

# Chatbot Based on Clinical Literature for Decision Support

Irene Sánchez-Montejo<sup>1</sup>, Carlos Teleria-Oriols<sup>1,2</sup>, Raquel Trillo Lado<sup>3</sup>

<sup>1</sup> Afiliación: Instituto Aragonés de Ciencias de la Salud  
e-mail: [isanchez.iacs@aragon.es](mailto:isanchez.iacs@aragon.es)

<sup>2</sup> Afiliación: Ciencia de Datos para la Investigación en Servicios y Políticas Sanitarias

<sup>3</sup> Afiliación: DisCo – Computación Distribuida  
Instituto de Investigación en Ingeniería de Aragón (I3A)  
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.  
Tel. +34-976762707

## Summary

We provide a thesis approach to evaluate different language models which are able to analyze documents to extract the essential information from each one, and develop a chatbot capable of answering according to clinical practice guidelines and biomedical literature, which provides the latest in diagnostic and treatment for each disease.

## Introduction

Natural Language Processing (NLP) allows a computer to understand natural texts while keeping its semantic meaning to perform multiple tasks, such as: text translation [1], sentiment analysis [2] or text summarization [3], among others. This has also enabled the creation of chatbots (virtual assistants).

Chatbots are based on the use of NLP, which allows them to understand and answer intelligently to queries made by a user, thus reproducing a conversation. Moreover, they use Large Language Models (LLM) that process and analyze complex texts such as large volumes of data (e.g. GPT-4 developed by OpenAI [4]).

In the healthcare field, these systems can be applied to process medical contents and help in the task of extracting relevant information from large and complex documents. There are already some chatbots available to give some help proposing a diagnostic to a patient according to their symptoms [5], thus reducing the communication gap between patients and healthcare professionals, and providing specific treatments.

Even more, NLP techniques and LLMs can be a very useful tool for clinical practice guidelines writers, in the process of reading, processing and

assessing scientific and clinical relevance of biomedical articles in the context of evidence based medicine. But all these uses of natural language technologies must be objectively evaluated and measured with regard to their efficiency and usefulness.

## Our proposal

With the objective of evaluating and propose specific solutions in the healthcare context based on NLP and LLM, we aim to develop different chatbots trained with clinical practice guidelines and other high-quality biomedical literature written or translated into Spanish. These chatbots will be used in real clinical contexts to evaluate their performance in terms of speed, resource consumption, accuracy and usefulness to make better clinical decisions.

Some previous works have studied how to build a chatbot trained on a medical corpus to provide information to the final user. When a chatbot is built, it is necessary to choose the NLP architecture that better fits the problem to solve and extracts the the most relevant information from the dataset. But, in most of the cases, these models have been trained using English as their core language. In our thesis, we plan to train models using Spanish as their core language.

Since LLMs and NLP techniques are constantly evolving, the first goal of our work is to gather different language models, analyze them and compare their performance. In our thesis, we plan to use a complete clinical dataset to train a more robust chatbot and use fine-tuning on different models, such as RoBERTa [6], LLaMa [7] or GPT4 [4], to retain their prior knowledge. These models will be used to develop different chatbots that will

be tested by doctors, patients and caregivers, given that there are different guidelines issues for different targets, and all of them can take benefit from these systems.

However, due to the inherent differences between the models, it is necessary to define the metrics that they will be compared under, such as the response time, resource consumption, feasibility to be integrated in actual healthcare systems, the quality and accuracy of the chatbot’s answer, and whether it is appropriate according to the ground truth, which also needs to be defined. So, a second goal in our work is to define these metrics and develop a methodology to evaluate the quality and utility of chatbots in a medical context.

An additional open line in this thesis is to use LLM models to assess the evidence level for any new biomedical article according to scientific criteria, as to include that article and its conclusions into the corresponding clinical practice guideline, enriching the guidelines in a continuous improvement process.

A final work line to explore is the feasibility of using patients’ clinical data in the models, maybe transforming this data into natural language expressions, so the answer of the chatbot will not depend only on its training and the formulated prompt, but also on the patient’s clinical conditions (see Fig 1).

## Conclusion

To sum up, starting from the well-known power of NLP techniques and LLMs in many real life

domains, and specifically in healthcare domain, and the possibility of having documents of proven scientific quality, like clinical practice guidelines, we aim to analyze and compare different models in their application to clinical support decision systems based on chatbots, to define objective and reproducible metrics to evaluate the quality and utility of these chatbots, and explore the feasibility of using LLMs as a way to facilitate the work of guidelines writers.

## REFERENCES

- [1]. Bird, S., Klein, E., and Loper, E. 2009. Natural language processing with Python: analyzing text with the natural language toolkit. ” O’Reilly Media, Inc.”.
- [2]. Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. The 3rd IEEE International Conference on data mining. IEEE
- [3]. Merchant, K., and Pande, Y. (2018). Nlp based latent semantic analysis for legal text summarization. In 2018 International conference on advances in computing, communications and informatics (ICACCI). IEEE.
- [4]. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774
- [5]. Xu, L., Sanders, L., Li, K., and Chow, J. C. 2021. Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. JMIR cancer, 7(4), e27850
- [6]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., ... and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
- [7]. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... and Lample, G. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

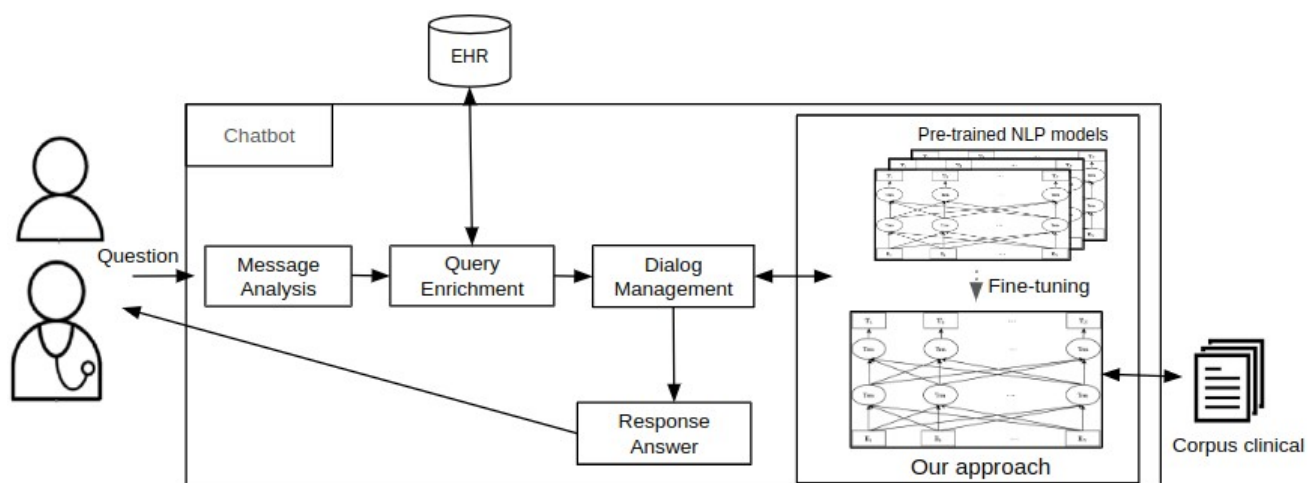


Fig 1: Overview of our proposal