

Conversational-Agent for Patient Information Leaflet

Riccardo Lunardi¹, Paolo Coppola

Abstract

Conversational agents offer a natural way to interact with users, providing a wide range of services in different fields. In the healthcare sector, conversational agents can be used to provide information about medications, diseases and treatments. In this paper, we present a conversational agent designed to provide information about Patient Information Leaflets (PIL), originated from the SeSaMo web service. The conversational agent is powered by a Large Language Model and uses a Retrieval-Augmented Generation (RAG) framework to generate the text. We present the preliminary results of the system, showing that the RAG framework can be used to generate high-quality text. The next steps will be to expand the architecture to handle questions about multiple medications and to provide information about the interactions between them, evaluating the system with a larger dataset of questions and answers.

Keywords


conversational agents, healthcare, patient information leaflet, retrieval augmented generation

1. Introduction

A conversational agent is a system engineered to exploit natural language, enabling text-based communication with users. Leveraging on chat-bot technologies significantly enhances platform accessibility [1], enabling elderly and disabled people to use natural language to interact with applications. The adoption of chat-bots has been growing in different settings, with development focused on enabling interaction with the public in different fields, such as in healthcare [2], education [3] and customer services [4]. With the recent breakthroughs in the NLP (Natural Language Processing) and then with the release of chat-bot services (e.g. ChatGPT [5]), the quality bar expected from these kind of agents has risen significantly, creating new challenges and opportunities for researchers to improve conversational agents. Especially in the healthcare sector, conversational agents offer promising applications, enabling citizens to access online services such as appointment booking and disease information through natural language interaction. Despite the challenges posed by the need for high adaptability and large data processing requirements, leveraging conversational agents in healthcare can significantly benefit the public by providing more efficient and simplified access to services.

IIR 2024: 14th Italian Information Retrieval Workshop, September 05-6, 2024, Udine, IT

 0009-0001-5550-317X (R. Lunardi)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Aims

In this paper, we present a early and experimental conversational agent based on LLMs, designed to provide information, upon questioning, about Patient Information Leaflets (PILs). A PIL is a document that provides information about a drug, such as its composition, its indications, its contraindications, its side effects and how to use it. At the time of writing, the Italian region Friuli-Venezia Giulia provides a web service called SeSaMo¹ that allows users to search for a medication’s PIL by providing its name. The presentation of the PIL content is in plain text, resulting in a long and difficult to read document. Given the importance of delivering clear information about medications [6], the conversation agent is designed to provide a user-friendly way for accessing the knowledge contained in the PIL. This empowers patients not only to ask questions about a single drug, but to seek information about interactions with other medications, too. The objective is inherently challenging, given that the conversational agent is powered by a Large Language Model, which they are know, while providing great interactivity and adaptability, to be prone to hallucinations and consequently generating incorrect or misleading information [7]. This pose the obstacle of designing a conversational agent capable of providing correct and reliable answers only when such proposed answers are correct, all while engaging with the user in a natural and compelling manner. This paper acknowledges the importance of UX design and scalability in conversational agents [8], but focuses on improving the accuracy and efficiency of information retrieval instead.

3. Related Work

The benefit of an easier way to access the information contained in the PIL has been widely studied in literature. An experiment conducted by Berry et al. [9] shows that the use of a more personalized style in the PIL presentation can significantly increase the user satisfaction and lower the ratings of likelihood of side effects occurring. Other works highlight the beneficial effects of having patients understanding the PIL’s content [10, 11].

Focusing on conversational agents in healthcare, systematic reviews such as Laranjo et al. [12] and Tudor Car et al. [13] report that the usability and satisfaction of conversational agents is generally high between users, while the effectiveness is mixed, underlining the need to improve the quality of the services provided by conversational agents. When implementing LLM-driven chat-bots, the necessity of elaborating substantial amount of data is well-established: to manage all of this information, while ensuring the correctness of the generated text, the literature advice to leverage on Retrieval-Augmented Generation (RAG) models [14]. Recent studies, like Asai et al. [15], have proposed methods to reduce hallucinations and inaccuracies in agents, including self-supervised information retrieval to enhance the quality of generated text. Similarly to the case study presented in this work, Sanna et al. [16] proposed a framework for constructing medical chat-bots. However, they do not focus on PILs, but on generically medical question answering.

¹<https://sesamo.sanita.fvg.it>

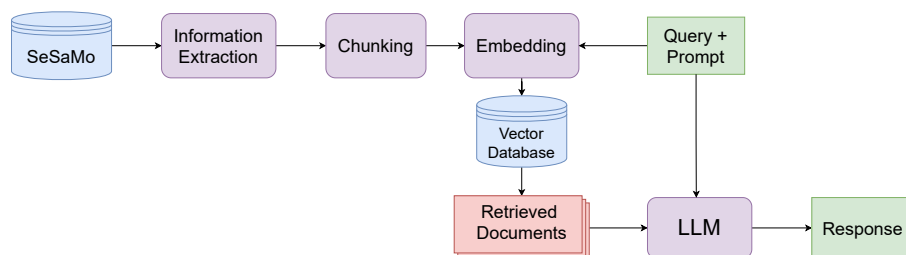


Figure 1: RAG framework for the conversational agent.

4. Preliminary Methodology

In our early work, we tested very simple RAG implementations to establish a solid baseline for evaluating the effectiveness of future versions of the conversational agent. The RAG system is composed as shown in Figure 1. First, the PILs are retrieved from the SeSaMo web service. Subsequently, the documents are processed by a series of algorithms to chunk them in smaller parts, which are then embedded using a text embedding model. During inference, the system embeds the user question and compares it with the embedded chunks of the PILs: the most similar are then used to generate the final answer using a LLM. The process follows the standard RAG pipeline, as described in the literature by Lewis et al. [14].

To implement the RAG, we used LangChain [17] in combination with LlamaCpp [18] to infer the LLMs. We systematically evaluated the quality of the RAG framework by defining a sequence of functions (e.g. chain) and slightly altered every step of it. Specifically, we varied the processing algorithms to chunk the input text, the text embedding models, the system prompts and the LLMs employed to generate the final answer. The basic version of the chunking algorithm simply splits the text after 200 characters, while the advanced version leverages the HTML structure of the PILs to divide the text in more meaningful portions, organizing them by chapter. Similarly, two versions of the prompt were employed: a basic version that prompted the model to answer based solely on the context provided and an advanced version that provided additional instructions on how the model should respond. The full prompts are included in the supplementary materials¹. The text embedding models employed are `all-mpnet-base-v2` [19] and `paraphrase-multilingual-MiniLM-L12-v2` [20], while the LLMs utilized are `Mistral-7B-Instruct-v0.2` [21] and `Llama-2-7b-hf` [22].

With a small dataset of handcrafted questions and answers, we evaluated the quality of the generated text by comparing it with the expected outputs. The comparison is accomplished by embedding both reference and prediction and subsequently calculating the cosine distance between them. For the sake of simplicity, we focused on a single drug in this preliminary work, deferring the evaluation of multiple medications to future work. The drug taken in consideration is the MOMENT² by Angelini Pharma. The results are

¹<https://osf.io/e6vwh>

²MOMENT 200 mg coated tablets - Ibuprofen

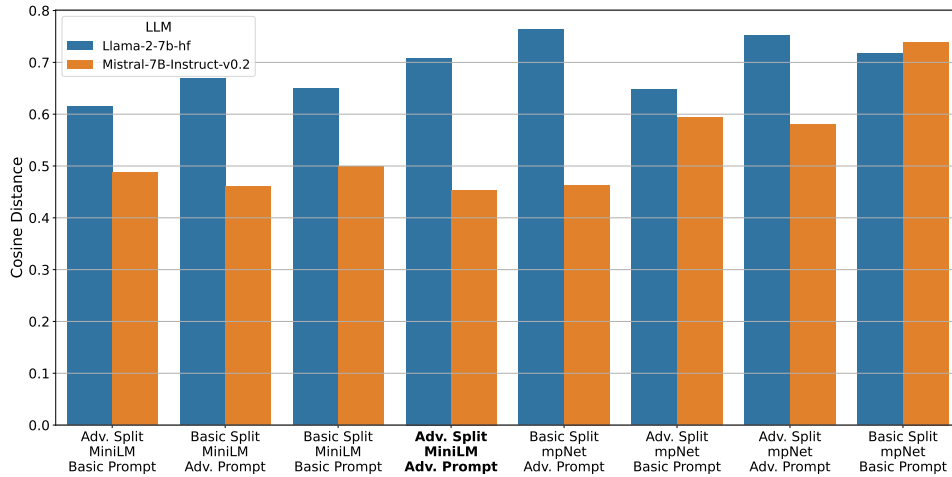


Figure 2: Cosine Distance Scores by Splitting Strategy, Embedding Model, Prompt and Large Language Model.

shown in Figure 2: the plot displays the distance scores obtained by the different splitting strategies, embedding models, prompts and LLMs. The lower the score, the better the quality of the generated text. In all the cases, except one, *Mistral-7B-Instruct-v0.2* scored consistently better than *Llama-2-7b-hf*. The best configuration is the one that uses *Mistral-7B-Instruct-v0.2*, *paraphrase-multilingual-MiniLM-L12-v2* and two advanced version of the prompt and splitting strategy.

5. Conclusions and Future Work

The preliminary results are promising, showing that a RAG framework can be used in a question-answering task about PILs. This framework allows users to ask for specific information they are looking for, enhancing accessibility and usability. The next steps will involve expanding the architecture to handle questions about multiple medications and provide information about their interactions. The framework must also recognize when it lacks an answer and respond appropriately, which is crucial in healthcare, where inaccurate information can have serious consequences. We also plan to evaluate the system with a larger dataset of questions and answers to better understand the quality of the generated text and identify areas that need improvement. Addressing the challenge of providing accurate information about drug interactions will be the primary focus.

6. Acknowledgments

This research has received partial funding from the "Department Strategic Plan of the University of Udine - Interdepartmental Project on Digital Governance and Public Administration". The initiative started in January 2021 and will end in December 2025.

References

- [1] K. S. Glazko, M. Yamagami, A. Desai, K. A. Mack, V. Potluri, X. Xu, J. Mankoff, An autoethnographic case study of generative artificial intelligence's utility for accessibility, in: Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: <https://doi.org/10.1145/3597638.3614548>. doi:10.1145/3597638.3614548.
- [2] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, E. Coiera, Conversational agents in healthcare: a systematic review, *Journal of the American Medical Informatics Association* 25 (2018) 1248–1258. URL: <https://doi.org/10.1093/jamia/ocy072>. doi:10.1093/jamia/ocy072.
- [3] B. Khosrawi-Rad, H. Rinn, R. Schlimbach, P. Gebbing, X. Yang, C. Lattemann, D. Markgraf, S. Robra-Bissantz, Conversational agents in education – a systematic literature review, 2022.
- [4] R. Bavaresco, D. Silveira, E. Reis, J. Barbosa, R. Righi, C. Costa, R. Antunes, M. Gomes, C. Gatti, M. Vanzin, S. C. Junior, E. Silva, C. Moreira, Conversational agents in business: A systematic literature review and future research directions, *Computer Science Review* 36 (2020) 100239. URL: <https://www.sciencedirect.com/science/article/pii/S1574013719303193>. doi:<https://doi.org/10.1016/j.cosrev.2020.100239>.
- [5] O. et al., Gpt-4 technical report, 2024. arXiv:2303.08774.
- [6] E. Bersellini, D. Berry, The benefits of providing benefit information in a patient information leaflet, *International Journal of Pharmacy Practice* 15 (2010) 193–199. URL: <https://doi.org/10.1211/ijpp.15.3.0006>. doi:10.1211/ijpp.15.3.0006.
- [7] H. Alkaiissi, S. I. McFarlane, Artificial hallucinations in chatgpt: implications in scientific writing, *Cureus* 15 (2023).
- [8] I. K. F. Haugeland, A. Følstad, C. Taylor, C. A. Bjørkli, Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design, *International Journal of Human-Computer Studies* 161 (2022) 102788. doi:<https://doi.org/10.1016/j.ijhcs.2022.102788>.
- [9] D. C. Berry, I. C. Michas, E. Bersellini, Communicating information about medication: the benefits of making it personal, *Psychology and Health* 18 (2003) 127–139.
- [10] E. Coudeyre, S. Poiraudau, M. Revel, A. Kahan, J. L. Drapé, P. Ravaud, Beneficial effects of information leaflets before spinal steroid injection, *Joint Bone Spine* 69 (2002) 597–603.
- [11] J. H. Barlow, C. Wright, Knowledge in patients with rheumatoid arthritis: a longer term follow-up of a randomized controlled study of patient education leaflets., *British Journal of Rheumatology* 37 (1998) 373–376.
- [12] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, E. Coiera, Conversational agents in healthcare: a systematic review, *Journal of the American Medical Informatics Association* 25 (2018) 1248–1258. URL: <https://doi.org/10.1093/jamia/ocy072>. doi:10.1093/

jamia/ocy072.

- [13] L. Tudor Car, D. A. Dhinakaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, R. Atun, Conversational agents in health care: Scoping review and conceptual analysis, *J Med Internet Res* 22 (2020) e17158. URL: <http://www.jmir.org/2020/8/e17158/>. doi:10.2196/17158.
- [14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. [arXiv:2005.11401](https://arxiv.org/abs/2005.11401).
- [15] A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi, Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023. [arXiv:2310.11511](https://arxiv.org/abs/2310.11511).
- [16] L. Sanna, P. Bellan, S. Magnolini, M. Segala, S. Ghanbari Haez, M. Consolandi, M. Dragoni, Building certified medical chatbots: Overcoming unstructured data limitations with modular RAG, in: D. Demner-Fushman, S. Ananiadou, P. Thompson, B. Ondov (Eds.), *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024*, pp. 124–130. URL: <https://aclanthology.org/2024.cl4health-1.15>.
- [17] H. Chase, Langchain, 2022. URL: <https://github.com/langchain-ai/langchain>.
- [18] G. Gerganov, llama.cpp, <https://github.com/ggerganov/llama.cpp>, 2024.
- [19] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, 2020. [arXiv:2004.09297](https://arxiv.org/abs/2004.09297).
- [20] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019. URL: <http://arxiv.org/abs/1908.10084>.
- [21] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. [arXiv:2310.06825](https://arxiv.org/abs/2310.06825).
- [22] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. [arXiv:2307.09288](https://arxiv.org/abs/2307.09288).