Smart Health 36 (2025) 100552

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Privacy-preserving LLM-based chatbots for hypertensive patient self-management

Sara Montagna [a,*], Stefano Ferretti [b], Lorenz Cuno Klopfenstein [a], Michelangelo Ungolo [a], Martino Francesco Pengo [c,d], Gianluca Aguzzi [b], Matteo Magnini [b]

[a] Department of Pure and Applied Sciences, University of Urbino, Urbino, 61029, Italy
[b] Department of Computer Science and Engineering, University of Bologna, Cesena, 47521, Italy
[c] University of Milano-Bicocca, Faculty of Medicine, Milan, Italy
[d] Istituto Auxologico Italiano IRCCS, Milan, Italy

ARTICLE INFO

Dataset link: https://github.com/cric96/chatbot-test-llm

Keywords:
Large Language Model
Medical chatbot
Patient self-management
Patient empowerment

ABSTRACT

Medical chatbots are becoming a basic component in telemedicine, propelled by advancements in Large Language Models (LLMs). However, LLMs' integration into clinical settings comes with several issues, with privacy concerns being particularly significant.

The paper proposes a tailored architectural solution and an information workflow that address privacy issues, while preserving the benefits of LLMs. We examine two solutions to prevent the disclosure of sensitive information: *(i)* a filtering mechanism that processes sensitive data locally but leverage a robust OpenAI's online LLM for engaging with the user effectively, and *(ii)* a fully local deployment of open-source LLMs. The effectiveness of these solutions is assessed in the context of hypertension management across various tasks, ranging from intent recognition to reliable and emphatic conversation. Interestingly, while the first solution proved to be more robust in intent recognition, an evaluation by domain experts of the models' responses, based on reliability and empathetic principles, revealed that two out of six open LLMs received the highest scores.

The study underscores the viability of incorporating LLMs into medical chatbots. In particular, our findings suggest that open LLMs can offer a privacy-preserving, yet promising, alternative to external LLM services, ensuring safer and more reliable telemedicine practices. Future efforts will focus on fine-tuning local models to enhance their performance across all tasks.

1. Introduction

Recent advances in generative AI have opened the door for the use of powerful deep learning techniques that could be used in health-related contexts and potentially support healthcare care providers and patients. However, when asking ChatGPT (GPT-4o)[1] "*which are the main issues in applying LLMs for building a chatbot supporting patients?*" this is one answer you can get (in 50 words):

* Corresponding author.
E-mail address: sara.montagna@uniurb.it (S. Montagna).
URL: https://www.uniurb.it/persone/sara-montagna (S. Montagna).

[1] OpenAI. (2024). ChatGPT (October 2024 version) [Large Language Model]. Available from https://chat.openai.com/.

https://doi.org/10.1016/j.smhl.2025.100552
Received 29 October 2024; Received in revised form 19 January 2025; Accepted 25 February 2025
Available online 7 March 2025

> Key issues include ensuring accuracy and reliability of medical information, protecting patient privacy and data security, addressing biases in responses, and handling sensitive topics with care. Additionally, the chatbot must integrate seamlessly with healthcare systems and maintain ongoing updates to reflect the latest medical knowledge and guidelines.

Even though there is a huge discussion around the application of LLMs in medicine, their application in the clinical context, especially for home-care, is far from being resolved. In fact, despite the capabilities and widespread diffusion of LLMs, their exploitation in the context of patient self-management as a key technology for chatbot development, still poses a number of issues that are worth to be discussed and explored for this approach to support the clinical practice. For instance, in addition to the risks of generating unfaithful or factually incorrect outputs, it is fundamental to establish mechanisms that prevent the transmission of sensitive medical data to external LLM systems, ensuring robust privacy safeguards in healthcare chatbot applications. This need comes not only from common sense, but also from the need to adhere to laws and regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) (Act of Congress enacted by the 104th United States Congress, 1996), or General Data Protection Regulation (GDPR) (European Parliament & Council of the European Union, 2016; Zichichi, Ferretti, D'Angelo, & Rodríguez-Doncel, 2022).

Although similar issues and considerations stand in all medical applications, in this paper we present the data workflow of an LLM-based chatbot specifically designed for hypertension self-management. According to Kurniawan, Handiyani, Nuraini, Hariyati, and Sutrisno (2024), the use of AI-based chatbots to support chronic patients is pervasive, contributing to chronic illness management. However, most of these efforts refer to work prior to the revolution of LLMs, thus do not account for the specific privacy issues raised (Xu, Sanders, Li, & Chow, 2021). Others, while discussing similar issues – such as unreliability (Bortoli et al., 2024) and privacy (Khalid, Qayyum, Bilal, Al-Fuqaha, & Qadir, 2023) –, do not provide generally applicable architectural solutions.

The workflow presented in this paper is meant to be applied in any healthcare scenario with similar requirements, and it is devised to provide data security while preserving the high performance of LLMs. The system comprises different components: a back-end for data storage and analysis and a front-end service for user-interaction, which is based on user query processing through Natural Language Processing (NLP) and answer generation via Natural Language Generation (NLG). The most critical part, the NLP&NLG module, is devised for managing patient queries by handling the above-mentioned privacy and security issues, as well as reducing the risk of misinformation by confining the conversation to sentences strictly related to general information on the specific disease the chatbot is designed for.

For this purpose, we proposed and tested two alternative solutions. The first solution aims to leverage the full potential of online reference LLM services. It consists of two modules. The first module processes patient input and extracts the intent behind the question. A machine learning pipeline, implemented using ML.NET 2.0,[2] is trained to recognise the user's intent and classify the input according to whether it contains sensitive data or not. If the input contains sensitive data, it is handled locally using classical methods to parse strings and extract data. If the input does not, it is forwarded to a third-party LLM (GPT-3.5 Turbo in the paper for legacy reasons) through an API call, with a properly fine-tuned prompt, in order to obtain a compelling answer (Montagna, Ferretti, Klopfenstein, Florio, & Pengo, 2023).

The second solution is based on the idea of having a local open-source LLM (like Llama2 Touvron et al., 2023 or Mistral Jiang et al., 2023) served through its HTTP API (like fastchat Zheng et al., 2023 or ollama[3]). This approach completely avoids sensitive data disclosure, since the model is running locally on-premises, but requires complex system instruction via ad-hoc prompts, Retrieval Augmented Generation (RAG) (Lewis et al., 2020) approaches, or model fine-tuning. In this paper, six open-source LLMs have been evaluated.

We performed an evaluation of the two approaches using a dataset of simulated patient queries. Experimental evaluations are performed across various tasks, encompassing intent recognition, parameter extraction and general conversation. The performances are compared according to well-established metrics and conducting a comprehensive evaluation by domain experts. Our preliminary results show how the first approach performs better than the second one in terms of intent recognition and parameter extraction. However, there are two major shortcomings of the first approach, that are: *(i)* the risk that some sensitive data can be passed on to the third-party LLM system, either due to errors in the classification module or due to the ambiguity of the input and *(ii)* the cost associated with third-party technologies, which affects how democratic the solution is. Moreover, in this paper, we identified two open LLMs, Jiang et al. (2024), whose responses received higher scores from domain experts. This indicates their potential as promising candidates for further refinement and enhancement in specific tasks through tailored fine-tuning efforts.

In summary, the paper contributes to the advancement of LLMs in healthcare by proposing a robust workflow and evaluating alternative strategies tailored specifically for patient self-management within the clinical care context. The emphasis on privacy, security, and accuracy underscores the significance of these advancements in improving healthcare delivery. With respect to our previous findings (Montagna et al., 2024), where the first strategy appeared to have no competitors, further evaluations with additional methods and the exploration of new emerging open source models present fresh prospects for the adoption of open models within this domain.

[2] https://learn.microsoft.com/en-us/dotnet/machine-learning/.
[3] https://github.com/jmorganca/ollama.

2. Background

LLMs hold the premise for a deep revolution in medicine due to their impressive ability to understand human language and to produce human-like conversations. Many alternatives exist, and the list of pre-trained models is continuously growing and updated. These models differ, for what it concerns the focus of this work, in accessibility and (open like Mistral, licensed like Llama2 or closed like GPT-X), supported languages (multilingual or a subset of specific languages). Moreover, existing work on medical domain-specific LLMs includes models such as Google's Med-PALM2.

The adoption of chatbots in the context of medicine, as a tool to support patient needs and caregivers in their work, is already well-established in literature (Preum et al., 2021). For instance, conversational agents are a well-known approach to implement personal cognitive assistants (Montenegro, da Costa, & da Rosa Righi, 2019; Sulis, Mariani, & Montagna, 2023). However, since the advent of LLMs, the interest of the scientific community has increased tremendously and the discussion around this topic is impressively lively (Cascella et al., 2024; Clusmann et al., 2023; Thirunavukarasu et al., 2023/08/01; Tian et al., 2024). Main experiences in the adoption of LLM-based chatbots in medicine, without claiming to be exhaustive, are devoted to design tools for:

1. Assisting physicians or nurses, during their clinical practice, in various areas of medicine. As an example, they may support clinical decisions, by abstracting key results from literature. Or, they can detect medical errors by identifying discrepancies between diagnosis and treatment.
2. Bootstrapping patient empowerment by providing trustworthy and emphatic answers to user queries. In this context, they must resemble a dialogue between the physician and the patient which is a key element to provide effective and compassionate care. Moreover, they should be able to proactively suggest actions, reasoning on tracked patient activities and vital signs dynamic.
3. Supporting basic research by automating certain tasks, such as data analysis, acquisition and interpretation, summarising information, paraphrasing text, scientific literature search for medical knowledge and related work extraction.
4. Sustaining medical education by providing teaching material and interacting tutoring. In this context, it should be noted that very good performance was demonstrated in passing medical examinations.

In particular, it should be clarified that incorporating and integrating an LLM in a new application that exploits its NLP/NLG abilities poses a higher level of complexity in modelling, designing and implementing the whole system, than querying an LLM-based chatbot via a web interface. Indeed, even though literature reports an increasing number of work in these areas, a realistic vision in this context foresees extensive validation and further development to overcome a set of issues that literature clearly highlights (Meskó & Topol, 2023; Wang et al., 2023):

- Ethical concerns, including risks of privacy and security (Haltaufderheide & Ranisch, 2024; Li et al., 2023): Third-party technology, such as OpenAI's GPTs, carries an inherent risk of compromising patient privacy, if patients enter test results, photos of their face, communication information, etc. All of this vital health information is collected and stored, potentially compromising patient privacy. Open LLMs, locally deployed, seem the most obvious choice to handle these concerns. Although the performances shown against benchmarks are impressive, further domain-specific evaluation is required to demonstrate their effectiveness in medicine, and ad-hoc fine-tuning to deal with inaccuracy, uncertainty and misinformation; this is true for each of the applications listed above, but especially if chatbots are meant to interact with patients without the intervention of a domain expert.
- Proposed integrated solutions have to take into account economic costs, hardware requirements and environmental impact in order to develop a democratic and sustainable technology.

However, the results reported so far in literature strongly encourage further research and evaluation.

3. Methods

3.1. Architecture

In this paper, we focus on an architecture to support patient empowerment by exploiting an LLM-based chatbot designed to fit patient needs in the context of chronic diseases. The system architecture is meant to fulfil two main requirements: *(i)* the system must support the acquisition of patient data, while ensuring privacy and security, and *(ii)* must provide trustworthy answers to a limited set of in-topic queries. The system's architecture is depicted in Fig. 1. It encompasses four main components: a Chatbot, an NLP&NLG module for understanding user inputs and reacting properly, a database to store data and a Data Processing Unit to provide some kind of data elaboration, for instance by statistic functions or data visualisation. In the following, we describe the requirements to be satisfied by each component, as well as how they interact.

3.1.1. Chatbot

The chatbot is the component in charge of the interaction with patients and, as such, it must be designed for interacting with users whose digital skills are the most varied. Accordingly, it must be multiplatform, and easy to install and use. Moreover, to motivate and engage patients, particular attention should be given to the user experience, through a suitable and effective user interface.

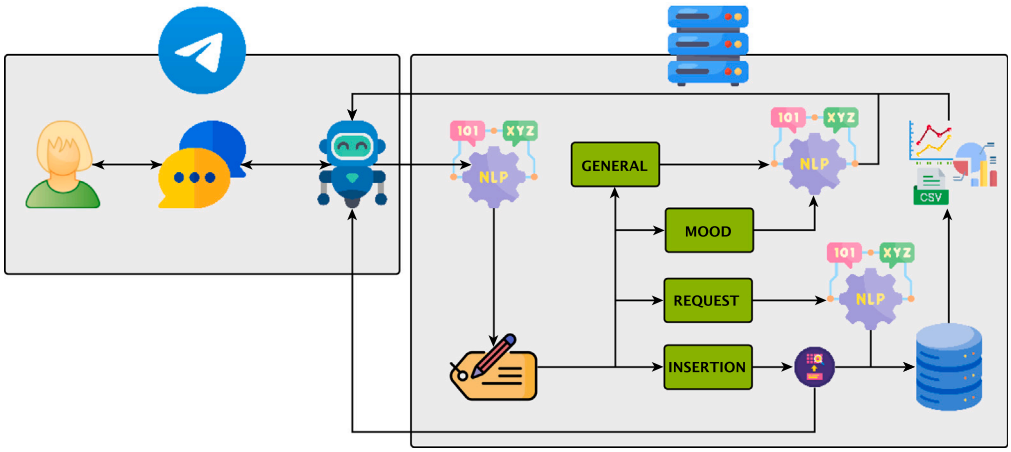


Fig. 1. System architecture. The patient interacts with the chatbot through an instant messaging application. Upon receiving a message, the chatbot forwards it to a server. The message is then categorised by an LLM into one of four categories: mood, request, insertion, or general—see 3.1.3 for more details. General and mood messages are immediately propagated to an LLM to generate the response. Insertion and request messages are post-processed by a parsing and LLM phase respectively to identify the parameters for the query to perform on the DB. After that, an acknowledgement message is sent to the user in the scenario of insertion. Instead, in the case of a request, the response is generated from the extracted data in the format specified by the user (e.g., graph, list, etc.).

3.1.2. Database

Patient data needs to be stored in order to allow for proper medical monitoring and diagnosis. Also, discussions, or some historical records of discussions made, might be important to improve the quality of the interactions with the user over time. Possible solutions range from a local database, which is the classic and straightforward solution to properly manage data in a client–server approach, to those already presented in our previous works (Montagna et al., 2023) leveraging distributed solutions which take into account also for data sovereignty requirements.

3.1.3. NLP&NLG

The component in charge of understanding the human language and generating answers with a human-like language is the most critical one. We expect patient to input almost every kind of sentence, which can span from general questions, regarding every aspect of their life, to specific questions related to their disease. Disease-related questions can vary from input data, request statistics and/or general information on their physical health state.

In this respect, there are several issues to deal with:

1. if NLP&NLG is provided on top of LLMs owned by private companies, data privacy and security are not granted: inputs containing sensitive data must be intercepted and managed accordingly;
2. the chatbot is typically devised for a specific medical domain. As such, it must not provide answers to every question the patient may input, but kindly remind the user of the tasks it is in charge of;
3. answers must be precise and emphatic, to improve user experience and trust, crucial elements for patient engagement and self-management;
4. the context of the prompt should be precisely fine-tuned in a way that answers are limited to the medical task at hand and no misinformation are produced.

With these requirements in mind, we identified four categories for the patient inputs, which represent the intents associated to each question. At each category, the execution of a different flow is associated.

Insertion The first category includes all the sentences that contain sensitive data, which can be vital signs, as well as general personal information, such as anthropometric measurements, anamnestic data, lifestyle, demographic information, and medical histories. Such information cannot be prompted to a third-party LLM but, since the interpretation must be reliable, an ad-hoc parser must be developed to extract relevant data from the sentence.

Request This category encompasses all the inputs specifically requesting to inspect patient data. This request does not typically contain sensitive information and an LLM may support the identification of the request parameters, such as which data the patient wants to inspect, since when and in which format. For instance, in response to the query: *Please provide me with a plot of the last week's values of my blood pressure*, the LLM must extract: `PRESSURE 7 GRAPH`. These parameters are used to opportunely invoke the data processing unit.

Mood Sentences referring to how the patient feels about their pathology must be specifically addressed by leveraging stored data analysis to inform patients about their condition, and possibly encourage and comfort them.

General It includes all the other questions. They can be managed by an external LLM-service via proper prompts: if they are out of scope, the chatbot will remind its tasks and will suggest to ask proper experts for an answer.

3.1.4. Data processing

The data processing unit is in charge of conducting a different set of computations over the data stored for each patient. It specifically responds to the requests identified by the NLP module (under the *request* category) by providing – over the period of time specified within the request by the user – the curve diagram of the blood pressure, or the mean or the list of all the stored values.

3.2. Prototype

This section presents a chatbot prototype designed for hypertensive patients, focusing on its implementation and the comparison of two distinct architectural solutions for data privacy and performance.

3.2.1. Chatbot for hypertensive patients

The adoption of chatbots in cardiovascular medicine holds immense potential, particularly in the context of cardiovascular prevention. Leveraging LLM-based chatbots can revolutionise patient care by providing timely, personalised, and empathetic information. For instance, chatbots can support patients by providing ad-hoc education, medication adherence and lifestyle modification guidance. They can also improve the quality of blood pressure measurement in the domiciliary setting which is often suboptimal according to a recent survey (Mancusi, Bisogni, et al., 2022). Even though they cannot substitute healthcare professionals, they can facilitate remote patient monitoring, enabling healthcare providers to track vital signs and adjust treatment plans as necessary. The efficiency of chatbots in collecting patient data and providing continuous support needs proper validation in observational studies but can potentially and significantly enhance the overall quality of care, making them a valuable tool for physicians seeking to optimise cardiovascular health outcomes. By integrating chatbots into the healthcare system, we can foster a more proactive and preventive approach to cardiovascular medicine, ultimately improving patient outcomes and reducing the burden on healthcare resources.

3.2.2. Prototype implementation

The prototype is set up as a Telegram chatbot that interacts with its users through a conversation within the messaging service's mobile app.[4] The chatbot is implemented as a .NET Core Web service and a MongoDB database for data storage. Both components run on-premises on a dedicated server in order to grant data security and isolation. The decisions we made enabled us to establish a practical prototype system to evaluate the feasibility of using LLM-based chatbots for chronic disease self-management. However, implementing this system in a real-world scenario presents several challenges. For example, while utilising local dedicated servers provides complete control over data and processing, it also necessitates resource allocation and ongoing maintenance. Furthermore, practical considerations such as scalability, backup and recovery procedures, and potential hardware failures must be addressed. These issues could potentially be mitigated by transitioning to a cloud-based service.

Basic chatbot functionality is implemented in the fashion of a simple state machine, handling basic conversation turns, while most requests are handled by a specific NLP&NLG module. The state machine enables us to control the conversation flow, understand the user's intentions, and handle sensitive data securely without disclosing it to third-party software. However, this approach can limit the interaction's flexibility and naturalness due to its reliance on predefined rules and responses. To enhance the level of naturalness where possible, we employ a separate module dedicated to natural language processing and generation.

In this paper, we devise and compare two possible solutions for the architectural model presented in the previous section. Both solutions are specifically meant to prevent sensitive data from being shared with third-party services. In particular, we propose *(i)* a solution that is grounded on two components, the first one trained to detect sentences with personal information, the second one that exploits GPT-3.5 Turbo to respond to those inputs that are not classified as potentially sensitive, and *(ii)* a solution based on open-source LLMs which are locally installed, so that personal data are not shared once processing user inputs. The first solution proposes a hybrid approach, integrating a custom classifier with a state-of-the-art LLM, to strike a balance between privacy and performance. However, this approach requires reliance on a third-party service, which could introduce some risks and limitations. The second solution operates under the assumption that open-source LLMs can deliver a satisfactory level of performance and quality for our chatbot, without compromising the user's privacy. However, exploiting open LLMs comes with a price, since downloading, installing and exploiting large models can be both costly and time-consuming. Additionally, these models may require fine-tuning to the specific domain and task, if they do not demonstrate good performances. Fig. 2 presents three snapshots of the chatbot, which were captured using the second solution with the Mixtral open LLM —one of the highest-rated models according to domain experts. These snapshots illustrate some of the key functionalities offered by the chatbot.

The purpose of this work is to compare and evaluate the feasibility of both solutions. The results of this comparison will be presented in Section 4.

[4] The bot, named *AI 4 HyperTension*, is available on Telegram.

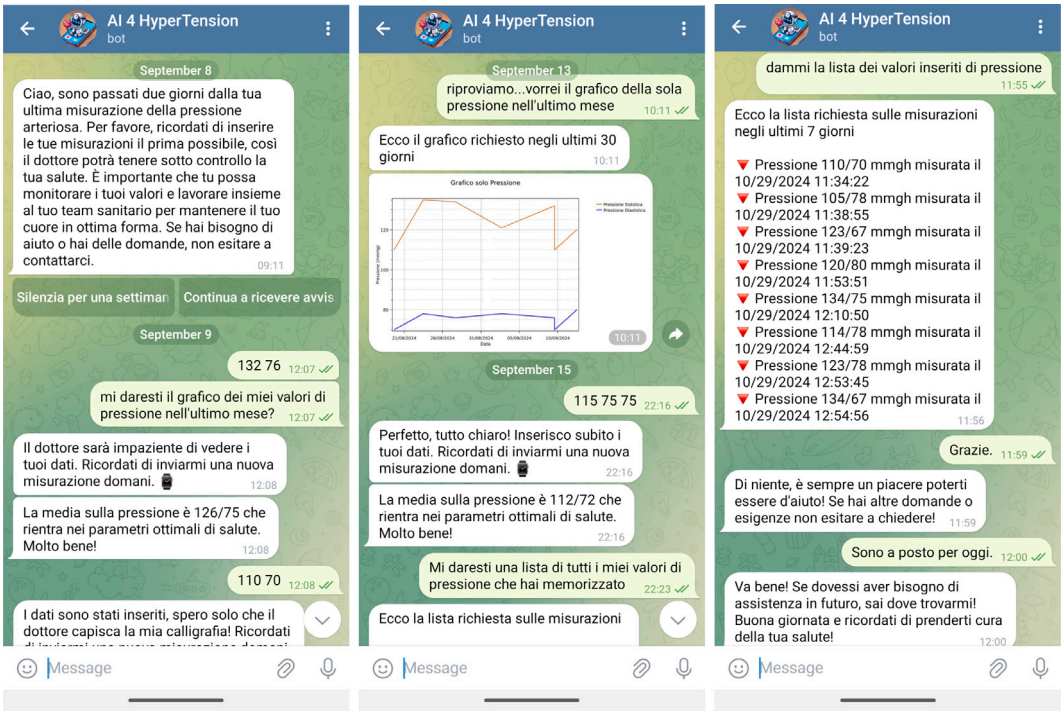


Fig. 2. Three snapshots of the *AI 4 HyperTension* chatbot are presented. The chatbot's language is Italian, as it is designed specifically for Italian patients. However, for clarity, we provide translations of the content. From left to right, the first snapshot displays a message in which the bot proactively encourages the user to acquire and provide a new blood pressure measurement. Once the measurement is entered, the system computes and provides the user with the average value of the blood pressure readings previously submitted. The second snapshot shows the plot of blood pressure measures of the last month. Moreover, the system is capable of handling a string of three values, with the last value representing the heart rate. In the third snapshot, we present an alternative data visualisation method: a list of all values stored in the database, including the date and time of acquisition.

3.2.3. ML.NET 2.0 and GPT-3.5 Turbo

This solution exploits the ML.NET 2.0 library, developed for C# application, to classify textual data and perform sentiment analysis to recognise user intent. As such, text classification can be performed through a variety of compressed models that do not suffer from big computational and memory costs. In particular, the developed code defines a ML pipeline that includes various data transformations and a multiclass classification trainer. The ML.NET 2.0 framework grounds on a pipeline that includes steps to feature text, concatenate and normalise features, and finally apply a one-versus-all multiclass classification thanks to a logistic regression binary estimator. This is the specific pipeline used:

- *FeaturiseText*: which transforms a text column into a featured vector that represents normalised counts of n-grams, this is essential to use the words as input to our regression model;
- *Concatenate*: which concatenates the obtained values into a single one;
- *Normalise*: which scales the features to a range between 0 and 1;
- *One Versus All Multiclass Classification*: that trains a set of binary classifiers, one for each class. Each predictor is trained to distinguish one class from all other classes. It uses L-BFGS logistic regression as the binary classifier.

The 80% of a set that simulates a variety of possible patient queries – defined by domain experts – is adopted for training a Logistic Regression model with the One Versus All strategy. Each sentence is labelled with one of the four intents defined above.

Downstream of the first classification module, those sentences that are not classified as containing sensitive data, are passed to GPT-3.5 Turbo with different prompts:

- Request: the prompt instructs the LLM to extract the parameters from the request.
- Mood: the LLM is instructed to generate short text, possibly reassuring the user or suggesting to redirect their queries to a medical expert.
- General: the LLM is instructed to limit the response only to the clinical condition it is devised for, reminding the user the tasks of the chatbot, giving the minimal safe set of medical information, in requests are pertinent (e.g., how can I measure my pressure? results in:

Table 1
Results of the classification phase for all messages. Models used in the experiments are reported in the first column. The next two macro columns – precision and recall – report the corresponding metric per single class (general, insertion, request and mood). The last column shows the overall accuracy of the models.

Model	Precision				Recall				Accuracy
	General	Insertion	Request	Mood	General	Insertion	Request	Mood	
Alfred (42b)	0.39	0.81	0.21	0.12	0.42	0.43	0.78	0.25	0.46
Llama2 (70b)	0.39	0.81	0.87	1.00	0.77	0.70	0.94	0.41	0.74
Llama2 (13b)	0.16	0.42	0.10	1.00	0.31	0.36	1.00	0.15	0.28
Llama2 (7b)	0.39	0.43	0.49	0.69	0.42	0.61	0.58	0.19	0.46
Mistral (7b)	0.40	0.71	0.45	0.19	0.38	0.61	0.85	0.08	0.51
Mixtral (8b)	0.61	0.73	0.84	0.81	0.66	0.68	0.95	0.54	0.74
ML.NET 2.0 framework	0.94	0.99	0.99	0.84	0.95	0.95	0.98	1.00	0.96

> To measure your blood pressure, sit calmly for 5 min. Place the cuff on your upper arm, aligning it with your heart. Press start on the monitor. The cuff inflates, measuring systolic and diastolic pressures. Record the values, noting any unusual readings. Consult a healthcare professional for interpretation and guidance.

3.2.4. Open-source LLMs

In this prototype, we utilise the REST API service named ollama, capable of hosting various LLMs. These range from smaller, like Mistral, to larger foundational models like Llama2 70 billion. Our choice of this service was influenced by its API, which aligns well with the GPT family, specifically liteLLM.[5] This ensures a seamless transition between the two solutions without necessitating any changes to our current codebase. We specifically assess the performance of Alfred, Llama2 (70, 13 and 7 billion), Mistral and Mixtral. This selection is driven by our goal to match the original GPT-3.5 Turbo performance while ensuring the application remains responsive. In particular, we focus on a smaller model that can load and respond swiftly.

Our primary methodology for instructing these LLMs to execute our specified tasks relies heavily on *prompt engineering*, thereby avoiding exposing users to prompt crafting, which is a complex task that may lead to suboptimal results. This technique involves creating *system prompts* (i.e., specifically, texts that are integrated into each query within the chat session) designed to generate specific text outputs. We developed specialised prompts for each function, such as sentence classification and response generation. These prompts are then refined using GPT-4, in line with recent advances in GPT solutions. To implement the architecture outlined in Fig. 1, we needed to utilise three distinct prompts, specifically:

- Sentiment analysis prompt: This text is used to categorise a sentence based on the described content.
- Request handling prompt: After selecting the request, there is a second phase where it is necessary to determine the type of request made, particularly by selecting the time range, the requested data (i.e., blood pressure or heartbeat), and the format (i.e., list, average graph). To achieve this, we created three prompts used in parallel to extract the required information.
- Response to mood and general inquiries prompt: Here, the language model should act as if it were a doctor, responding in a concise yet clear and reassuring manner. We also tried to configure the bot to respond in this way.

Given these prompts, the maximum number of requests per message is four, as a) one is to understand the category and (b) in cases where the category is a request, there is a scope for an additional three calls to the model. However, since the responses are typically short (a word), the collective performance does not suffer. Even the largest model on a server machine can produce two words per second, even when run on a CPU.

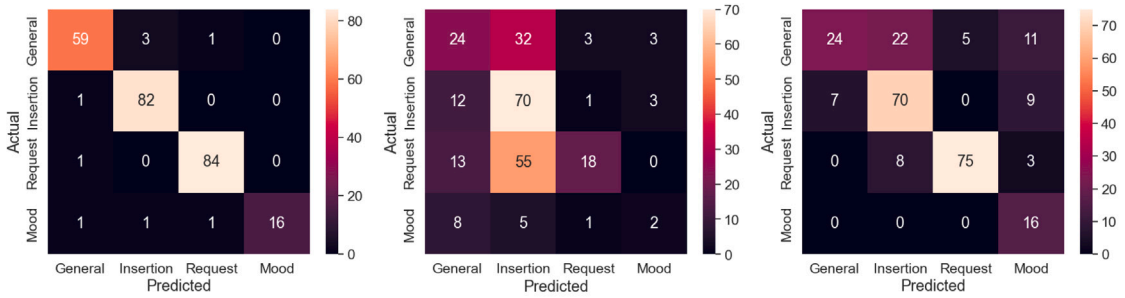
4. Results

The evaluation[6] and comparison of the two architectures is based on three main tests, the first is related to the component related to the intent analysis, the second one is the one related to request handling and the third one is related to the semantics of the responses. For the sake of reproducibility, the specific prompts used to query the LLMs, as well as the model hyperparameters and the dataset used, are available in the linked repository.

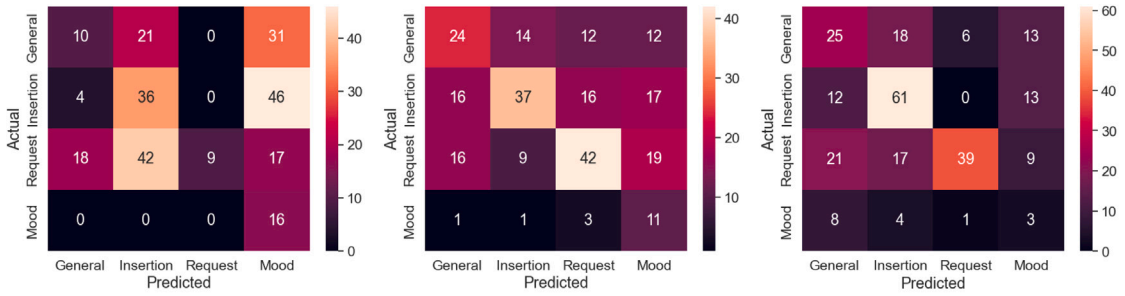
4.1. Intent recognition evaluation

To assess the effectiveness of the component dedicated to intent recognition, we generate performance metrics for the trained model and the instructed open LLMs with the sentiment analysis prompt by comparing the output labels with our ground truth. The comparative analysis, as illustrated in Fig. 3 and Table 1, reveals that the trained ML.NET 2.0 models significantly outperform

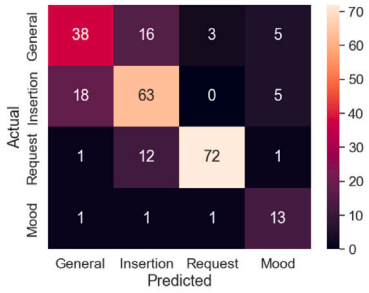
[5] https://github.com/BerriAI/litellm.
[6] Code available at https://github.com/cric96/chatbot-test-llm.



(a) ML.NET 2.0 framework (b) Alfred (c) Llama2 70b



(d) Llama2 13b (e) Llama2 7b (f) Mistral



(g) Mixtral

Fig. 3. Confusion matrices for the classification of the patient input into the four categories we identified, which correspond to the intents associated with each question. Interestingly, the ML.NET 2.0 framework, that we trained on a variety of labelled patient queries, exhibits superior performance, for this specific task, compared to other LLMs that were not fine-tuned for this purpose.

Table 2
Results of the analysis phase for request messages. The first column describes the model used in the experiments (116 queries in total). The following four columns report the accuracy for the measure, the quantity, the format and for all combined.

Model	Accuracy			
	Measure	Quantity	Format	Overall
Alfred (42b)	0.59	0.59	0.72	0.32
ChatGPT3.5 (?b)	0.77	0.79	0.96	0.62
Llama2 (70b)	0.71	0.76	0.71	0.45
Llama2 (13b)	0.56	0.52	0.82	0.37
Llama2 (7b)	0.35	0.47	0.75	0.23
Mistral (7b)	0.48	0.56	0.67	0.28
Mixtral (8b)	0.15	0.26	0.58	0.13

other (LLMs). This superior performance can be attributed to the fine-tuning phase performed with the ML.NET 2.0 framework. Additionally, it is noteworthy that smaller models demonstrate limitations in responding accurately in our target language, Italian,

Table 3
Evaluation of LLMs responses via BERTScore: this analysis presents the outcomes of assessing open LLM's performances using BERTScore, based on a set of 128 questions and using as reference GPT-3.5. The report includes average values of precision, F1 score, and recall as calculated through BERTScore metrics.

Model	Bert score		
	Precision	Recall	F1
Alfred (42b)	0.70	0.73	0.72
Llama2 (70b)	0.68	0.73	0.71
Llama2 (13b)	0.67	0.73	0.70
Llama2 (7b)	0.67	0.72	0.69
Mistral (7b)	0.67	0.72	0.69
Mixtral (8b)	0.67	0.72	0.69

suggesting a correlation between model size and language proficiency. Despite that, the biggest models – Llama2 70b and Mixtral – give consistent results and have a good accuracy of 74%. This insight emphasises the importance of model scale and training in achieving high linguistic accuracy, especially in language-specific applications.

4.2. Data extraction evaluation

The second verification carried out involves the extraction of data from user requests after the categorisation phase. In this case, we compared the responses from GPT-3.5 Turbo against those from our open-source models. Here, GPT-3.5 Turbo utilised a single prompt, in contrast to the three prompts used by our LLMs. These findings are summarised in Table 2. This analysis clearly shows the GPT-based solution consistently outperforms the alternative. Although the 70-billion parameter model shows similar performance in terms of selecting the measure and time range, it falls short in finding the right format. This is often due to the difficulty in classifying this type of information. For example, the word "visualise" could either mean a request to graphically represent data or simply to present it in a textual format. Furthermore, the smaller models perform worse than the reference model, highlighting a correlation between model size and its capability to handle complex data extraction and interpretation tasks.

4.3. Semantic evaluation

We compare the semantics between the response with *bertscore* (Zhang, Kishore, Wu, Weinberger, & Artzi, 2020). This is a metric for evaluating machine-generated text, based on the transformer architecture like BERT. It calculates the semantic similarity between reference and generated text, using vector representations of words. It is used to evaluate automatic translations and other natural language generation tasks. In this case, since we did not have a ground truth (i.e., the correct expected answers for each question, which may be, for instance, those provided by a domain expert), we decided to use the response from GPT-3.5 Turbo as a reference and check how much the models differed from the given response. This assumption is based on the findings reported in the previous section, especially those of Table 1, which identifies GPT-3.5 Turbo as the best-performing model. Moreover, this assertion is permissible given the objective of this evaluation, which pertains to the degree of semantic similarity among responses rather than their qualitative efficacy.

The results are shown in Table 3. The six open models demonstrate similar performance but fall significantly short of GPT-3.5 Turbo. However, it is important to emphasise that this comparison does not verify that the responses are contextually accurate.

To further assess and compare the responses provided by different chatbot models, we involved human reviewers, who were able to judge the context suitability of the responses. In this phase, we include only domain experts, namely physicians with specific specialities in Internal Medicine. The primary objective of this evaluation was to discern which model offered the most reasonable answers for an expert in the given domain, addressing a variety of questions, even those deemed off-topic. 30 responses to mood and general questions, for each of the seven models we are testing, for a total of 210 question-answer sessions, have been submitted for evaluation hiding the generating model. Domain experts provided an integer score for each of them in the range 1 (very bad) - 5 (very good). The evaluation focused on several criteria to gauge the coherence of the chatbot responses. These criteria included accuracy and relevance to the query, contextual understanding, logical reasoning, and the ability to provide informative and coherent responses that align with the expectations of an expert in the field. By emphasising these criteria, the evaluation aimed to identify the model that consistently demonstrated a high level of proficiency in delivering accurate and contextually appropriate answers across diverse topics, reflecting the responses a chatbot might provide to a patient.

In Fig. 4, we present the outcomes of this assessment. While not intended to be statistically rigorous, it aims to offer insights into the user experience of physicians seeking accurate responses and effective handling of user queries, essential for adoption within clinical settings. Even though, GPT-3.5 Turbo outperforms all the Llama2 versions and Mistral, Alfred and Mixtral provides excellent performances according to medical domain knowledge, requirements and user experience.

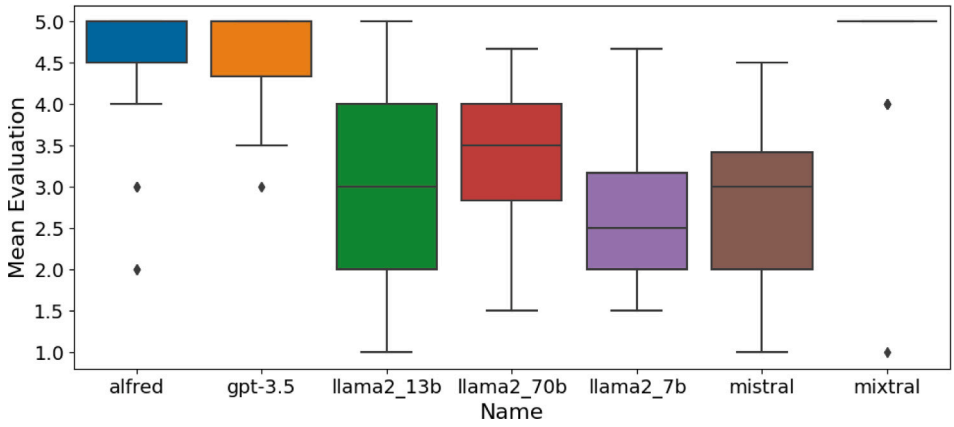


Fig. 4. Domain experts evaluated the user experience of 210 responses, generated equally by seven models, on a scale of 1 to 5. GPT-3.5, Alfred, and Mixtral received high scores, while Llama and Mistral received only sufficient evaluations. These results highlight that Llama models and Mistral may require further optimisation to meet expert expectations.

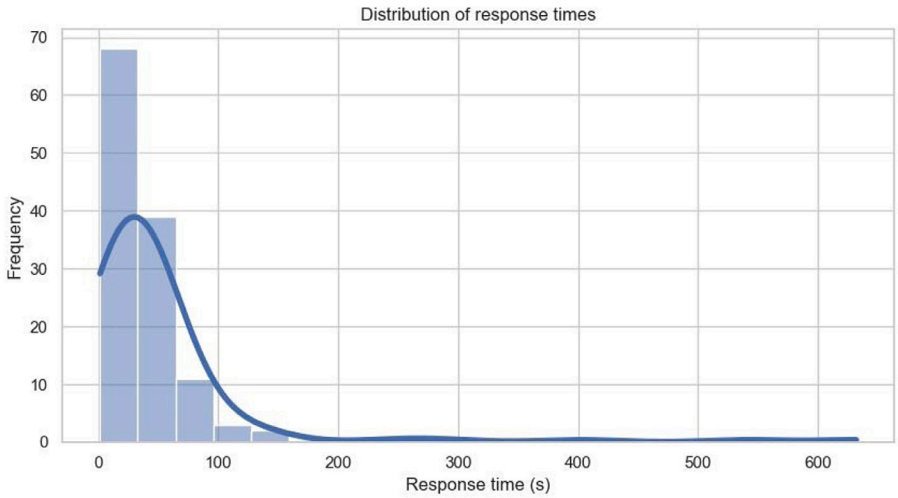


Fig. 5. Distribution of response times of the chatbot to users' requests. The number of responses at the time of writing – 14th January 2025 – is 150. The first recorded user's message is dated 3rd October 2024.

4.4. User satisfaction

To provide a first evaluation of user satisfaction once using the chatbot, we focused our analysis on usability of the framework in terms of response time. Accordingly, we collected statistics from the chatbot. Fig. 5 shows the response times of the chatbot to users' requests. Most requests are answered in less than a minute. We notice that very few responses took many minutes. This can be justified by network connection issues. The fastest response took 1.45 s, the slowest 631.93 s, while the average response time is 51.08 s.

5. Discussion

In this paper, we are taking concrete steps to explore the opportunities and limitations of LLM-based chatbots in medicine, in particular in the context of chronic disease self-management. The peculiar characteristics of the context force us to deal with a set of issues that impose caution once delegating NLP&NLG to an LLM that directly interacts with a patient, without the mediation of an expert. Particular attention should be given to the reliability and accuracy of the LLM's answers and to sensitive data protection. Accordingly, specific solutions must be envisioned.

In this paper, we devised and tested two alternative solutions. Exploiting a model of the GPT family seemed the most obvious choice, since they demonstrated the best performances in all tasks, compared to the other pre-trained LLMs.[7] Properly prompted, it can easily lead to providing the minimum set of safe responses. to the patient, to avoid the disclosure of possible misinformation, still resulting empathetic, reassuring, and reliable. However, some considerations can be made about this solution:

1. The risk of disclosing sensitive data and of information-leakage still remains if the first filtering module fails: in particular also general questions may contain sensitive information in sentences not included in our training dataset;
2. The call of a GPT-based model comes with the payment of a subscription. For this reason, in this paper we used one of the cheapest OpenAI's models;
3. The free plan is limited in the number of overall tokens exchanged, in the number of requests/per minute and in the number of interactions/per day. As such, the solution does not scale with the number of users;
4. The dependence from third party services is a risk to be considered.

About the first point, from a privacy perspective, the principal concern arises when a category requiring primarily local processing (namely in the case of the Insertion and Mood category) is incorrectly classified as general. An information leak occurs in those cases because the query is unnecessarily transmitted to the remote GPT service. Using our content recognition filtering we observed a single error out of 83 insertions and one out of 19 mood classifications. While these error rates are low, we acknowledge that even infrequent misclassifications can be problematic. Given that, in principle, any incorrectly classified data is unacceptable within this context, the privacy implications remain a significant concern, and this solution, despite its trade-offs, cannot be used in this scenario.

The adoption of open models locally deployed resolves all these issues by avoiding sharing any information with external services, but results in a significant loss of response quality, as demonstrated in Fig. 3, Tables 1 and 2. The ability of open models to correctly classify questions and properly extract parameters is strongly compromised in the solution that leverages open models, possibly requiring a fine-tuning to addressing these specific tasks. However, the landscape of open models is continuously evolving, and new models with improved performance may be trained. However, Fig. 4 brings forth a new perspective, where Alfred and Mixtral emerge as two open models whose responses to general questions are highly endorsed by domain experts. Moreover, this second solution, which does not require any specific training, opens to new perspectives. For instance, if the chatbot is to be extended to a multilingual version, it offers significant advantages. While the first solution would require training the intent-recognition module on new set of sentences in other languages, the second solution would handle this extension automatically, as the models themselves are trained to manage multiple languages.

The paper findings thus highlight the delicate balance between privacy and improved patient interactions, essential for secure and informed healthcare communication, but at the same time that new open models deserve attention in future research aimed at enhancing their efficacy, particularly in areas where GPT-3.5 Turbo demonstrated superior performance in this study.

Threats to validity. This study has potential limitations – due to legacy requirements within our project – related to the LLMs employed: the closed and open LLMs used (e.g., GPT-3 and LLaMa 2) have been replaced by more recent state-of-the-art models like GPT-4 and LLaMa 3. Even though the findings of this study about the performance may not be directly applicable to these newer models, the core discussion regarding the trade-offs between these two remains relevant (i.e., privacy vs. performance) as similar issues persist with newer models.

Future work. Accordingly, further investigations can consider a fine-tuning phase for each distinct task (i.e., sentence categorisation and general response), thereby enhancing the model's performance capabilities. This approach also offers a potential solution to language inconsistencies we encountered in Open-LLMs responses. Furthermore, RAG can be considered to leverage the dataset that we collected for our comparison/training of ML.NET 2.0 framework, enabling the model to generate responses more in line with pre-existing outputs. Additionally, advanced prompt engineering strategies, such as the "chain of thoughts" (Wei et al., 2022) technique, could significantly improve the model's response accuracy. Lastly, exploiting larger models, such as LlaMa 3.1 405b,[8] may provide enhancements similar to those observed in GPT-3.5 Turbo, presenting a promising direction for future research, even though the findings of this study underscore that model size alone does not guarantee enhanced performance.

It is worth mentioning that, to enhance the practical relevance of the chatbot, future evaluations will be necessary to collect feedback from patients and better assess the system's usability and its impact on the patient experience. Additionally, a clinical pilot study will further support this work by providing insights into whether the adoption of such a system can improve patient adherence and demonstrate that LLM-based chatbots could be powerful healthcare solutions for supporting chronic illness management.

CRediT authorship contribution statement

Sara Montagna: Writing – review & editing, Writing – original draft, Validation, Supervision, Conceptualization. **Stefano Ferretti:** Writing – review & editing, Conceptualization. **Lorenz Cuno Klopfenstein:** Software, Conceptualization. **Michelangelo Ungolo:** Software. **Martino Francesco Pengo:** Writing – review & editing, Validation, Conceptualization. **Gianluca Aguzzi:** Writing – original draft, Validation, Software, Investigation. **Matteo Magnini:** Writing – original draft, Validation, Software, Investigation, Conceptualization.

[7] https://lmsys.org/blog/2023-06-22-leaderboard/.

[8] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Statements of ethical approval

This study has been conducted involving no patient, no volunteer, and no animal. No personal or sensitive data has been exploited.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GitHub's Copilot in order to speed up their writing and OpenAI's ChatGPT to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Fundings

The authors declare that there was no financial support for this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Dataset and code is available on the GitHub repository https://github.com/cric96/chatbot-test-llm.

References

Act of Congress enacted by the 104th United States Congress (1996). Health Insurance Portability and Accountability Act of 1996 (HIPAA). URL https://www.congress.gov/bill/104th-congress/house-bill/3103/text. 191. Public law 104.

Bortoli, M., Fiore, M., Tedeschi, S., Oliveira, V., Sousa, R., Bruschi, A., et al. (2024). GPT-based chatbot tools are still unreliable in the management of prosthetic joint infections. *Musculoskeletal Surgery*, *108*(4), 459–466. http://dx.doi.org/10.1007/s12306-024-00846-w.

Cascella, M., Semeraro, F., Montomoli, J., Bellini, V., Piazza, O., & Bignami, E. (2024). The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. *Journal of Medical Systems*, *48*(1), 22. http://dx.doi.org/10.1007/s10916-024-02045-3.

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J.-N., Laleh, N. G., et al. (2023). The future landscape of large language models in medicine. *Communications Medicine*, *3*(1), 141. http://dx.doi.org/10.1038/s43856-023-00370-1.

European Parliament, & Council of the European Union (2016). Regulation (EU) 2016/679 of the European parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679. Official Journal of the European Union.

Haltaufderheide, J., & Ranisch, R. (2024). The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). *NPJ Digital Medicine*, *7*(1), 183. http://dx.doi.org/10.1038/s41746-024-01157-x.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., et al. (2024). Mixtral of experts. http://dx.doi.org/10.48550/ARXIV.2401.04088.

Jiang, A. Q., et al. (2023). Mistral 7B. http://dx.doi.org/10.48550/ARXIV.2310.06825, CoRR abs/2310.06825. arXiv:2310.06825.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, *158*, Article 106848. http://dx.doi.org/10.1016/j.compbiomed.2023.106848.

Kurniawan, M. H., Handiyani, H., Nuraini, T., Hariyati, R. T. S., & Sutrisno, S. (2024). A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. *Annals of Medicine*, *56*(1), Article 2302980. http://dx.doi.org/10.1080/07853890.2024.2302980.

Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). 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*. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Li, H., Moon, J. T., Purkayastha, S., Celi, L. A., Trivedi, H., & Gichoya, J. W. (2023). Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, *5*(6), e333–e335. http://dx.doi.org/10.1016/S2589-7500(23)00083-3.

Mancusi, C., Bisogni, V., et al. (2022). Accuracy of home blood pressure measurement: the accurapress study – a proposal of Young investigator group of the Italian hypertension society (società Italiana dell'ipertensione arteriosa). *Blood Pressure*, *31*(1), 297–304. http://dx.doi.org/10.1080/08037051.2022.2137461.

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Medicine*, *6*, http://dx.doi.org/10.1038/S41746-023-00873-0.

Montagna, S., Aguzzi, G., Ferretti, S., Pengo, M. F., Klopfenstein, L. C., Ungolo, M., et al. (2024). LLM-based solutions for healthcare chatbots: a comparative analysis. In *2024 IEEE international conference on pervasive computing and communications workshops and other affiliated events* (pp. 346–351). http://dx.doi.org/10.1109/PerComWorkshops59983.2024.10503257.

Montagna, S., Ferretti, S., Klopfenstein, L. C., Florio, A., & Pengo, M. F. (2023). Data decentralisation of LLM-based chatbot systems in chronic disease self-management. In *Proceedings of the 2023 ACM conference on information technology for social good* (pp. 205–212). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3582515.3609536.

Montenegro, J. L. Z., da Costa, C. A., & da Rosa Righi, R. (2019). Survey of conversational agents in health. *Expert Systems with Applications*, *129*, 56–67. http://dx.doi.org/10.1016/j.eswa.2019.03.054.

Preum, S. M., Munir, S., Ma, M., Yasar, M. S., Stone, D. J., Williams, R. D., et al. (2021). A review of cognitive assistants for healthcare: Trends, prospects, and future directions. *ACM Computing Surveys*, *53*(6), 130:1–130:37. http://dx.doi.org/10.1145/3419368.

Sulis, E., Mariani, S., & Montagna, S. (2023). A survey on agents applications in healthcare: Opportunities, challenges and trends. *Computer Methods and Programs in Biomedicine*, *236*, http://dx.doi.org/10.1016/j.cmpb.2023.107525.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023/08/01). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. http://dx.doi.org/10.1038/s41591-023-02448-8.

Tian, S., Jin, Q., Yeganova, L., Lai, P.-T., Zhu, Q., Chen, X., et al. (2024). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*, *25*(1), bbad493. http://dx.doi.org/10.1093/bib/bbad493.

Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. http://dx.doi.org/10.48550/ARXIV.2307.09288, CoRR abs/2307.09288. arXiv:2307.09288.

Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical Considerations of Using ChatGPT in Health Care. *Journal of Medical Internet Research*, *25*, Article e48009. http://dx.doi.org/10.2196/48009.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), *Advances in neural information processing systems 35: annual conference on neural information processing systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Xu, L., Sanders, L., Li, K., & Chow, J. C. L. (2021). Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. *JMIR Cancer*, *7*(4), Article e27850. http://dx.doi.org/10.2196/27850.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, URL https://openreview.net/forum?id=SkeHuCVFDr.

Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), *Advances in neural information processing systems 36: annual conference on neural information processing systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Zichichi, M., Ferretti, S., D'Angelo, G., & Rodríguez-Doncel, V. (2022). Data governance through a multi-DLT architecture in view of the GDPR. *Cluster Computing*, *25*(6), 4515–4542. http://dx.doi.org/10.1007/s10586-022-03691-3.