

A CHATBOT APPLICATION WITH VOICE COMMUNICATION FOR THE HEALTHCARE SYSTEM

Thai-Son Nguyen¹, Thai-Toan Nguyen², Ngoc-Chien Tran^{3*}, Thanh-Hai Ta⁴,
Dang-Khoa Huynh⁵, Hoang-Phuong Nguyen⁶

Abstract – *The increasing demands of patient care necessitate innovative solutions to alleviate the workload on healthcare providers. One such solution is the implementation of chatbots to address common patient inquiries before intervention by medical staff. This paper presents a chatbot system designed to enhance patient interactions by leveraging the question-answering capabilities of the Llama 3 open-source artificial intelligence (AI) model. Llama 3 plays a central role in generating accurate and contextually appropriate responses from stored data, ensuring that patients receive timely and relevant information. Voice communication between the chatbot and users is facilitated, with a focus on the seamless integration of Llama’s advanced artificial intelligence question-answer function. Chainlit is utilized to create an intuitive interface, further optimizing patient-system interactions. This study outlines the development, integration, and evaluation of the chatbot system within a hospital environment, demonstrating its effectiveness in reducing the burden on healthcare providers while enhancing patient engagement through voice-based Artificial Intelligence interactions.*

Keywords: *Chainlit, Chatbots, Google Speech-to-Text, Google Text-to-Speech, Llama3, Qdrant.*

I. INTRODUCTION

Chatbot is known as a computer program or software being capable of conversing with humans using text or natural language. Chatbot has

quickly become an effective assistant, providing health advice and supporting hospitals to take good care of patients. Chatbot interacts with people through automatic message responses, by building conversation scenarios using machine learning, artificial intelligence (AI),... thereby providing appropriate answers. In hospitals, chatbots can be used to provide patients with information about common diseases, medical examinations, and treatment procedures.

In the chatbot system, natural language is used to enter questions and return results verbally while also providing text information on a web interface. User input is processed to produce contextual responses based on data provided by hospital information sources. Recently, Google Text-to-Speech (TTS) is used to convert text into speech to feedback results to the patient and Google Speech-to-Text (STT) is used to convert the user’s voice into text for the system to process and to give answers. Qdrant is a vector similarity search engine used to store hospital information to provide to patients. From the data source stored in Qdrant, the open-source AI model Llama 3 synthesizes the complete answer.

In the last five years, many research projects on chatbots have been undertaken in the healthcare field [1–5]. According to Badlani et al. [1], a healthcare chatbot is proposed aiming at addressing the complex healthcare challenges in rural India. This chatbot is multilingual and utilizes Natural Language Processing (NLP) for user interaction. It also integrates machine learning algorithms for disease prediction, with the Random Forest Classifier achieving 98.43% accuracy. The solution emphasizes cost-effective access to information and diagnosis, overcoming language barriers.

^{1,2,3,4,5}Tra Vinh University, Vietnam

⁶Mien Tay Construction University, Vietnam

*Corresponding author: tnchien@tvu.edu.vn

Received date: 27th June 2024; Revised date: 23rd August 2024; Accepted date: 28th August 2024

To provide accurate information to users across various domains, including medicine, a medical chatbot is designed to interact with users in a manner resembling conversations with medical professionals [2]. Utilizing Artificial Intelligence Mark-up Language (AIML), it analyzes user messages to identify potential medical issues and offers direct advice and solutions. Later, Atmauswan et al. [3] proposed a chatbox using DialogFlow and NLP to offer consolidated, accurate information to users through live chat. It gathers knowledge from frequently asked questions, enabling users to search for answers and receive assistance with admission procedures, study programs, and scholarships. The goal is to provide a user-friendly platform on the university's website for students to effortlessly find answers to their queries.

Shinde et al. [4] introduced a healthcare chatbot powered by AI to engage with users and address basic health issues before consulting a doctor. The main objective is to analyze user symptoms and provide medical suggestions to reduce time and costs. Utilizing NLP for communication, the chatbot employs techniques such as TF-IDF, Stemming, n-grams, and cosine similarity for calculations like ranking and sentence similarity. Machine learning algorithms are utilized to resolve healthcare queries and specify symptoms. Serving as an advisor for primary healthcare and minor health concerns, the chatbot provides support to those in urgent need. Users can identify diseases by inputting symptoms and receive corresponding solutions, preventive measures, and treatments. The AI chatbot is proposed to engage in natural conversations with users and to address their queries using a knowledge base [5]. The knowledge base, stored in real-time data in JSON format, undergoes preprocessing to enable further processing. Information is communicated in both speech and text formats, with the chatbot offering navigation links and predicting issues through symptom diagnosis. Additionally, it recommends consulting a doctor, taking immediate measures, and providing advanced information on diagnostics.

Tra Vinh University Hospital (Tra Vinh Province, Vietnam) was established and developed based on the General Clinic of Tra Vinh University with a scale of 50 beds. Since its establishment in February 2019, the hospital has continuously improved the quality of medical services, expanded the scope of expertise, and strived to bring quality and effective medical examination and treatment opportunities to the community, people in the locality, and surrounding areas. Currently, the number of patients coming for medical examination and treatment is increasing, to about 1,000 per day [6]. Most patients are people with low and middle income. Therefore, hospital services provided at an affordable cost are considered. However, with the increasing number of patients, to provide better care while still ensuring medical examination and treatment costs are at an acceptable level. The paper proposed chatbots aiming to contribute to patient care and consultation. To ensure ease and flexibility in operation, STT and TTS have been applied. Experimental results showed that the chatbots met the requirements, providing consistent and useful information to patients.

The remainder of the paper is organized as follows: Section 2 summarizes related approaches. Section 3 details the main idea and implementation steps of the proposed methodology. Section 4 outlines the evaluation process for the method's effectiveness. Section 5 covers the conclusion.

II. THEORETICAL BACKGROUND

A. *Google Speech-To-Text (STT), Google Text-To-Speech (TTS)*

Google STT is a powerful and flexible technology that enables the conversion of speech into text with high accuracy and can be easily integrated into various types of applications and services. It also supports 125 languages and variants. This service finds applications in numerous industries, including media, telecommunications, healthcare, and the like, offering solutions like captioning videos, generating transcripts for meetings, and enabling voice command functionality in applications. It is one of the leading tech-

nologies in the field of NLP and is widely used in various applications and services worldwide [7].

Google TTS is a powerful tool that converts text into natural-sounding speech. This technology is integrated into various Google services like Google Assistant and Google Translate, offering a seamless user experience across devices and applications. Its application extends to providing voice guidance as a chatbot, reading notifications, and text messages on Android devices. It allows developers to create natural-sounding speech using over 100 different voices [8].

B. Qdrant – Vector database

Qdrant is an open-source platform for searching and querying large vector datasets. This vector similarity search engine offers a production-ready service, complete with an easy-to-use API. It allows for the storage, search, and management of vectors along with additional payloads. These payloads are extra pieces of information that enhance search capabilities and deliver valuable insights to users. Qdrant is designed to work with complex vector data, including text, images, audio, and video. It’s a useful tool for applications such as similar content discovery in social media platforms, and various applications related to processing and analyzing vector data [9].

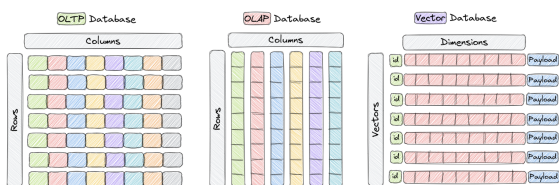


Fig. 1: Structural differences between online transactional processing (OLTP), online analytical processing (OLAP), and vector databases [9]

Vector databases are a specific type of database optimized for storing and querying high-dimensional vectors, which differ from traditional databases that organize data in rows and columns

as shown in Figure 1. These databases are essential for applications such as image recognition and recommendation systems, where data is typically represented in vector form. These databases optimize storage and querying using techniques like Hierarchical Navigable Small World (HNSW) and Product Quantization, enabling efficient similarity search based on distance metrics like Cosine Similarity, Dot Product, and Euclidean Distance [9].

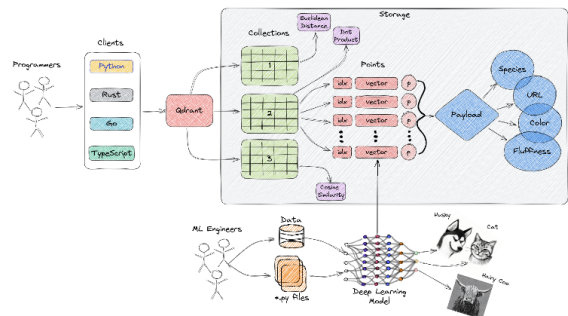


Fig. 2: High-level overview of Qdrant’s architecture [9]

Figure 2 provides a high-level overview of some key components of Qdrant. Below are the key terminologies:

- Collections: Defined as a named group of points, each collection is composed of vectors carrying associated data, intended for search purposes. All vectors in a collection share the same dimensionality and are evaluated based on a uniform metric.
- Named vectors enable the inclusion of several vectors at a single point, each possessing unique dimensions and metric specifications.
- Distance metrics are employed to assess the similarities between vectors. They should be chosen concurrently with the creation of a collection.
- Points: Central entities within Qdrant, consisting of a vector, an optional ID, and a payload.
- Storage: Qdrant offers two storage options:
 - + In-memory storage keeps all vectors directly in RAM, allowing for rapid operations and only using disk access to ensure data persistence.
 - + Memmap storage establishes a virtual address space that links directly to a disk file.

- Clients: Programming languages used to interface and integrate with Qdrant for seamless interaction [9].

C. Large language model Meta artificial intelligent 3 (Llama 3)

Llama 3, the next iteration of the open-access Llama family, is the latest large language model (LLM) from Meta. It can be used to create generative AI, including chatbots that can respond in natural language to a wide variety of queries. The model is available in two key versions: an 8-billion and a 70-billion-parameter model. A version with 400B+ parameters is currently being trained. Llama 3 was trained on over 15 trillion tokens, offering a dataset seven times larger than that used for Llama 2. These models are both pre-trained and instruction-tuned to address various tasks effectively [10].

The model is designed to be efficient, able to run even on consumer-grade GPUs by using lower precision and integrates with frameworks like Hugging Face and PyTorch to make deployment simple [11].

Even with 70 billion parameters, Meta claims Llama 3 is still competitive with much larger models [11].

Figure 3 shows Meta Llama 3 instruct model performance across various benchmarks.

Meta Llama 3 Instruct model performance

	Meta Llama 3 8B	Gemini 1.5 Flash	Mistral 7B Instruct
MMLU (0-100)	68.4	55.3	58.4
GPQA (0-100)	34.2	21.4	26.3
HumanEval (0-100)	62.2	30.5	36.6
CodeBLEU (0-100)	79.6	30.6	39.9
MATH (0-100)	30.0	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Flash	Claude 3 Sonnet
MMLU (0-100)	82.0	81.9	79.0
GPQA (0-100)	39.5	41.5	38.5
HumanEval (0-100)	81.7	71.9	75.0
CodeBLEU (0-100)	93.0	91.7	92.3
MATH (0-100)	50.4	58.5	40.5

Fig. 3: Meta Llama 3 instruct model performance [12]

Figure 4 illustrates a safety architecture for large language models (LLMs) in Meta’s Llama 3, emphasizing various safeguards to ensure responsible AI deployment.

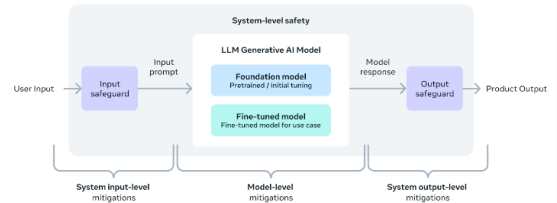


Fig. 4: A safety architecture for large language models [12]

D. Hugging Face

Hugging Face provides a platform and community dedicated to machine learning (ML) and data science. It enables users to efficiently build, deploy, and train machine learning models. This platform serves as an infrastructure for showcasing and deploying AI in real-world applications, while also facilitating exploration of models and datasets uploaded by others. Hugging Face is a Transformers Python library, this library simplifies downloading and training machine learning models [13].

E. Chainlit

Chainlit is an open-source Python package designed for creating production-ready conversational AI. It can be seamlessly integrated into an existing codebase or used to start a new project from scratch within minutes [14].

III. PROPOSED SYSTEM

In this section, a detailed description of the proposed system is provided. Figure 5 shows the system architecture of the chatbot.

When a user provides a question as input to the system. The input data is a natural-sounding speech from the user. The system will convert natural-sounding speech into text. Then the query is processed by an embedding model – Vietnamese-SBERT to transform into a vector that encodes the semantic meaning of the query. The vector of the query is used to search a vector database for K-similar chunks, which are semantically aligned with the query. Qdrant database

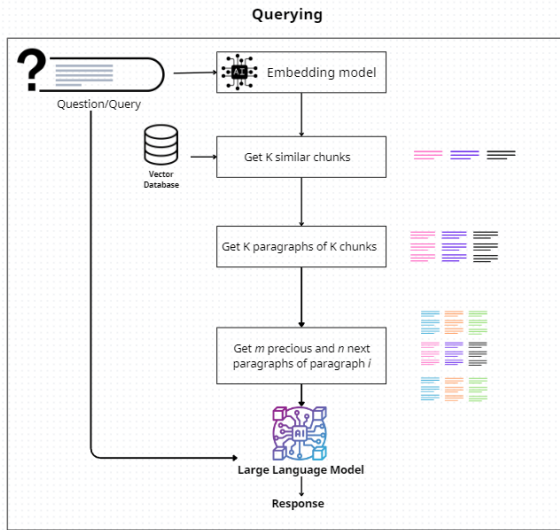


Fig. 5: The system architecture of the chatbot

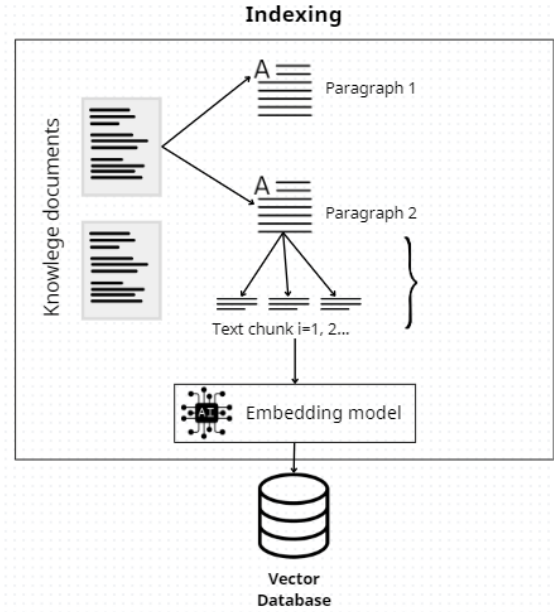


Fig. 6: Document parsing to vector database storage

is optimized for fast similarity searches to ensure the most relevant chunks are retrieved. To enhance comprehension and provide additional information, the system retrieves previous and subsequent paragraphs for each relevant paragraph. This information ensures that the response isn't isolated and provides more depth. Llama 3 analyzes the question and synthesizes the data to produce a comprehensive and accurate response. The response generated by the Llama 3 model is delivered back to the user, answering the original question clearly and completely. Finally, the system converts answers from text into natural-sounding speech.

A. Data processing

Figure 6 illustrates the process of indexing knowledge documents into a vector database for effective information retrieval.

Knowledge documents

The source material that will be used as the knowledge base for the chatbot is collected directly at Tra Vinh University Hospital in PDF file format. These documents are whole files that haven't been parsed or segmented into smaller components.

Paragraph extraction

The documents are parsed to extract individual paragraphs. This step involves identifying logical boundaries between different sections. Breaking documents into paragraphs makes subsequent processing more manageable and allows retrieval at a finer granularity.

Text chunking

Paragraphs are further divided into manageable text chunks to enhance the retrieval process. This helps improve the relevance of search results. Each paragraph is further split into smaller chunks. Chunking is often done to keep each chunk within a certain character limit. Chunking enables embeddings to be generated more efficiently and ensures the system can handle each chunk individually. This is especially important if paragraphs are long or contain complex ideas.

Embedding model

The Vietnamese-SBERT model is a sentence-transformers framework designed to represent sentences and paragraphs in a 768-dimensional vector space. This model is particularly useful for clustering and conducting semantic searches

within the Vietnamese language [15]. It processes each text chunk to convert it into a numerical vector, mapping similar chunks close together in the vector space. It allows meaningful comparisons between different chunks using similarity measures.

Vector database

The resulting vectors are then stored in a vector database (Qdrant) for fast similarity searches. This database can be used to quickly retrieve relevant chunks of information during querying. This process aims to turn knowledge documents into indexed vectors that can be searched quickly. By segmenting documents into paragraphs and chunks, and then encoding them using an embedding model, relevant information can be quickly retrieved from a vector database. This indexing system is very powerful for various search and retrieval applications, especially for unstructured data.

B. Prompt engineering

Prompt engineering is the process of designing instructions that a generative AI model can understand and follow. A prompt is a segment of natural language text that outlines the particular task the AI is supposed to carry out [16]. The prompt can be a word, a phrase, or an entire paragraph of text. Prompt acts as input or query that the user provides to suggest the AI model to give specific responses.

There are many approaches to creating a prompt engineering. The Figure 7 is the structure of prompt engineering used to ask the chatbot to provide information about the hospital.

```

custom_prompt_template = """
Bạn là một chatbot của bệnh viện Trường Đại học Trà Vinh. Bạn cung cấp thông tin và bệnh viện để hỗ trợ bệnh nhân.
Hãy dựa trên [thông tin] được cung cấp, viết câu trả lời có giải thích (nếu có thể) cho [câu hỏi].
Nếu không có thông tin đầy đủ, hãy trả lời: 'Tôi chưa có thông tin đầy đủ để trả lời'.
(TUỶỆ ĐỐI phải dựa vào thông tin được cung cấp, không dùng kiến thức của bạn để trả lời)

[Thông tin]: {context}
[Câu hỏi]: {question}

LƯU Ý: Hãy căn cứ vào thông tin được cung cấp để trả lời. Chú ý các dịch vụ khám bệnh của bệnh viện.
Nếu họ chào bạn, hãy chào lại họ. Hãy luôn thân thiện và trả lời ngắn gọn. Trả lời bằng Tiếng Việt.
"""
    
```

Fig. 7: Prompt engineering is used to ask the chatbot

Figure 7 has the following meaning:
 custom_prompt_template = """

Description: You are a chatbot for the Tra Vinh University Hospital. You provide hospital-related information to assist patients. Please base your responses on the [information] provided, and write concise and clear answers (if possible) to the [question]. If there is not enough information available, respond: 'I don't have enough information to answer this.' (ABSOLUTELY base your response only on the information provided, do not use your own knowledge to respond.)

[Information]: context

[Question]: question

NOTE: Please rely solely on the provided information to answer. Pay attention to the hospital's medical services. If someone greets you, respond accordingly. Always be friendly and give short, concise answers. Respond in Vietnamese.

C. Synthesizing answers with Llama 3

Llama 3 operates as a text generation model. It is trained to predict and generate text answers based on input prompts. The model processes this input and generates a sequence of tokens representing the answer. Once the model has generated a sequence of tokens representing the answer, it undergoes some post-processing before being presented to the user. This could involve removing special tokens, applying formatting, or truncating the output to a reasonable length.

D. User interface

Chainlit is used to design the chatbot interface. The interface of the chatbot supports providing medical examination and treatment information in the hospital as shown in Figure 8.

IV. EXPERIMENTAL RESULTS

A. Speech recognition process

The speech recognition capabilities of a chatbot to accurately convert natural-sounding speech to text is crucial for a seamless user experience. Under normal testing conditions, the development team found that the chatbot recognized speech accurately even in a light-noise environment.

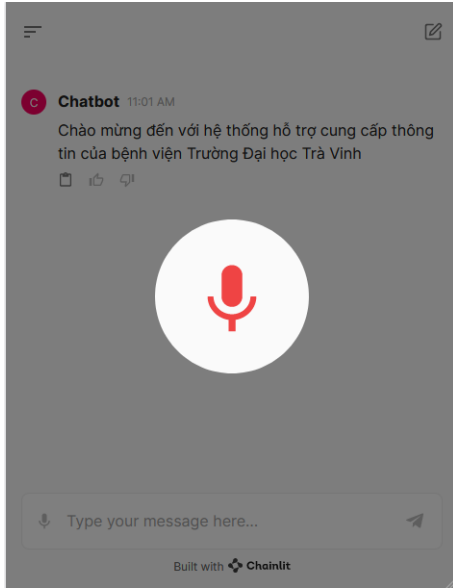


Fig. 8: Chatbot user interface

B. Evaluate the chatbot’s answer

The experimental setup was designed by using data sourced from Tra Vinh University Hospital to evaluate the chatbot’s performance. A dataset was collected from Tra Vinh University Hospital illustrates in Figure 9, including common questions and their corresponding answers. This dataset serves as a baseline for constructing the questions and answers to be used in the evaluation. From the collected dataset, a set of 100 questions representative of typical patient concerns and inquiries is obtained. Through the chatbot’s interface, each of the 100 questions was tested by the answer from the chatbot. Then, the answer was compared with the expected answers whether the chatbot’s responses align with the expected answer or not.

The results showed that the chatbot answered 84 questions correctly, the remaining 16 questions were incorrect because some information was missing in the desired answer. According to our records, the questions that the chatbot answered incorrectly due to keywords overlapping with the results of other answers, led to the chatbot misinterpreting the answer as the other contents.

Question	Answer	Chatbot Response
1. Ai được ưu tiên làm thủ tục khám bệnh trước?	- Bệnh nhân trên 80 tuổi, trẻ dưới 6 tuổi, phụ nữ mang thai, người tàn tật.	Theo quy định chung tại Khoa Khám bệnh, người bệnh có nhu cầu khám dịch vụ sẽ được ưu tiên làm thủ tục khám bệnh trước. Tại khu vực đăng ký và tiếp nhận của khu khám bệnh, người bệnh đến Quay bất số khám bệnh được hướng dẫn cách lấy số thứ tự đăng ký khám bệnh. Sau đó người bệnh chờ nhận gọi số lượt học xem hiển thị số đăng ký trên màn hình của phòng tiếp nhận để làm thủ tục. Cần mang theo những giấy tờ sau khi đi khám bệnh ngoài trú.
2. Làm thế nào để lấy số thứ tự khám bệnh?	- Đến quầy đăng ký, lấy số thứ tự và chờ nghe gọi số.	
3. Cần mang theo những giấy tờ gì khi đi khám bệnh ngoài trú?	- Sổ khám bệnh có BHYT, giấy tờ tùy thân có ảnh.	* Phiếu khám bệnh * Thẻ BHYT (if applicable) * Sổ khám bệnh Các giấy tờ này sẽ giúp các bác sĩ và kĩ thuật viên chẩn đoán tình hình thực hiện các thủ thuật và kĩ thuật đặc biệt một cách an toàn và đúng quy định. Quy trình khám bệnh tại phòng khám chuyên khoa Phụ Sản như sau: 1. Bệnh nhân nhập viện và được hướng dẫn đến phòng khám Phụ Sản. 2. Tại phòng khám, bệnh nhân sẽ được bác sĩ khám và thực hiện các cận lâm sàng cần thiết theo chỉ định của Bác sĩ (nếu có). 3. Sau khi được Bác sĩ chẩn đoán bệnh, kê đơn thuốc, bệnh nhân đến phòng thu viện phí đồng đồng chi trả (nếu có) sau đó đến quầy lễ tân thuốc BHYT. 4. Nếu cần thiết, bệnh nhân sẽ được nhập viện điều trị nội trú khi có chỉ định của Bác sĩ hoặc nhận giấy chuyển tuyến (nếu có). Quy trình này đảm bảo rằng bệnh nhân sẽ được thăm dò và chẩn đoán chính xác, sau đó được điều trị phù hợp với nhu cầu của người bệnh.
4. Quy trình khám bệnh tại phòng khám chuyên khoa diễn ra như thế nào?	- Chờ khám theo số thứ tự, vào khám khi được gọi, thực hiện cận lâm sàng (nếu có) chỉ định, chờ nhận kết quả cận lâm sàng, quay lại phòng khám và nộp kết quả cho bác sĩ, sau khi được chẩn đoán bệnh và kê đơn thuốc, đến phòng thu viện phí đồng đồng chi trả (nếu có) và làm thuốc BHYT.	
5. Nhiệm vụ của điều dưỡng trong quy trình khám bệnh ngoài trú?	- Mời bệnh nhân vào phòng khám theo thứ tự, lấy dấu hiệu sinh tồn, cân nặng, thông báo cho bác sĩ nếu phát hiện bệnh nhân có dấu hiệu bất thường, hướng dẫn bệnh nhân thực hiện cận lâm sàng, nhận kết quả cận lâm sàng từ bệnh nhân, kiểm tra toa thuốc, và số khám bệnh, tư vấn dùng thuốc, hướng dẫn đóng viện phí và làm thuốc BHYT.	Theo quy định, Điều dưỡng có trách nhiệm: Chỉ số khám bệnh chuyên khoa. Đưa người bệnh vào khoa điều trị, ban giao đầy đủ hồ sơ bệnh án, tình trạng người bệnh với Điều dưỡng trưởng khoa hoặc Điều dưỡng hành chính khoa.

Fig. 9: The dataset serves as a baseline for constructing the questions and answers used in the evaluation

Table 1 presents the dataset serves as a baseline for constructing the questions and answers used in the evaluation in English version.

V. CONCLUSION

To reduce the workload of medical staff, the implementation of a chatbot system in Tra Vinh University Hospital is introduced in this paper. By integrating cutting-edge technologies such as Google Text-to-Speech and Speech-to-Text, Llama 3, Qdrant, and Chainlit, the chatbot can offer efficient, accurate, and consistent responses to patient inquiries. This system not only provides preliminary answers to common questions but also reduces wait times for patients. As a result, this chatbot system is a crucial step towards modernizing healthcare services and optimizing the delivery of essential medical information reliably and consistently. Further development will ensure its adaptability and effectiveness in a wider range of healthcare facilities.

Table 1: English version of the dataset serves as a baseline for constructing the questions and answers used in the evaluation

Question	Answer	Chatbot Response
1. Who is given priority when completing the medical procedures first?	- Patients over 80 years old, children under 6 years old, pregnant women, and the disabled.	According to the general regulations of the Examination Department, patients in need of medical examination services will be given priority for procedures.
2. How to get the registration number for a medical examination?	- Go to the registration counter, get a number, and wait for your number to be called.	At the registration and reception area of the examination zone, patients go to the number counter and are instructed on how to get a registration number for their medical examination. After that, they sit and wait until their number is called (in combination with checking the number displayed on the screen of the reception area) to proceed with the procedures.
3. What documents need to be brought for outpatient examination?	- Health insurance card, ID card, or other identification documents with a photo.	Please bring the following documents when you go for an outpatient examination: + Medical examination form + Health insurance card (if applicable) + Medical record book These documents will help doctors and imaging technicians perform procedures and specialized techniques safely and in compliance with regulations.
4. What is the process for examination in specialized clinics?	- Wait for your turn based on the registration number, enter the examination room when called, undergo any indicated laboratory tests, wait for the test results, return to the clinic, and submit the results to the doctor. After being diagnosed and receiving a prescription, go to the payment counter to pay the fees (if any) and collect medication with your health insurance card.	The process of medical examination at the Obstetrics and Gynecology specialized clinic is as follows: 1. The patient is admitted and guided to the Obstetrics and Gynecology clinic. 2. At the clinic, the patient is examined by a doctor and undergoes necessary laboratory tests as prescribed by the doctor (if any). 3. After being diagnosed and receiving a prescription, the patient goes to the payment counter to pay the fees (if any) and then proceeds to the pharmacy to collect the medication covered by health insurance. 4. If necessary, the patient may be admitted for inpatient treatment as prescribed by the doctor or based on a referral letter (if any). This process ensures that the patient receives accurate examinations and diagnoses, leading to appropriate treatment that meets the patient's needs.
5. What are the duties of the nurses in the outpatient examination department ?	- Call patients into the examination room in order, check vital signs and weight, inform the doctor if any abnormalities are detected, guide the patient to perform laboratory tests, receive laboratory results from the patient, check the prescription and medical record, provide medication counseling, and guide the patient on how to pay fees and collect medication covered by health insurance.	According to regulations, the nurse is responsible for Recording the specialized clinic's medical records. Escorting the patient to the treatment department and fully handing over the medical records and the patient's condition to the head nurse or the department's administrative nurse.

REFERENCES

- [1] Badlani S, Aditya T, Dave M, Chaudhari S. Multilingual healthcare chatbot using machine learning. In: *2021 2nd International Conference for Emerging Technology (INCET)*. 21st-23rd May 2021; Belagavi, India. IEEE; 2021. p.1-6. <https://doi.org/10.1109/INCET51464.2021.9456304>.
- [2] Rarhi K, Bhattacharya A, Mishra A, Mandal K. Automated medical chatbot. *SSRN Electronic Journal*. 2017;10. <https://doi.org/10.2139/ssrn.3090881>.
- [3] Atmauswan P, Abdullahi A. Intelligent chatbot for university information system using natural language approach. *Albukhary Social Business Journal*. 2022;3(6): 59-64. <https://doi.org/10.55862/asbjV3I2a007>.
- [4] Shinde NV, Akhade A, Bagad P, Bhavsar H, Wagh SK, Kamble A. Healthcare chatbot system using artificial intelligence. In: *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. 03rd-05th June 2021; Tirunelveli, India. IEEE; 2021. p.1-8. <https://doi.org/10.1109/ICOEI51242.2021.9452902>.
- [5] Dammavalam SR, Chandana N, Rao TR, Lahari A, Aparna B. AI based chatbot for hospital management system. In: *2022 3rd International Conference on Computing, Analytics and Networks (ICAN)*. 18th-19th November 2022; Rajpura, Punjab, India. IEEE; 2022. p.1-5. <https://doi.org/10.1109/ICAN56228.2022.10007105>.
- [6] Tra Vinh online. *Tra Vinh University Hospital: Gradually Enhancing the Quality of Medical Examination and Treatment [Bệnh viện Trường Đại học Trà Vinh: Từng bước nâng cao chất lượng khám, chữa bệnh]*. <https://www.baotravinh.vn/xa-hoi/benh-vien-truong-dai-hoc-tra-vinh-tung-buoc-nang-cao-chat-luong-kham-chua-benh-27005.html>. [Accessed 1st May 2024].
- [7] Google. *Turn speech into text using Google AI*. <https://cloud.google.com/speech-to-text?hl=en> [Accessed 20th April 2024].
- [8] Google. *Text-to-Speech AI*. <https://cloud.google.com/text-to-speech?hl=vi> [Accessed 30th April 2024].
- [9] Qdrant. *Introduction Qdrant*. <https://qdrant.tech/documentation/overview> [Accessed 21st April 2024].
- [10] Meta. *Llama 3 Cheat Sheet: A Complete Guide for 2024*. <https://www.techrepublic.com/article/what-is-llama-3>. [Accessed 30th April 2024].
- [11] Meta. *Welcome Llama 3 - Meta's new open LLM*. <https://huggingface.co/blog/llama3> [Accessed: 30th April 2024].
- [12] Meta. *Introducing Meta Llama 3: The most capable openly available LLM to date*. <https://ai.meta.com/blog/meta-llama-3> [Accessed 30th April 2024].
- [13] Hugging Face. *What is Hugging Face?* <https://www.techtargget.com/whatis/definition/Hugging-Face> [Accessed 30th April 2024].
- [14] Chainlit. *Overview*. <https://docs.chainlit.io/get-started/overview> [Accessed 1st May 2024].
- [15] Hugging Face. *Vietnamese-sbert*. <https://huggingface.co/keepitreal/vietnamese-sbert> [Accessed 1st May 2024].
- [16] Microsoft. *Overview of prompts*. <https://learn.microsoft.com/en-us/ai-builder/prompts-overview> [Accessed 1st May 2024].

