI infrastructure that accelerates access to information across the organization, supports processes holistically, and improves the employee experience.

4. Discussion

This study clearly demonstrates that specialized LLMs can create an effective and reliable information access infrastructure in areas requiring high accuracy, consistency, and regulatory compliance, such as banking. The findings demonstrate that LLM-based systems not only offer question-answering capabilities but also contribute to the standardization of operational information flow, the automation of processes, and the sustainable digitization of institutional memory. The architecture developed within this framework is considered to have the potential to form the basis for future, more advanced AI applications in audit automation, regulatory gap analysis, regulatory compliance monitoring, early detection of process risks, and other institutional domains with high textual data density. In addition, such systems appear to offer significant opportunities to strengthen decision support processes and increase institutional efficiency by reducing human dependency in areas requiring specialized knowledge.

5. Conclusion

This study clearly demonstrates that specialized LLMs can create an effective and reliable information access infrastructure in areas requiring high accuracy, consistency, and regulatory compliance, such as banking. The findings demonstrate that LLM-based systems not only offer question-answering capabilities but also contribute to the standardization of operational information flow, the automation of processes, and the sustainable digitization of institutional memory. The architecture developed within this framework is considered to have the potential to form the basis for future, more advanced AI applications in audit automation, regulatory gap analysis, regulatory compliance monitoring, early detection of process risks, and other institutional domains with high textual data density. In addition, such systems appear to offer significant opportunities to strengthen decision support processes and increase institutional efficiency by reducing human dependency in areas requiring specialized knowledge.

6. Acknowledgements

This study was carried out using the infrastructure of Vakıf Katılım and KFT Bilişim Sistemleri. On this occasion, we would like to thank Vakıf Katılım and KFT Bilişim Sistemleri for the valuable support and infrastructure they provided.

7. References

[1] A. S. N. P. N. U. J. J. L. G. A. N. .. &. P. I. Vaswani, "Attention is all you need," Advances in neural information processing systems, p. 30, 2017.

[2] R. W. (. A. White, “dvancing the search frontier with AI agents,” Communications of the ACM, pp. 67(9), 54-65., 2024.

[3] C. &. S. M. Bhattacharya, “The role of artificial intelligence in banking for leveraging customer experience,” Australasian Accounting, Business and Finance Journal, p. 16(5)., 2022.

[4] B. I. M. F. A. &. F. A. Fazlija, “ Implementing financial regulations using large language models,” Available at SSRN, 2024.

[5] X. L. S. K. V. K. A. T. H. H. L. Z. .. &. P. C. Cao, “Empowering financial futures: Large language models in the modern financial landscape.,” 2024.

[6] S. V. S. J. R. L. T. N. I. &. Y. N. (. Veturi, “ Rag based question-answering for contextual response prediction system.,” p. arXiv preprint arXiv:2409.03708., 2024.

[7] O. K. D. K. V. A. D. L. M. B. C. .. &. S. O. Lithgow-Serrano, “Assessing RAG System Capabilities on Financial Documents,” Assessing RAG System Capabilities on Financial Documents. In Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing, pp. pp. 124-147, 2025, November.

[8] B. V. P. &. S. M. Lajčinová, “ Intent Classification for Bank Chatbots through LLM Fine-Tuning,” arXiv , p. arXiv preprint arXiv:2410.04925, 2024.

[9] X. J. T. J. M. L. T. J. X. P. S. C. G. Y. X. C. D. H. T. .. &. C. C. S. Chua, “Banking Done Right: Redefining Retail Banking with Language-Centric AI.,” In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. pp. 646-658, 2025, November.

[10] H. P. Kothandapani, “Ai-driven regulatory compliance: Transforming financial oversight through large language models and automation.,” Emerging Science Research, 2025.

[11] S. Joshi, “Bridging the AI skills gap: Workforce training for financial services.,” Available at SSRN 5206490., 2025.

[12] H. Padmanaban, “ Revolutionizing regulatory reporting through AI/ML: Approaches for enhanced compliance and efficiency.,” Journal of Artificial Intelligence General science , pp. (JAIGS) ISSN: 3006-4023, 2(1), 71-90., 2024.

[13] M. M. A. M. I. E. N. K. N. P. K. &. S. S. Ferdaus, “Towards trustworthy ai: A review of ethical and robust large language models,” arXiv preprint , p. arXiv:2407.13934., 2024.