Article

Not peer-reviewed version

Two-Phase RAG-Based Chatbot for Italian Funding Application Assistance

Tommaso Boccato *, Matteo Ferrante *, Nicola Toschi

Posted Date: 28 June 2024

doi: 10.20944/preprints202406.1999.v1

Keywords: natural language processing; large language models; retrieval-augmented generation

Article

Two-Phase RAG-Based Chatbot for Italian Funding Application Assistance †

Tommaso Boccato [1,‡,*], Matteo Ferrante [1,‡] and and Nicola Toschi [2]

[1] Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy
[2] A.A. Martinos Center for Biomedical Imaging and Harvard Medical School, Boston, USA
* Correspondence: tommaso.boccato@uniroma2.it

† This research activity is part of "Invisible Business" (Exponential technology for invisible businesses), an initiative project by Links Management and Technology S.p.A. co-financed by Regione Puglia through the "PIA - Programmi Integrati di Agevolazione" call for proposals. Invisible Business is a research and development project aimed at experimenting with innovative technologies in the banking sector, with particular reference to the areas of credit and digital retail and corporate banking.

‡ Equal contributions–surnames in alphabetical order.

Abstract: Securing funding is a critical yet complex task for organizations and individuals. This study presents an innovative chatbot designed to streamline the process using advanced Natural Language Processing (NLP) techniques and, specifically, a Retrieval-Augmented Generation (RAG) pipeline optimized for real-world applications. Our chatbot assists users in identifying suitable public tenders for financial support through natural language queries and a comprehensive public data database. The chatbot operates in a two-stage interaction model, initially providing summarized tender information for exploratory brainstorming, followed by detailed data upon user selection. A custom filtering mechanism ensures that the user interface elements responsible for swapping the interaction stages are consistent with the responses generated by the conversational agent. Human-evaluation tests demonstrated an average accuracy of 90.4% in document retrieval, with an average of 2.11 interactions required to find a specific tender. User satisfaction, rated on a scale of 1 to 5, averaged 3.14 ($\pm$1.73), indicating generally positive user experience with room for improvement. This approach addresses challenges of relevance, accuracy, and conversational flow, resulting in a reliable chatbot that simplifies the process of finding funding opportunities.

Keywords: natural language processing; large language models; retrieval-augmented generation

1. Introduction

In the landscape of modern entrepreneurship and research, securing funding is a critical yet complex challenge. Organizations and individuals often find themselves navigating a labyrinthine process to identify suitable funding opportunities, understand intricate application guidelines, and craft compelling proposals. To address these challenges, we propose the development of an innovative chatbot designed to streamline this process using advanced Natural Language Processing (NLP) techniques.

Our project focuses on building a language-specific chatbot application aimed at assisting users in finding appropriate public tenders for financial support through natural language queries and a comprehensive database of public data. Unlike many contemporary approaches that seek technical advancements, we leverage a classical Retrieval-Augmented Generation (RAG) pipeline, optimized for real-world applications.

The primary goal of our chatbot is to emulate the consultative process typically carried out by human experts. This is achieved through a two-stage interaction model. In the first stage, the chatbot is limited to accessing previously generated summaries of tenders, which contain key information. This stage facilitates an exploratory phase where users can brainstorm and refine their search criteria, guided by the chatbot's responses based on summarized data. To ensure consistency between the graphical elements that trigger the second conversational stage, in which the end-user is allowed to query the system for more detailed information, and the content returned by the chatbot, we

 doi:10.20944/preprints202406.1999.v1

implemented a filtering mechanism. This mechanism compares the semantic relevance of retrieved documents with the chatbot's responses, displaying recommendations (buttons) that are aligned with the guidance provided by the chatbot. Once the user identifies a potentially interesting tender, they can access comprehensive information about that specific case. During this second phase, indeed, the chatbot shifts focus to detailed and specific information, minimizing the risk of presenting mixed or inaccurate data from multiple sources.

We validated our approach through human-evaluation achieving an average accuracy of 86% in document retrieval within the first two interactions with the chatbot. Additionally, user evaluations of the conversations, on a scale of 1 to 5, resulted in an average rating of 4, highlighting the effectiveness and user satisfaction with our system.

To reproduce the fluency and brainstorming process provided by a human consultant, with the idea to build a software useful for real-life scenarios to help both consultant and users, we need to use natural language as a layer of interaction with documents. Incorporating elements such as language modeling via large pretrained language models, the chatbot is adept at comprehending and responding to user queries with a high degree of effectiveness. This advanced NLP capability is crucial for interpreting user inquiries accurately and providing contextually relevant responses [1].

Further enhancing the chatbot's utility is the integration of extensive domain knowledge. The system is meticulously equipped with detailed information on a plethora of funding sources, encompassing diverse application processes and sector-specific requirements. This repository of knowledge is not static; it can continually be updated and expanded, ensuring coverage across a wide array of industries and research fields. As the system interacts with users, it progressively offers bespoke advice tailored to individual user needs.

The architecture of the system is strategically designed for accessibility and efficiency. The User Interface (UI)–Appendix A–is crafted to be intuitive and user-friendly, compatible across various platforms (i.e., desktop and mobile), thereby catering to a diverse range of users. At the core of the chatbot is a comprehensive knowledge database, which serves as the foundation, storing vital information about funding opportunities, including eligibility criteria, deadlines, and detailed application procedures. The retrieval augmented system, another critical component, analyzes user queries, correlating them with relevant information from the database to generate precise and helpful responses and recommendations through the use of a powerful Large Language Model (LLM).

1.1. Related Work

Natural Language Processing (NLP). NLP is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and human languages. It encompasses the development of algorithms and systems that enable computers to understand, interpret, and respond to human language in a valuable and intelligent way [2]. NLP combines computational linguistics–rule-based modeling of human language–with statistical, machine learning, and deep learning models. These models enable computers to process human language in the form of text or voice data and to "understand" its full meaning, complete with the speaker or writer's intent and sentiment.

Transformers. In the development of our funding application assistance chatbot, we employ LLMs, particularly those based on the transformer architecture, to model language and effectively solve the task at hand. Understanding these concepts is key to appreciating the capabilities and the innovative approach of our chatbot. The transformer architecture, introduced by Vaswani et al. [3], represents a significant shift from previous sequence-to-sequence models that relied heavily on recurrent neural networks (RNNs) [4] or convolutional neural networks (CNNs). The key innovations of transformers are their *scaled dot-product* attention mechanism, which allows models to weight the importance of different parts of the input data, and their parallelizability.

Providing domain-specific knowledge to LLMs is essential for specialized tasks, such as our project on developing a chatbot for funding application assistance. There are primarily three techniques to impart this domain-specific expertise to LLMs: Supervised Fine-Tuning (SFT) [5–9], Low-Rank

 doi:10.20944/preprints202406.1999.v1

Adaptation (LoRA) [10], and Retrieval-Augmented Generation (RAG) [11,12]. Each of these methods has unique characteristics and applications.

Supervised Fine-tuning (SFT). SFT involves retraining a pre-existing, general-purpose LLM on a dataset that is meticulously curated and rich in domain-specific content. This method equips the model with the nuanced understanding and specialized knowledge necessary to excel in specific applications, such as providing accurate, relevant assistance in the complex domain of funding applications. However, this was not an alternative in our case due to the extensive dataset size required for maintaining language elaboration capabilities.

Low-Rank Adaptation (LoRA). LoRA is an efficient technique for customizing LLMs for specific tasks or domains by integrating small, low-rank matrices into the transformer layers of a pre-trained LLM. This targeted adaptation allows the model to acquire new patterns and knowledge pertinent to the particular domain without the need to retrain the entire model. However, fine-tuning a model with LoRA was not feasible for our project due to its effectiveness being demonstrated mainly on larger problems [10].

Retrieval-Augmented Generation (RAG). RAG combines robust language generation capabilities of LLMs with dynamic information retrieval from external databases or knowledge repositories. This method significantly enhances the model's ability to provide accurate, informed, and up-to-date responses, making it particularly useful for our domain-specific application [12].

1.2. Contributions

RAG approaches are gaining popularity in applications that require accurate, up-to-date information retrieval from external knowledge bases, particularly in domain-specific chatbots. RAG offers significant advantages by combining robust language generation with dynamic information retrieval. However, several challenges arise when applying this technique to real-life problems. Our work aims to address some of the key challenges in the funding application assistance field through the following contributions:

- **Two-Phase conversation workflow**. We implement a two-phase conversation system to mimic the workflows of real expert consultants and mitigate the limitations of RAG. Due to document chunking, traditional RAG approaches risk feeding the conversational engine incomplete or misleading information[1]. This can lead to hallucinations or suboptimal matches. Our solution involves splitting the conversational flow into two phases. In the first phase, the chatbot searches for information in a set of tender summary cards, allowing the system to evaluate multiple funding initiatives comprehensively. The LLM's reasoning abilities then filter out irrelevant documents, ensuring the user receives the most accurate and relevant funding opportunities.
- **Italian speaking capabilities**. We enhance our chatbot's effectiveness by deploying state-of-the-art LLMs pre-trained on large multi-language text corpora, ensuring robust Italian language support.
- **Pertinent visual recommendations**. We incorporate an ad-hoc user interface powered by an LLM-based filter with *function-calling* capabilities. This feature triggers a second conversation phase where users can delve into the details of the most promising tenders.

By addressing these challenges, we developed a reliable and effective Italian-speaking chatbot that assists users in finding funding sources, thereby simplifying the often complex and time-consuming process of securing financial support. Figure 1 provides a visual overview of our proposal.

---

[1] This claim is correct under the assumption that knowledge base documents exceed the context length of the deployed LLM.

 doi:10.20944/preprints202406.1999.v1

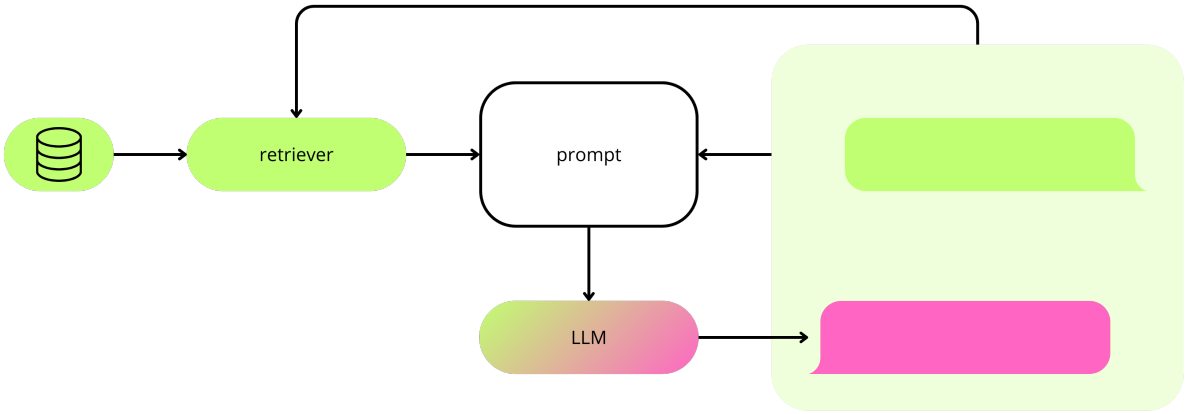

Figure 1. Overview of the chatbot architecture. The architecture consists of an LLM for text generation, an information retrieval system made of a vector database and a retriever, and a conversational interface where the end user interacts with the system.

2. Material And Methods

2.1. Chatbot Architecture

To attain the objectives outlined for the virtual assistant, we devised a bespoke Retrieval-Augmented Generation (RAG) architecture meticulously crafted to suit the specified use-case. The success of this architecture hinges on the seamless integration of various systems.

2.2. 2-Phase Conversation

The conversational flow of the chatbot has been devised drawing inspiration from the working methods of human consultants operating in the relevant sector of the considered use case (i.e., matching funding opportunities to clients), while also considering the technical characteristics of the models and systems usable in the chatbot's implementation. The two-phase conversation model is designed to mimic the typical interactions between a consultant and a client, structured to optimize the chatbot's performance and user experience.

In the first phase, the chatbot introduces itself and presents the end user with a questionnaire aimed at understanding their characteristics and objectives. The questionnaire takes the form of a web form, seamlessly integrated into the chatbot's conversational interface. Once the user's information is received, the chatbot, through its information retrieval system, retrieves the top-k summary cards of the most significant calls for the client, and based on this data, provides its suggestions. If the suggestions prove inadequate or if it is necessary to delve into general elements, the user can continue to interact with the chatbot through natural language. Alternatively, the user may choose to proceed to the second phase of the conversation, prompted by the conversational agent, and through a specific interface where a series of buttons, associated with the most relevant calls retrieved up to that specific point in the conversation, are displayed.

By clicking on a button, the client can then decide to focus the conversation on a specific call, triggering the second phase of the conversation. This process modifies the conversational agent's knowledge base by removing summary cards related to other credit initiatives and adding all documents related to the selected call. The user can then ask specific questions, and the chatbot, relying on complete information, will provide detailed, relevant, and exhaustive answers. If the client notices a misalignment between their objectives and the credit initiative under consideration, they can return to the first phase of the conversation, refining the call-client match iteratively.

2.3. Technical Motivations

The need to divide the conversational flow into two distinct phases is also motivated by the technical characteristics of the LLMs used in the chatbot architecture for text generation. The context

 doi:10.20944/preprints202406.1999.v1

window, expressed in tokens, is a crucial parameter, describing the number of basic textual units that a model can accept as input for text generation. At the time of writing, deployable LLMs in production have maximum context window sizes of 4k tokens for models executable on local machines (e.g., Llama models [13]) or 128k tokens for cloud models callable through APIs (e.g., OpenAI GPTs). Considering these characteristics, an LLM cannot fully process hundreds or thousands of different calls in a single model call, as potentially required by the use case. The use of summary cards for calls constitutes an initial mitigation of this issue. Following a RAG philosophy, only the top-k most relevant summary cards end up in the model's input prompt, further reducing the number of tokens occupied by text from the chatbot's knowledge base, and avoiding the saturation of the context window even in long conversations.

The functioning modes of information retrieval systems also influenced the design of the conversational flow. These systems perform searches for semantic similarity between a user's query and representations of texts within the knowledge base. Without constraints to documents belonging to the same call, there is a risk that the retriever will provide documents from different credit initiatives, potentially confusing the LLM or failing to retrieve all required information to answer the client's question.

2.4. Back-end Architecture

The back-end of the developed application executes all the necessary NLP tasks essential for sustaining a coherent conversation with the end-user. This system comprises three main architectural components: an LLM, an information retrieval system, and a vector database.

The LLM serves as the linchpin, generating responses for end-users by "reasoning" on an enriched query derived from processing the user's input. This prompt is constructed by merging the user's question, the historical conversation context, and pertinent information necessary for effectively addressing the user's query.

The information retrieval system contributes to the context by efficiently extracting the most significant documents from the vector database, which initially contains a pool provided by the service provider. These documents shape the context and enable the system to aptly address user queries.

2.5. Large Language Model (LLM)

The LLM is a cornerstone within the overall architecture, significantly influencing the quality of conversations facilitated by the virtual assistant. Typically, the LLM adopts a transformer-based model characterized by an extensive parameter count, often reaching up to 10^{12} parameters, and is trained on substantial data volumes. The training process typically unfolds in two phases: initially in a self-supervised manner, followed by fine-tuning for specific downstream tasks such as state-less chat-like conversations.

The LLM generates text in accordance with a designated input prompt, a formatted string where system-level instructions, the user's query, the chat history, and contextual information are denoted by specific tags. The exact prompt format is crucial since it ensures the same configuration used in model training (Appendix B for the English version):

```
Sei un assistente bancario utile e rispettoso. Rispondi sempre nel modo più utile possibile.

Le tue risposte non dovrebbero includere contenuti dannosi, non etici, razzisti, sessisti, tossici,
    pericolosi o illegali. Assicurati che le tue risposte siano socialmente imparziali e positive.

Se una domanda non ha senso o non è coerente dal punto di vista fattuale, spiega il motivo invece di
    rispondere con informazioni non corrette. Se non conosci la risposta a una domanda, per favore
    evita di condividere informazioni false.

Il tuo obiettivo è trovare la migliore risposta possibile per il cliente.

Di seguito una conversazione di esempio, da cui puoi prendere spunto per rispondere alle domande e
    chiedere informazioni aggiuntive.
```

 doi:10.20944/preprints202406.1999.v1


```
Fai questa domande al cliente per raccogliere informazioni.

1) Quale è la dimensione della tua impresa?

Esempi di risposta:
- Start Up Innovativa
- Piccola
- Media
- Grande

2) Qual è il tuo Codice ATECO?

2a) Qual è il settore in cui vuoi investire ?

Esempi di risposta:
agricoltura
sanità
aerospazio
turismo e beni culturali

3) Quali tipologie di attività  vorresti finanziare? / Quali sono gli obiettivi principali della
     richiesta di finanziamento?

Esempi di risposta:
ricerca e sviluppo
formazione
innovazione dei processi e dell'organizzazione
attivi materiali:
immobili
immobilizzazioni materiali (es. HW)
immobilizzazioni immateriali (es. SW)

4) Hai una sede operativa (o intendi attivarla) in una delle seguenti regioni ? (es. Regioni di
     convergenza: Puglia, Calabria, Sicilia, Campania, ecc) (Localizzazione dell'investimento)

5) Qual è il volume dell'investimento? (Facoltativo)

6) Quale è la tipologia di contributo desiderato? Forma dell'agevolazione (Facoltativo)

Esempi di risposta:
- a fondo perduto
- finanziamento agevolato

Rispondi esclusivamente in Italiano. il testo in uscita dovrà  essere formattato in html per rispettare
     una giusta gestione dei paragrafi.

Qualora non trovassi bandi pertinenti, chiedi maggiorni informazioni all'utente senza arrenderti subito,
      indicando di essere più specifico e dare maggiori dettagli, in modo da trovare la soluzione
     migliore per il cliente.
```

In our architecture, the LLM block can be implemented using a pretrained local model, such as LLama2, or a cloud service like the OpenAI APIs, levaraging GPT4. Deploying a local model requires a high-performance computing platform and various software optimizations, such as parameter quantization and computational graph compilation. This approach provides direct control over system prompts, enabling fine-tuning of the model's behavior and mitigating concerns related to potential changes or discontinuation of external APIs. Conversely, opting for public APIs could enhance conversation quality but comes with reduced control over underlying models and usage fees.

The final choice of the LLM used in the chatbot architecture (OpenAI GPT-4) was based on qualitative and quantitative evaluations. We relied on literature reviews comparing the performance of major LLMs across a wide range of academic benchmarks. GPT-4 emerged as the state-of-the-art

 doi:10.20944/preprints202406.1999.v1

model, surpassing competitors on most datasets [14]. Additionally, GPT-4's suitability for the Italian language, crucial for our use case, was validated by its performance on the multi-language MMLU benchmark, showing minimal accuracy drop compared to English [15].

2.6. Vector Database

A vector database stores data in the form of high-dimensional vectors, which represent features or attributes. In our application, these vectors can represent document summaries or non-overlapping document chunks. FAISS (Facebook AI Similarity Search) was chosen for its open-source nature, efficiency, and comprehensive documentation. FAISS's search-time/recall graph shows superior performance compared to other solutions, making it ideal for our needs [16].

2.7. Summarizer

To acquire document summaries, we implemented a system leveraging GPT-4. The LLM is prompted with the content of the document for summarization, accompanied by custom instructions guiding the model to retain the most important points relevant to the funding criteria (Appendix B for the English version):

```
Genera una sintesi del testo fornito in calce, restituendo inoltre le seguenti informazioni obbligatorie.
     Il riassunto sarà  utilizzato in un motore di ricerca semantico pertanto dovrà  essere ricco di
    informazioni cruciali e semanticamente importanti.

INFORMAZIONI ESSENZIALI:

Titolo del bando

Finalità  del bando

Inserire, se disponibili, gli interventi ammissibili ed obiettivi che le aziende intendono perseguire
    aderendo al bando

I dettagli delle agevolazioni previste dal bando, dando importanza, se presente, all'importo minimo di
    investimento

Inserire l'indirizzo web pubblico del sito istituzionale del bando, se noto

Data apertura e data chiusura bando

Se questo è un Bando Nazionale/Bando regionale (selezionare regione)/ Bando camerale (selezionare una
    provincia)

Area geografica Provincia + Regione (la regione puoi dedurla anche se non presente ma è importante che
    siano presenti)

Codici ATECO, sii molto preciso qui, riporta esattameente tutti i codici ammessi ed i codici esclusi,
    fai un elenco dei codici numerici

Spese finanziate

Se il bando è valido fino ad esaurimento fondi

Tipo di concessione bonus-fiscale, contributo-a-fondo-perduto, finanziamento-a-tasso-agevolato,
    finanziamento-a-tasso-zero, garanzia

tipo di spesa: finanziata affitto-locali, assunzione, attrezzature-e-macchinari, avvio-attivita,
    certificazioni, consulenze, digitalizzazione, formazione, hardware-software, innovazione-ricerca-e-
    sviluppo, marchi-brevetti-disegni, opere-edili-e-impianti,
 patrimonializzazione, pubblicita-promozione-marketing, risparmio-energetico, servizi, sostegno-alla-
     liquidita, spese-generali
```

 doi:10.20944/preprints202406.1999.v1


```
beneficiari: associazione-consorzio, associazione-ente-non-profit-terzo-settore-impresa-sociale, centro-
    di-ricerca, cooperativa, ente-di-formazione, ente-pubblico, grande-impresa, libero-professionista,
    micro-impresa, persona-fisica-aspirante-imprenditore,
 pmi, startup, startup-innovative

È fondamentale compilare tutti i campi, tuttavia non è ammesso invetare o inserire contenuti di fantasia
    , in caso di informazioni assenti semplicemente non riportarle, sii il più completo possibile nella
     compilazione di questo riassunto, non importa se lungo.

Non utilizzare markdown e formatta tutto come un paragrafo non come un documento strutturato, la sintesi
     deve cominciare con il nome del bando.
```

Summaries were generated once per document, ensuring cost-effectiveness. Figure 2 shows a scheme of our summarization procedure to reduce complexity for the first phase.

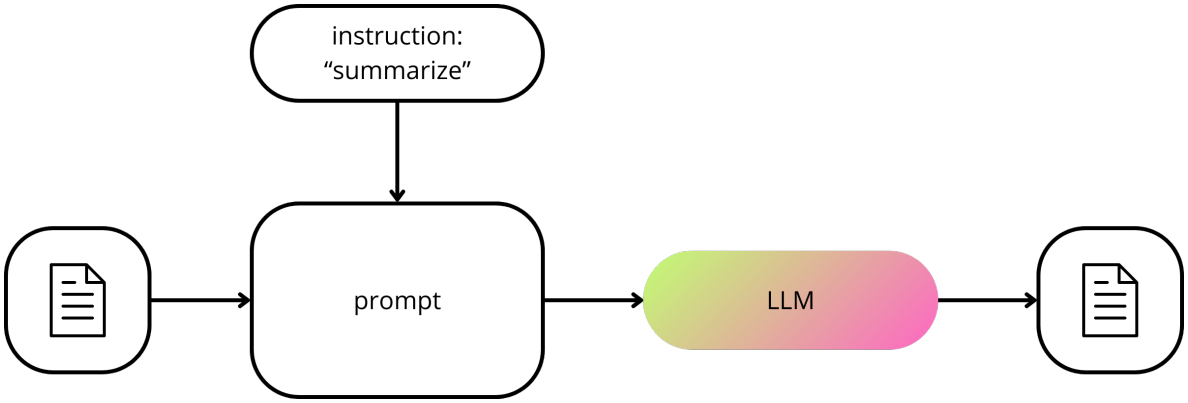

Figure 2. A scheme of the summarizer used to reduce complexity in the first phase of conversations

In cases where tenders exceeded the LLM's context window, summaries were generated hierarchically. High-dimensional vector representations (1536-dimensional) of the summaries were created using the OpenAI `text-embedding-3-small` model and inserted into the vector database.

2.8. Knowledge Base Structure

The knowledge base used in phase one of the conversation consists of summary sheets generated from tenders. For phase two, the knowledge base includes the summary sheet and all associated documents of a single subsidized credit initiative, divided into manageable chunks if necessary. All texts are processed into embeddings and stored in the vector database. A total of 22 publicly available tenders, in Italian, were chosen as a benchmark to assess quality of this solution.

2.9. Information Retrieval System

The vector database enables rapid and precise similarity search based on vector distance or similarity. The retriever, implemented using FAISS, operates by taking a representation of the chat history and user query, and performing a similarity search within the vector database. The search results, ranked by relevance, form the context for the LLM to generate responses.

The process involves preparing the input query, performing the similarity search, and filtering the results. The filtered contents are concatenated and inserted into the LLM's input prompt for generating responses.

2.10. Challenges and Solutions

While the RAG approach offers significant advantages, several challenges can arise when applying this technique to real-life problems. Ensuring the relevance and accuracy of the retrieved information is critical. This problem has been solved by using the most capable LLMs available.

 doi:10.20944/preprints202406.1999.v1

Another challenge is the seamless integration of multilingual capabilities, particularly for creating a reliable Italian-speaking chatbot. Our solution involves employing advanced NLP techniques and large pretrained language models specifically fine-tuned for the Italian language, ensuring natural and coherent responses.

Furthermore, the dynamic nature of funding opportunities requires continuous updates to the chatbot's knowledge base. We integrated mechanisms that support easy updates and expansions, ensuring comprehensive coverage across various industries and research fields.

By addressing these challenges, our approach aims to create a reliable and effective Italian-speaking chatbot that assists users in finding funding sources, thereby simplifying the often complex and time-consuming process of securing financial support.

Figure 3 shows a scheme of the inner working of our augmented RAG procedure.

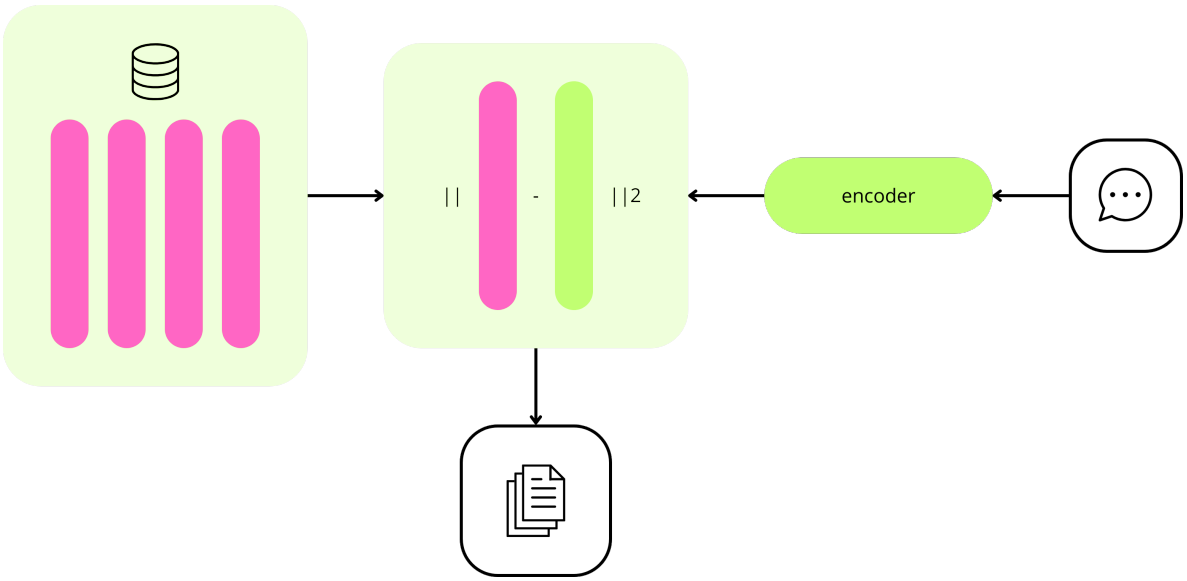

Figure 3. The information retrieval system, consisting of a vector store, an encoder for queries, and a similarity search algorithm.

2.11. Evaluation

Objectives. The evaluation of the chatbot aims to assess the performance of the developed system according to three different metrics, both objective and subjective:

1. **Accuracy** in identifying the tender that best matches the characteristics of the end user (objective).
2. **Average perceived quality of each response**, based on a scale containing five different levels of satisfaction (subjective).
3. **Average perceived quality of an entire conversation**, based on a scale containing five different levels of satisfaction (subjective).

Protocol. To obtain results for the aforementioned metrics, the following evaluation protocol has been implemented:

1. Each evaluator will briefly study the summary sheet (about a paragraph of text) of the assigned tender.
2. The evaluator will then start a conversation with the chatbot, simulating the behavior of a user whose characteristics exactly match the requirements of the tender in question. Specifically, the evaluator will correctly and as completely as possible fill out the form shown by the web application's interface.
3. The chatbot will generate an initial response, which will be evaluated through the specific interface (5 emojis) provided by the application. All subsequent responses will also be evaluated individually in the same manner.

 doi:10.20944/preprints202406.1999.v1

4. The evaluator will continue to interact with the chatbot for up to 3 iterations, until the following condition occurs: the chatbot identifies the optimal tender, presents it in the response, and displays it in the follow-up buttons. When this condition is met, the evaluator notes the successful match, along with the number of iterations performed (including the form – thus +1), in a designated spreadsheet and proceeds by pressing the button associated with the identified tender. If the condition is not met after the maximum number of iterations, the evaluator notes the failure of the tender-client match in the spreadsheet, and the conversation is considered ended.
5. If entering a possible second phase of the conversation, the evaluator continues to interact with the chatbot for up to 3 iterations, asking more detailed questions about the tender in question (e.g., application requirements, support in submitting the funding application).
6. At the end of the conversation, the evaluator, using the interface (5 emojis) within the sidebar of the application, provides an overall satisfaction rating, possibly attaching a summary comment on the service experience. The correct saving of the evaluation is done by clicking the "Submit" button, which triggers a visual confirmation feedback.

This structured protocol ensures comprehensive assessment and facilitates the collection of consistent and reliable data for evaluating the chatbot's performance.

3. Results

The evaluation of our chatbot was conducted to assess its performance in accurately identifying relevant tenders, and the quality of interactions. The evaluation focused on two primary aspects: the quality of individual responses and the overall quality of the conversation. The evaluation was conducted by a group of internal and external people on a total of 52 tenders. The average satisfaction level was 3.14 ± 1.73 and the number of iterations required to find a specific tender was on average 2.11, suggesting that the augmented RAG pipeline is able to quickly guide users to the right documents. The accuracy measured for this retrieval task was 90.4%[2]. Results are summarized in Figures 4 and 5.

Figure 4 presents the distribution of user satisfaction scores for entire conversations. The scores are distributed across five levels of satisfaction, represented by emojis ranging from "very dissatisfied" (left) to "very satisfied" (right).

As shown in Figure 4, the satisfaction levels of the chatbot reveal a bimodal distribution, with peaks at both very high and low ratings, likely reflecting two distinct profiles among the evaluators. Approximately 50% of users rated their conversations with the chatbot using either the highest or second-highest satisfaction emoji. The lower satisfaction levels also had relatively high values, which could be attributed to varying user profiles and expectations that emerged during subsequent discussion phases. Beyond the protocol, some users, more experienced with funding applications, found it extremely important to locate very subtle details rather than just the right documents. These users impersonated highly experienced individuals looking for crucial information. The absence of such detailed information negatively impacted the overall evaluation of the model. Besides that, these results indicate that the chatbot was generally effective in maintaining user satisfaction throughout the interactions.

Figure 5, instead, illustrates the distribution of user satisfaction scores for individual responses. As in the previous evaluation, scores are distributed across five levels of satisfaction. Results show that the chatbot's individual responses were well-received by users. The highest satisfaction level was the most frequent, representing about 50% of the responses. Explanations are similar to the previous case.

The evaluation results demonstrate that the chatbot performed well in both identifying relevant tenders and providing satisfactory interactions with users. The high levels of user satisfaction in both individual responses and entire conversations highlight the effectiveness of the chatbot in fulfilling its

[2] The last two metrics were computed on conversations for which the required data was available

 doi:10.20944/preprints202406.1999.v1

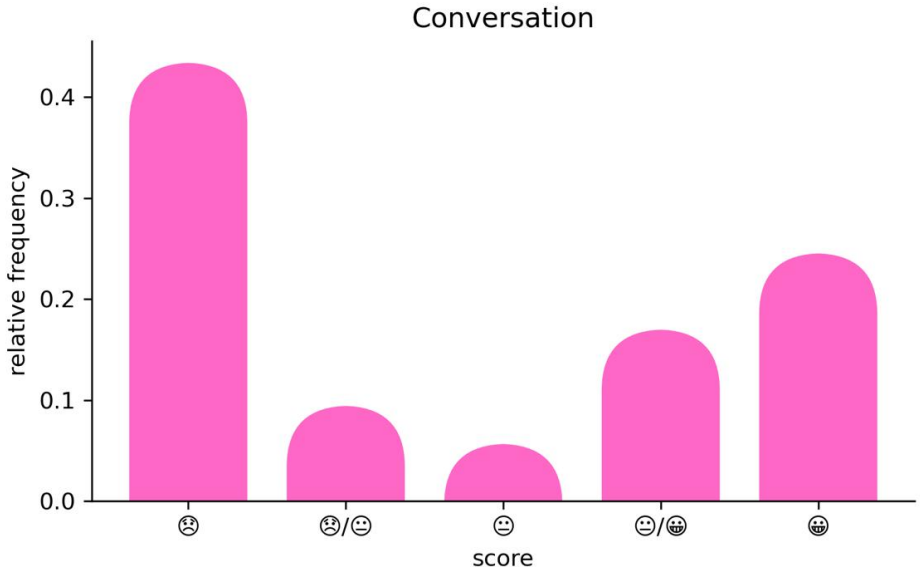

Figure 4. User satisfaction distribution for entire conversations.

design purpose. Additionally, the dissatisfaction scores indicate that any issues encountered by users could be attributed to a mismatch between what the users are specifically looking for and the current phase of the chatbot (e.g., users seeking fine-grained details about funding applications during the first phase, when the chatbot can only access high-level information). These issues could potentially be resolved with a better summarizer.

Overall, results support the conclusion that our chatbot is a reliable and effective tool for assisting users in finding suitable funding opportunities. Future improvements can focus on addressing minor dissatisfaction causes to further enhance user experience.

4. Discussion

The results of our evaluation demonstrate that the developed chatbot effectively assists users in identifying relevant funding opportunities. High levels of user satisfaction, both in terms of individual responses and overall conversation quality, suggest that the chatbot successfully fulfills its intended purpose.

One notable aspect of the evaluation is the preference for higher satisfaction ratings, indicating that the chatbot's responses were not only accurate but also perceived as helpful and relevant by users. This is particularly significant given the complexity of matching user characteristics with appropriate tenders. The chatbot's ability to consistently deliver high-quality responses is a testament to the robustness of the underlying RAG architecture and the meticulous design of the conversational flow.

Despite the positive feedback, there were instances of lower satisfaction scores. These outliers point to potential areas for improvement. For example, the relatively small proportion of users who rated their experience as less satisfactory might have encountered issues related to the specificity of the tenders or the clarity of the chatbot's explanations. Addressing these concerns could involve refining the summarization and retrieval processes to ensure even greater relevance and clarity in the responses provided.

The implementation of a two-phase conversation model has proven effective in managing the context window limitations of the LLM and enhancing the relevance of the information presented to users. This approach not only mirrors the consultative process of human experts but also optimizes the chatbot's performance within the constraints of current technology. The positive feedback from users reinforces the validity of this design choice.

 doi:10.20944/preprints202406.1999.v1

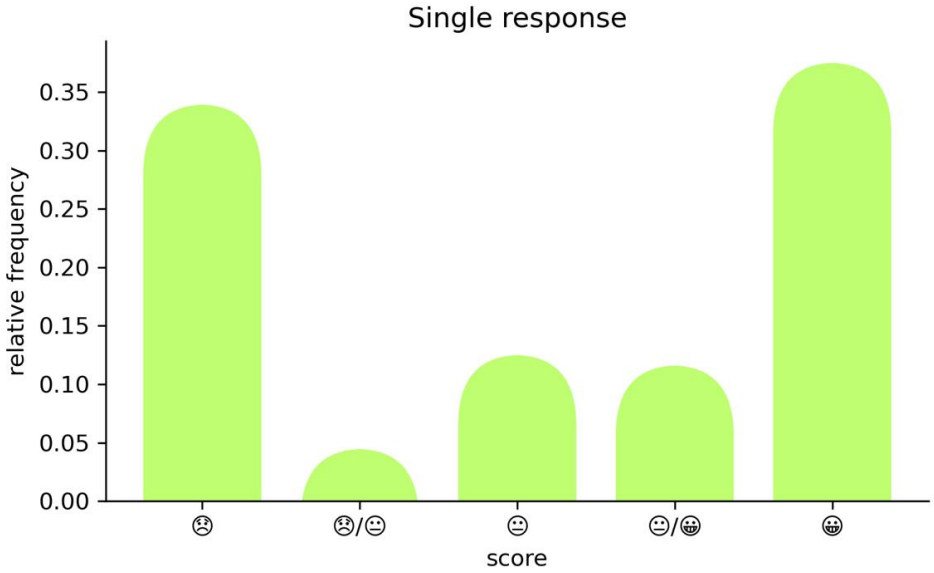

Figure 5. User satisfaction distribution for individual responses.

In terms of future work, several directions can be pursued to further improve the chatbot. New filtering mechanisms could ensure even higher precision in the relevance of retrieved documents. Additionally, expanding the chatbot's capabilities to handle more complex queries and provide more detailed guidance on application procedures could further enhance user satisfaction.

5. Conclusion

In conclusion, the evaluation results confirm that our chatbot is a reliable and effective tool for assisting users in identifying suitable funding opportunities in Italian. The high levels of user satisfaction, both in individual responses and overall conversations, indicate that the chatbot successfully meets the needs of its users. The two-phase conversation model, coupled with a robust RAG architecture, has proven to be an effective strategy for managing context limitations and delivering relevant information. While the overall performance of the chatbot is commendable, there remains room for improvement, particularly in addressing the minor instances of dissatisfaction. Future developments should focus on refining the summarization and retrieval processes and expanding the chatbot's capabilities to handle more complex queries and provide more detailed guidance.

 doi:10.20944/preprints202406.1999.v1

Appendix A Chatbot's User Interface

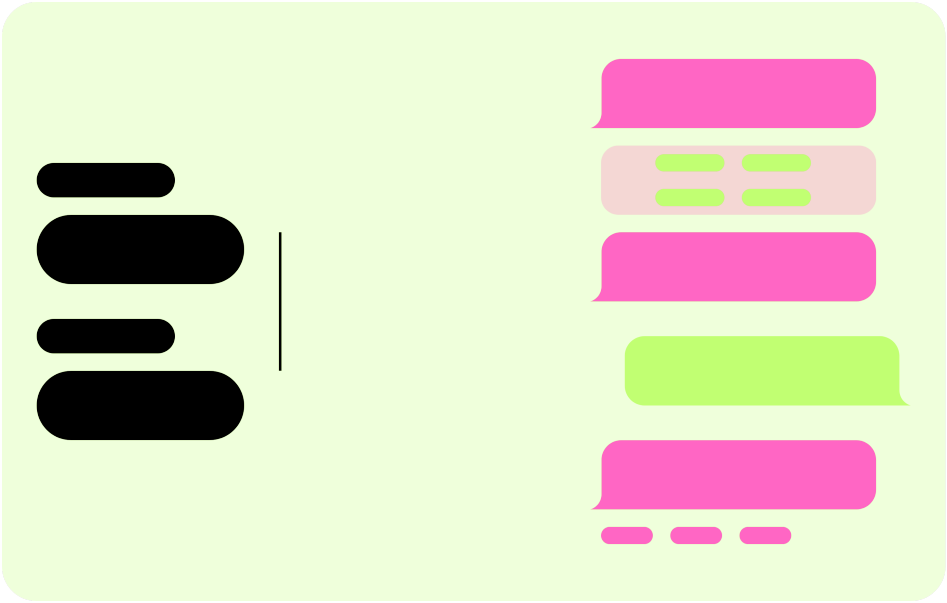

Figure A1. Mock-up of the developed UI. The chatbot's messages are displayed in pink while the user's ones are green. The sidebar is located on the left.

Figure A1 shows a mock-up of the developed UI. Initially, the UI displays a welcome message where the virtual assistant introduces itself. Next, the chatbot presents the user with a form to disclose information about their company and financial objectives in the context of funding applications. At this point, the system begins to guide the user toward the most appropriate funding opportunity. Additionally, at the end of each chatbot response, the system provides recommendations in the form of buttons, allowing the user to continue the conversation focusing on the details of a specific tender. Finally, the UI includes a sidebar where the user can find a list of documents used by the system to generate responses.

Appendix B Prompts' Translations

We report below the translations of the used "system" and "summarizer" prompts:

```
You are a helpful and respectful banking assistant. Always respond in the most helpful way possible.

Your responses should not include harmful, unethical, racist, sexist, toxic, dangerous, or illegal
    content. Ensure that your responses are socially unbiased and positive.

If a question does not make sense or is factually incoherent, explain why instead of responding with
    incorrect information. If you do not know the answer to a question, please avoid sharing false
    information.

Your goal is to find the best possible answer for the customer.

Below is a sample conversation from which you can take inspiration to answer questions and ask for
    additional information.

Ask these questions to the customer to gather information.

1) What is the size of your business?

Examples of answers:
- Innovative Start-Up
- Small
```

 doi:10.20944/preprints202406.1999.v1


```
- Medium
- Large

2) What is your ATECO Code?

2a) What is the sector you want to invest in?

Examples of answers:
agriculture
healthcare
aerospace
tourism and cultural heritage

3) What types of activities would you like to finance? / What are the main objectives of the funding
     request?

Examples of answers:
research and development
training
process and organizational innovation
tangible assets:
real estate
tangible fixed assets (e.g., HW)
intangible fixed assets (e.g., SW)

4) Do you have an operational headquarters (or intend to activate one) in one of the following regions?
     (e.g., Convergence regions: Puglia, Calabria, Sicily, Campania, etc.) (Location of the investment)

5) What is the investment volume? (Optional)

6) What type of contribution is desired? Form of facilitation (Optional)

Examples of answers:
- non-repayable grant
- subsidized loan

Respond exclusively in Italian. The output text should be formatted in HTML to ensure proper paragraph
     management.

If you do not find relevant calls for proposals, ask the user for more information without giving up
     immediately, indicating to be more specific and give more details, in order to find the best
     solution for the customer.
```

```
Generate a summary of the text provided below, also returning the following mandatory information. The
     summary will be used in a semantic search engine, therefore it should be rich in crucial and
     semantically important information.

ESSENTIAL INFORMATION:

Title of the call for proposals

Purpose of the call for proposals

Include, if available, the eligible interventions and objectives that companies intend to pursue by
     adhering to the call for proposals

Details of the benefits provided by the call for proposals, emphasizing, if present, the minimum
     investment amount

Include the public web address of the institutional site of the call for proposals, if known

Opening date and closing date of the call for proposals
```

 doi:10.20944/preprints202406.1999.v1


```
Indicate if this is a National Call/Regional Call (select region)/Chamber Call (select a province)

Geographical area Province + Region (you can infer the region if not present but it is important that
    both are present)

ATECO codes, be very precise here, report exactly all the allowed codes and excluded codes, list the
    numerical codes

Funded expenses

If the call for proposals is valid until funds are exhausted

Type of concession tax-bonus, non-repayable-grant, subsidized-rate-financing, zero-rate-financing,
    guarantee

type of expense: financed premises-rent, hiring, equipment-and-machinery, business-start-up,
    certifications, consulting, digitization, training, hardware-software, innovation-research-and-
    development, trademarks-patents-designs, building-works-and-systems,
 capitalization, advertising-promotion-marketing, energy-saving, services, liquidity-support, general-
     expenses

beneficiaries: association-consortium, association-non-profit-third-sector-social-enterprise, research-
    center, cooperative, training-entity, public-entity, large-enterprise, freelancer, micro-enterprise,
     individual-aspiring-entrepreneur, SMEs, startup, innovative-startups

It is essential to fill in all the fields, however it is not allowed to invent or insert imaginary
    content, if information is missing simply do not report it, be as complete as possible in filling
    out this summary, it does not matter if it is long.

Do not use markdown and format everything as a paragraph not as a structured document, the summary must
    begin with the name of the call for proposals.
```

References

1. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *arXiv* **2024**.
2. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*; 2024.
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**.
4. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Physica D: Nonlinear Phenomena* **2020**, *404*, 132306.
5. Li, Z.; Li, X.; Liu, Y.; Xie, H.; Li, J.; lee Wang, F.; Li, Q.; Zhong, X. Label Supervised LLaMA Finetuning. *arXiv* **2023**.
6. Lv, K.; Yang, Y.; Liu, T.; Gao, Q.; Guo, Q.; Qiu, X. Full Parameter Fine-tuning for Large Language Models with Limited Resources. *arXiv* **2023**.
7. Tian, K.; Mitchell, E.; Yao, H.; Manning, C.D.; Finn, C. Fine-tuning Language Models for Factuality. *arXiv* **2023**.
8. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; et al., P.C. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**.
9. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *arXiv* **2020**.
10. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**.
11. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; et al., Y.D. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**.
12. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; et al., H.K. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2021**.

 doi:10.20944/preprints202406.1999.v1

13. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; et al., N.B. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**.
14. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; et al., H.H. Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. *Meta-Radiology* **2023**, *1*, 100017.
15. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; et al., F.L.A. GPT-4 Technical Report. *arXiv* **2024**.
16. Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.E.; Lomeli, M.; Hosseini, L.; Jégou, H. The Faiss library. *arXiv* **2024**.