

Improving Preliminary Diagnosis in Italian Through a Retrieval-Augmented Medical Chatbot

Mariano Barone^{1,2}, Gian Marco Orlando^{1,2}, Marco Perillo^{1,2}, Giuseppe Riccio^{1,2,*}, Antonio Romano^{1,2}, Diego Russo^{1,2,4}, Ferdinando Tammaro^{1,2}, Valerio La Gatta³, Marco Postiglione³ and Vincenzo Moscato^{1,2}

¹*University of Naples Federico II, Department of Electrical Engineering and Information Technology (DIETI), Via Claudio, 21 - 80125 - Naples, Italy*

²*Consorzio Interuniversitario Nazionale per l'Informatica (CINI) - ITEM National Lab, Complesso Universitario Monte S. Angelo, Naples, Italy*

³*Northwestern University, Department of Computer Science, McCormick School of Engineering and Applied Science, 2233 Tech Dr, Evanston, IL 60208, United States*

⁴*University of Bergamo, Department of Management, Information and Production Engineering, Via Pasubio 7b, Dalmine (BG), 24044, Italy*

Abstract

The integration of Big Data and Artificial Intelligence in healthcare offers significant potential to address the growing need for supporting clinical medical consultation systems. However, many existing solutions struggle to effectively utilize unstructured medical data and provide contextually relevant responses to user queries. This paper addresses this gap by presenting the architecture of an AI-driven medical chatbot based on Retrieval-Augmented Generation. The system leverages data from Italian medical forums and encyclopedias to offer preliminary diagnoses and treatment suggestions. Our approach integrates a retrieval mechanism with a large language model, enhanced by query expansion techniques for improving retrieval accuracy and reranking methods to prioritize the most relevant information. The effectiveness of the system is demonstrated through both qualitative and quantitative evaluations, showcasing improvements in user experience and response precision. We publicly release our code on GitHub: <https://github.com/PRAISELab-PicusLab/RAGMedicalChatbot>.

Keywords

NLP for Healthcare, Medical Assistant, Clinical Decision Support System, Big Data in Healthcare

1. Introduction

The advent of artificial intelligence (AI) and Big Data has profoundly transformed the healthcare landscape, offering new ways to process and analyze large amounts of unstructured data from medical forums, research papers, and medical records. However, current AI systems, in particular large language models (LLMs), have significant limitations, including the risk of generating hallucinated or inaccurate information, which undermines their reliability, or the problems of interoperability and transparency that are crucial in critical sectors such as healthcare [1].

This paper introduces a medical chatbot system that leverages retrieval-augmented generation (RAG) techniques to address these shortcomings. By integrating natural language processing (NLP) with RAG, the chatbot generates personalized responses based on user-reported symptoms while retrieving

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

*Corresponding author.

 mariano.barone@unina.it (M. Barone); gianmarco.orlando@unina.it (G. M. Orlando); marc.perillo@studenti.unina.it (M. Perillo); giuseppe.riccio3@unina.it (G. Riccio); antonio.romano5@unina.it (A. Romano); diego.russo@unibg.it (D. Russo); fer.tammaro@studenti.unina.it (F. Tammaro); valerio.lagatta@northwestern.edu (V. L. Gatta); marco.postiglione@northwestern.edu (M. Postiglione); vincenzo.moscato@unina.it (V. Moscato)

 <https://github.com/csmariano> (M. Barone); <https://github.com/giammy677dev> (G. M. Orlando); <https://github.com/giuseppericcio> (G. Riccio); <https://github.com/LaErre9> (A. Romano); <https://github.com/diegorusso95> (D. Russo); <http://wpage.unina.it/vmoscato/> (V. Moscato)

 0009-0004-0744-2386 (M. Barone); 0009-0004-7136-1804 (G. M. Orlando); 0009-0009-0649-4068 (M. Perillo); 0009-0002-8613-1126 (G. Riccio); 0009-0000-5377-5051 (A. Romano); 0009-0007-1095-5168 (D. Russo); 0009-0003-6989-8365 (F. Tammaro); 0000-0002-5941-4684 (V. L. Gatta); 0000-0001-6092-940X (M. Postiglione); 0000-0002-0754-7696 (V. Moscato)

 © 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

relevant, verified medical data from external sources. This dual approach not only enhances response accuracy but also grounds the system’s outputs in factual content, significantly reducing the risk of misinformation [2]. In particular, the system incorporates trusted Italian medical knowledge bases such as MedicItalia¹ and the Humanitas² encyclopedia. Despite the promising results of RAG-based solutions in various domains [3], their use in the Italian medical context remains underexplored. Moreover, most existing systems fail to integrate domain-specific retrieval with language generation while maintaining traceability and source citation, especially in non-English clinical data. By addressing these gaps, the proposed chatbot not only delivers preliminary diagnostic suggestions but also provides explicit references to authoritative medical sources, thereby enhancing both transparency and user confidence.

The system offers practical benefits for both patients and healthcare professionals. Patients gain immediate access to reliable, tailored medical insights, while doctors can use the chatbot as a preliminary assessment tool to streamline decision-making process [4]. By reducing diagnostic errors and improving healthcare efficiency, this approach demonstrates the potential of AI to personalize healthcare delivery and support informed medical decisions.

2. Related Work

Large Language Models (LLMs) have shown significant versatility across various domains, including legal document retrieval [5], knowledge graph construction from geopolitical corpora [6], and agent-based simulations with generative reasoning capabilities [7]. Recent advancements have extended LLM-based generative agents to applications in social simulation and decision support, such as modeling the Friendship Paradox in online social networks [8], detecting insider threats through agent coordination [9], and supporting fact-checking via diverse and structured agent collectives [10].

In the biomedical domain, LLM integration has yielded systems like PIE-Med [11], which combines generative reasoning with graph inference to deliver explainable and interpretable medical recommendations. Other approaches have leveraged medical entity recognition and generative summarization to improve the accessibility and structure of clinical records [12]. Complementary to LLM-based solutions, diversity-aware recommender mechanisms have been proposed to mitigate over-specialization and promote exploratory recommendations in health-related contexts [13].

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm for improving factual grounding in healthcare NLP systems by integrating external retrieval with generative models. However, challenges related to domain specificity, hallucination, and contextual accuracy remain prominent. General limitations of RAG—including grounding instability—are analyzed in [14], while the hallucination risks of healthcare-focused chatbots are specifically examined in [15]. Addressing these issues, the present study focuses on Italian clinical discourse, incorporating domain-adapted retrieval and semantic query expansion to improve factual accuracy and linguistic appropriateness.

Several works have explored architectural strategies to overcome RAG limitations. Clinical data extraction from electronic health records (EHRs) has been attempted using RAG pipelines, though limited domain integration remains a bottleneck [16]. Reranking-based refinement methods have been introduced to enhance answer precision [17], though performance in open-ended clinical settings continues to face constraints. Contextual retrieval mechanisms have also been explored to narrow the relevance gap, yet scalability remains a concern [18].

Multilingual adaptability constitutes a critical research axis in medical NLP. Results from [19] demonstrate that language-specific fine-tuning and data augmentation significantly enhance disorder identification performance in non-English clinical corpora. These findings support the relevance of localization strategies, as adopted in this work.

In summary, although prior research has advanced the integration of LLMs and RAG in clinical applications, key challenges persist around scalability, reliability, and language adaptation. This study addresses these limitations by combining semantic query expansion, source-linked summarization, and

¹<https://www.medicalitalia.it/>

²<https://www.humanitas.it/>

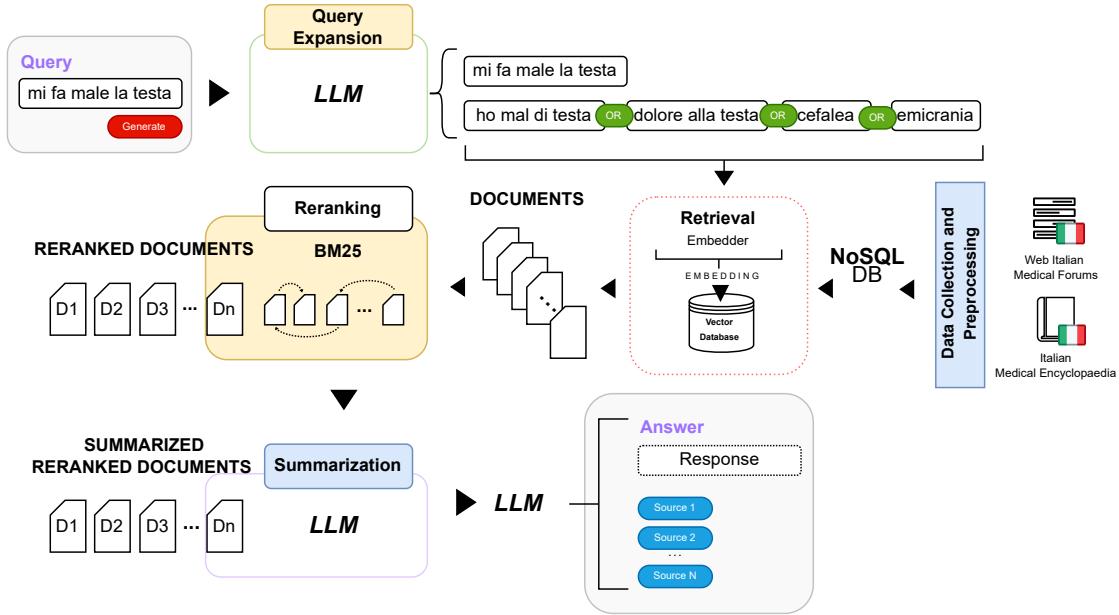


Figure 1: The workflow outlines the key stages of the Italian Medical Assistant. In the **Query Expansion** phase, the user’s initial question is expanded with synonyms and related terms to improve the retrieval of relevant documents. During **Retrieval**, documents are initially collected and preprocessed from medical forums and medical encyclopedias, then selected and compared with the query using embedding models. In the **Reranking** phase, documents are reordered based on relevance using the **BM25** algorithm. In **Summarization**, the selected documents are synthesized to produce concise responses. Finally, in the **answer**, the summaries are integrated to produce a comprehensive response, accompanied by source citations to ensure accuracy and reliability.

reranking to support trustworthy and context-aware response generation in Italian-language clinical scenarios.

3. Methodology

The medical chatbot developed in this project utilizes RAG and consists of several key components: starting from data collection of Italian medical forums and Italian encyclopedia, to the generation of responses based on contextually relevant information retrieved from a knowledge base. As shown in Figure 1 the system architecture consists of five main phases: (1) query expansion, (2) document retrieval, (3) reranking, (4) summarization, and (5) response generation. The Figure shows how the user query is first expanded with synonyms to improve recall. Relevant documents are retrieved from a vector database using cosine similarity and reranked via the BM25 algorithm. These documents are then summarized and used to construct the final diagnostic prompt, ensuring contextual relevance and citation transparency.

3.1. Data Collection and Preprocessing

The system collects data from Italian medical forums such as MedicItalia, where users post health-related questions and licensed doctors provide responses, as well as from the Italian certified medical encyclopedia Humanitas. This unstructured data was collected via scraping for research purposes only. To ensure transparency and traceability, each extracted Question/Answer pair includes its original source URL. The collected data is systematically categorized and transformed into a chatbot-compatible format using an ETL (Extract, Transform, Load) pipeline, designed for scalable management of large text corpora typical of Big Data environments. Subsequently, forum interactions are structured into a Question/Answer format. Long questions are segmented into smaller units to enhance processing

efficiency. This organization enables the creation of embeddings, which are essential for the retrieval of relevant content and the generation of accurate responses. Finally, document embeddings are generated and stored in a vector database to optimize real-time retrieval during response generation.

3.2. Retrieval-Augmented Generation (RAG)

The RAG technique integrates document retrieval with language generation. User queries are first transformed into embeddings. To optimize the retrieval process, we evaluated several embedding strategies, such as embedding the concatenated string of both the question and the answer (the entire forum post) versus embedding only the user's question.

We ultimately chose to embed only the question to minimize context loss, which can occur when using concatenated strings that may be too long, include technical jargon from doctors' responses, or contain irrelevant information in the answers. The retrieval phase leverages cosine similarity to compare the user query embedding with pre-stored document embeddings in the vector database, returning the top five relevant documents for optimal system performance.

3.3. Query Expansion and Reranking

In order to enhance the retrieval process, the user's query undergoes an expansion process using a generative language model that introduces synonyms and contextually related terms. During this Query Expansion step, the model analyzes the query to identify key concepts and semantic relationships, generating alternative phrases that improve the chances of capturing relevant documents. For example, if the user queries "symptoms of diabetes", the model may expand it to include terms like "signs of diabetes", "glucose levels", or "insulin resistance". This step significantly increases the likelihood of retrieving accurate medical responses. Once the relevant documents are retrieved, a reranking step is applied to order the top 15 documents based on the BM25 scoring algorithm. BM25 computes a relevance score (1) for each document by considering factors such as term frequency (TF), inverse document frequency (IDF), and document length normalization.

$$\text{score}(D, Q) = \sum_{n=1}^N \text{IDF}(q_n) \cdot \frac{f(q_n, D) \cdot (k_1 + 1)}{f(q_n, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \quad (1)$$

Where query Q contains the terms q_1, q_2, \dots, q_n ; document D : the document to be evaluated; frequency: $f(q_n, D)$ represents the frequency of the term q_n within document D ; constant parameters: k_1 and b are parameters that control the model's behavior; average document length: avgdl is the average length of documents in the collection. The Inverse Document Frequency (IDF) is calculated as:

$$\text{IDF}(q_n) = \ln \left(\frac{N - n(q_n) + 0.5}{n(q_n) + 0.5} \right) + 1 \quad (2)$$

where N is the total number of documents in the collection; $n(q_n)$ is the number of documents containing the term q_n .

This allows the algorithm to evaluate how well each document matches the expanded query. The system then selects the top five most relevant documents, ensuring that the highest-quality responses are prioritized for the user.

3.4. Summarization and Response Generation

To further improve the quality of responses, the retrieved medical documents are summarized using a large language model that extracts the most critical parts of the doctors' responses. This summarization reduces the length of the response while correcting syntactic and grammatical issues often present in forum posts, as well as eliminating irrelevant or redundant content. Despite the computational overhead, this step plays a pivotal role in improving the relevance and fluency of the generated responses. It is

crucial to note that, during the response generation, the sources utilized are clearly indicated. This allows users to immediately access the original documents from which the information is derived, ensuring transparency and reliability in the responses provided by the chatbot.

Following summarization, the documents are reformatted into a standardized schema that separates the original user question and the corresponding medical response, as illustrated in Figure 2.

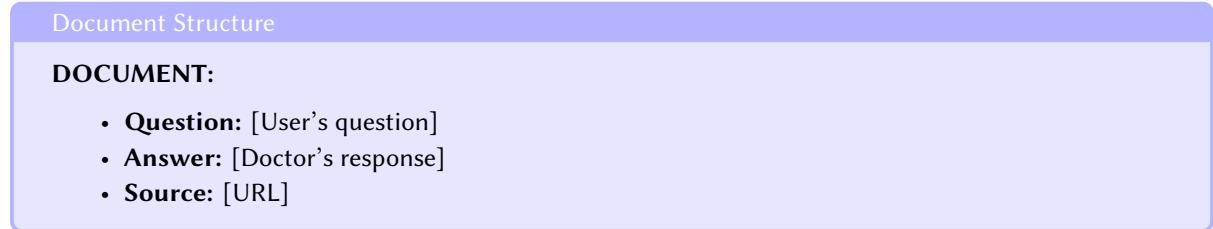


Figure 2: Schema adopted for the representation of summarized documents. Each entry is structured as a question-answer pair extracted from medical forums.

As shown in Figure 3, the final prompt, to be provided to the language model, is constructed by concatenating the processed user query with the standardized documents, which contain relevant answers given by physicians in similar cases.

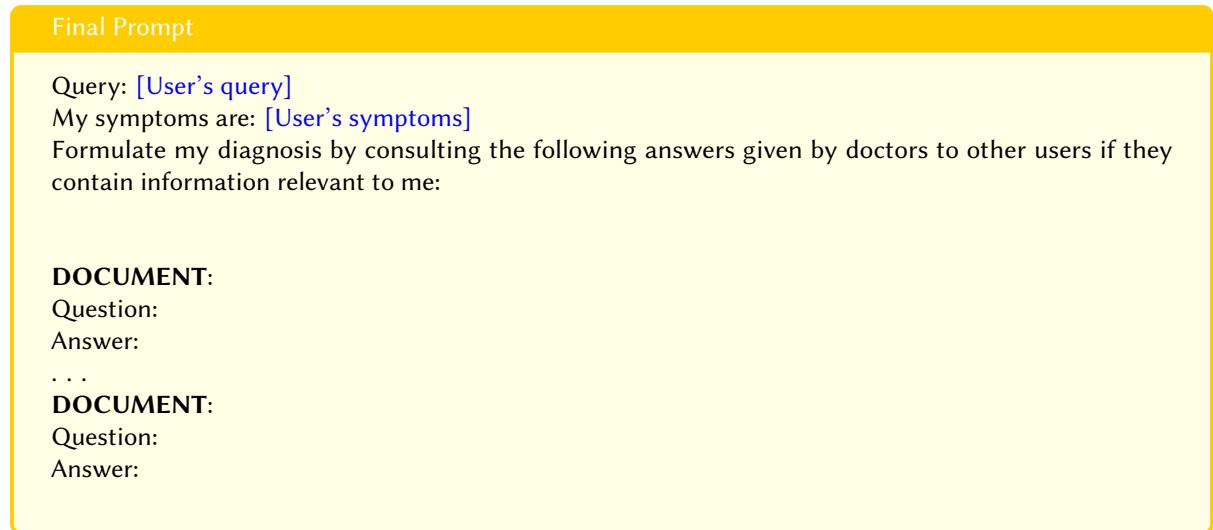


Figure 3: Example of the final prompt sent to the generative model. It includes the user's processed query and multiple relevant question-answer documents for contextual retrieval.

This structured approach, which combines summarization and the sourcing of medical information, not only enhances the chatbot's response accuracy but also fosters trust with users by ensuring transparency.

4. Experiments

4.1. Experimental setup

All experiments were conducted on a local machine with an NVIDIA RTX 3070 GPU (8 GB VRAM). Due to hardware limitations, we selected generative "instruct" models with fewer than 8 billion parameters in their quantized versions, as well as embedding models with fewer than 600 million parameters.

For the embedding task, we used the multilingual-e5-large model³, a multilingual model with 560

³<https://huggingface.co/intfloat/multilingual-e5-large>

million parameters that supports 94 languages, including Italian. This model has been found to be particularly effective in capturing context during information retrieval from information sources.

For query expansion, summarization and text generation, we employed the LLaMA-3-8B-Instruct model⁴ [20] with top_k = 0.9 and temperature = 0.3, which provided coherent and contextually rich responses in Italian despite its 8k token context window. The system architecture integrated ChromaDB⁵ for efficient embedding storage and retrieval, while the NoSQL database MongoDB⁶ (used in Python) is used to persistently store conversations extracted from relevant forums and encyclopaedias.

4.2. Dataset

The dataset used as the knowledge base for the RAG phase was collected from Italian medical forums with a total of 268,019 conversations between physicians and patients. In total there are 65 medical categories to which discussions on these forums belong, and the most covered are certain areas such as Psychology, Gastroenterology and Digestive Endoscopy, and Infectious Diseases. In addition, articles from Italian medical encyclopedias have also been collected, with a total of 2,981 articles, most of which cover the field of general medicine.

4.3. Results

4.3.1. Quantitative Evaluation

The chatbot’s performance was quantitatively evaluated using a diverse test suite consisting of realistic patient queries modeled after typical health-related interactions. Three generation strategies were analyzed:

- S_1 : No_RAG — baseline with no external document retrieval or augmentation;
- S_2 : R+Q+RR — Retrieval-Augmented Generation enhanced with Query Expansion and BM25-based Reranking;
- S_3 : R+Q+RR+S — the previous strategy with an additional Summarization step.

The test set was constructed manually by taking 50 samples at random from the original dataset. The informativeness of generated responses was first evaluated using the TF-IDF metric, which quantifies term frequency normalized by document relevance. As reported in Table 1, strategy S_3 achieved the highest score (10.7269), indicating the greatest lexical density and coverage of medically relevant content. Strategy S_2 followed closely (10.4144), while the baseline S_1 obtained the lowest score (9.2093).

Table 1
Performance Scores of Different Strategies

Strategy	TF-IDF	BERTScore (F1)
No_RAG	9.2093	0.8421
R+Q+RR	10.4144	0.8783
R+Q+RR+S	10.7269	0.8914

Table 2
Informative Differences Between Strategies

Compared To	R+Q+RR	R+Q+RR+S
No_RAG	-1.054	-1.3252
R+Q+RR	-	-0.2377

To better characterize the differences in informativeness between strategies, we introduce a novel metric—the *Informative Difference Matrix* (IDM)—presented in Table 2. This metric adjusts the raw difference in TF-IDF scores by incorporating textual similarity via ROUGE-L [21], defined as:

$$Q_{i,j} = (1 - \text{ROUGE-L}_{i,j}) \cdot (S_{\text{tfidf},i} - S_{\text{tfidf},j}) \quad (3)$$

Here, $Q_{i,j}$ quantifies the net gain in informativeness from strategy S_i over S_j , penalized by overlapping structure. Positive values suggest that S_i offers more distinctive and rich content compared to S_j . Manual

⁴<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

⁵<https://www.trychroma.com/>

⁶<https://www.mongodb.com/>

inspection of high-scoring $Q_{i,j}$ pairs confirmed that higher values correlated with more context-specific and clinically informative answers.

To complement these lexical evaluations, we employed BERTScore [22], a semantic similarity metric based on transformer embeddings. Using the multilingual model paraphrase-multilingual-MiniLM-L12-v2 [23], we computed average F1 scores across all outputs. Results indicate that strategies S_2 and S_3 offer significantly better alignment with expected clinical answers, with S_3 showing the highest semantic relevance (0.8914).

These findings confirm that the proposed enhancements—query expansion, reranking, and summarization—substantially improve both the surface-level informativeness and deeper semantic coherence of the generated content.

Key performance metrics are summarized as follows:

- **Response Time:** Average of 30 seconds per query, ranging from 25s (R+Q+RR) to 37s (R+Q+RR+S).
- **Retrieval Accuracy:** 86% of test queries successfully returned at least one medically relevant document.
- **Structural Similarity:** Measured using ROUGE-L [21] and BLEU [24] against reference phrasing.
- **Semantic Preservation:** Assessed with BERTScore [22], highlighting improved contextual alignment.

4.3.2. Qualitative Evaluation

Subject matter experts evaluated the chatbot’s responses, finding that RAG-generated replies provided more precise diagnostic suggestions than those without RAG. The inclusion of medical references boosted user trust. Queries were tested in three modes: Base LLM, LLM with RAG, and LLM with full processing. RAG improved relevance through query expansion and reranking. Responses without RAG were often vague, while RAG-enabled responses offered preliminary diagnoses and treatments. Response times ranged from 25 to 45 seconds, with query expansion adding 2 to 4 seconds and summarization involving up to 7 seconds per step. To balance speed and accuracy, the chatbot offers two modes: Fast (no summarization) and Precise (full processing).

4.3.3. System Demonstration

Figure 4 presents a demonstration of the system and the patient interaction, illustrating its core components: *Question*, *Answer*, and *Sources*.

5. Ethical Considerations

While AI-driven medical chatbots offer numerous advantages, they also raise significant ethical concerns that must be addressed. One of the most critical issues is the risk of misdiagnosis or misinformation, particularly when users rely solely on the system’s recommendations without consulting a healthcare professional. Data privacy is another key concern, given the sensitive nature of the medical information being processed. To mitigate these risks, the system must clearly inform users that it provides only preliminary guidance and is not a substitute for professional diagnosis. By also attaching sources to its responses, the system enhances transparency, allowing users to verify the credibility of the information and fostering greater trust through the ability to explore reliable medical sources. To ensure accountability, each response includes links to original medical forum threads and encyclopedia articles. Additionally, the chatbot explicitly states it is not a replacement for professional medical consultation and does not store user data.

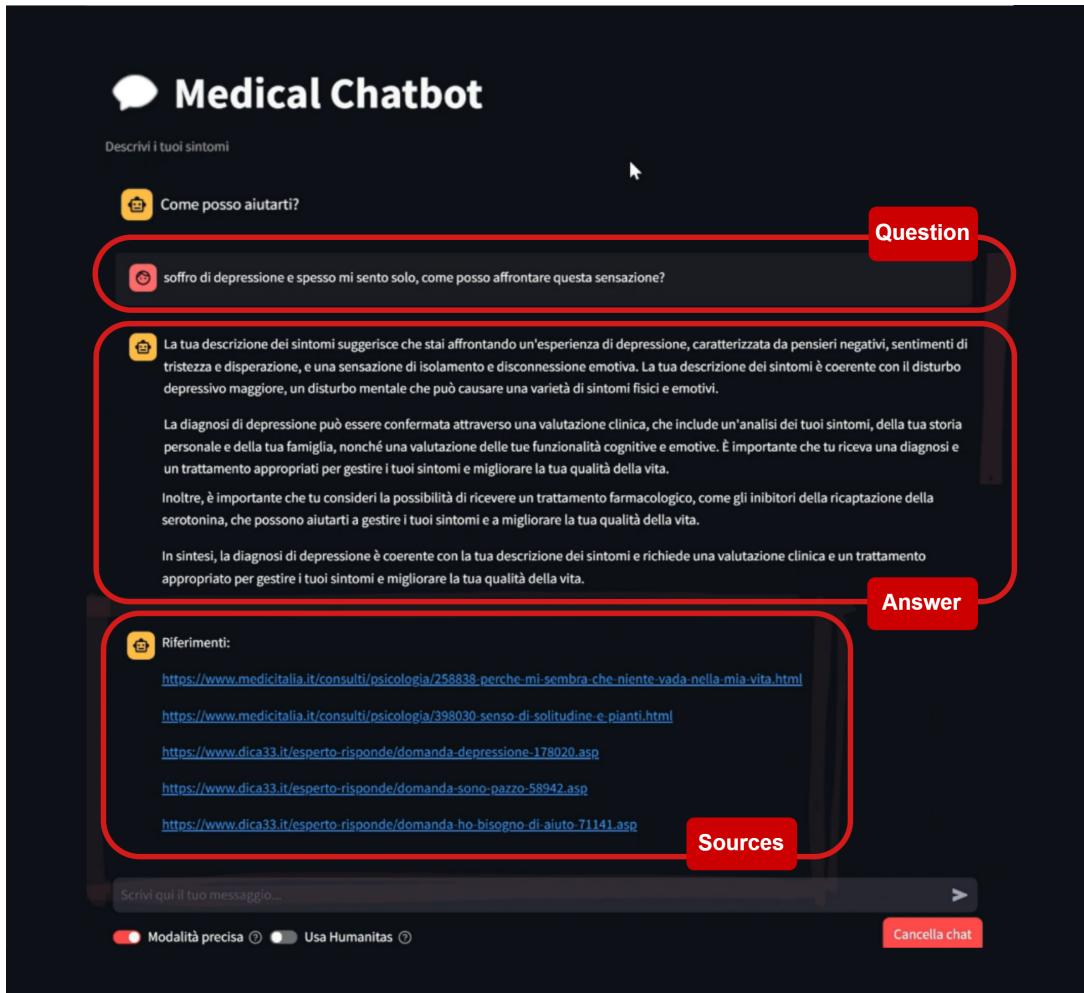


Figure 4: Demonstration of the Italian Medical Assistant System

6. Conclusion and Future direction

This paper presented the implementation of an AI-based medical chatbot utilizing RAG techniques and Big Data technologies. The system's ability to provide relevant medical responses, supported by reliable sources, makes it a valuable tool for both preliminary consultations and patient education, helping guide users towards appropriate medical professionals. Future work will focus on improving the chatbot's consultation accuracy by integrating larger and more diverse datasets from certified, high-quality sources, as well as adopting more advanced generative models. Additionally, a promising direction for further development is the integration of real-time medical data from wearable devices and electronic health records, with the aim of making the chatbot not only a useful resource for preliminary consultations but also a supportive diagnostic tool for healthcare professionals.

Acknowledgments

This work was conducted with the financial support of (1) the PNRR MUR project PE0000013-FAIR and (2) the Italian ministry of economic development, via the ICARUS (Intelligent Contract Automation for Rethinking User Services) project (CUP: B69J23000270005).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and DeepL in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] M. A. Ahmad, I. Yaramis, T. D. Roy, Creating trustworthy llms: Dealing with hallucinations in healthcare AI, CoRR abs/2311.01463 (2023). URL: <https://doi.org/10.48550/arXiv.2311.01463>. doi:10.48550/ARXIV.2311.01463. arXiv:2311.01463.
- [2] A. Bora, H. Cuayáhuitl, Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications, Machine Learning and Knowledge Extraction 6 (2024) 2355–2374. URL: <https://www.mdpi.com/2504-4990/6/4/116>. doi:10.3390/make6040116.
- [3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020, pp. –. URL: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- [4] S. G. Haez, M. Segala, P. Bellan, S. Magnolini, L. Sanna, M. Consolandi, M. Dragoni, A retrieval-augmented generation strategy to enhance medical chatbot reliability, in: J. Finkelstein, R. Moskovich, E. Parimbelli (Eds.), Artificial Intelligence in Medicine - 22nd International Conference, AIME 2024, Salt Lake City, UT, USA, July 9-12, 2024, Proceedings, Part I, volume 14844 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 213–223. URL: https://doi.org/10.1007/978-3-031-66538-7_22. doi:10.1007/978-3-031-66538-7\22.
- [5] R. Russo, D. Russo, G. M. Orlando, A. Romano, G. Riccio, V. L. Gatta, M. Postiglione, V. Moscato, Europeanlawadvisor: an open source search engine for european laws, in: W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, X. Hu (Eds.), IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024, IEEE, 2024, pp. 4751–4756. URL: <https://doi.org/10.1109/BigData62323.2024.10826025>. doi:10.1109/BIGDATA62323.2024.10826025.
- [6] D. Russo, G. M. Orlando, A. Romano, G. Riccio, V. L. Gatta, M. Postiglione, V. Moscato, Scaling llm-based knowledge graph generation: A case study of italian geopolitical news, in: W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, X. Hu (Eds.), IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024, IEEE, 2024, pp. 3494–3497. URL: <https://doi.org/10.1109/BigData62323.2024.10825937>. doi:10.1109/BIGDATA62323.2024.10825937.
- [7] A. Ferraro, A. Galli, V. L. Gatta, M. Postiglione, G. M. Orlando, D. Russo, G. Riccio, A. Romano, V. Moscato, Agent-based modelling meets generative AI in social network simulations, in: L. M. Aiello, T. Chakraborty, S. Gaito (Eds.), Social Networks Analysis and Mining - 16th International Conference, ASONAM 2024, Rende, Italy, September 2-5, 2024, Proceedings, Part I, volume 15211 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 155–170. URL: https://doi.org/10.1007/978-3-031-78541-2_10. doi:10.1007/978-3-031-78541-2\10.
- [8] G. M. Orlando, V. La Gatta, D. Russo, V. Moscato, Can generative agent-based modeling replicate the friendship paradox in social media simulations?, in: Proceedings of the 17th ACM Web Science Conference 2025, 2025, pp. 510–515.
- [9] A. Ferraro, G. M. Orlando, D. Russo, Generative agent-based modeling with large language models for insider threat detection, Engineering Applications of Artificial Intelligence 157 (2025) 111343.
- [10] L. Costabile, G. M. Orlando, V. La Gatta, V. Moscato, Assessing the potential of generative agents in crowdsourced fact-checking, arXiv preprint arXiv:2504.19940 (2025).

- [11] A. Romano, G. Riccio, M. Postiglione, V. Moscato, Pie-med: Predicting, interpreting and explaining medical recommendations, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V*, volume 15576 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 6–12. URL: https://doi.org/10.1007/978-3-031-88720-8_2. doi:10.1007/978-3-031-88720-8\2.
- [12] G. Riccio, A. Romano, A. Korsun, M. Cirillo, M. Postiglione, V. L. Gatta, A. Ferraro, A. Galli, V. Moscato, Healthcare data summarization via medical entity recognition and generative AI, in: N. Bena, B. D. Martino, A. Maratea, A. Sperduti, E. D. Nardo, A. Ciaramella, R. Montella, C. A. Ardagna (Eds.), *Proceedings of the 2nd Italian Conference on Big Data and Data Science (ITADATA 2023), Naples, Italy, September 11-13, 2023, volume 3606 of CEUR Workshop Proceedings, CEUR-WS.org*, 2023, pp. –. URL: <https://ceur-ws.org/Vol-3606/paper47.pdf>.
- [13] A. Ferraro, A. Galli, V. La Gatta, M. Postiglione, D. Russo, G. M. Orlando, G. Riccio, A. Romano, V. Moscato, From explanation to exploration: Promoting diversity in recommendation systems, in: *International Workshop on Recommender Systems for Sustainability and Social Good*, Springer, 2024, pp. 135–150.
- [14] S. Zhao, Y. Yang, Z. Wang, Z. He, L. Qiu, L. Qiu, Retrieval augmented generation (RAG) and beyond: A comprehensive survey on how to make your llms use external data more wisely, *CoRR* abs/2409.14924 (2024). URL: <https://doi.org/10.48550/arXiv.2409.14924>. doi:10.48550/ARXIV.2409.14924. arXiv: 2409.14924.
- [15] S. Ranasinghe, D. D. Silva, N. Mills, D. Alahakoon, M. Manic, Y. Lim, W. Ranasinghe, Addressing the productivity paradox in healthcare with retrieval augmented generative AI chatbots, in: *IEEE International Conference on Industrial Technology, ICIT 2024, Bristol, UK, March 25-27, 2024, IEEE*, 2024, pp. 1–6. URL: <https://doi.org/10.1109/ICIT58233.2024.10540818>. doi:10.1109/ICIT58233.2024.10540818.
- [16] M. Alkhafaf, P. Yu, M. Yin, C. Deng, Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records, *Journal of Biomedical Informatics* 156 (2024) 104662. URL: <https://www.sciencedirect.com/science/article/pii/S1532046424000807>. doi:<https://doi.org/10.1016/j.jbi.2024.104662>.
- [17] S. Murali, S. S., S. R., Remag-kr: Retrieval and medically assisted generation with knowledge reduction for medical question answering, in: X. Fu, E. Fleisig (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Student Research Workshop, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics*, 2024, pp. 140–145. URL: <https://aclanthology.org/2024.acl-srw.13>.
- [18] J. Bayarri-Planas, A. K. Gururajan, U. C. D. Garcia-Gasulla, Boosting healthcare llms through retrieved context, *CoRR* abs/2409.15127 (2024). URL: <https://doi.org/10.48550/arXiv.2409.15127>. doi:10.48550/ARXIV.2409.15127. arXiv: 2409.15127.
- [19] A. Romano, G. Riccio, M. Postiglione, V. Moscato, Identifying cardiological disorders in spanish via data augmentation and fine-tuned language models, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of CEUR Workshop Proceedings, CEUR-WS.org*, 2024, pp. 207–222. URL: <https://ceur-ws.org/Vol-3740/paper-19.pdf>.
- [20] AI@Meta, Llama 3 model card, Meta (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- [21] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text summarization branches out*, 2004, pp. 74–81.
- [22] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675* (2019).
- [23] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4512–4525.
- [24] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine

translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.