Eric Tachdjian

Gabriel Rocha Pinto Santos

Pedro Munhoz de Souza Rivero

Rafael Lupovici Moritz

Chatbot for assisting documentary research using natural language

SÃO PAULO
2025

Eric Tachdjian

Gabriel Rocha Pinto Santos

Pedro Munhoz de Souza Rivero
Rafael Lupovici Moritz

Chatbot for assisting documentary research using natural language

Final Course Project submitted to the Institute of Technology and Leadership (INTELI), to obtain a bachelor's degree in Information System.

Advisor: Prof. José Romualdo

SÃO PAULO
2025

Cataloging in Publication
Library and Documentation Service
Institute of Technology and Leadership (INTELI)
Data entered by the author.

Brief

Tachdjian, Eric; Santos, Gabriel Rocha Pinto; Rivero, Pedro Munhoz de Souza; Moritz, Rafael Lupovici. **Chatbot for assisting documentary research using natural language**. 2025. 23 pages. TCC (Graduation) – Information System, Instituto de Tecnologia e Liderança, São Paulo, 2025.

The objective of the project was to build a chatbot capable of reliably answering questions based on documents provided by users, using Retrieval-Augmented Generation (RAG) techniques to ensure accuracy and traceability of responses. To achieve this, a system was developed that can receive PDFs, HTML pages, and links, extract their content, normalize it, split it into smaller chunks, and convert them into semantic vectors stored in the Qdrant vector database. Based on this foundation, the chatbot is able to search for the most relevant chunks and use them as context to generate clear answers aligned with the content of the documents.

During development, essential features for user adoption were also implemented, such as role-based authentication (ensuring that each user has access only to what they are permitted), automatic version control to prevent duplication when the same document is uploaded again, and context expansion to increase query accuracy. The final result is a robust chatbot capable of interpreting complex documents and providing contextualized and auditable answers, reducing the need for manual reading and facilitating access to information. It is concluded that the system achieves its goals of usability, security, and reliability, representing a solid foundation for corporate applications that depend on large volumes of documents.

ABSTRACT

Tachdjian, Eric; Santos, Gabriel Rocha Pinto; Rivero, Pedro Munhoz de Souza; Moritz, Rafael Lupovici. **Chatbot for assisting documentary research using natural language.** 2025. 23 pages. TCC (Graduation) – Information System, Instituto de Tecnologia e Liderança, São Paulo, 2025.

The objective of this project was to develop a chatbot capable of reliably answering users' questions based on the documents provided to the system, employing Retrieval-Augmented Generation (RAG) techniques to ensure the accuracy and traceability of the generated responses. To this end, we built a system capable of receiving PDFs, HTML pages, and web links, extracting their content, normalizing the text, splitting it into smaller segments, and converting these segments into semantic vectors stored in the Qdrant vector database.

With this structured foundation, the chatbot retrieves the most relevant chunks and uses them as context to generate clear answers aligned with the original content. During development, additional essential features for real-world use were implemented, including role-based authentication (ensuring that each user has access only to authorized information), automatic version control to prevent duplication when the same document is uploaded multiple times, and context expansion to increase retrieval accuracy.

The final result is a robust chatbot capable of interpreting complex documents and providing contextualized and auditable responses, reducing the need for manual reading and facilitating access to information. It is concluded that the system meets its objectives of usability, security, and reliability, serving as a solid foundation for corporate applications that rely on large document collections.

List of Illustrations

Summary

1 Introduction

Currently, internal processes and procedures within Operations Brazil are documented in internal portals and repositories, but searching for this information is not intuitive and is often inefficient. Employees waste time looking for manuals, workflows, and policies, and there is no integrated mechanism that provides up-to-date answers adapted to the context of each area and to Compliance and Information Security rules **(Gupta et al., 2011; Valentine et al., 2018)**.

The goal is to develop a solution that enables the centralization and continuous updating of information related to processes, access, training, standards, and guidelines that must be considered by employees, allowing the integration of multiple sources and content types. The solution should also include a corporate chatbot that, using on-premises technologies, enables the retrieval of previously centralized and updated information through natural language queries made by users **(Chandar et al., 2017; Stoeckli et al., 2019; Akkiraju et al., 2024)**.

1.1. **Partner Company Context:**

The partner company operates in the financial sector, offering credit services, regulatory analysis, and operational support. It is a large organization that deals daily with a high volume of complex documents, such as Central Bank regulations and internal policies. The area most impacted by the project is regulatory analysis, which is responsible for interpreting legal guidelines and disseminating accurate information to other teams, facing challenges related to agility and standardization **(Cao & Feinstein, 2024; Adeniran et al., 2024)**.

The solution brings significant value to the company by transforming a process that was previously slow and manual—searching for information in lengthy documents—into something fast, standardized, and secure. With the creation of the chatbot, teams no longer rely on reading regulatory documents and instead obtain direct answers that are always grounded in official excerpts. This reduces rework, increases the accuracy of analyses, and minimizes the risk of misinterpretation. In

addition, automation frees analysts to focus on higher-impact tasks, improves information governance, and accelerates strategic decision-making in an increasingly dynamic regulatory environment. The system also contributes greatly to the onboarding of new employees, allowing them to learn processes and regulations more quickly, in a guided and reliable manner, reducing the learning curve and accelerating productivity. **(Zhong et al., 2020; Moharana et al., 2023; Patel et al., 2025)**.

1.2. **Problem Definition (Corporate Pain Point):**

The central problem faced by the company is the difficulty in accessing accurate and up-to-date information within extensive documents, such as internal regulations, resolutions, and corporate policies. Regulatory texts are inherently complex, frequently updated, and often contain implicit structures and ambiguous language, which makes information retrieval and interpretation particularly challenging **(Sapkota et al., 2012; Rayo et al., 2025)**.

Currently, analysts need to manually read large volumes of text, interpret complex rules, and validate specific excerpts in order to respond to operational and regulatory questions. This process is slow, requires a high level of concentration and prior knowledge, and is highly susceptible to divergent interpretations, especially in environments where documents evolve continuously and may contain inconsistencies or contradictions **(Schumann & Gómez, 2024)**.

As a result, operational bottlenecks arise, along with frequent rework, dependence on more experienced professionals, and the risk of inconsistencies in critical analyses. Manual information retrieval significantly reduces productivity, increases response times, and hinders the standardization of operations, while conflicts and inconsistencies across regulatory documents may lead to errors, delays, and increased compliance risks **(Schumann & Gómez, 2024)**.

Before the implementation of the chatbot, the time required to locate and confirm information in regulatory documents was considerably higher due to the volume,

complexity, and constant evolution of these materials. This scenario also resulted in high variability in the quality of the answers provided and substantial rework, particularly for recurring questions, reinforcing the need for a more efficient and reliable approach to regulatory information access **(Sapkota et al., 2012; Rayo et al., 2025)**.

1.3. Proposed Solution and Expected Contribution:

The proposed solution is a chatbot that uses Retrieval-Augmented Generation (RAG) techniques to search for information within the company's documents and answer questions accurately, always based on official excerpts. By combining information retrieval mechanisms with large language models, RAG-based approaches improve the accuracy and reliability of document-based question answering systems, mitigating issues such as hallucinations and outdated knowledge **(Muludi et al., 2024; Guo et al., 2025)**.

The system automates steps such as text extraction, content organization, embedding creation, and storage in a vector database (Qdrant), enabling the rapid identification of the most relevant excerpts for each query. Embedding-based semantic search allows the system to retrieve contextually relevant information from large and heterogeneous document collections, which is essential in enterprise environments **(Muludi et al., 2024)**.

Important features were also implemented, including version control, access permissions, and context expansion, aiming to ensure response consistency, information reliability, and secure access to sensitive corporate content. Although the adoption of RAG in enterprise contexts presents implementation challenges, proper architectural design and evaluation frameworks enable its effective use in complex organizational settings **(Bruckhaus, 2024)**.

The project aims to significantly reduce the time teams spend searching for information in lengthy documents, increase the accuracy of responses, and decrease interpretation errors. As a result, it is expected to improve daily operational efficiency and reduce rework by providing fast, consistent, and reliable answers to recurring operational and regulatory questions.

1.4. **Business Objectives:**

The business objectives include accelerating access to critical information, reducing misinterpretation errors in regulatory documents, decreasing the time teams spend on manual searches, and increasing overall productivity. In addition, the project aims to strengthen information governance, improve the onboarding of new employees, and support faster and more secure decision-making within the company.

1.5. **Structure of the thesis/dissertation:**

This work is organized into chapters that progressively present the development of the solution. The first chapter introduces the project context, its purpose, and the relevance of automation in document querying. Next, the second chapter describes in detail the problem faced by the partner company and the identified requirements. The third chapter presents the theoretical foundation, addressing concepts such as RAG, vector databases, embeddings, and language models. The fourth chapter details the adopted methodology, including the stages of information ingestion, processing, storage, and retrieval. The fifth chapter describes the implementation of the solution, covering the architecture, technical decisions, and version control and permission mechanisms. The sixth chapter presents the results obtained and discusses how the system meets the proposed objectives. Finally, the last chapter provides the overall conclusions and suggestions for improvements and future developments of the solution.

2 Solution Development

This Solution Development section brings details about the timeline to develop the solution.

2.1 Applied Rationale

The following section is dedicated to explaining the reasons why certain decisions were made throughout the development process:

2.1.1 **Business Area Rationale:**

The project aims to develop a solution capable of centralizing and facilitating access to knowledge related to the execution of internal processes, while also enabling its continuous updating through a corporate chatbot. The project seeks to improve the consultation of documents, regulations, workflows, and institutional materials in a structured manner, ensuring that employees obtain accurate responses aligned with the organization's guidelines.

The proposed solution includes several essential functionalities. The first is natural language search, allowing employees to ask questions such as "What access do I need to work in Cash Operations?" and receive answers based on the specific rules of the area and the Compliance standards previously registered in the centralized knowledge base.

In addition, the solution considers an adaptive profile model, allowing responses to vary according to the user's area of activity, role, and access level, ensuring both accuracy and security in information sharing.

2.1.2 **Technological rationale for the solution:**

Qdrant is an open-source vector database optimized for similarity search and widely used in applications that rely on semantic retrieval. It was selected for this project because it offers high performance even when handling thousands of embeddings, native support for filters such as access roles and timestamps, and straightforward deployment on internal servers without the need for additional external

dependencies. This combination meets the performance, security, and isolation requirements expected in corporate environments.

The embedding model used, all-MiniLM-L6-v2, was chosen because it is lightweight and efficient for local execution, does not require a GPU, delivers strong semantic performance, provides full support for Portuguese, and has an open license, allowing unrestricted corporate use. This model is responsible for transforming text into numerical vectors that enable semantic comparison between documents and queries.

For the document extraction and normalization stage, the solution incorporates the pypdf library for reading PDFs, while Selenium together with Chrome DevTools is used to convert HTML pages into PDFs when necessary. After extraction, textual normalization is applied to remove noise, accents, and repetitive patterns, ensuring content consistency prior to vectorization. Next, a chunking step is performed, which splits lengthy documents into smaller parts to improve the retrieval of relevant passages.

The system uses Text Generation Inference (TGI) as the LLM engine, enabling local execution of language models with low latency, full control over inference parameters, support for open-source models such as Qwen and Mistral, and fully on-premises operation, ensuring compliance with security and privacy requirements.

At the backend layer, the project uses FastAPI, which is responsible for managing document ingestion, implementing JWT-based authentication, controlling role-based access permissions (RBAC), and orchestrating calls to Qdrant and the LLM. The framework was chosen for its speed, strong typing, and ease of maintenance.

On the frontend, Streamlit was selected because it enables rapid interface development, is highly intuitive for non-technical users, is lightweight and compatible with on-premises environments, and does not require complex web servers to run.

Finally, the entire solution was containerized using Docker, including the API, frontend, vector database, and LLM. This approach ensures reproducibility, ease of internal deployment, isolation between components, and a fully controlled and secure environment. This architecture simplifies maintenance, standardizes environments,

and ensures that the system operates stably across different corporate infrastructures.

2.1.3 **Fundamentals of Management and Development Methods:**

The development of the project was guided by three main pillars: agile methodologies, DevOps principles, and PMBOK fundamentals. This combination made it possible to organize the workflow, maintain a consistent delivery pace, and ensure technical quality at all stages.

Within the agile methodologies pillar, Scrum and Kanban practices were adopted to structure the work into short cycles. Weekly sprints, daily meetings, and frequent reviews facilitated continuous adaptation to project changes, while the use of a Kanban board made it possible to prioritize activities and clearly track task progress. This set of practices directly contributed to faster deliveries aligned with the partner's needs.

Regarding DevOps principles, the project incorporated practices focused on environment stability and reproducibility. Containerization with Docker ensured that the API, LLM, and vector database ran consistently on any machine; isolated environments enabled controlled testing; and version control with Git ensured traceability and change management. These elements reduced configuration-related errors and facilitated integration among the solution's components.

Finally, PMBOK fundamentals were applied primarily to structure project governance. Clear scope definition, early risk identification, organized communication with stakeholders, and systematic quality control of deliverables helped maintain alignment between the team and the partner.

The integration of agile methods, DevOps practices, and PMBOK principles provided an effective balance between organization and flexibility, supporting a consistent and well-structured development of the chatbot.

2.2 Specification and Development

In this section, all technical specifications should be presented and must be aligned with the standards and needs of the partner company:

2.2.1 Requirements and Specifications:

Functional Requirements

The system must allow the upload of documents in different formats, such as PDFs and HTML pages, as well as the submission of individual or batch URLs. The chatbot must be able to process these documents, extract and normalize their content, generate semantic embeddings, and store them in a vector database. It must also perform contextual searches, retrieve the most relevant excerpts, and generate responses based exclusively on this content. In addition, the system must include role-based authentication, ensuring that each user can view only information compatible with their access level, as well as automatic document version control.

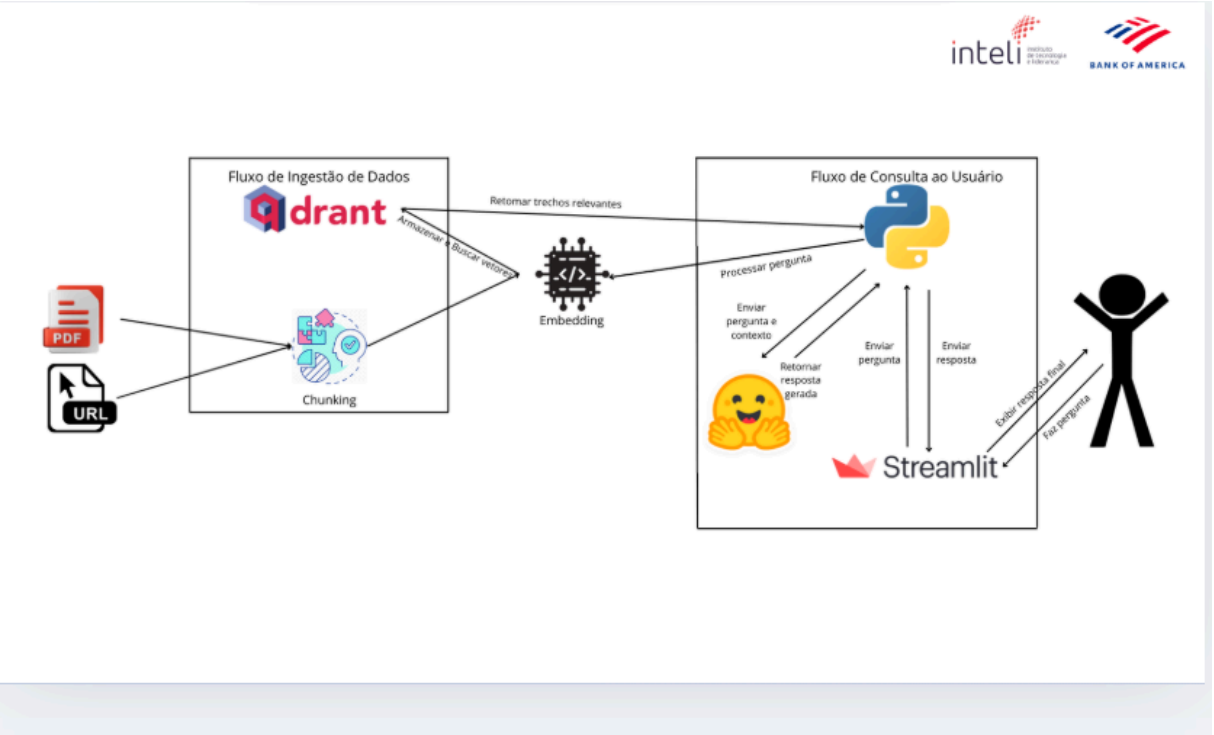
Non-Functional Requirements

The solution must ensure secure access to information, adequate performance even with large volumes of documents, traceability of generated responses, and ease of maintenance. It is also required that the system be scalable, reliable, and compatible with on-premises environments, in compliance with the company's internal infrastructure and compliance policies.

User Specifications and Use Cases

The main users of the system are analysts, managers, and administrators. Use cases include uploading documents for ingestion, querying the chatbot to clarify questions, viewing responses based on reliable sources, and managing access permissions. The system was designed to support both experienced users and new employees, facilitating access to information and reducing the learning curve.

2.2.2 Architecture and Technology:



○ *Image 1 - Solution Design*

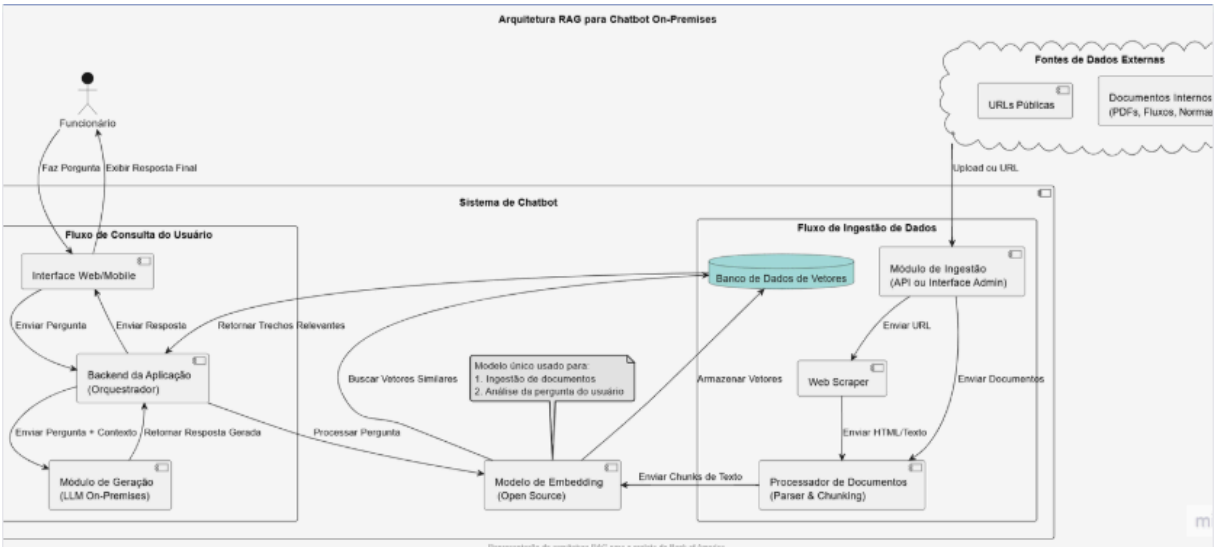


Image 2 - Detailed Solution Design

Above, examples of the architecture developed for the solution is presented. The model was designed to ensure compatibility with the company's IT ecosystem, integrating computationally efficient tools that are well suited to corporate needs.

Among these, the use of Qdrant as the vector database stands out, as it offers high performance for semantic search operations, enabling fast information processing and significantly reducing operational costs.

The architecture includes a data ingestion flow responsible for processing documents in various formats, such as PDFs and content obtained from URLs. This flow performs chunking and embedding generation steps, allowing data to be structured and stored in an optimized manner for contextual queries.

In parallel, the user query flow uses technologies such as Python and Streamlit, which enable natural language interaction by processing questions, retrieving relevant excerpts, and returning answers in an intuitive and efficient way. This integration ensures a smooth user experience and meets the usability, security, and performance requirements demanded by the corporate environment.

2.2.3 **Development and Implementation (MVP):**

The development was divided into two phases, each consisting of five incremental sprints, for a total of ten sprints. The breakdown is as follows.

The first phase focused on data persistence. In the first sprint, the project plan document was finalized. The second sprint was dedicated to designing the project architecture, including the database, local server, and supporting tools. In the third sprint, processing of the partner's data began, with the definition of algorithms and technical solutions. The fourth sprint involved vectorizing the data and storing it on the local server. In the fifth sprint, the vectorization process was refined and made accessible through an API.

The second phase focused on developing the chatbot and its interface. In the first sprint, a POST endpoint was created to submit user questions, apply the embedding model to transform the text into vectors, retrieve the most relevant chunks from Qdrant, and return the top-k closest results through the API, with tests implemented. In the second sprint, different LLMs were evaluated and the RAG prompt template was developed, with the expectation of already returning responses generated by the

LLM. The third sprint focused on developing a minimal front-end interface and completing the end-to-end communication flow, resulting in an MVP for the final project presentation. In the fourth sprint, both the front-end and the LLM model were improved. The fifth sprint addressed error handling and hyperparameter tuning, and finalized the documentation and user manual.

The project is delivered through a public repository and includes a usage manual that can be adapted to the partner's environment.

2.2.4 **Testing and Technical Evaluation:**

The solution was primarily tested by validating the pipeline's operation from document ingestion through chunk retrieval and response generation. During development, we analyzed logs, printed the values used for metrics to confirm that each step was correct, and manually evaluated the chatbot's responses to ensure they were coherent and accurately reflected the content of the documents.

The tests showed that the system executes the entire workflow in a stable manner, retrieves the most relevant excerpts, and meets the technical criteria defined for the project.

2.3 Assessment of Impact and Contribution to the Business

In this section, the return on investment in terms of time and resources for implementing the solution should be measured.

2.3.1 **Defining Corporate Success Metrics:**

The success of the project is evaluated through indicators such as the reduction in time spent searching for information in documents, the increase in the accuracy of the answers provided by the chatbot, and the decrease in the need for manual

consultation of regulations. As proposed by **Wang, Kuo, and Zhou (2022)**, the model performance against benchmarks is also considered. The tool's usage rate by users is also considered, indicating its level of adoption within the corporate environment.

The measurement of results is carried out by comparing the previous scenario, based on manual and non-standardized searches, with the scenario after the chatbot's implementation. This comparison is made using system usage logs, average response time, and user feedback, making it possible to verify the efficiency and reliability gains achieved by the solution.

2.3.2 **Results and Impact Analysis:**

The developed solution shows clear gains compared to the previous process, which relied on manual reading of lengthy documents. Although exact numerical metrics were not collected, there is a significant potential reduction in time spent on queries and a decrease in rework, since the chatbot delivers direct answers grounded in official excerpts from the documents.

From a qualitative perspective, the tool increases team agility, improves response consistency, and reduces the risk of misinterpretation. It also strengthens information governance and supports the onboarding of new employees by facilitating the understanding of regulations and internal procedures.

From a cost-benefit standpoint, the on-premises solution eliminates licensing expenses and leverages the company's existing infrastructure. The main costs are related to development and maintenance, while the benefits include continuous productivity gains, knowledge standardization, and scalability without significant cost increases. Thus, even without direct financial figures, the project demonstrates a positive return and a tendency toward fast payback as it becomes regularly used.

2.3.3 **Critical Success Factors and Lessons Learned:**

The project worked well because the problem was clearly defined, there was constant communication with the partner, and deliveries were made incrementally, allowing for rapid validation. The choice of technologies—especially RAG and Qdrant in an on-premises environment—was also decisive, ensuring security, efficiency, and the possibility of future expansion. According to **Reimers and Gurevych (2019)**, BERT has set a new state-of-the-art performance on sentence-pair regression tasks like semantic textual similarity (STS), and SBERT uses it in a context with sentences.

We learned that the size and quality of the chunks have a strong influence on the chatbot's accuracy, requiring continuous adjustments. Small configurations of the similarity mechanism also directly impact the results. In addition, it became evident that user adoption depends on a simple and objective interface, reinforcing the importance of prioritizing usability from the beginning.

3 Conclusion

It is concluded that the objectives proposed in this project were fully achieved. The developed chatbot proved capable of reliably answering questions based on documents provided by users, ensuring information traceability and compliance with role-based access permissions defined for each employee. The solution meets the partner's needs by transforming a manual and error-prone process into a more agile, standardized, and secure experience.

From a business perspective, the impacts are positive, particularly in reducing operational effort, increasing team productivity, and improving information governance. Similar results have been observed in enterprise environments adopting AI-driven chatbots, which have demonstrated improvements in productivity, decision-making, and operational efficiency when strategically implemented **(Durach & Gutierrez, 2024; Akkiraju et al., 2024)**.

The tool also proves to be valuable for onboarding new employees, facilitating the understanding of complex documents and accelerating adaptation to the corporate

environment, as supported by recent studies highlighting the effectiveness of RAG-based chatbots in onboarding processes **(Frischen & Fiebig, 2025)**.

After delivering the solution and all technical documentation, the next step for the partner is to incorporate the chatbot into the corporate environment and begin using it in real scenarios involving internal document queries. At this stage, the internal team will be able to validate the tool's operation with real users, collect feedback, and adjust parameters to ensure that the solution is effectively used and aligned with the company's needs and reality.

Finally, the adopted architecture allows the solution to be maintained and evolved independently by the partner's team, facilitating future adjustments, model updates, and adaptations to business needs without compromising system stability—an important factor for the long-term scalability and sustainability of enterprise AI solutions **(Akkiraju et al., 2024)**.

References

Book:

[1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[2] J. Wang, Y. Kuo, and D. Zhou, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," *arXiv preprint arXiv:2212.03533*, 2022. [Online]. Available: https://arxiv.org/abs/2212.03533

[3] Z. Cao and Z. Feinstein, "Large Language Model in Financial Regulatory Interpretation," *IEEE Conference on Computational Intelligence for Financial Engineering & Economics*, 2024.
[Online]. Available: https://ieeexplore.ieee.org/document/10408478

[4] I. A. Adeniran et al., "Strategic risk management in financial institutions: Ensuring robust regulatory compliance," *Finance & Accounting Research Journal*, 2024.
[Online]. Available: https://doi.org/10.51594/farj.v6i1.721

[5] B. Zhong et al., "A building regulation question answering system: A deep learning methodology," *Advanced Engineering Informatics*, vol. 45, 2020.
[Online]. Available: https://doi.org/10.1016/j.aei.2020.101123

[6] J. Moharana et al., “Revolutionizing Search: Artificial Intelligence and Machine Learning's Impact on Information Retrieval,” *International Journal For Multidisciplinary Research*, 2023.
[Online]. Available: https://www.ijmr.net.in

[7] J. Patel et al., “A Survey: Information Search Time Optimization Based on RAG (Retrieval Augmented Generation) Chatbot,” *Paripex – Indian Journal of Research*, 2025.
[Online]. Available: https://www.worldwidejournals.com/paripex

[8] S. Chandar et al., “Towards Intelligent Question Answering Systems,” 2017.
[Online]. Available: https://arxiv.org/abs/1705.02025

[9] S. Stoeckli et al., “How Assistive AI Improves Knowledge Work,” 2019.
[Online]. Available: https://hbr.org/2019/07/how-assistive-ai-improves-knowledge-work

[10] R. Akkiraju et al., “Enterprise-Grade Conversational AI Systems,” 2024.
[Online]. Available: https://www.ibm.com/blog/enterprise-conversational-ai

[11] T. Bruckhaus, “RAG Does Not Work for Enterprises,” *arXiv preprint*, 2024.
[Online]. Available: https://arxiv.org/abs/2401.02684

[12] Y. Guo, L. Yan, J. Niu, D. Gao, and X. Yuan, “Integrating Retrieval-Augmented Generation (RAG) and Knowledge Augmented Generation (KAG) Frameworks to Build Accurate Enterprise Question Answering Systems,” *IEEE Advanced Information Technology, Electronic and Automation Control Conference*, 2025.
[Online]. Available: https://ieeexplore.ieee.org

[13] K. Muludi, K. M. Fitria, J. Triloka, and Sutedi, “Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model,” *International Journal of Advanced Computer Science and Applications*, 2024.
[Online]. Available: https://doi.org/10.14569/IJACSA.2024.01523

[14] J. Rayo, R. de la Rosa, and M. Garrido, “A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts,” *COLING Workshops*, 2025.
[Online]. Available: https://aclanthology.org

[15] G. Schumann and J. Marx Gómez, “Detection of Conflicts, Contradictions and Inconsistencies in Regulatory Documents: A Literature Review,” *International Conference on Intelligent Data Science Technologies and Applications*, 2024.
[Online]. Available: https://link.springer.com

[16] K. Sapkota, A. Aldea, M. Younas, D. Duce, and R. Bañares-Alcántara, “Extracting meaningful entities from regulatory text: Towards automating regulatory compliance,” *International Workshop on Requirements Engineering and Law*, 2012.
[Online]. Available: https://doi.org/10.1109/REL.2012.6295357

[17] B. Zhong, W. He, Z. Huang, P. Love, and J. Tang, “A building regulation question answering system: A deep learning methodology,” *Advanced Engineering Informatics*, vol. 45, 2020.
[Online]. Available: https://doi.org/10.1016/j.aei.2020.101123

[18] J. Moharana, A. Bawangade, Y. Samarth, T. Ghate, and P. Gomase, “Revolutionizing Search: Artificial Intelligence and Machine Learning’s Impact on Information Retrieval,” *International Journal For Multidisciplinary Research*, 2023.
[Online]. Available: https://www.ijmr.net.in

[19] J. Patel, A. Malhotra, A. Pande, and P. Caire, “A Survey: Information Search Time Optimization Based on RAG (Retrieval Augmented Generation) Chatbot,” *Paripex – Indian Journal of Research*, 2025.
[Online]. Available: https://www.worldwidejournals.com/paripex

[20] S. Chandar et al., “Towards Intelligent Question Answering Systems,” *arXiv preprint*, 2017.
[Online]. Available: https://arxiv.org/abs/1705.02025

[21] S. Stoeckli, C. Dremel, F. Uebernickel, and W. Brenner, “How Assistive AI Improves Knowledge Work,” *Harvard Business Review*, 2019.
[Online]. Available: https://hbr.org/2019/07/how-assistive-ai-improves-knowledge-work

[22] R. Akkiraju et al., “Enterprise-Grade Conversational AI Systems,” *IBM Research*, 2024.
[Online]. Available: https://www.ibm.com/blog/enterprise-conversational-ai

[23] C. Durach and L. Gutierrez, “ ‘Hello, this is your AI co-pilot’: Operational implications of artificial intelligence chatbots,” *International Journal of Physical Distribution & Logistics Management*, 2024.
[Online]. Available: https://doi.org/10.1108/IJPDLM-2023-0456

[24] R. Akkiraju et al., “FACTS About Building Retrieval Augmented Generation-based Chatbots,” *arXiv preprint*, 2024.
[Online]. Available: https://arxiv.org/abs/2403.01307

[25] L. Frischen and M. Fiebig, “A Perfect Start with Retrieval-Augmented Generation: Building a Chatbot to Support the Onboarding Process in SMEs,” *Studies in Health Technology and Informatics*, 2025.
[Online]. Available: https://doi.org/10.3233/SHTI250123