sensors

MDPI

Article

eHealth Assistant AI Chatbot Using a Large Language Model to Provide Personalized Answers through Secure Decentralized Communication

Iuliu Alexandru Pap [1,*] and Stefan Oniga [1,2]

[1] Department of Electric, Electronic and Computer Engineering, Technical University of Cluj-Napoca, North University Center of Baia Mare, 430083 Baia Mare, Romania; stefan.oniga@ieec.utcluj.ro

[2] Department of IT Systems and Networks, Faculty of Informatics, University of Debrecen, 4032 Debrecen, Hungary

* Correspondence: iuliu.pap@ieec.utcluj.ro

Abstract: In this paper, we present the implementation of an artificial intelligence health assistant designed to complement a previously built eHealth data acquisition system for helping both patients and medical staff. The assistant allows users to query medical information in a smarter, more natural way, respecting patient privacy and using secure communications through a chat style interface based on the Matrix decentralized open protocol. Assistant responses are constructed locally by an interchangeable large language model (LLM) that can form rich and complete answers like most human medical staff would. Restricted access to patient information and other related resources is provided to the LLM through various methods for it to be able to respond correctly based on specific patient data. The Matrix protocol allows deployments to be run in an open federation; hence, the system can be easily scaled.

Keywords: eHealth; mHealth; telehealth; telemedicine; remote patient monitoring; Internet of Things; artificial intelligence; large language model; Matrix open communication protocol

check for updates

Citation: Pap, I.A.; Oniga, S. eHealth Assistant AI Chatbot Using a Large Language Model to Provide Personalized Answers through Secure Decentralized Communication. *Sensors* **2024**, *24*, 6140. https://doi.org/10.3390/s24186140

Academic Editors: Tayeb Lemlouma, Sébastien Laborie, Yevgeniya Kovalchuk and Abderrezak Rachedi

Received: 18 May 2024
Revised: 26 August 2024
Accepted: 20 September 2024
Published: 23 September 2024

1. Introduction

In a post-pandemic scenario, the effort of relieving some of the stress that the entire healthcare system had to cope with has led to a growing amount of research regarding implementations incorporating artificial intelligence. As one of the technologies that has a significant impact on multiple sectors, we believe that this proof-of-concept implementation could help facilitate the first steps needed to advance similar healthcare systems towards better performance, fluid human–machine interaction, improved patient support, actual doctor–patient time gains, and reduced costs.

The scope of this paper is to propose an architecture for a modular system capable of interconnecting existing cross-platform chat applications with AI large language models and previously recorded physiological or medical patient data. This study focuses on designing a suitable system architecture that can accomplish the scope and implementing an initial proof of concept to demonstrate its technical feasibility. With this work, we do not intend to improve one specific field of eHealth but to provide a foundation architecture on which specialized architectures can be built.

The two main objectives of our study are the following:

1. Ensure secure communications for the entire system;
2. Provide an AI chatbot with user-specific data isolation.

To accomplish our first objective, we utilize a modern, encrypted, and decentralized communication platform for all types of messaging between medical staff, the AI chatbot, and patients that also allows users to send or receive prescriptions, medical information, instructions, or any type of messages (text, image, audio, video, or file).

The second objective consists of implementing an eHealth AI chatbot that can be accessed only through the aforementioned communication platform that safely allows patients to ask the chatbot any question, even about their own physiological data, prescriptions, pharmaceutical instructions, etc. To achieve this, we customize the context of each question with specific user information so the chatbot will provide helpful and personalized answers based on the actual patient's data, and if there is not enough information, it will answer accordingly, without misleading the user.

The decision of complementing our previously built eHealth system [1,2] with an artificial intelligence large language model capable of delivering unparalleled functionality through a secure messaging protocol was made after considering all the pros and cons of mixing such technologies with healthcare applications, as reviewed in our previous work [3]. We decided that our current work should be entirely independent of our previously built eHealth data acquisition system so that by using the same method, we could adapt our eHealth AI chatbot to any data source.

1.1. Related Work and Challenges

Through the following paragraphs, we present the most relevant eHealth implementation scenarios we found in the works we studied, highlighting some research gaps, pros, and cons, to clarify our motivations for the current work.

We will focus on the following research gaps:

1. Works relevant to ours are based on ChatGPT, which is a paid third-party AI service; hence, it can increase costs and decrease medical information privacy;
2. Security of communications between users and AI is controlled by third parties;
3. We were unable to find any eHealth AI chatbots that use a mature, open protocol to secure communications;
4. There is no seamless communication between patients, medical staff, and AI;
5. There are no local AI chatbots that use Retrieval-Augmented Generation based on the patient's medical information to offer relevant answers;
6. AI chatbots used in eHealth applications do not restrict the AI to a user-specific context, hence delivering more imprecise or wrong answers;
7. There are no interchangeable AI open model implementations in eHealth;
8. Very few eHealth LLM applications use exclusively open-source software.

Bringing AI technologies into healthcare can be undertaken for several reasons, among which the review in [4] presents the following:

1. Medical notes: medical professionals can use ChatGPT in writing medical records, clinical notes, and related reports;
2. Education: ChatGPT can give medical professionals and students access to additional information and resources;
3. Medical consultation: initial medical consultations, patient information and test results can be summarized to help physicians;
4. Patient triage: ChatGPT can help with patient triage by asking patients about their symptoms and previous medical conditions, attempting to determine the priority of their medical condition;
5. Virtual assistants: with rising telemedicine popularity, creating a virtual assistant that can help patients with making appointments, administering treatment, and handling health records, all from the comfort and safety of their homes, can be very appealing in distant locations;
6. Clinical use: this includes augmenting medical robots, offering dietary advice, or even explaining the dangers behind smoking and other harmful habits.

In trying to build a comprehensive eHealth AI application, we explored related works and searched for solutions to the problems most eHealth applications face. From reviews such as [5], before the launch of ChatGPT, we can observe the variety of natural language processing (NLP) approaches, the difficulties of building applications with natural language understanding (NLU), and how performant pre-trained artificial intelligence language models are compared to other embedding algorithms while taking into account their large size and the expensive training costs. Health literacy, remote assistance, and triage are some of the aspects that can be improved when adopting ChatGPT in low- to medium-income countries (LMICs), but the process needs to take into account several factors like cultural, ethical, and illiteracy rates [6]. We believe that employing a third-party service in the healthcare system of LMICs can pose additional concerns like compromising privacy, liability, and ethics for the sake of having access to AI services. These challenges made us consider implementing remote model access and finding suitable pre-trained models to reduce running cost and increase adoption rate.

One of the first studies [7] that examined ChatGPT's ability to provide answers to disease-specific questions, evaluating the correctness and the reproducibility of its replies to frequently asked questions, concluded that ChatGPT was able to identify the patient's emotional response, offered correct first steps for newly diagnosed patients and proactive measures for treatment, and recommended caregivers to help patients find support groups while underlining the importance of mental and physical health for both patients and caregivers. The study [7] excluded user-specific questions and only focused on general answers, which is why we aim to provide the means of asking user-specific questions through our architecture's implementation.

A study of 54 teleconsultants in healthcare [8] examined how their operations and services were affected by the use of ChatGPT, with positive aspects including increased efficiency, reduction of cost, and diagnostic assistance, while some of the negative impacts included misdiagnosis, reduced medical context and knowledge, and security and privacy concerns related to the fact that sensitive medical information needed to be sent to ChatGPT. Sending sensitive medical information to an online third-party service would raise many concerns, so self-hosting the LLM would allow us to benefit from ChatGPT-like services that bring increasing advantages to healthcare implementations and have their place in future healthcare applications, even if currently, they cannot surpass human execution for certain tasks, as noted in [9]. Pharmacy-related questions can imply greater complexity, and according to [10], even if ChatGPT showed a 65% and a remarkable 85% performance rate for certain categories, it failed with 30–35% rates for drug interaction and dosage categories, highlighting the much-needed improvements the model requires. Additionally, in the world of telepharmacy, both ChatGPT 3.5 and 4 show great performance when asked to act as a pharmacist, offering responses that were relatively safe [11]. Public health dentistry is another field where ChatGPT can have beneficial effects, but as stated in [12], it cannot replace a dentist because it is not able to assist beyond recommending a diagnosis, so in our implementation we need to steer the model towards stating it is not a healthcare professional and refusing to answer without valid context data. To accomplish this, we needed to find an alternative to ChatGPT that we could have more control over. If we do not take these steps, by integrating large language models into chatbots, users could mistakenly think they are presented with specialist medical information, while in reality, AI products are still unable to provide reliable health-related knowledge [13]. ChatGPT is not designed to admit it does not know the answer. In order to avoid such confusion and spreading misinformation, we need to guide our local LLM to offer responsible answers. The latest version of ChatGPT, GPT-4, proved its potential in extracting relevant information from patients in non-English conversations by scoring on average 72.4% in the Chinese Medical Licensing Examination's Clinical Knowledge [14]. For some queries, all of the tested chatbots (ChatGPT 3.5, ChatGPT 4, BARD, and Claude 2) provided answers that were too complicated for the general public to understand, and only one (BARD) listed sources for later verification [15]. Another research gap is highlighted by the importance of providing context to AI questions, whether it is age, education level, or medical information. In order to safely utilize ChatGPT or similar large language models, we need to take into account the dangers of spreading misinformation, privacy issues, and the possibility of bias [16]. ChatGPT shows evident language bias, but it manages to perform similarly regardless of the language in which the question was formulated [17]. Since the training of LLMs falls outside the scope of this research, the alternative would be to replace the LLM with a less biased one, which in the case of most eHealth applications is not possible, so our solution to this issue was to use an open-source LLM manager.

Our eHealth AI chatbot could be used for theoretical questions, because when used for medical exams, ChatGPT performed well regardless of the type of question, so medical students can trust it for academic aid; however, it was unable to comprehend questions that contained images, to exhibit critical thinking, or to gain insight [18]. Also in healthcare academia, a meaningful ratio of students do not possess enough information about ChatGPT and have a positive opinion about it; hence, they have the disposition of using it for academic purposes, emphasizing the need to better understand the risks involved in using large language models by both students and educators [19]. While taking into account remarkable accomplishments such as ChatGPT's responses being often indistinguishable from human ones and its accuracy on public health exams well surpassing human accuracy, we should not ignore the risks this entails for the entire educational process, especially in the context of public health education [20]. Accuracy-wise, there is a considerable difference between the responses that ChatGPT generates and the actual information obtained from official sources, even if from a relevance standpoint they can outperform at times [21]. To prevent the spread of misinformation in situations where patients try to self-diagnose, the LLM should only answer questions related to the data we provide and control.

The main use case for large language models in medicine is the AI assistant tool that is intended to interpret symptoms and offer initial steps for diagnosis, treatment, and management; however, as the medical questions become more complex, the model starts facing more difficulties, so further considerations and development of large language models are important in potentially reducing the risks involved in their adoption and increasing accountability [22]. When creating lists of questions destined to help in prenatal counseling, ChatGPT can be a useful resource, but all the generated information should be checked by experts prior to including them in the training materials [23]. In the case of occupational medicine, ChatGPT's responses did not reach the accuracy of physicians' answers but managed to attain the same level of completeness as doctors' answers only when context was included along with the question, emphasizing the importance of providing additional information for the model through text embeddings [24]. In some situations, as in [14], an intermediary between patient and physician can be beneficial to overcoming some of the physicians' challenges like consulting large numbers of patients, which limits the time allocated for each patient to a few minutes; therefore, in a real-world implementation, the AI would need access to medical information with regard to data privacy. Hence, our eHealth AI chatbot must provide secure communications to safeguard sensitive patient information and potentially fix the lack of interoperability we see between eHealth applications.

Among the most problematic challenges encountered in implementing AI applications is the matter of finding a balance between technological advancement and ethics, with their multiple ramifications (legal, humanistic, algorithmic, and information ethics). Some of the most challenging parts in the development and implementation of ChatGPT in healthcare are as follows [25]:

- Following regulations like the Health Insurance Portability and Accountability Act (HIPAA) for the United States of America;
- Evaluating the potential impact of ChatGPT on human relationships as well as communication between medical staff, caregivers, and patients;
- Monitoring how ChatGPT influences empathy and trust in the healthcare setting;
- Identifying the biases in ChatGPT's language processing;
- Investigating how ChatGPT can increase or reduce systemic inequalities in healthcare;
- Checking the accuracy, reliability, and transparency of the information supplied by ChatGPT;
- Mitigating concerns about data privacy and security when handing sensitive patient information to ChatGPT for processing.

A similar approach to our RAG implementation is achieved in [26] through prompt engineering and augmenting the LLM with additional knowledge, thus creating a Chat-Orthopedist based on ChatGPT that manages to avoid hallucinations but, compared to our solution, does not provide answers based on user-specific medical information. Keeping medical information private is also the focus of [27], where they use ChatGPT to generate synthetic data to avoid sending private medical information to the ChatGPT API, this being one of the reasons we opted to use local LLMs in our architecture's implementation.

The novelty of our approach is supported by the lack of research regarding the applications of open-source LLM management software in the context of eHealth and is further strengthened by using non-ChatGPT LLMs. Following a modular approach in integrating LLMs in eHealth applications can allow healthcare solutions to seamlessly adapt to newer and better LLMs, as the number models is ever increasing.

In eHealth, the matter of information security is essential, so works like [28] offer a different approach in securing healthcare systems by using blockchain technology to implement smart contracts responsible for managing device access to data. Federated learning, the paradigm based on the distributed learning concept, can also benefit from blockchain security features [29]. Security and privacy are discussed in the context of healthcare through the metaverse [30], where the packaging of patient's medical information into a Non-Fungible Token (NFT) could improve the sharing of health records. In an attempt to improve cloud security, ref. [31] introduces a new encryption framework that is a combination of four different encryption algorithms. Another encryption method is proposed in [32], where they manage to decrease encryption/decryption times and reduce healthcare costs by using less energy. While these implementations focus on improving medical information security, one of their limitations is the fact that they do not present a clear template on which other researchers could build upon. Many eHealth researchers need to implement their own secure communications because the state of the art lacks architectural templates to support this. Hence, as a contribution for our architecture's implementation, we avoid building a security system from scratch, so we aim to integrate a tried and tested open secure communication platform.

To conclude our motivations for the current work, we list the most relevant aspects we aim to improve through our proof-of-concept implementation, proving the novelty of our work:

1. Running costs: by providing remote model access to suitable pre-trained models, we avoid spending enormous sums on training models and become independent from third-party online AI services;
2. AI answers: we instruct the LLM to state it is not a healthcare professional and to avoid answering questions without our specific context information, improving its reliability;
3. Replicability: by exclusively using open-source software components, we allow others to freely replicate our solution by simply copying and personalizing it;
4. Scalability: our system's architecture could be implemented in an entire network of healthcare facilities because the communications platform we use has this functionality built in;
5. Privacy: we avoid sending sensitive medical information to an online third-party service by self-hosting the LLM, so the privacy of the medical information is protected from online third parties;
6. Security: the vast majority of related implementations use ChatGPT or other third-party AI services, entrusting them with the security of their communications. In our solution, we self-host the LLM and utilize a secure open communications protocol to keep everything under our control;
7. Interoperability: by using an open communication protocol, we could allow other eHealth systems to communicate with our entire federation of servers.

1.2. Implementation Choices

The avalanche of large language model applications began after the success of ChatGPT, and since its unveiling, numerous businesses and institutions have started experimenting with chatbots. Among these diverse applications, some started experimenting with healthcare implementations. AI applications range from frivolous to vital ones like emergency situations [33–35], and with every additional functionality, there are many risk factors that need to be taken into consideration, like inaccuracies, wrong information, large language model hallucinations, and data privacy and security when handling patient data. With these risks in mind, knowing we need our AI model to have direct access to patient data, we determined that hosting and running a large language model locally would eliminate these security and privacy concerns, because this way, no sensitive data ever leave our system, so no third-party service has access to our data.

Self-hosting an open-source large language model addresses the security and privacy concerns regarding third-party services, so to overcome potential data privacy and communication security issues, we decided to use a decentralized open communication protocol called Matrix. After studying the protocol [36–38], we researched its adoption, which revealed the following:

1. The chat platform for officials and citizens (Luxchat4Gov) in Luxembourg is based on the Matrix open protocol [39];
2. The French government messaging service Tchap is also based on the Matrix open protocol [40];
3. Germany's united armed forces (Bundeswehr) started testing a Matrix implementation called BwMessenger [41];
4. Most importantly, the German healthcare sector's digital solutions provider, Gematik, started rolling out the TI Messenger communications platform in the last few years [42].

The Matrix open protocol offers a plethora of functionalities, and here are some of the advantages it brings to our implementation:

1. Communication encryption support;
2. The communication server can be self-hosted either locally or in a federation of private servers;
3. Multiple server software choices, all open source (Synapse, Dendrite, Conduit, Conduwuit, Construct, etc.);
4. Multiple client applications for desktop, mobile, and web (Element, Element X, FluffyChat, Quadrix, etc.);
5. Support for bridges that can allow users to be notified even on other messaging networks;
6. Advanced user management and access restrictions to conversations, rooms, and resources;
7. Possibility of managing all patient contact through application features like text conversations, voice and video calls, file attachments, and more.

2. Materials and Methods

In the following paragraphs, we offer an overview of how all the parts of our system are tied together and how it all works. It consists of multiple components that could be implemented in a variety of ways, meaning that some parts could be hosted on the same machine. Due to the nature of the Matrix protocol, we could scale the system to multiple locations and servers, providing access to our eHealth bot in any scenario.

The main components of our system are the following:

1. IoT eHealth data acquisition system [1] (optional);
2. Matrix server;
3. Large language model;
4. eHealth AI chatbot.

An overview of the system can be seen in Figure 1, where we can observe the Matrix protocol related communications in green and all the LLM-related communications in blue.

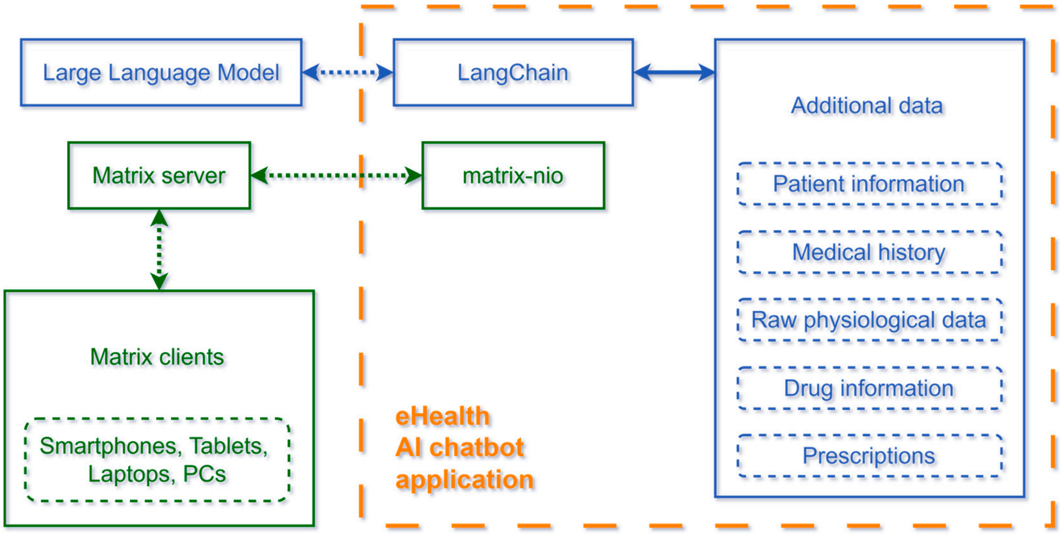


Figure 1. Overview of the entire system.

Both the LLM and the Matrix server can run on machines separate from the eHealth AI chatbot application, but this is not a requirement. For fast, almost instant chatbot replies, the LLM must run on a machine that has a modern GPU with enough memory to load the entire model in VRAM, otherwise the performance of the LLM is severely diminished.

Architecturally, we tried to keep the system modular and have a clear data flow, as described below:

1. Matrix communication: All users chat with the system through Matrix clients, which communicate with the eHealth AI chatbot through Matrix servers;
2. LLM communication: The main Python application (eHealth AI chatbot) is the only component with access to sensitive patient information. This component prepares the user's information and combines it with the user's message, forming chains of prompts that are sent to the LLM. The LLM's response is sent back to the user through the same Matrix conversation.

One of the main challenges in implementing this eHealth AI chatbot was the context window, the number of tokens our large language models can process at once (mostly 2048 tokens), especially for the ones we had enough resources to host and run locally.

In our tests, we wanted to analyze how accurate and reliable the answers of our eHealth AI chatbot were, so we extracted relevant physiological data recorded through the eHealth data acquisition system from our previous work. Because of the context window restriction and to better test our chatbot, we carefully selected and anonymized test data samples to avoid any privacy issues and be able to better evaluate the performance of our eHealth AI chatbot, creating a sample single-user test dataset.

Below, we describe each component in detail with all the essential characteristics, configurations, and approaches.

2.1. IoT eHealth Data Acquisition System

Our previously implemented physiological data acquisition system allows us to record, store, and stream raw sensor data in real time. We used this system to collect data like blood pressure, temperature, galvanic skin response, and air flow. The eHealth data acquisition system [1] is running on a Raspberry Pi 3 and can be connected directly to the eHealth AI chatbot application. In our test, we use anonymized data samples, but in a production deployment we would connect the chatbot directly to the recorded raw data.

Because our system is modular, the eHealth data acquisition system could be replaced with any other physiological or medical information data source.

2.2. Matrix Server

There are a significant number of servers available in the Matrix ecosystem.

Matrix servers [43] come in a variety of development stages (stable, beta, and alpha), licenses, and programming languages, the most important ones being, at the time of writing this paper, as follows:

1. Synapse: Python, stable;
2. Conduwuit: Rust, beta;
3. Construct: C++, beta;
4. Conduit: Rust, beta;
5. Dendrite: Go, beta;
6. Telodendria: C, alpha.

Because of the nature of our eHealth application, needing to take every security and privacy measure into account, we had to have total control over our Matrix server, so we installed and configured the Dendrite server in a Docker container on a computer running Linux. We chose Dendrite because it is still under active development and in our tests, it proved to be stable and needing fewer resources than the main Synapse server.

In a multilocation scenario, we could install several Matrix servers, and they could communicate with each other, this being useful in a nationwide implementation, where each hospital or region could have its own Matrix server to prevent any downtime.

As the Matrix protocol is open and designed with encryption in mind, we can address the issue of secure communications in eHealth by directing all our communications through the Matrix server. Also, by allowing all users (human or AI) to chat through this Matrix platform, we ensure seamless communication between patients, medical staff, and the AI chatbot.

2.3. Large Language Model

Having security and privacy restrictions, we could not use services like ChatGPT or any other third-party AI chatbot Application Programming Interfaces (APIs), because we needed to have full control over the handling of sensitive patient data.

We recognize the ethical concerns [44] regarding LLM bias and fairness, so in our modular system architecture, we opted for a plug-and-play approach to LLMs, meaning that by integrating with Ollama [45], we provided the ability to easily switch from one LLM to another, without affecting patient data, Matrix server integrations, or any other settings. While we do not evaluate or mitigate LLM ethical issues, the system administrators could effortlessly switch to another LLM that provides a higher degree of fairness.

There are multiple tools that allow us to run LLMs locally, but one of the most stable ones is Ollama. It serves an API on a specific port, so our chatbot software can connect to it similar to ChatGPT's API. Ollama supports downloading several models from their repository and using them through its API or Command Line Interface (CLI), this being our solution to the issue of interchangeable AI models. Another side to this decision is the contribution to the state of the art, as there is a shortage of non-ChatGPT eHealth AI applications, especially based on local and open LLMs, which can solve privacy issues and decrease costs.

Context sizes of LLMs can vary, but for our hardware (NVIDIA GeForce RTX 4060 Ti 16GB), most models had a context window of 2048 tokens. With the help of Ollama's API, we were able to cycle through multiple models, assessing their accuracy, speed, and consistency, determining which LLM was best suited for our chatbot. LLM performance falls outside the scope of this paper, but we had to ensure our eHealth AI chatbot would perform as expected. In our tests, we concluded that the following worked best for our application:

- Llama 3 8B from Meta;

- Phi-3 3.8B from Microsoft;
- Mistral 7B from Mistral AI.

2.4. eHealth AI Chatbot

Having such a limited context window, LLMs cannot receive substantial amounts of input along with their prompt. To personalize their response, we had to use a method called Retrieval-Augmented Generation (RAG) [46–48], implemented using LangChain, which allows the model to access relevant information in addition to its trained knowledgebase. Along with [47], we can also review the comparison between fine-tuning a model and using the RAG method, as the study highlights RAG's suitability in healthcare applications. LangChain has been used in other privacy-focused implementations like MedAide [49], where the previous version of Llama3, the Llama2 7B model, was chosen as the most performant.

After we generate embeddings for the additional data, the model accesses and indexes the embeddings with FAISS (Facebook AI Similarity Search from Meta), then allows LangChain to perform similarity searches on the data, or in our case to convert the newly created vector store into a retriever-class object, which can be used in other methods supported by LangChain.

The eHealth AI chatbot software component is written in Python (version 3.11.8) and uses two essential Python packages:

1. LangChain [50]: Can "chain" multiple components together to accomplish complex tasks. LangChain can use multiple vector stores. We use LangChain to populate the vector store, connect to the Ollama API, invoke a response from the LLM, and allow Matrix-nio to transmit it to the specific room or chat it was requested in.
2. Matrix-nio [51]: With the help of this Python package we were able to connect our chatbot software to the Matrix server, verify identity, join rooms, initiate chats, respond to questions asked by users, etc.

Using LangChain, we can create loaders for different types of information, including webpages, documents, or practically any type of information that we can access locally or remotely.

By using RAG with carefully tailored, user-specific information, our solution not only provides relevant AI answers but also eliminates LLM hallucinations and most imprecise answers.

In our tests, we provided the following types of information to the retriever:

1. JSON files: These include measurements for blood pressure, pulse, and weight, patient information (first name, last name, birthdate, patient ID, email, etc.), patient's medical conditions (condition name, onset date, confirmation date, etc.), and the prescriptions of currently administered treatments;
2. PDF files: To better control the answers about medication, we included entire leaflets of instructions about the drugs the respective patient was taking. For this scenario, we could later implement an online search component that could download these instructions from official databases and cache them for later use;
3. Extra information: We tried to remedy some of the basic flaws of LLMs, like not knowing what date it is. One of our most successful attempts was to always supply the chatbot with additional information, telling it what date and time it was right before starting to generate an answer.

3. Results

Since the focus of our work was the design and implementation of an architecture for an eHealth system capable of fulfilling our main objectives, the performance of the chosen communication platform and AI large language models were not the focus of our tests, so we only analyzed the functionality of the extra features we built in our proof-of-concept application.

We managed to implement our main objectives by integrating the tried and tested Matrix protocol and the Ollama project into a novel eHealth AI chatbot application, bridging features from these projects to solve most security, privacy, and financial concerns that prevent similar eHealth applications from being implemented successfully on a larger scale. Here we highlight the most important features of our eHealth AI chatbot system:

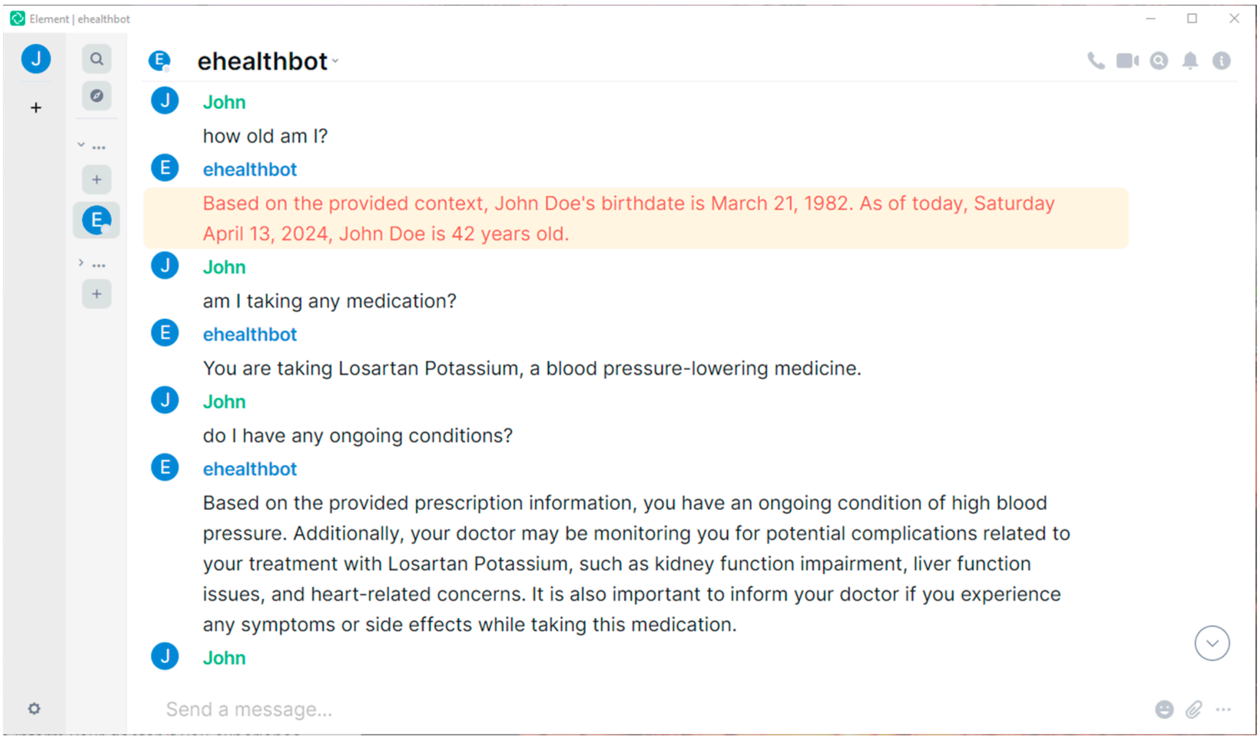
1. The Matrix communication platform, LangChain, as well as the Ollama project, are open-source;
2. The Matrix ecosystem already contains several server and client implementations, so the system administrators can choose which ones to use according to their hardware specifications. The freedom of choice also applies to users, allowing them to choose their web, desktop, or mobile application according to the device they are using;
3. The hardware necessary to run the entire system can be limited to one consumer-grade PC, and as the number of users increases, it can be easily scaled, so this implementation can be deployed even in low-income regions;
4. Matrix ensures that all the sensitive information transferred from one user to another is encrypted, thus solving an important eHealth issue;
5. Having a decentralized communication platform can provide additional benefits, like security, privacy, data ownership, resilience, independence, and federation;
6. Patients have access to their medical information, prescriptions, and medicine leaflets through an AI chatbot that is securely hosted and only communicates with its users through the Matrix protocol;
7. The AI chatbot offers personalized health information through a chat interface in a more natural and efficient way. If the requested information is missing, the chatbot will answer accordingly.

In the following paragraphs, we present our results, providing screenshots taken from the Element cross-platform application. The screenshots include question–answer pairs taken directly from the eHealth AI chatbot room, highlighting only the features corresponding to our main objectives, without including the plethora of functionalities inherited from the included software projects (Matrix, Ollama, etc.).

In Figure 2 we observe a basic interaction, where the eHealth AI chatbot recalls the patient's name from the context file provided and manages to calculate his age. For the second question, the answer is extracted from another JSON file, "prescription", where the treatment is described in detail. While the last answer is correct, the model omitted the fact that this patient also suffers from another condition for which he does not have any treatment assigned. We observe how information from a medication leaflet PDF file is intertwined with the patient's treatment data to form a comprehensive answer.

As we see in Figure 3, the eHealth AI chatbot extracted the relevant date about the start of an ongoing treatment, failed in finding the date of the initial diagnosis, but managed to correctly recommend the user against taking another medication that, according to the leaflet, can cause adverse effects when taken along with his current treatment.

Figure 4 shows the potential of LLMs, searching tens of files of blood pressure recordings and being able to extract meaningful information while formulating these findings in a syntactically correct manner, close to a human being.



John

how old am I?

ehealthbot

Based on the provided context, John Doe's birthdate is March 21, 1982. As of today, Saturday April 13, 2024, John Doe is 42 years old.

John

am I taking any medication?

ehealthbot

You are taking Losartan Potassium, a blood pressure-lowering medicine.

John

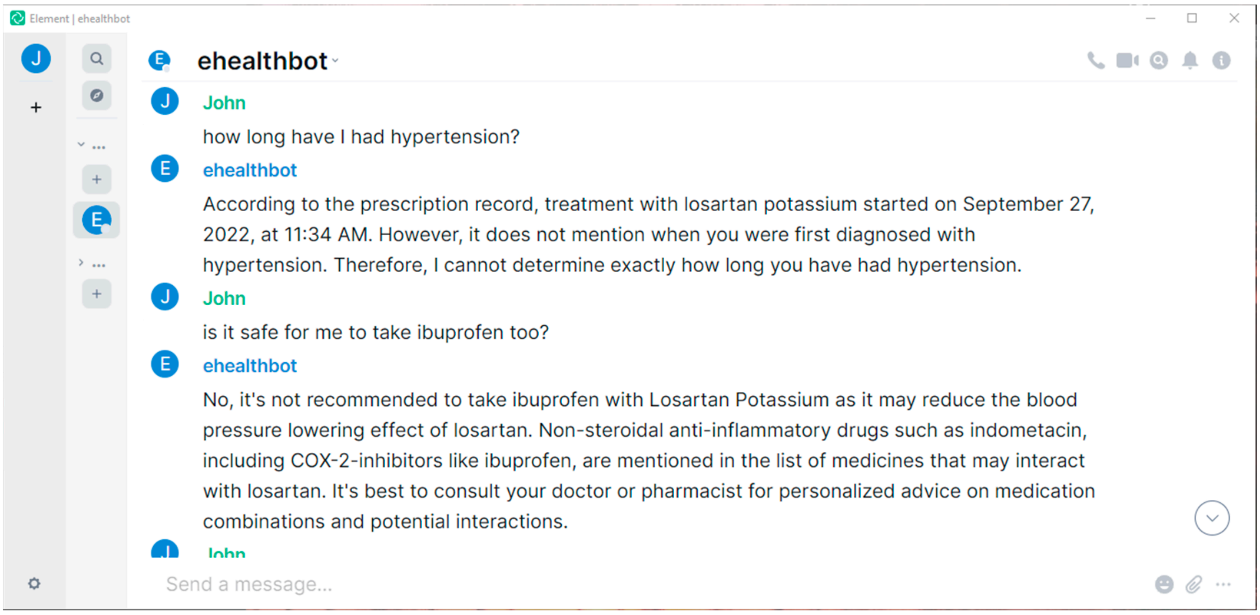
do I have any ongoing conditions?

ehealthbot

Based on the provided prescription information, you have an ongoing condition of high blood pressure. Additionally, your doctor may be monitoring you for potential complications related to your treatment with Losartan Potassium, such as kidney function impairment, liver function issues, and heart-related concerns. It is also important to inform your doctor if you experience any symptoms or side effects while taking this medication.

John

Figure 2. A patient's conversation with the eHealth AI chatbot showing how the LLM managed to recall the patient's name, calculate his age, locate treatment information, find ongoing medical conditions, and provide potential medication complications directly from the PDF leaflet.



John

how long have I had hypertension?

ehealthbot

According to the prescription record, treatment with losartan potassium started on September 27, 2022, at 11:34 AM. However, it does not mention when you were first diagnosed with hypertension. Therefore, I cannot determine exactly how long you have had hypertension.

John

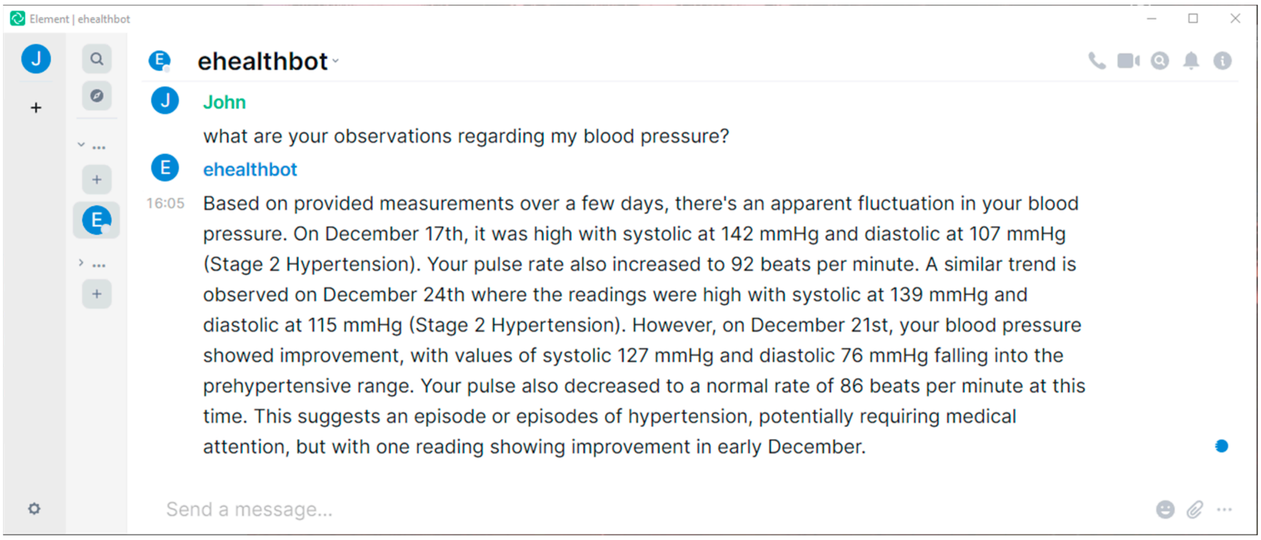
is it safe for me to take ibuprofen too?

ehealthbot

No, it's not recommended to take ibuprofen with Losartan Potassium as it may reduce the blood pressure lowering effect of losartan. Non-steroidal anti-inflammatory drugs such as indometacin, including COX-2-inhibitors like ibuprofen, are mentioned in the list of medicines that may interact with losartan. It's best to consult your doctor or pharmacist for personalized advice on medication combinations and potential interactions.

John

Figure 3. The eHealth AI chatbot managed to find the start of the treatment, failed in finding the first diagnosis, but successfully advised against mixing incompatible medications.



John

what are your observations regarding my blood pressure?

ehealthbot

16:05 Based on provided measurements over a few days, there's an apparent fluctuation in your blood pressure. On December 17th, it was high with systolic at 142 mmHg and diastolic at 107 mmHg (Stage 2 Hypertension). Your pulse rate also increased to 92 beats per minute. A similar trend is observed on December 24th where the readings were high with systolic at 139 mmHg and diastolic at 115 mmHg (Stage 2 Hypertension). However, on December 21st, your blood pressure showed improvement, with values of systolic 127 mmHg and diastolic 76 mmHg falling into the prehypertensive range. Your pulse also decreased to a normal rate of 86 beats per minute at this time. This suggests an episode or episodes of hypertension, potentially requiring medical attention, but with one reading showing improvement in early December.

Figure 4. A conversation with the eHealth AI chatbot where it excels in extracting meaningful blood pressure recordings from tens of files and clearly describing its findings.

By using open-source software exclusively, we managed to reduce costs to a minimum and ensure a high level of replicability, since our architecture could be copied by others at no additional cost. Of course, we are not considering other use cases, locations, and variables that fall outside the scope of this paper.

In our tests, we found that the very nature of LLMs [52] can be the greatest limitation of our study because the answers can be unpredictable; hence, the reproducibility of the conversation is one of the largest issues. Even though switching LLMs could lead to biased results, we took the decision to design the system this way because the ability to use specific LLMs for certain situations can be extremely useful. The accuracy or performance of LLMs is not within our focus, but we integrated the mechanisms necessary to guide LLMs into providing precise answers based on useful custom data. The first step in stopping the LLM's hallucinations was to force it, through specific prompts, to only answer if it is provided with the relevant information. Due to the nature of how LLMs work, their responses have great variability, and this leads to unclear or even wrong answers, showing that before adopting such technologies in real-world scenarios, more research is needed.

4. Discussion

The authors' contributions consist of writing the eHealth AI chatbot Python application; preparing anonymized test samples from previous works to avoid privacy issues; configuring, hosting, and testing all the components of the system including the Matrix server and clients and LLM server; and partially adapting the previously built eHealth IoT data acquisition system.

The most important contribution of our research to the state of the art is the architecture itself, ensuring the system has great potential to grow. We believe that most of the works we studied could benefit from implementing our architecture, even partially, by continuing to use ChatGPT or their preferred AI service instead of a local LLM while adopting the rest of our architecture.

By basing all our communications on the Matrix open protocol, we only secure communications between Matrix chat applications (users and medical staff) and the eHealth AI chatbot, but the data acquisition, user management, and server administration should also be secured locally to prevent potential vulnerabilities.

During development, we found that our approach of integrating the eHealth AI chatbot into a well-established communications platform opened the door to a plethora of possibilities, among which are the following:

- Every user could interact with the eHealth AI chatbot, meaning that not only patients but also medical staff could benefit from its features. They could ask questions about one or more of their patients, improving search speed, finding connections between related cases or even asking details about upcoming procedures;
- Communicating through the Matrix platform could increase remote collaboration between doctors by allowing them to discuss problematic cases or by responding to patient questions, all from the same application;
- The AI chatbot could be added to any room, assisting multiple doctors in discussing specific cases;
- The type of communication could fit the needs of different patients or cases. In urgent situations or when the patient has a disability, communicating through voice or video might be preferred;
- Messages could include images, audio, video, and other files, which could save a lot of time for both patients and doctors.

In addition to the above possibilities, our application was able to connect information from multiple different sources and to offer a precise answer that would prevent the user from having adverse reactions caused by adding new medication to the existing treatment. These findings indicate the importance of integrating modern communication methods with AI and eHealth, giving healthcare applications more possibilities to improve our quality of life.

Compared to one of the closest implementations we found, our current research is more focused on providing a secure, encrypted, and versatile communications platform (Matrix) in contrast to the limited user interface (Streamlit) used in [49], and using a different LangChain approach where the model can access user-specific medical data instead of a collection of online databases. Still, allowing our system to access these databases could improve AI answers, especially in situations where it does not have enough information about a certain user.

Future work or potential enhancements based on this proposed foundation could include the following:

- In a production deployment, we could implement multiple answers based on several LLMs, so users could evaluate the answers they receive and provide useful feedback for mitigating LLM ethical issues;
- Following the Matrix protocol model, we could analyze if the eHealth AI chatbot software should be hosted on each eHealth IoT Data Acquisition System or choose a centralized approach;
- To better serve each patient, regardless of which minority they belong to, we could either use multilingual or entirely different language models based on the specific Matrix server's location. This could be a preference the patient can configure in their account profile;
- Implementing prescription functionality on top of our system would be straightforward because the communication infrastructure supports several data types;
- Disease detection based on computed tomography scans and bloodwork could be implemented by adding specialized classification models trained to detect certain medical conditions. This could prove helpful especially in the situation where doctors must examine huge numbers of patients and the time allotted to each patient is too short. Hence, the highlighting of potential issues doctors need to take into consideration when examining a patient could save lives;
- Access to official medical information repositories could be implemented through the LangChain API, enabling the eHealth AI chatbot to find reliable data when needed.

While navigating all the challenges regarding ethical, privacy, and security concerns, we believe that by curating medical information access, future implementations will also be able to improve disease prevention and treatment. We have come to believe that the proportion and complexity of AI healthcare implementations will continue to rise, so it is of paramount importance to safeguard the rights of patients and to adjust the legal and social frameworks. As the sophistication of such systems increases, so does the potential for vulnerabilities and errors.

5. Conclusions

With our proposed architecture, through our proof-of-concept implementation, we manage to provide a comprehensive eHealth system that provides solutions to several problems like privacy, security, and LLM answer relevance and correctness. The proposed system aims to solve some of the most important problems in healthcare applications while providing a more natural conversational interface and maintaining costs at a minimum.

In conclusion, our work delivers a novel approach that can be employed by other eHealth applications and addresses relevant research gaps by showing the plethora of possibilities AI can bring to the healthcare system just by gaining access to the patient's data and by using modern communication platforms. The results showcase the unrealized potential of providing much improved healthcare services with privacy, security, and convenience in mind.

Author Contributions: Conceptualization, I.A.P. and S.O.; methodology, I.A.P. and S.O.; software, I.A.P.; validation, S.O.; investigation, I.A.P. and S.O.; resources, I.A.P. and S.O.; data curation, I.A.P.; writing—original draft preparation, I.A.P.; writing—review and editing, I.A.P. and S.O.; visualization, I.A.P.; supervision, S.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request due to restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pap, I.A.; Oniga, S.; Orha, I.; Alexan, A. IoT-Based eHealth Data Acquisition System. In Proceedings of the 2018 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 24–26 May 2018. [CrossRef]
2. Pap, I.A.; Oniga, S.; Alexan, A. Machine Learning EEG Data Analysis For eHealth IoT System. In Proceedings of the 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 21–23 May 2020; pp. 143–146. [CrossRef]
3. Pap, I.A.; Oniga, S. A Review of Converging Technologies in eHealth Pertaining to Artificial Intelligence. *Int. J. Environ. Res. Public. Health* **2022**, *19*, 11413. [CrossRef] [PubMed]
4. Wojcik, S.; Rulkiewicz, A.; Pruszczyk, P.; Lisik, W.; Pobozy, M.; Domienik-Karlowicz, J. Beyond ChatGPT: What Does GPT-4 Add to Healthcare? The Dawn of a New Era. *Cardiol. J.* **2023**, *30*, 1018–1025. [CrossRef] [PubMed]
5. Zhou, B.; Yang, G.; Shi, Z.; Ma, S. Natural Language Processing for Smart Healthcare. *IEEE Rev. Biomed. Eng.* **2024**, *17*, 4–18. [CrossRef]
6. Wang, X.; Sanders, H.M.; Liu, Y.; Seang, K.; Tran, B.X.; Atanasov, A.G.; Qiu, Y.; Tang, S.; Car, J.; Wang, Y.X.; et al. ChatGPT: Promise and Challenges for Deployment in Low- and Middle-Income Countries. *Lancet Reg. Health-W. Pac.* **2023**, *41*, 100905. [CrossRef] [PubMed]
7. Yeo, Y.H.; Samaan, J.S.; Ng, W.H.; Ting, P.-S.; Trivedi, H.; Vipani, A.; Ayoub, W.; Yang, J.D.; Liran, O.; Spiegel, B.; et al. Assessing the Performance of ChatGPT in Answer-Ing Questions Regarding Cirrhosis and Hepatocellu- Lar Carcinoma. *Clin. Mol. Hepatol.* **2023**, *29*, 721–732. [CrossRef]
8. Alanzi, T.M. Impact of ChatGPTon Teleconsultants in Healthcare: Perceptions of Healthcare Experts in Saudi Arabia. *J. Multidiscip. Healthc.* **2023**, *16*, 2309–2321. [CrossRef]
9. Williams, S.C.; Starup-Hansen, J.; Funnell, J.P.; Hanrahan, J.G.; Valetopoulou, A.; Singh, N.; Sinha, S.; Muirhead, W.R.; Marcus, H.J. Can ChatGPT Outperform a Neurosurgical Trainee? A Prospective Comparative Study. *Br. J. Neurosurg.* **2024**. [CrossRef]
10. Salama, A.H. The Promise and Challenges of ChatGPT in Community Pharmacy: A Comparative Analysis of Response Accuracy. *Pharmacia* **2024**, *71*, e116927. [CrossRef]
11. Bazzari, F.H.; Bazzari, A.H. Utilizing ChatGPT in Telepharmacy. *Cureus J. Med. Sci.* **2024**, *16*, e52365. [CrossRef]
12. Tiwari, A.; Kumar, A.; Jain, S.; Dhull, K.S.; Sajjanar, A.; Puthenkandathil, R.; Paiwal, K.; Singh, R. Implications of ChatGPT in Public Health Dentistry: A Systematic Review. *Cureus J. Med. Sci.* **2023**, *15*, e40367. [CrossRef]
13. Ulusoy, I.; Yilmaz, M.; Kivrak, A. How Efficient Is ChatGPT in Accessing Accurate and Quality Health-Related Information? *Cureus J. Med. Sci.* **2023**, *15*, e46662. [CrossRef] [PubMed]

14. Zhu, Z.; Ying, Y.; Zhu, J.; Wu, H. ChatGPT's Potential Role in Non-English-Speaking Outpatient Clinic Settings. *Digit. Health* **2023**, *9*, 20552076231184091. [CrossRef] [PubMed]
15. Ghanem, Y.K.; Rouhi, A.D.; Al-Houssan, A.; Saleh, Z.; Moccia, M.C.; Joshi, H.; Dumon, K.R.; Hong, Y.; Spitz, F.; Joshi, A.R.; et al. Dr. Google to Dr. ChatGPT: Assessing the Content and Quality of Artificial Intelligence-Generated Medical Information on Appendicitis. *Surg. Endosc.* **2024**, *38*, 2887–2893. [CrossRef]
16. Kalam, K.T.; Rahman, J.M.; Islam, M.R.; Dewan, S.M.R. ChatGPT and Mental Health: Friends or Foes? *Health Sci. Rep.* **2024**, *7*, e1912. [CrossRef]
17. Tong, W.; Guan, Y.; Chen, J.; Huang, X.; Zhong, Y.; Zhang, C.; Zhang, H. Artificial Intelligence in Global Health Equity: An Evaluation and Discussion on the Application of ChatGPT, in the Chinese National Medical Licensing Examination. *Front. Med.* **2023**, *10*, 1237432. [CrossRef]
18. Sumbal, A.; Sumbal, R.; Amir, A. Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT&apos;s Performance in Academic Testing. *J. Med. Educ. Curric. Dev.* **2024**, *11*, 23821205241238641. [CrossRef] [PubMed]
19. Pallivathukal, R.G.; Soe, H.H.K.; Donald, P.M.; Samson, R.S.; Ismail, A.R.H. ChatGPT for Academic Purposes: Survey Among Undergraduate Healthcare Students in Malaysia. *Cureus J. Med. Sci.* **2024**, *16*, e53032. [CrossRef]
20. Davies, N.P.; Wilson, R.; Winder, M.S.; Tunster, S.J.; McVicar, K.; Thakrar, S.; Williams, J.; Reid, A. ChatGPT Sits the DFPH Exam: Large Language Model Performance and Potential to Support Public Health Learning. *BMC Med. Educ.* **2024**, *24*, 57. [CrossRef]
21. Wang, G.; Gao, K.; Liu, Q.; Wu, Y.; Zhang, K.; Zhou, W.; Guo, C. Potential and Limitations of ChatGPT 3.5 and 4.0 as a Source of COVID-19 Information: Comprehensive Comparative Analysis of Generative and Authoritative Information. *J. Med. Internet Res.* **2023**, *25*, e49771. [CrossRef]
22. Harskamp, R.E.; De Clercq, L. Performance of ChatGPT as an AI-Assisted Decision Support Tool in Medicine: A Proof-of-Concept Study for Interpreting Symptoms and Management of Common Cardiac Conditions (AMSTELHEART-2). *Acta Cardiol.* **2024**, *79*, 358–366. [CrossRef]
23. Gray, M.; Baird, A.; Sawyer, T.; James, J.; Debroux, T.; Bartlett, M.; Krick, J.; Umoren, R. Increasing Realism and Variety of Virtual Patient Dialogues for Prenatal Counseling Education Through a Novel Application of ChatGPT: Exploratory Observational Study. *JMIR Med. Educ.* **2024**, *10*, e50705. [CrossRef] [PubMed]
24. Padovan, M.; Cosci, B.; Petillo, A.; Nerli, G.; Porciatti, F.; Scarinci, S.; Carlucci, F.; Dell'Amico, L.; Meliani, N.; Necciari, G.; et al. ChatGPT in Occupational Medicine: A Comparative Study with Human Experts. *Bioengineering* **2024**, *11*, 57. [CrossRef] [PubMed]
25. Wang, C.; Liu, S.; Yang, H.; Guo, J.; Wu, Y.; Liu, J. Ethical Considerations of Using ChatGPT in Health Care. *J. Med. Internet Res.* **2023**, *25*, e48009. [CrossRef] [PubMed]
26. Shi, W.; Zhuang, Y.; Zhu, Y.; Iwinski, H.J.; Wattenbarger, J.M.; Wang, M.D. Retrieval-Augmented Large Language Models for Adolescent Idiopathic Scoliosis Patients in Shared Decision-Making. In *BCB '23: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*; Assoc Computing Machinery: New York, NY, USA, 2023; Article 14; pp. 1–10. [CrossRef]
27. Tang, R.; Han, X.; Jiang, X.; Hu, X. Does Synthetic Data Generation of LLMs Help Clinical Text Mining? *arXiv* **2023**. [CrossRef]
28. Azbeg, K.; Ouchetto, O.; Andaloussi, S.J. BlockMedCare: A Healthcare System Based on IoT, Blockchain and IPFS for Data Management Security. *Egypt. Inform. J.* **2022**, *23*, 329–343. [CrossRef]
29. Rani, S.; Kataria, A.; Kumar, S.; Tiwari, P. Federated Learning for Secure IoMT-Applications in Smart Healthcare Systems: A Comprehensive Review. *Knowl.-Based Syst.* **2023**, *274*, 110658. [CrossRef]
30. Letafati, M.; Otoum, S. On the Privacy and Security for E-Health Services in the Metaverse: An Overview. *Ad. Hoc Netw.* **2023**, *150*, 103262. [CrossRef]
31. Vellore Pichandi, K.; Janarthanan, V.; Annamalai, T.; Arumugam, M. Enhancing Healthcare in the Digital Era: A Secure e-Health System for Heart Disease Prediction and Cloud Security. *Expert. Syst. Appl.* **2024**, *255*, 124479. [CrossRef]
32. Almalawi, A.; Khan, A.I.; Alsolami, F.; Abushark, Y.B.; Alfakeeh, A.S. Managing Security of Healthcare Data for a Modern Healthcare System. *Sensors* **2023**, *23*, 3612. [CrossRef]
33. Salazar, G.Z.; Zuniga, D.; Vindel, C.L.; Yoong, A.M.; Hincapie, S.; Zuniga, A.B.; Zuniga, P.; Salazar, E.; Zuniga, B. Efficacy of AI Chats to Determine an Emergency: A Comparison Between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. *Cureus J. Med. Sci.* **2023**, *15*, e45473. [CrossRef]
34. Petersson, L.; Vincent, K.; Svedberg, P.; Nygren, J.M.; Larsson, I. Ethical Considerations in Implementing AI for Mortality Prediction in the Emergency Department: Linking Theory and Practice. *Digit. Health* **2023**, *9*, 20552076231206588. [CrossRef] [PubMed]
35. Brandao-de-Resende, C.; Melo, M.; Lee, E.; Jindal, A.; Neo, Y.N.; Sanghi, P.; Freitas, J.R.; Castro, P.V.I.P.; Rosa, V.O.M.; Valentim, G.F.S.; et al. A Machine Learning System to Optimise Triage in an Adult Ophthalmic Emergency Department: A Model Development and Validation Study. *EClinicalMedicine* **2023**, *66*, 102331. [CrossRef] [PubMed]
36. Jacob, F.; Grashöfer, J.; Hartenstein, H. A Glimpse of the Matrix (Extended Version): Scalability Issues of a New Message-Oriented Data Synchronization Middleware. *arXiv* **2019**. [CrossRef]
37. Jacob, F.; Becker, L.; Grashöfer, J.; Hartenstein, H. Matrix Decomposition: Analysis of an Access Control Approach on Transaction-Based DAGs without Finality. In Proceedings of the 25th ACM Symposium on Access Control Models and Technologies, Barcelona, Spain, 10–12 June 2020; pp. 81–92. [CrossRef]

38. Schipper, G.C.; Seelt, R.; Le-Khac, N.-A. Forensic Analysis of Matrix Protocol and Riot.Im Application. *Forensic Sci. Int. Digit. Investig.* **2021**, *36*, 301118. [CrossRef]
39. Karhu, J. Luxembourg Launches Open Source Chat for Officials and Citizens. 2023. Available online: https://joinup.ec.europa.eu/node/706091 (accessed on 17 May 2024).
40. Dussutour, C. French Government Launches In-House Developed Messaging Service, Tchap. 2020. Available online: https://joinup.ec.europa.eu/node/702563 (accessed on 17 May 2024).
41. Hillenius, G. German Armed Forces Testing Open Source Chat. 2020. Available online: https://joinup.ec.europa.eu/node/702455 (accessed on 17 May 2024).
42. Pätsch, S. German Health Professionals Will Communicate with Each Other through the Open Source Matrix Protocol. 2021. Available online: https://joinup.ec.europa.eu/node/704580 (accessed on 17 May 2024).
43. Matrix. Dendrite server version: 0.13.4+317b101. Available online: https://matrix.org (accessed on 17 May 2024).
44. Anthis, J.; Lum, K.; Ekstrand, M.; Feller, A.; D'Amour, A.; Tan, C. The Impossibility of Fair LLMs 2024. *arXiv* **2024**. [CrossRef]
45. Ollama. Version: 0.1.38. Available online: https://ollama.com (accessed on 17 May 2024).
46. Singh, A.; Ehtesham, A.; Mahmud, S.; Kim, J.-H. Revolutionizing Mental Health Care through LangChain: A Journey with a Large Language Model. In Proceedings of the 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 8–10 January 2024; pp. 0073–0078. [CrossRef]
47. Ke, Y.; Jin, L.; Elangovan, K.; Abdullah, H.R.; Liu, N.; Sia, A.T.H.; Soh, C.R.; Tung, J.Y.M.; Ong, J.C.L.; Ting, D.S.W. Development and Testing of Retrieval Augmented Generation in Large Language Models—A Case Study Report. *arXiv* **2024**. [CrossRef]
48. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**. [CrossRef]
49. Basit, A.; Hussain, K.; Hanif, M.A.; Shafique, M. MedAide: Leveraging Large Language Models for On-Premise Medical Assistance on Edge Devices. *arXiv* **2024**. [CrossRef]
50. LangChain. Versions: Langchain 0.1.9, Langchain-Community 0.0.24, Langchain-Core 0.1.27. Available online: https://langchain.com (accessed on 17 May 2024).
51. Matrix-Nio. Version: 0.24.0. Available online: https://github.com/matrix-nio/matrix-nio (accessed on 17 May 2024).
52. Kosch, T.; Feger, S. Risk or Chance? Large Language Models and Reproducibility in Human-Computer Interaction Research. *arXiv* **2024**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.