information

Article

Towards Reliable Healthcare LLM Agents: A Case Study for Pilgrims during Hajj

Hanan M. Alghamdi [1,*] **and Abeer Mostafa** [2]

[1] Department of Computers, College of Engineering and Computing Al Qunfidhah, Umm Al-Qura University, Makkah 24382, Saudi Arabia
[2] Department of Computer Science and Engineering, Egypt-Japan University of Science and Technology, Alexandria 21934, Egypt; abeer.mostafa@ejust.edu.eg
* Correspondence: hmhghamdi@uqu.edu.sa

Abstract: There is a pressing need for healthcare conversational agents with domain-specific expertise to ensure the provision of accurate and reliable information tailored to specific medical contexts. Moreover, there is a notable gap in research ensuring the credibility and trustworthiness of the information provided by these healthcare agents, particularly in critical scenarios such as medical emergencies. Pilgrims come from diverse cultural and linguistic backgrounds, often facing difficulties in accessing medical advice and information. Establishing an AI-powered multilingual chatbot can bridge this gap by providing readily available medical guidance and support, contributing to the well-being and safety of pilgrims. In this paper, we present a comprehensive methodology aimed at enhancing the reliability and efficacy of healthcare conversational agents, with a specific focus on addressing the needs of Hajj pilgrims. Our approach leverages domain-specific fine-tuning techniques on a large language model, alongside synthetic data augmentation strategies, to optimize performance in delivering contextually relevant healthcare information by introducing the HajjHealthQA dataset. Additionally, we employ a retrieval-augmented generation (RAG) module as a crucial component to validate uncertain generated responses, which improves model performance by 5%. Moreover, we train a secondary AI agent on a well-known health fact-checking dataset and use it to validate medical information in the generated responses. Our approach significantly elevates the chatbot's accuracy, demonstrating its adaptability to a wide range of pilgrim queries. We evaluate the chatbot's performance using quantitative and qualitative metrics, highlighting its proficiency in generating accurate responses and achieve competitive results compared to state-of-the-art models, in addition to mitigating the risk of misinformation and providing users with trustworthy health information.

Keywords: text generation; deep learning; large language models (LLMs); healthcare chatbot; Hajj

Citation: Alghamdi, H.M.; Mostafa, A. Towards Reliable Healthcare LLM Agents: A Case Study for Pilgrims during Hajj. *Information* **2024**, *15*, 371. https://doi.org/10.3390/info15070371

Academic Editors: Arkaitz Zubiaga and Katsuhide Fujita

Received: 26 April 2024
Revised: 27 May 2024
Accepted: 21 June 2024
Published: 26 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Research around building a healthcare agent for Hajj pilgrims is essential due to the healthcare challenges faced during the event. Hajj draws millions of Muslims from around the world, leading to a massive congregation in a confined space. This congregation presents significant health challenges, including the risk of infectious diseases spreading rapidly due to close proximity and crowded conditions, as well as the potential for heat-related illnesses, accidents, and other medical emergencies [1–5].

A healthcare agent customized for Hajj pilgrims can address several critical needs. Firstly, it will provide pilgrims with accurate and up-to-date information on preventive measures, such as vaccination requirements, hygiene practices, and crowd management strategies, to minimize the risk of disease transmission. Secondly, it will offer guidance on managing common health issues encountered during the pilgrimage, such as dehydration, heatstroke, and musculoskeletal injuries. Additionally, the healthcare agent will facilitate access to medical assistance by providing information on nearby healthcare facilities, emergency contacts, and virtual consultations with healthcare professionals. This highlights the importance of a virtual agent in this context. Given the sheer scale of pilgrims and their diverse backgrounds and languages, a virtual agent offers a scalable and easy-to-access solution to deliver healthcare information and support. Pilgrims can access the virtual agent via their smartphones or other devices, allowing for widespread dissemination of critical health-related guidance. This accessibility is particularly valuable in emergency situations when immediate access to reliable healthcare information and assistance can be life-saving.

The advent of LLMs has revolutionized various natural language processing (NLP) tasks, ranging from text generation to question answering. With the growing complexity of human–computer interactions, there is an increasing demand for LLMs that are not only powerful but also finely tuned to specific domains and trustworthy. LLMs have shown immense value in the medical field. Their uses span from medical writing and documentation to medical education. With their advanced capabilities, they can analyze data thoroughly, assisting in translational medicine and drug development. Additionally, LLMs improve tasks like medical reporting, diagnostics, and treatment planning, resulting in a better overall patient experience [6,7]. Following the global pandemic, significant advancements have been observed in the implementation of medical chatbots as conversational agents for patients. Traditionally, these chatbots were designed to handle specific tasks such as answering user queries on specific medical subjects by using a pre-defined database and incorporating user feedback to enhance their responses [8].

GPT-3 is one of the LLMs that is built upon the Transformer architecture, which was introduced by Vaswani et al. in the paper "Attention is All You Need" [9]. This architecture relies on self-attention mechanisms to capture long-range dependencies in sequences efficiently. GPT employs an unaltered Transformer decoder, distinguishing itself by the absence of an encoder attention component [10]. This distinction is evident in the visual representations provided in the diagrams above. Unlike BERT, which utilizes Transformer encoder blocks, GPT, GPT-2, and GPT-3 are constructed using Transformer decoder blocks. Notably, GPT-3 underwent training with extensive Internet text datasets totaling 570 GB, marking it as the most substantial neural network upon its release, boasting an impressive 175 billion parameters, a hundredfold increase from GPT-2. GPT-3 comprises 96 attention blocks, each housing 96 attention heads. As Transformers do not inherently understand the order of tokens, positional embeddings are added to the token embeddings to give the model information about the position of each token. This helps the model understand the order of tokens in a sequence. The attention mechanism which is used in Transformers is called scaled dot-product attention.

The retrieval-augmented generation (RAG) module can be utilized for uncertainty validation. The RAG module serves two primary functions:

- Knowledge retrieval: When the model encounters uncertain or ambiguous input, the RAG module retrieves relevant knowledge from specific external resources. This retrieval process enables the model to augment its understanding of the topic at hand and generate more informed responses.
- Validation of uncertain text: After retrieving relevant knowledge, the RAG module validates the uncertain text generated by the GPT-3.5 Turbo model against the retrieved information. By cross-referencing the model's output with external knowledge sources, the RAG module assesses the accuracy and credibility of the generated text, identifying and correcting any inaccuracies or inconsistencies before finalizing the response.

The knowledge retrieval process within the RAG module involves several key steps. Firstly, the module analyzes the input text, identifying keywords, entities, and contextual cues that signal the need for additional information or clarification. Next, it formulates structured queries based on the identified context, aiming to retrieve pertinent knowledge vectors from the inherited databases. These queries are carefully crafted to extract relevant information aligned with the topic at hand, ensuring that the retrieved knowledge is both informative and contextually appropriate.

Research conducted in the field of health promotion and communication consistently underscores the effectiveness of strategic messaging and communication campaigns in driving awareness and promoting positive health behavior among different populations [11]. By leveraging technology like a medical chatbot, healthcare providers can efficiently manage the influx of inquiries and deliver prompt assistance to those in need. This automated system can contribute to enhancing the overall healthcare experience during mass gatherings like Hajj, ensuring a safer and more organized environment. When individuals have access to a reliable and credible source of information, they are more likely to trust and utilize the health-related information provided by that source. This, in turn, increases their likelihood of effectively reducing potential health threats. A thorough investigation was carried out, involving 280 pilgrims from 28 different countries, in order to assess the perception of health risks associated with the Hajj pilgrimage [12]. The results of this study demonstrate a decline in the awareness level among pilgrims when it comes to these risks [13]. Hence, it emphasizes the immediate need for a thorough strategy that includes increasing awareness, implementing surveillance systems, enforcing hygiene standards, providing healthcare services, and promoting international cooperation [14].

The objective of this study is to present a medical chatbot that specifically caters to the needs of Hajj pilgrims. By using real-world data and integrating a synthetic dataset, an AI-powered multilingual chatbot is developed. This chatbot effectively interacts with pilgrims, providing them with medical advice and addressing their common inquiries. For pilgrims, having access to accurate and current medical information is essential for providing reliable suggestions and treatment options. The data used are sourced from trustworthy sources and are consistently updated. This not only saves time for healthcare providers but also improves the overall experience for pilgrims. The LLM's advanced natural language understanding capabilities enable it to offer pilgrims highly accurate and relevant information. The contributions of the presented work can be summarized as follows:

- Domain-specific fine-tuning of LLM: We fine-tune a large language model (LLM) specifically for the domain of healthcare and cultural sensitivities relevant to Hajj pilgrims. This fine-tuning process ensures that the model is capable of understanding and generating relevant responses within the context of healthcare conversations during the pilgrimage.
- Introducing the HajjHealthQA dataset: To facilitate the development and evaluation of our healthcare chatbot, we introduce the HajjHealthQA dataset. This dataset contains a diverse collection of questions, answers, and conversations relevant to healthcare issues faced by Hajj pilgrims. We also employ synthetic data augmentation techniques (https://github.com/AbeerMostafa/HajjHealthQA-Dataset (accessed on 1 March 2024)).
- RAG module for uncertainty validation: We add a retrieval-augmented generation (RAG) module to validate uncertain information provided by the chatbot. This mechanism enhances the reliability and accuracy of the chatbot's responses by cross-referencing generated text with external knowledge sources.
- Training a secondary AI agent on the HealthVer dataset: We train two separate models as part of our framework, one on the HajjHealthQA dataset for Hajj-specific healthcare inquiries and another on the HealthVer dataset for medical information verification. The latter is used to verify that the medical information generated by our chatbot is supported by medical evidence.
- Prompt engineering for case study specifics: We employ prompt engineering techniques tailored to the specific case study of building a healthcare chatbot for Hajj pilgrims. This ensures that the chatbot's responses are optimized for relevance, accuracy, and cultural appropriateness within the context of Hajj-related healthcare scenarios.
- Multilingual support: To accommodate the linguistic diversity of Hajj pilgrims, our chatbot offers multilingual support, allowing users to interact in their preferred language.

2. Related Work

The annual Hajj pilgrimage, the largest gathering in the Islamic world, involves millions of pilgrims facing considerable physical and mental challenges. Two studies, conducted during different time frames, contribute to understanding the common health problems encountered by Hajj pilgrims and emphasize the need for further research to address these challenges effectively.

2.1. Health Challenges Faced by Hajj Pilgrims

The study in [15] conducted a comprehensive analysis of common health problems (CHPs) faced by Hajj pilgrims between 1998 and 2013. The research aimed to identify patterns and types of illnesses through a review of articles retrieved from various databases. The analysis of 27 studies involving 17,753 respondents highlighted respiratory diseases (76.2%) as the predominant health issue, followed by skin diseases (7.4%), meningococcal disease (3.7%), and heat stroke (3.7%). The findings, while significant, underscored a limitation in available studies, hindering the formation of definitive conclusions. The study advocated for more research to accurately report the incidence of respiratory diseases among Hajj pilgrims and a big need for methods to help pilgrims overcome these diseases.

The authors in [2] examine the health challenges faced by 2–3 million pilgrims during the Hajj pilgrimage every year from 2008 to 2016. This study focused on diseases and emergency incidents. Respiratory diseases, including pneumonia, influenza, and asthma, were identified as the primary health problems (73.33%), followed by non-communicable diseases such as heat stroke or heart attack (16.67%) and cardiovascular diseases (10%). The study also highlighted the scarcity of research in this domain, emphasizing the need for further investigations to address the diverse health problems experienced by pilgrims during Hajj.

In conclusion, these studies collectively underscore the critical need for further research to comprehensively overcome and address the health challenges faced by Hajj pilgrims. The synthesis of findings from both papers emphasizes the importance of a holistic approach to healthcare assistance and intervention during the Hajj pilgrimage, considering both communicable and non-communicable diseases and the most recent technological approaches to help pilgrims during this significant religious event.

2.2. Medical Q&A

Medical-question-answering chatbots have emerged as a major breakthrough in the field of medical science. These chatbots utilize neural networks and natural language processing to provide personalized healthcare information and advice to users. They can answer user queries related to various health issues, ranging from psychological counseling to domain-specific medical knowledge [16,17]. These advancements in medical-question-answering chatbots have the potential to revolutionize healthcare delivery and improve access to information for patients.

One such chatbot called MedBot aims to identify illnesses, provide necessary information, and promote a healthy lifestyle, reducing the need for hospital visits and healthcare costs [18]. The authors contribute the largest Arabic Healthcare Question and Answer (Q&A) dataset, named MAQA. Three distinct models, namely, long short-term memory (LSTM), bidirectional LSTM (Bi-LSTM), and Transformers, are employed to experiment with the Q&A system. Evaluation metrics such as cosine similarity and BLeU score reveal that the Transformer model surpasses traditional deep learning models, emphasizing its efficacy in Arabic healthcare information processing.

The authors in [19] introduce ChatENT, a specialized knowledge question-and-answer platform. The authors systematically gather OHNS-relevant data from open-access internet sources, indexing it into a database. Retrieval-augmented language modeling (RALM) is employed to recall information, which is integrated into ChatGPT 4.0, creating ChatENT. This paper marks the inception of the first specialty-specific large language model (LLM) in the medical field, demonstrating improved Q&A performance through domain-specific knowledge augmentation.

Focused on biomedical reasoning and classification, the authors in [20] evaluate the performance of various language models, comparing large language models (LLMs) with classic machine learning (ML) approaches. Using the HealthAdvice dataset for classification and CausalRelation for reasoning, the authors compare LLMs with logistic regression models using bag-of-words representations and fine-tuned BioBERT models. Notably, fine-tuning BioBERT yields the best results for both classification and reasoning, emphasizing the importance of tailored approaches in biomedical tasks.

In reference to [21], the paper highlights the innovative use of synthetic data in question and answer generation to enhance the accuracy of question-answering (QA) models, addressing the challenge of limited human-labeled data. The study explores various factors such as model size, pre-trained model quality, scale of synthesized data, and algorithmic choices to bridge the gap between synthetic and human-generated question–answer pairs. Notably, the researchers achieve higher accuracy on the SQUAD1.1 question-answering task using solely synthetic questions and answers compared to using the SQUAD1.1 training set questions alone. The absence of access to real Wikipedia data prompts the synthesis of questions and answers from an 8.3 billion-parameter GPT-2 model, demonstrating that state-of-the-art question-answering networks can be trained on entirely model-generated data with impressive results, achieving an 88.4 exact match (EM) and 93.9 F1-score on the SQUAD1.1 dev set. The methodology is extended to SQUAD2.0, showcasing a 2.8 absolute gain on EM score compared to prior work using synthetic data. This emphasizes the pivotal role synthetic data play in significantly improving the accuracy and performance of QA models.

2.3. Use of Synthetic Data

Reference [22] from Google DeepMind supports the use of synthetic data in mitigating undesirable behaviors exhibited by language models, such as sycophancy. The study explores how models tend to tailor their responses to align with a human user's viewpoint, even when that viewpoint may not be objectively correct. In response, the researchers propose a synthetic-data intervention to reduce sycophantic behavior. The study conducts evaluations on sycophancy tasks and observes that model scaling and instruction tuning exacerbate sycophancy for large language models. The evaluation is extended to statements that are objectively incorrect, revealing that language models may agree with these statements to align with user opinions. The synthetic-data intervention, presented as a straightforward fine-tuning step using public NLP tasks, is shown to significantly reduce sycophantic behavior on held-out prompts. This supports the broader idea that synthetic data can serve as a valuable tool in addressing and mitigating undesirable behaviors in language models, further emphasizing its positive impact on model performance and behavior.

2.4. Hajj Q&A

In terms of the Hajj Q&A system, various studies have been conducted to address questions related to Hajj rituals. An example of such an approach is the Hajj-QAES, proposed by Sulaiman et al. [23]. This expert system utilizes knowledge-based techniques to offer real-time responses to all inquiries regarding the Hajj pilgrimage. In addition to addressing queries, this system plays a crucial role in educating pilgrims by encoding pertinent facts and rules to effectively address their questions during the preparation phase. Sharef et al. [24] worked towards enabling self-guided education for pilgrims. They proposed a semantic-based question-and-answer system that utilizes ontology for knowledge representation. Additionally, M-Hajj [25] introduced the concept of the mobile decision support system, which utilizes case-based reasoning (CBR) and decision trees to offer optimized answers.

As we can observe, all AI agents and apps concerning Hajj and Umrah primarily focus on providing general instructions and logistical assistance for pilgrims (such as Nusuk [26], Mecca WABot [27]). However, they lack specific features designed to deliver healthcare information and support the customized medical needs of pilgrims. There is a notable gap in the availability of apps specifically dedicated to addressing healthcare concerns during Hajj and Umrah, and that is what we propose in this research.

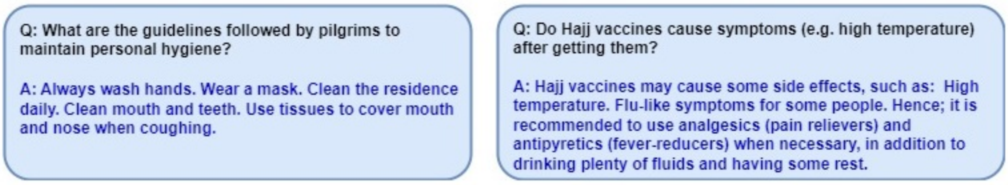
3. HajjHealthQA Dataset

The success of any healthcare chatbot hinges on the quality and diversity of its dataset. In the context of assisting pilgrims during Hajj, a critical and challenging event that draws millions of people annually, it is paramount to have a robust dataset that encompasses real-world medical scenarios and synthetic data to enhance the chatbot's performance. In this section, we provide a detailed overview of the two datasets we collected and used in our study, highlighting their sources, characteristics, and the rationale behind their inclusion. The datasets are publicly available at (https://github.com/AbeerMostafa/HajjHealthQA-Dataset (accessed on 1 March 2024)).

The HajjHealthQA dataset has been obtained from three primary sources: the Ministry of Health (MOH) in the Kingdom of Saudi Arabia (KSA) [28], the World Health Organization (WHO) [29], and the Ministry of Hajj and Umrah (MOHU) in KSA [30]. These sources were accessed on 1 November 2023, as mentioned in the References section. Of the data, 60% are retrieved Q&A from the MOH, 20% from the WHO, and the remaining 20% from the MOHU. The HajjHealthQA dataset includes frequently asked and common questions from users, along with corresponding authoritative answers retrieved from these reputable resources. The MOH resource is the official website of the Ministry of Health in Saudi Arabia, a governmental organization subsidized and funded by the Saudi government. The MOHU portal is a verified website for the Ministry of Hajj and Umrah, a government agency responsible for facilitating the procedures for performing the rituals of Hajj and Umrah in KSA. The WHO, on the other hand, is the United Nations agency that connects nations, partners, and people to promote health, keep the world safe, serve the vulnerable, and spearhead international public health efforts.

The primary foundation of our dataset is comprised of real-world medical questions and answers sourced from reputable platforms, including the official website of the Ministry of Health in Saudi Arabia [28] and various other trusted healthcare websites [29,31,32]. This dataset is a reflection of the actual health concerns and inquiries that pilgrims may encounter during their Hajj journey. To ensure the authenticity and reliability of the data, we employed a systematic approach to curate the real medical questions and answers. Web scraping techniques using Python version 3.10.12 were applied to extract information from the official Ministry of Health website and other credible health platforms. Only verified and authoritative sources were used to compile a diverse range of questions related to common health issues faced by pilgrims during Hajj.

The real-world medical dataset comprises a vast array of health-related queries, covering topics such as preventive measures, vaccination requirements, common ailments, and emergency protocols [28,33,34]. Each question is paired with its corresponding authoritative answer, often sourced directly from healthcare professionals or government health agencies. The inclusion of real-world medical data serves to ground the chatbot's knowledge in the practical concerns of pilgrims. By drawing on actual questions posed by individuals preparing for or participating in Hajj, the chatbot can offer relevant and accurate information tailored to the specific health challenges associated with this religious pilgrimage. Figure 1 shows examples of the collected real-world Q&A data.


Q: What are the guidelines followed by pilgrims to maintain personal hygiene?

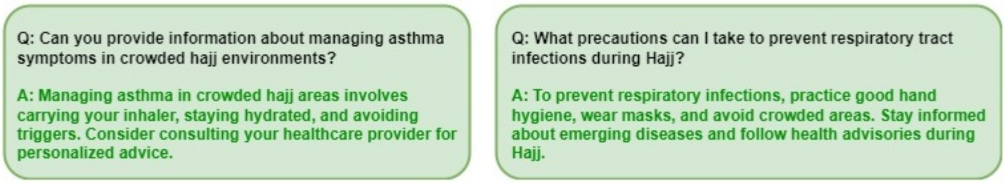
A: Always wash hands. Wear a mask. Clean the residence daily. Clean mouth and teeth. Use tissues to cover mouth and nose when coughing.

Q: Do Hajj vaccines cause symptoms (e.g. high temperature) after getting them?

A: Hajj vaccines may cause some side effects, such as: High temperature. Flu-like symptoms for some people. Hence; it is recommended to use analgesics (pain relievers) and antipyretics (fever-reducers) when necessary, in addition to drinking plenty of fluids and having some rest.

Figure 1. Q&A examples from real-world data.

In addition to real-world data, we incorporated a synthetic dataset generated using ChatGPT, a powerful language model developed by OpenAI. This dataset was designed to supplement the real medical questions and answers, providing a broader spectrum of potential queries and responses that may not be covered by the real dataset alone. Our synthetic dataset was created by prompting ChatGPT with healthcare-related questions specific to the context of Hajj. We first extracted the topics and general themes from the real questions asked by users. Based on these insights, we then formulated synthetic data that accurately reflected these themes. The model's responses were then used to generate a diverse set of synthetic Q&A pairs. Figure 2 shows examples of synthetic Q&A data. This process allowed us to explore hypothetical scenarios, address niche concerns, and anticipate questions that may not have been explicitly addressed in the real medical dataset. The synthetic dataset enhances the chatbot's versatility by introducing a wide range of hypothetical medical scenarios, preventive measures, and nuanced inquiries that pilgrims might have. It complements the real dataset, providing a more comprehensive knowledge base for the chatbot to draw upon when assisting users.


Q: Can you provide information about managing asthma symptoms in crowded hajj environments?

A: Managing asthma in crowded hajj areas involves carrying your inhaler, staying hydrated, and avoiding triggers. Consider consulting your healthcare provider for personalized advice.

Q: What precautions can I take to prevent respiratory tract infections during Hajj?

A: To prevent respiratory infections, practice good hand hygiene, wear masks, and avoid crowded areas. Stay informed about emerging diseases and follow health advisories during Hajj.

Figure 2. Q&A examples from synthetic data.

Our curated dataset comprehensively addresses a spectrum of health concerns prevalent during the Hajj pilgrimage, drawing insights from research papers identifying the most common diseases [1–5]. It encompasses detailed Q&A pairs covering respiratory diseases, pneumonia, influenza, asthma, sunlight effects, cardiovascular diseases, heart diseases, heat strokes, skin diseases, and meningococcal diseases. Recognizing the crowded conditions and environmental factors during Hajj, our dataset provides valuable information on respiratory diseases, pneumonia, and influenza, offering guidance on prevention and intervention. For asthma, tailored advice is given to manage and prevent attacks in the pilgrimage setting. The dataset also delves into the impact of prolonged sunlight exposure and addresses associated health risks. With a focus on cardiovascular diseases, heart diseases, and the potential for heat strokes, the dataset equips pilgrims with knowledge on prevention and management. Skin-related queries cover sun protection, hygiene practices, and managing skin conditions during the pilgrimage. Lastly, preventive measures against meningococcal diseases, including vaccination information and symptom awareness, are integrated into the dataset. This collective information ensures that the healthcare chatbot is well equipped to provide nuanced guidance on a wide array of health challenges faced by pilgrims during Hajj.

4. Methodology

In crafting a robust methodology for the development of our healthcare chatbot tailored to the needs of Hajj pilgrims, a strategic approach was undertaken. In this section, we describe the methodology employed in our research, detailing the process of fine-tuning

GPT-3.5 Turbo and Llama 3 for domain-specific healthcare applications, utilizing the RAG module for uncertainty validation, evaluating text generation using quality metrics, and training a different model on a specialized dataset for medical information verification.

Our methodology workflow, as illustrated in Figure 3, begins with the step of model fine-tuning, where we adapt the LLM to the domain of healthcare communication customized for Hajj pilgrims. This process involves iterative cycles of hyperparameter tuning to optimize the model's performance. Following model fine-tuning, we address the aspect of medical information validation within the generated text. Initially, we add the retrieval-augmented generation (RAG) module, which retrieves knowledge from databases inherited from reputable websites such as the World Health Organization (WHO), the Ministry of Health in the Kingdom of Saudi Arabia (KSA), and the Ministry of Hajj and Umrah. This ensures that our responses are grounded in reliable medical information. Subsequently, we employ another chatbot trained on the HealthVer dataset [35] to verify the accuracy and consistency of our responses against established medical evidence. This dual-validation approach enhances the reliability and trustworthiness of our chatbot's output. Additionally, we incorporate prompt engineering techniques to further refine and optimize the relevance and specificity of our responses within the context of Hajj-related healthcare inquiries. Through this comprehensive workflow, we aim to develop a robust and dependable healthcare chatbot capable of providing accurate and evidence-based support to pilgrims.

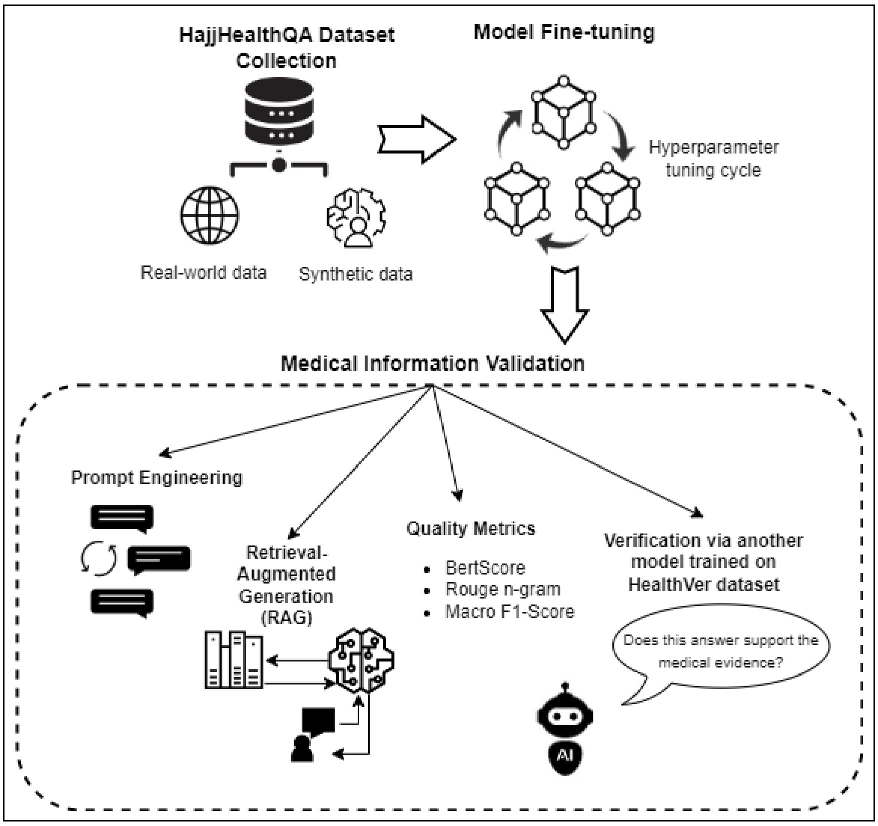

HajjHealthQA Dataset Collection
Real-world data
Synthetic data
Model Fine-tuning
Hyperparameter tuning cycle
Medical Information Validation
Prompt Engineering
Retrieval-Augmented Generation (RAG)
Quality Metrics
BertScore
Rouge n-gram
Macro F1-Score
Verification via another model trained on HealthVer dataset
Does this answer support the medical evidence?
AI

Figure 3. Proposed system architecture.

4.1. Model Fine-Tuning

Central to our methodology is the utilization of the GPT-3.5 Turbo model. It is considered a fine-tuned version of GPT3. GPT-3.5 stands out for its notably superior performance compared to any open-source LLMs. This performance advantage underscores its effectiveness in various tasks and domains, making it a preferred choice for many applications requiring advanced natural language processing capabilities. The model operates on a highly parallelized architecture, enabling efficient processing of vast amounts of data during both pre-training and fine-tuning phases. The fine-tuning process involves modifying the model's pre-trained parameters to specialize in healthcare queries relevant to the pilgrimage context.

GPT-3 follows a pre-training strategy where it is initially trained on a large corpus of diverse text data. During this phase, the model learns to predict the next word in a sentence, given the context of preceding words. This process enables the model to acquire a rich understanding of grammar, syntax, and semantics. After pre-training, GPT-3 leverages the power of transfer learning. The model, once pre-trained, can be fine-tuned for specific downstream tasks, such as language translation, question answering, summarization, and more. This flexibility makes it a versatile tool for a wide array of natural language processing applications.

The model exhibits remarkable zero-shot and few-shot learning capabilities. Zero-shot learning involves making predictions on tasks the model has never seen during training, while few-shot learning involves providing the model with a small amount of task-specific information. It can perform surprisingly well in both scenarios. In our case, we employ a few-shot learning strategy to make the model capture the context of our custom chatbot application. GPT-3.5 allows fine-tuning on specific tasks, enabling users to customize the model for their particular application. This feature enhances the model's adaptability and ensures better performance on domain-specific tasks. There is already evidence that fine-tuning with a few samples is sufficient to improve results from GPT in health domains [36]. We conducted domain-specific fine-tuning using a combination of real-world and synthetic data obtained from the HajjHealthQA dataset. The fine-tuning process involves exposing the model to 150 samples of text data related to healthcare queries, responses, and conversations pertinent to the Hajj pilgrimage. This enables the model to learn domain-specific nuances, language patterns, and cultural sensitivities relevant to healthcare interactions during Hajj. While we tested and evaluated the model's performance on 33 data samples, experiments were performed on real data alone, synthetic data alone, and a combination of real and synthetic data. Details of our experiments are further discussed in Section 5.

In addition, we utilized the Llama3 model, an advanced language model renowned for its enhanced capabilities over its predecessors in handling complex natural language processing tasks. The primary objective was to assess its performance against the GPT-3.5 model, specifically in the context of the same dataset. The fine-tuning process involved extensive training on our dataset, ensuring that both Llama3 and GPT-3.5 were optimized for the same tasks under identical conditions. By maintaining consistent training parameters and evaluation metrics, we aimed to establish a fair and rigorous comparative analysis.

4.2. Retrieval-Augmented Generation

The RAG module stands as a cornerstone of our methodology, serving as a powerful mechanism for enhancing the accuracy, relevance, and reliability of generated text within our healthcare chatbot framework. This module integrates advanced techniques in knowledge retrieval from vector databases inherited from esteemed websites, including the World Health Organization (WHO) [29], the Ministry of Health in the Kingdom of Saudi Arabia (KSA) [28], and the Ministry of Hajj and Umrah [30]. These sources contain a wealth of information spanning diverse healthcare topics, including disease prevention, treatment guidelines, health advisories, and cultural considerations specific to the Hajj pilgrimage.

Using the RAG module, the vector database is prepared as follows. We first collect and load our data from the previously mentioned sources. Afterwards, we split these documents into smaller chunks of data. Lastly, we embed and store these chunks of data using vector embeddings to enable semantic search across the text chunks. All this process is implemented using langchain. As an example, if a user poses a query regarding the management of a specific medical condition during the Hajj pilgrimage, the RAG module retrieves relevant knowledge from the inherited databases, such as WHO's recommendations for managing infectious diseases in crowded settings or the

Ministry of Health's guidelines for providing medical care to Hajj pilgrims. This retrieved knowledge is then used to validate the accuracy of the generated response, ensuring that it aligns with authoritative medical information and cultural considerations specific to the Hajj pilgrimage.

4.3. Evidence-Based Verification

Using an LLM to evaluate another LLM enhances the reliability and objectivity of the assessment process by leveraging advanced language understanding capabilities [37]. In this stage, we conducted evidence-based verification to ensure the accuracy and reliability of the responses generated by our healthcare chatbot. Our approach involved fine-tuning the GPT-3.5 Turbo model on the HealthVer dataset [35], which comprises medical questions paired with claims and corresponding evidence. This dataset served as a valuable resource for training a model to discern whether the provided evidence supports, refutes, or is neutral to the given claim. By fine-tuning the model on this dataset, we aimed to imbue it with the ability to evaluate the validity and credibility of medical claims based on supporting evidence.

During the fine-tuning process, we presented the model with medical questions accompanied by claims and associated evidence, prompting it to classify the evidence as supporting, refuting, or neutral in relation to the claim. This iterative training procedure enabled the model to learn to recognize patterns and correlations between medical claims and evidence, thereby enhancing its ability to make evidence-based judgments.

Once trained, we integrated this verification model into our healthcare chatbot framework for testing purposes. When generating responses to inquiries using our original model, which was fine-tuned on the HajjHealthQA dataset, we employed the verification model to assess the validity of the generated answers. After obtaining a response from our original model, we presented it to the verification model and asked whether the provided evidence supported, refuted, or was neutral to the reference answer. An example of this workflow is illustrated in Figure 4.

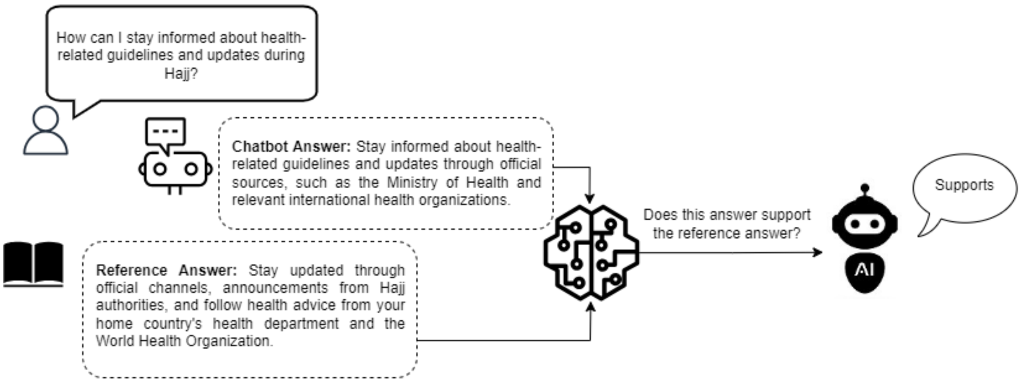

How can I stay informed about health-related guidelines and updates during Hajj?
Chatbot Answer: Stay informed about health-related guidelines and updates through official sources, such as the Ministry of Health and relevant international health organizations.
Reference Answer: Stay updated through official channels, announcements from Hajj authorities, and follow health advice from your home country's health department and the World Health Organization.
Does this answer support the reference answer?
AI
Supports

Figure 4. Example of verification via another AI agent.

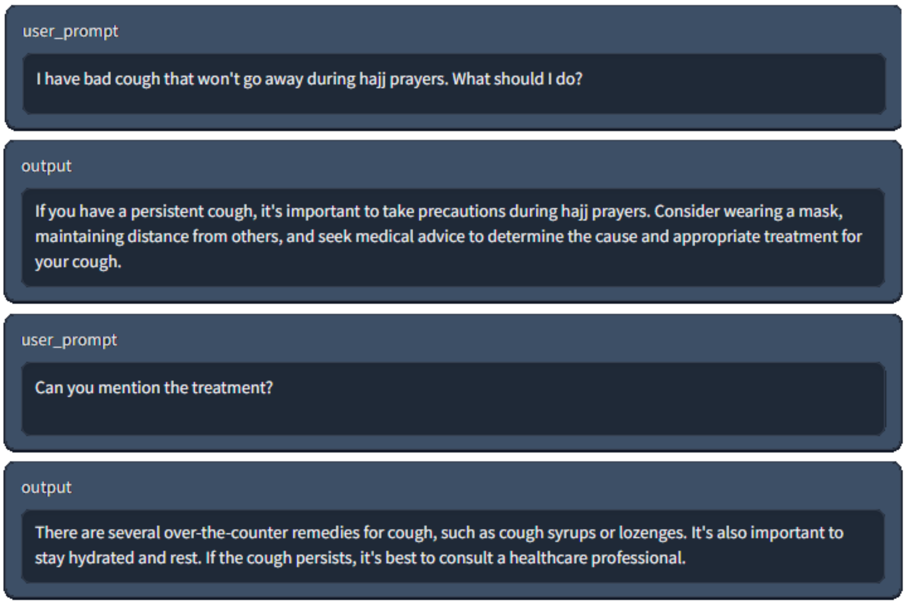
This verification step served as a crucial quality control measure, allowing us to validate the accuracy and reliability of the responses generated by our chatbot. By leveraging the verification model trained on the HealthVer dataset, we were able to evaluate the consistency and alignment of the generated answers with established medical evidence. Responses that were deemed to be supported by the evidence provided a higher level of confidence in their accuracy, while those refuted by the evidence signaled potential inaccuracies or inconsistencies that required further refinement. This rigorous validation process underscores our commitment to upholding standards of evidence-based practice in healthcare communication, ultimately contributing to the provision of reliable and trustworthy support for Hajj pilgrims and other users seeking medical information.

4.4. Prompt Engineering

Prompt engineering plays a pivotal role in the effectiveness and user experience of conversational agents. In the context of our healthcare chatbot, the process of crafting prompts is particularly crucial to ensure accurate, relevant, and context-aware responses. Here, we explore the strategies and considerations involved in prompt engineering for our specialized chatbot. This prompt engineering process is mainly considered role-based as we assign a specific role to the conversational agent to stick to the specified topics and provide only relevant responses.

- Task-specific prompts

  The healthcare chatbot serves a range of purposes, from providing medical information to offering guidance on common health concerns during the pilgrimage. Task-specific prompts are necessarily formulated to elicit relevant information from users. For instance, a pilgrim seeking advice on managing heat-related illnesses might receive prompts that inquire about their current location, symptoms, and recent activities. We made sure that all aspects of potential prompts were included in our dataset. The Q&A datasets include prompts about protection, identifying symptoms, healthcare tips and advice, vaccination, and handling emergency situations. Figure 5 shows an example of a conversation about dealing with symptoms.



I have bad cough that won't go away during hajj prayers. What should I do?

If you have a persistent cough, it's important to take precautions during hajj prayers. Consider wearing a mask, maintaining distance from others, and seek medical advice to determine the cause and appropriate treatment for your cough.

Can you mention the treatment?

There are several over-the-counter remedies for cough, such as cough syrups or lozenges. It's also important to stay hydrated and rest. If the cough persists, it's best to consult a healthcare professional.

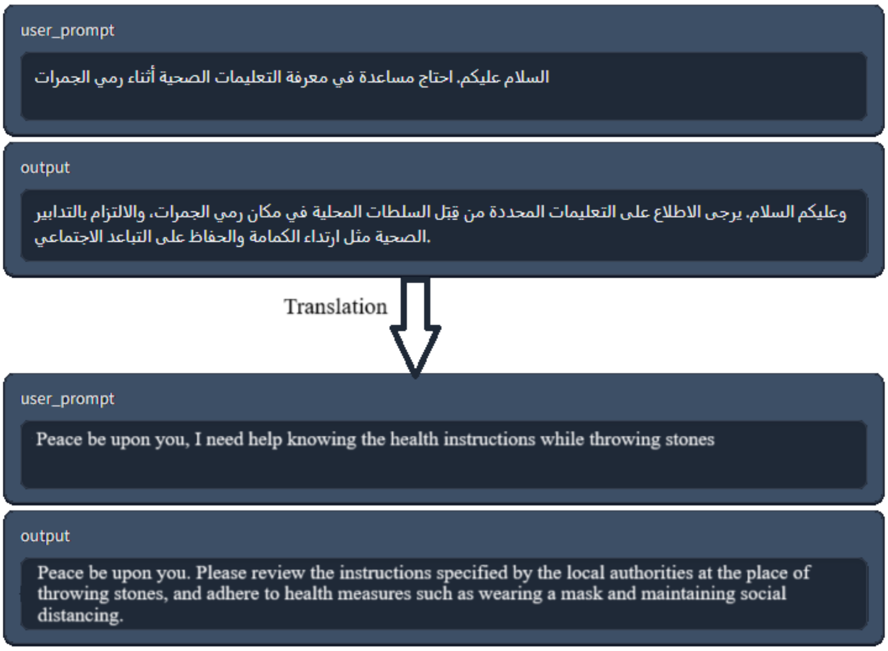
Figure 5. Example of task-specific prompts.

- Multilingual support

  Hajj attracts a diverse group of pilgrims from various linguistic backgrounds. The chatbot's prompt system is designed to handle multiple languages to accommodate the linguistic diversity of the users. Pilgrims can interact with the chatbot seamlessly in their preferred language, ensuring inclusivity and accessibility. We can see this when using Arabic language in the prompt illustrated in Figure 6.

- Customization for cultural sensitivity

  Understanding the cultural and religious context of Hajj is paramount in providing meaningful interactions. Prompts are carefully designed to align with the sensitivities and nuances associated with the pilgrimage. This involves incorporating appropriate greetings, addressing pilgrims respectfully, and avoiding content that may be deemed culturally insensitive. Figure 6 illustrates a prompt to emphasize culture sensitivity, where we used Arabic language, Islamic greetings, and a question related to the Hajj rituals.



user_prompt

السلام عليكم. احتاج مساعدة في معرفة التعليمات الصحية أثناء رمي الجمرات

output

وعليكم السلام. يرجى الاطلاع على التعليمات المحددة من قِبَل السلطات المحلية في مكان رمي الجمرات، والالتزام بالتدابير الصحية مثل ارتداء الكمامة والحفاظ على التباعد الاجتماعي.

user_prompt

Peace be upon you, I need help knowing the health instructions while throwing stones

output

Peace be upon you. Please review the instructions specified by the local authorities at the place of throwing stones, and adhere to health measures such as wearing a mask and maintaining social distancing.

Figure 6. Customization for cultural sensitivity via prompts.

- Contextual awareness and follow-up prompts

  To enhance the continuity of conversations, the chatbot is equipped with contextual awareness. It can recall previous interactions and responses, allowing for more coherent and personalized conversations. Follow-up prompts are carefully crafted to seek additional details or provide further assistance based on the context of the ongoing conversation. This feature ensures that the chatbot can address complex healthcare inquiries in a more dynamic and iterative manner. Figure 5 shows an example of a user asking a follow-up question which the chatbot could handle easily.
- Iterative improvement through user feedback

  Prompt engineering is an ongoing process, and user feedback is invaluable for refining and enhancing the effectiveness of the chatbot. Pilgrim interactions are analyzed to identify areas of improvement in prompt design, language clarity, and overall user satisfaction. This iterative feedback loop allows for continuous optimization of the prompt system to better serve the healthcare needs of Hajj pilgrims.

5. Experimental Setup

In this section, we detail the experiments conducted to fine-tune and evaluate our healthcare chatbot, leveraging the GPT-3.5 Turbo model. The primary objective was to optimize the model's performance for delivering reliable information to pilgrims. The fine-tuning experiments focused on three key hyperparameters: the number of epochs, batch size, and learning rate multiplier. We iteratively fine-tuned the model, systematically adjusting hyperparameters to achieve optimal performance.

5.1. Hyperparameter Tuning

The effectiveness of our healthcare chatbot heavily relies on the fine-tuning process of the GPT-3.5 Turbo model. Central to this process is the optimization of hyperparameters, which are crucial configuration settings influencing the learning dynamics of the model.

5.1.1. Number of Epochs

The number of training epochs, representing the complete pass through the entire training dataset, is a critical hyperparameter. Too few epochs may result in underfitting, while too many epochs may lead to overfitting. Through a systematic approach, we determined an appropriate number of epochs that ensures the model captures the intricacies of healthcare-related queries specific to Hajj pilgrims without compromising on generalization. Figure 7 illustrates the accuracy achieved by our experiments with trial and error techniques.

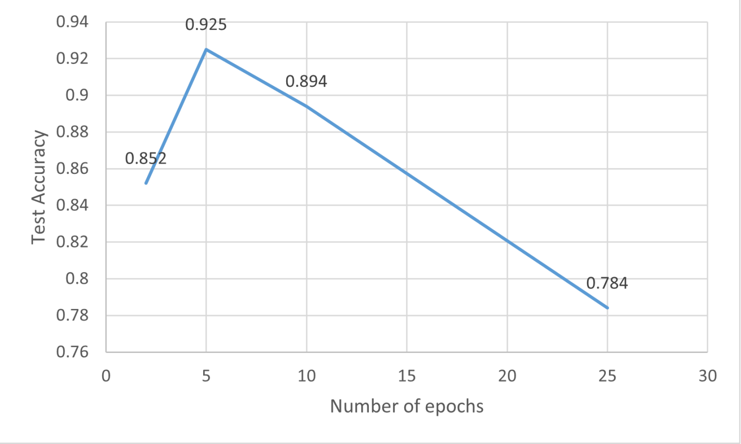


Figure 7. Number of epochs vs. test accuracy.

5.1.2. Batch Size

The choice of batch size, representing the number of training samples processed in a single iteration, significantly affects the model's training efficiency. A carefully selected batch size ensures efficient memory utilization and accelerates convergence. Our experimentation involved varying batch sizes to identify an optimal setting that balances computational efficiency and training stability; Figure 8 illustrates this process.

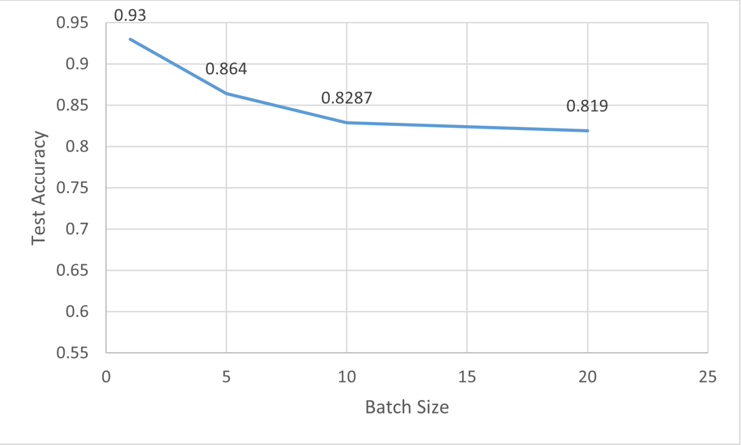


Figure 8. Batch size vs. test accuracy.

5.1.3. Learning Rate Multiplier

The learning rate is a hyperparameter that determines the step size at each iteration while moving toward a minimum of a loss function. It influences how much the model's weights should be updated during training. The learning rate multiplier is a factor applied to the learning rate for a specific set of parameters or layers in a neural network. It allows you to adjust the learning rate differently for different parts of the model. This can be useful when certain parameters require more or less aggressive updates than others. For example, in transfer learning, you might want to fine-tune only the last few layers of a pre-trained model with a smaller learning rate to avoid overfitting to the new data. Tuning the learning rate multiplier is important in fine-tuning pre-trained language models as it facilitates the adaptation of the model to a specific task or domain. This hyperparameter adjustment is instrumental in preventing catastrophic forgetting, where the model loses previously acquired knowledge during fine-tuning, by enabling selective updates to different layers. It contributes to stabilizing training by achieving a balance between stable convergence and effective task-specific adaptation. The learning rate multiplier also plays a pivotal role in controlling the rate of change, ensuring that the model efficiently converges within a reasonable computational budget. Figure 9 visually represents the relationship between the learning rate multiplier and test accuracy during the fine-tuning process.

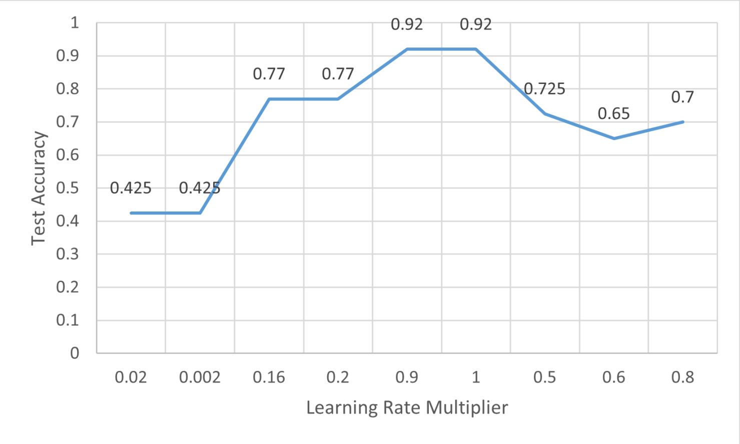


Figure 9. Learning rate multiplier vs. test accuracy.

5.2. Evaluation Metrics

In assessing the efficacy of our healthcare chatbot tailored for pilgrims, a meticulous evaluation strategy was deployed, employing a diverse set of metrics to gauge the model's performance across multiple dimensions. The fundamental criterion for evaluation, accuracy, served as a foundational metric to measure the model's ability to provide correct responses to healthcare queries posed by pilgrims during their journey. This metric was particularly crucial, considering the critical nature of healthcare information dissemination in the context of pilgrimage.

The evaluation framework further incorporated the ROUGE score (Recall-Oriented Understudy for Gisting Evaluation) [38]. ROUGE is a set of metrics used for the automatic evaluation of machine-generated text, particularly in the context of natural language processing tasks such as text summarization and machine translation. ROUGE focuses on assessing the quality of summaries or translations by comparing them to reference (human-generated) summaries using various measures. In the context of our healthcare chatbot for pilgrims, we use ROUGE as an evaluation metric to measure how well the responses generated by our chatbot align with reference responses or desired information. Adapted to our healthcare chatbot, ROUGE enabled us to measure the recall-oriented overlap between the model's responses and the reference data, providing a nuanced perspective on the comprehensiveness of the generated information. One fundamental aspect of ROUGE is its computation of scores based on n-gram overlap, where "n" represents the size of the contiguous word sequences. The formula for ROUGE-N involves calculating the count of overlapping n-grams in the generated text and reference, normalized by the total count of n-grams in the reference, as shown in Equation (1).

$$\mathrm{ROUGE-N} = \frac{\mathrm{Count\,of\,overlapping\,n-grams}}{\mathrm{Total\,count\,of\,n-grams\,in\,reference}} \quad (1)$$

The third evaluation metric we considered was precision. Precision assesses the accuracy of the chatbot's responses by measuring the proportion of correctly generated responses among all responses produced by the chatbot. As illustrated in Equation (2), precision answers the question: "Of all the responses generated by the chatbot, how many are actually correct?". A high precision indicates that the chatbot tends to provide accurate responses, minimizing the likelihood of producing incorrect or irrelevant information.

$$\text{Precision} = \frac{\text{Count of overlapping n} - \text{grams}}{\text{Total count of n} - \text{grams in generated text}} \tag{2}$$

F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both accuracy and completeness. It is particularly useful when there is an uneven class distribution between correct and incorrect responses. The F1-score formula is given by Equation (3). In this study, we use the macro F1-score. This metric is particularly valuable in text generation tasks where multiple dimensions of quality need to be considered, such as coherence, relevance, and fluency. However, it is important to note that the macro F1-score does not account for the inherent diversity in generated text and may not fully capture nuanced aspects of quality.

$$\text{F1} - \text{Score} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}} \tag{3}$$

Lastly, we employ BERTScore [39], which is an automatic evaluation metric to asses the quality of generated text. It is rooted in contextual embeddings derived from pre-trained BERT models. Unlike traditional metrics that rely on exact matches, BERTScore computes similarity scores between tokens in the generated text and tokens in the reference text using contextual embeddings obtained from BERT models. This approach enables BERTScore to capture semantic similarity and contextual relevance, providing a more nuanced evaluation of text generation quality. The method exhibits strong correlation with human judgments and better performance compared to existing metrics. Additionally, BERTScore demonstrated greater robustness to challenging examples in an adversarial paraphrase detection task. In our evaluation, we leveraged BERTScore to assess the similarity and relevance of generated responses to reference answers, providing a more comprehensive and contextually informed measure of text generation quality.

These metrics collectively contribute to a comprehensive evaluation of the language model's ability to generate responses that are both accurate and comprehensive. In the context of our chatbot, achieving a high precision ensures that the generated responses are correct, while a high recall ensures that a significant portion of relevant responses is captured. Striking the right balance between these metrics is pivotal, aligning with the specific goals of accurate and contextually relevant response generation. This multifaceted evaluation framework guides iterative improvements, fostering the development of a highly effective and context-aware healthcare chatbot for pilgrims during Hajj.

6. Results and Discussion

In this section, we delve into the results obtained by our proposed methodology, which highlight the performance metrics of our healthcare chatbot across different training data scenarios. The metrics include accuracy, ROUGE n-gram, precision, macro F1-score, and BERTScore. These metrics provide a comprehensive view of the chatbot's effectiveness in generating responses. It is noted that according to the latest improvements in OpenAI's API version 1.3.0 (as of 20 May 2024), there is a good boost in the reported accuracy.

6.1. Accuracy Analysis

Figure 10 explains the GPT-3.5 Turbo fine-tuning progression during the downstream training task steps exclusively with real-world data of the HajjHealthQA dataset. The right chart exhibits the dynamic evolution of training and validation loss, revealing a steady decline in both metrics, indicating successful convergence. The left chart details mean token accuracy for training and validation sets, showcasing the model's improving proficiency in accurate token-level predictions as training progresses.

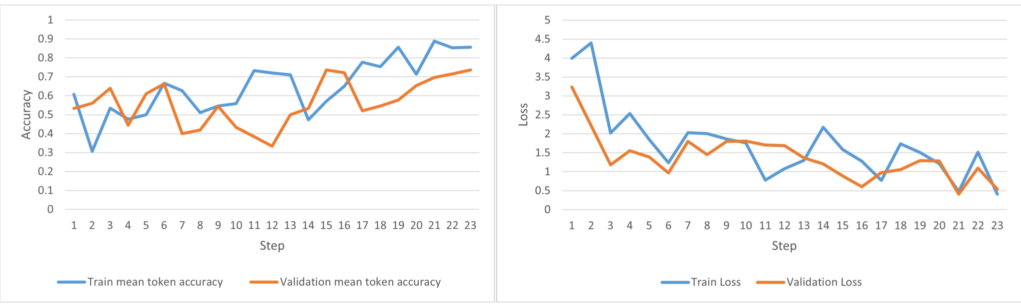


Figure 10. Model performance when using real-world data only.

Figure 11 illustrates GPT-3.5 Turbo training dynamics using synthetic data only. The right chart demonstrates a notable decrease over steps, indicating effective convergence. The left chart reveals the model's capacity to accurately predict tokens during training. Comparing Figure 11 to Figure 10, the synthetic data exhibit faster convergence in terms of loss and a slightly different trend in token accuracy. This suggests the unique impact of synthetic data on the model's learning dynamics, emphasizing their role in shaping the model's performance characteristics.

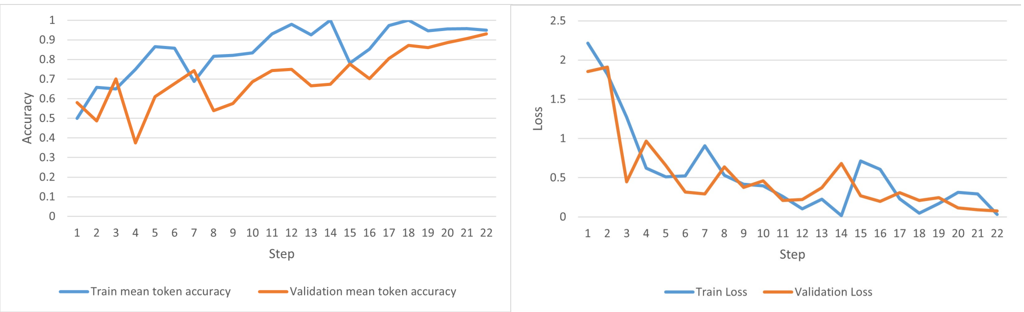


Figure 11. Model performance when using synthetic data only.

Figure 12 demonstrates the influence of different synthetic-to-real data ratios on model performance. The results show that increasing the proportion of synthetic data generally maintains higher accuracy levels, particularly in validation performance. When synthetic data dominate, the model exhibits strong validation accuracy, suggesting that synthetic data are more effective in training the model. Conversely, as the proportion of real data increases, there is a noticeable decline in validation accuracy. This trend highlights the critical role of synthetic data in achieving better overall performance in both training and validation phases.

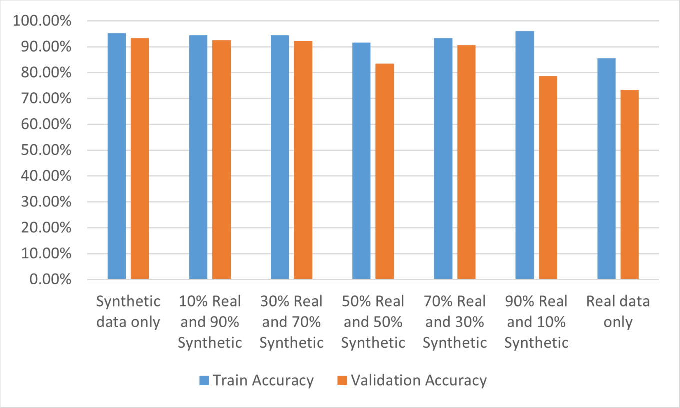


Figure 12. Mean token accuracy when using different portions of real/synthetic data.

In Table 1, we present the accuracy achieved with different model components in our methodology, including base GPT-3.5 Turbo, fine-tuning on the HajjHealthQA dataset, and fine-tuning combined with the RAG module. We observe that incorporating fine-tuning techniques leads to improved accuracy across all datasets compared to using the base GPT-3.5 Turbo model alone. Specifically, fine-tuning on synthetic data yields the highest accuracy, highlighting the effectiveness of synthetic data augmentation in enhancing model performance and the importance of domain-specific fine-tuning. Furthermore, integrating the RAG module with the fine-tuned model further improves accuracy by almost 5%. This underscores the great effect of knowledge retrieval in improving the accuracy of the healthcare chatbots and also enhancing reliability by providing validated information.

Table 1. Mean token accuracy achieved with different model components.

Dataset	GPT-3.5 Turbo	Fine-Tuning	Fine-Tuning + RAG
Real data only	68.1%	73.3%	79.8%
Synthetic data only	86.6%	93.3%	97.4%
Real and synthetic 50/50	76.4%	83.5%	89%

6.2. Quality Metrics

In Table 2, we present the results obtained on the HajjHealthQA test dataset, evaluating the performance of our healthcare chatbot using quality evaluation metrics such as ROUGE, precision, and F1-score. We observe that using synthetic data alone yields slightly higher scores compared to using real data only, indicating the effectiveness of synthetic data augmentation in improving model performance. However, combining real and synthetic data leads to comparable results, suggesting that a balanced approach incorporating both types of data can maintain performance while benefiting from the diversity provided by synthetic data.

Table 2. Results on HajjHealthQA test data.

Dataset	ROUGE	Precision	F1-Score
Real data only	0.78	0.76	0.76
Synthetic data only	0.92	0.89	0.9
Real and synthetic 50/50	0.87	0.84	0.84

Table 3 showcases the results of our model evaluation using BERTScore. Here, we observe consistently high recall, precision, and F1-score across different datasets. Specifically, using synthetic data alone yields the highest scores, indicating that our chatbot generates responses with high relevance and semantic similarity to reference answers when trained on synthetic data. Additionally, combining real and synthetic data maintains high performance, further validating the effectiveness of our approach in enhancing the quality of generated responses.

Table 3. Results using BERTScore evaluation on GPT-3.5 Turbo.

Dataset	Recall	Precision	F1-Score
Real data only	0.873	0.844	0.86
Synthetic data only	0.93	0.91	0.92
Real and synthetic 50/50	0.91	0.9	0.898

The BERTScore results are relatively higher because it is designed to capture semantic similarity between generated responses and reference answers. This results in more accurate and meaningful evaluations of text generation quality. Additionally, the effectiveness of BERTScore may be attributed to its ability to consider the entire context of the generated response and reference text, leading to higher scores that reflect a closer alignment between the model-generated responses and human references.

Table 4 highlights the competitive performance of Llama3 compared to GPT-3.5 Turbo, especially in handling synthetic and mixed datasets. Llama3 shows robust and balanced metrics across different data types, indicating its strong generalization capabilities. However, GPT-3.5 Turbo slightly outperforms Llama3 in scenarios involving synthetic and combined datasets, achieving higher precision and F1-scores. This suggests that while Llama3 is highly effective and reliable, GPT-3.5 Turbo may have a slight edge in generating responses with higher semantic relevance and accuracy when evaluated using BERTScore. Despite this, Llama3's consistency across various datasets underscores its potential for applications requiring diverse data handling and robust performance.

Table 4. Results using BERTScore evaluation on Llama3.

Dataset	Recall	Precision	F1-Score
Real data only	0.87	0.85	0.86
Synthetic data only	0.89	0.88	0.88
Real and synthetic	0.88	0.86	0.86

The incorporation of synthetic data has played a pivotal role in enhancing the overall performance of our healthcare chatbot for pilgrims. This strategic integration of simulated data into the training set has yielded improvements in key metrics, particularly accuracy, ROUGE scores, and precision, showcasing the potential benefits of leveraging synthetic data augmentation in the development of conversational agents.

For the quantitative quality measurements conducted for this chatbot, the evaluation process involved obtaining 50 responses from individuals, each of whom assessed one of the 10 questions posed and the corresponding answer generated by the bot. It is notable that 60% of the recipients were planning to undertake the Hajj pilgrimage, while the remaining 40% had already completed the Hajj. This diversity in the sample population provides valuable insights into the bot's performance in addressing the needs and inquiries of individuals at different stages of their Hajj-related medical journey. The response data indicate that the majority of recipients, 68%, were very satisfied with the answers provided by the bot. An additional 20% expressed satisfaction, while 10% were not satisfied, and 2% were very unsatisfied. This distribution of user satisfaction levels suggests that the bot's performance was generally well received, with a significant proportion of users finding the responses informative and helpful.

In addition, our evidence-based fact-checking module performed very well. We have tested it with all the synthetic dataset generated by ChatGPT to ensure the trustworthiness of the generated text. From all 150 answers, we received only 2 answers that were not supported by the medical evidence; therefore, these 2 answers were removed from the dataset. This validation ensures that all synthetic data align with trustworthy medical information, maintaining the high standards of our ground truth data. In Table 5, we provide examples of prompting the validation AI agent for fact-checking. Each prompt presents a claim generated by the healthcare chatbot and asks the validation AI agent to determine if the claim is supported by the provided evidence. This gives insights into the reliability and accuracy of the generated responses. By leveraging fact-checking with another AI agent, we ensure the credibility and trustworthiness of the healthcare chatbot's responses, validating the effectiveness of our methodology in providing evidence-based information to users.

Table 5. Examples of prompting the second AI agent for fact-checking. The general template for the prompt begins with: "You are a healthcare chatbot who is responsible for validating health information by checking if a health evidence supports or refutes a specific health claim. You should only reply with the word (Supports) or the word (Refutes) or the word (Neutral).", where the claim is the chatbot's answer and the evidence is the reference answer.

Prompt	Output
Please decide if the following claim supports the evidence. **Claim:** Engage in light to moderate physical activities, such as walking, and avoid strenuous exercises. Rest when needed to prevent overexertion. **Evidence:** Engage in light exercises, such as walking, and pace yourself during rituals. Listen to your body, take breaks, and avoid strenuous activities that may lead to exhaustion.	SUPPORTS
Please decide if the following claim supports the evidence. **Claim:** Elderly pilgrims should consult with their healthcare provider to ensure they are physically able to participate in Hajj. They should also take precautions to prevent heat-related illnesses and stay hydrated. **Evidence:** Elderly pilgrims should undergo a thorough medical evaluation before Hajj. Consider factors such as mobility, medication management, and the overall impact on their health.	SUPPORTS
Please decide if the following claim supports the evidence. **Claim:** Yes, there are medical facilities available during Hajj to provide emergency care. **Evidence:** Yes, medical facilities are set up along the Hajj route, and hospitals are equipped to handle emergencies.	SUPPORTS

6.3. Comparison with Benchmark Results

Comparing the benchmark results presented in Table 6 to our results, our healthcare chatbot exhibits noteworthy performance (these results were obtained on 27 December 2023 and is the most recent benchmark as of our knowledge). While the authors in [40] focus on datasets like MMLU in knowledge-based question-answering tasks, our chatbot targets the unique context of providing health-related information to pilgrims during Hajj. We should also mention that GPT 4 is not yet available for public use. However, our chatbot's accuracy, especially when trained on synthetic data, outperforms most of the benchmark results, showcasing its effectiveness in generating accurate responses.

Table 6. LLM benchmark on knowledge-based QA tasks [40].

Dataset	Gemini Pro	GPT-3.5 Turbo	GPT 4 Turbo	Mixtral
MMLU (5-shot)	65.22	67.75	80.48	68.81
MMLU (CoT)	62.09	70.07	78.95	59.57

In conclusion, our healthcare chatbot demonstrates competitive performance compared to benchmark results, in addition to more reliability. The integration of the RAG module and synthetic data has played a crucial role in achieving these results, suggesting their effectiveness in enhancing healthcare chatbots' robustness, language understanding, and responsiveness.

7. Data Privacy and Ethical Considerations

Ethical considerations are crucial in the development and deployment of any AI system, especially in the healthcare domain. We provide a discussion on the ethical considerations related to the development of our healthcare chatbot for pilgrims.

Respecting user privacy and ensuring the security of sensitive medical information are paramount. The dataset used for training the chatbot was carefully anonymized and stripped of personally identifiable information to safeguard the privacy of individuals. Additionally, protective measures will be considered within the deployment stage to protect any user data collected during interactions with the chatbot. Users interacting with the chatbot will be provided with clear and concise information regarding the nature of the chatbot, the purpose of data collection, and the handling of their queries. Informed consent mechanisms are considered to ensure that users are aware of and agree to the chatbot's functionalities. Transparency in communication is prioritized to build trust and empower users to make informed decisions about their interactions with the chatbot.

Given the diverse backgrounds of pilgrims, special attention was given to incorporating cultural sensitivity into the chatbot's responses. The model was fine-tuned to recognize and respect cultural nuances in medical queries, ensuring that responses were not only accurate from a medical standpoint but also culturally appropriate. Furthermore, the chatbot was designed to provide explanations and elaboration for its responses, especially in cases where medical advice or information was provided. Ensuring transparency in the decision-making process of the model fosters user understanding and trust.

By integrating these ethical considerations into the development process, the healthcare chatbot aimed to prioritize user well-being, privacy, and trust, fostering a responsible and ethical deployment within the pilgrim community.

8. Conclusions

In conclusion, our study presents a comprehensive methodology designed to enhance the reliability and effectiveness of healthcare conversational agents, particularly customized to meet the needs of Hajj and Umrah pilgrims. Through experimentation and evaluation, we have demonstrated the efficacy of our approach in addressing key challenges faced by current healthcare conversational agents.

Our methodology incorporates domain-specific fine-tuning of the GPT-3.5 Turbo and Llama3 models with the help of our collected HajjHealthQA dataset. An essential component of our methodology is addressing medical information validation and reliability. We implement evidence-based validation mechanisms, such as the RAG module and secondary AI agents for fact-checking. These mechanisms play a crucial role in ensuring the accuracy and credibility of the generated responses. By retrieving knowledge from reputable sources and cross-verifying responses with established medical evidence, we mitigate the risk of misinformation.

Our evaluation results across multiple metrics demonstrate the effectiveness of our methodology in delivering reliable healthcare assistance by achieving competitive results against state-of-the-art models. The significant improvements observed in these metrics highlight the success of our approach in filling critical gaps in current healthcare conversational agents.

In future work, we aim to further refine and extend our methodology to enhance the scalability of healthcare conversational agents for Hajj and Umrah pilgrims. This could involve integrating real-time data sources to enable dynamic updates and personalized recommendations based on individual profiles and current environmental conditions. Furthermore, collaboration with healthcare professionals will be essential to ensure the relevance of the information provided by the conversational agents. Additionally, integrating multimodal capabilities, such as support for text, voice, and image inputs, could broaden the range of interactions and accommodate diverse user preferences and accessibility needs.

Author Contributions: Conceptualization, H.M.A. and A.M.; methodology, A.M.; software, A.M.; validation, A.M. and H.M.A.; formal analysis, H.M.A.; investigation, H.M.A.; resources, A.M.; data curation, A.M.; writing—original draft preparation, A.M. and H.M.A.; writing—review and editing, H.M.A. and A.M.; visualization, A.M. 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: The original data presented in the study are openly available in https://github.com/AbeerMostafa/HajjHealthQA-Dataset (accessed on 1 March 2024).

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

References

1. Abdelmoety, D.; El-Bakri, N.; Almowalld, W.; Turkistani, Z.; Bugis, B.; Baseif, E.; Melbari, M.H.; AlHarbi, K.; Abu-Shaheen, A. Characteristics of Heat Illness during Hajj: A Cross-Sectional Study. *BioMed Res. Int.* **2018**, *2018*, 5629474. [CrossRef] [PubMed]
2. Al-Masud, S.M.R.; Bakar, A.A.; Yussof, S. Determining the types of diseases and emergency issues in Pilgrims during Hajj: A literature review. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 87–94.
3. Razavi, S.M.; Mardani, M.; Salamati, P. Infectious diseases and preventive measures during hajj mass gatherings: A review of the literature. *Arch. Clin. Infect. Dis.* **2018**, *13*, e62526. [CrossRef]
4. Salmon-Rousseau, A.; Piednoir, E.; Cattoir, V.; de La Blanchardiere, A. Hajj-associated infections. *MEdecine Mal. Infect.* **2016**, *46*, 346–354. [CrossRef] [PubMed]
5. Yezli, S.; Yassin, Y.; Mushi, A.; Almuzaini, Y.; Khan, A. Pattern of utilization, disease presentation, and medication prescribing and dispensing at 51 primary healthcare centers during the Hajj mass gathering. *BMC Health Serv. Res.* **2022**, *22*, 143. [CrossRef] [PubMed]
6. Shen, Y.; Heacock, L.; Elias, J.; Hentel, K.D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and other large language models are double-edged swords. *Radiology* **2023**, *307*, e230163. [CrossRef] [PubMed]
7. Javaid, M.; Haleem, A.; Singh, R.P. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCounc. Trans. Benchmarks Stand. Eval.* **2023**, *3*, 100105. [CrossRef]
8. De Angelis, L.; Baglivo, F.; Arzilli, G.; Privitera, G.P.; Ferragina, P.; Tozzi, A.E.; Rizzo, C. ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Front. Public Health* **2023**, *11*, 1166120. [CrossRef] [PubMed]
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
10. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 5 February 2024).
11. Glik, D.C. Risk communication for public health emergencies. *Annu. Rev. Public Health* **2007**, *28*, 33–54. [CrossRef]
12. Almehmadi, M.; Pescaroli, G.; Alqahtani, J.S.; Oyelade, T. Investigating health risk perceptions during the Hajj: Pre-Travel advice and adherence to preven-tative health measures. *Afr. J. Respir. Med.* **2021**, *16*, 1–6.
13. Alqahtani, A.S.; Tashani, M.; Heywood, A.E.; Booy, R.; Rashid, H.; Wiley, K.E. Exploring Australian Hajj Tour Operators' Knowledge and Practices Regarding Pilgrims' Health Risks: A Qualitative Study. *JMIR Public Health Surveill.* **2019**, *5*, e10960. [CrossRef] [PubMed]
14. Aljohani, A.; Nejaim, S.; Khayyat, M.; Aboulola, O. E-government and logistical health services during Hajj season. *Bull. Natl. Res. Cent.* **2022**, *46*, 112. [CrossRef]
15. Dzaraly, N.D.; Rahman, N.I.A.; Simbak, N.B.; Ab Wahab, S.; Osman, O.; Ismail, S.; Haque, M. Patterns of communicable and non-communicable diseases in Pilgrims during Hajj. *Res. J. Pharm. Technol.* **2014**, *7*, 12.
16. Abdelhay, M.; Mohammed, A.; Hefny, H.A. Deep learning for Arabic healthcare: MedicalBot. *Soc. Netw. Anal. Min.* **2023**, *13*, 71. [CrossRef] [PubMed]
17. Singh, S.; Susan, S. Healthcare Question–Answering System: Trends and Perspectives. In Proceedings of the International Health Informatics Conference: IHIC 2022, Cuttack, India, 17–19 May 2022; Springer: Cham, Switzerland, 2023; pp. 239–249.
18. Pal, V.K.; Singh, S.; Sinha, A.; Shekh, M.S. Medical Chatbot using AI and NLP. *i-Manag. J. Softw. Eng.* **2022**, *16*, 46.
19. Long, C.; Subburam, D.; Lowe, K.; Santos, A.d.; Zhang, J.; Hwang, S.; Saduka, N.; Horev, Y.; Su, T.; Cote, D.; et al. ChatENT: Augmented Large Language Model for Expert Knowledge Retrieval in Otolaryngology-Head and Neck Surgery. *medRxiv* **2023**, 2023-08. . [CrossRef] [PubMed]
20. Chen, S.; Li, Y.; Lu, S.; Van, H.; Aerts, H.J.W.L.; Savova, G.K.; Bitterman, D.S. Evaluating the ChatGPT family of models for biomedical reasoning and classification. *J. Am. Med. Inform. Assoc. JAMIA* **2024**, *31*, 940–948. [CrossRef]
21. Puri, R.; Spring, R.; Shoeybi, M.; Patwary, M.; Catanzaro, B. Training Question Answering Models From Synthetic Data. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5811–5826. [CrossRef]
22. Wei, J.; Huang, D.; Lu, Y.; Zhou, D.; Le, Q.V. Simple synthetic data reduces sycophancy in large language models. *arXiv* **2023**, arXiv:2308.03958.
23. Sulaiman, S.; Mohamed, H.; Arshad, M.R.M.; Rashid, N.A.A.; Yusof, U.K. Hajj-QAES: A Knowledge-Based Expert System to Support Hajj Pilgrims in Decision Making. In Proceedings of the 2009 International Conference on Computer Technology and Development, Kota Kinabalu, Malaysia, 13–15 November 2009; Volume 1, pp. 442–446. [CrossRef]
24. Sharef, N.M.; Murad, M.A.; Mustapha, A.; Shishechi, S. Semantic question answering of umrah pilgrims to enable self-guided education. In Proceedings of the 2013 13th International Conference on Intellient Systems Design and Applications, Salangor, Malaysia, 8–10 December 2013; pp. 141–146. [CrossRef]
25. Mohamed, H.H.; Arshad, M.R.H.M.; Azmi, M.D. M-HAJJ DSS: A mobile decision support system for Hajj pilgrims. In Proceedings of the 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 15–17 August 2016; pp. 132–136. [CrossRef]
26. Nusuk: Your Official Guide to Makkah and Madinah. Available online: https://www.nusuk.sa/ (accessed on 1 November 2023).

27. Mecca WABot: Smart System Makes Hajj and Umrah Pilgrims Easy to Worship. Available online: https://kumparan.com/beritaanaksurabaya/mecca-wabot-sistem-pintar-mudahkan-jemaah-haji-dan-umrah-beribadah-20f6faH8EMZ/2 (accessed on 1 November 2023).
28. Ministry of Health in the Kingdom of Saudi Arabia. Available online: https://www.moh.gov.sa/en/ (accessed on 1 November 2023).
29. WHO Chronic Respiratory Diseases. Available online: https://www.who.int/health-topics/chronic-respiratory-diseases#tab=tab_3 (accessed on 1 November 2023).
30. Ministry of Hajj and Umrah in the Kingdom of Saudi Arabia. Available online:https://www.haj.gov.sa/Home (accessed on 1 November 2023).
31. CGD Society—FAQ Lung Issues. Available online: https://cgdsociety.org/living-with-cgd/managing-cgd/common-problems/lung-problems/faqs-lung-issues/ (accessed on 1 November 2023).
32. Top Doctors—Frequently Asked Questions about Lung Diseases. Available online: https://www.topdoctors.co.uk/medical-articles/frequently-asked-questions-about-lung-diseases# (accessed on 1 November 2023).
33. Hajj and Umrah Health Requirements. Available online: https://www.saudiembassy.net/hajj-and-umrah-health-requirements. (accessed on 1 November 2023).
34. Health Requirements for Hajj. Available online: https://www.moh.gov.sa/en/HealthAwareness/Pilgrims_Health/Pages/default.aspx (accessed on 1 November 2023).
35. Sarrouti, M.; Abacha, A.B.; M'rabet, Y.; Demner-Fushman, D. Evidence-based fact-checking of health-related claims. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 3499–3512.
36. Phatak, A.; Mago, V.K.; Agrawal, A.; Inbasekaran, A.; Giabbanelli, P.J. Narrating Causal Graphs with Large Language Models. *arXiv* **2024**, arXiv:2403.07118.
37. Gao, M.; Hu, X.; Ruan, J.; Pu, X.; Wan, X. LLM-based NLG Evaluation: Current Status and Challenges. *arXiv* **2024**, arXiv:2402.01383.
38. Saadany, H.; Orăsan, C. BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-Oriented Text. In Proceedings of the Translation and Interpreting Technology Online Conference, Online, 5–7 July 2021; pp. 48–56.
39. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
40. Akter, S.N.; Yu, Z.; Muhamed, A.; Ou, T.; Bäuerle, A.; Cabrera, Á.A.; Dholakia, K.; Xiong, C.; Neubig, G. An In-depth Look at Gemini's Language Abilities. *arXiv* **2023**, arXiv:2312.11444.

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.