




Please cite the Published Version

Mohammad, Rasheed , Alkhnbashi, Omer S  and Hammoudeh, Mohammad  (2024) Optimizing Large Language Models for Arabic Healthcare Communication: A Focus on Patient-Centered NLP Applications. *Big Data and Cognitive Computing*, 8 (11). 157 ISSN 2504-2289

DOI: <https://doi.org/10.3390/bdcc8110157>

Publisher: MDPI AG

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/637242/>

Usage rights:  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

Additional Information: This is an open access article which first appeared in *Big Data and Cognitive Computing*, published by MDPI

Data Access Statement: These data were derived from the following resources available in the public domain <https://shorturl.at/dR46Y> (accessed on 12 November 2024).

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Article

Optimizing Large Language Models for Arabic Healthcare Communication: A Focus on Patient-Centered NLP Applications

Rasheed Mohammad ^{1,*} , Omer S. Alkhnabshi ^{2,3,4,*}  and Mohammad Hammoudeh ^{2,5} ¹ College of Computing, Birmingham City University, Birmingham B5 5JU, UK² Information and Computer Science Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia; mohammad.hammoudeh@kfupm.edu.sa³ Center for Applied and Translational Genomics (CATG), Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai Healthcare City, Dubai P.O. Box 505055, United Arab Emirates⁴ College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai Healthcare City, Dubai P.O. Box 505055, United Arab Emirates⁵ School of Computing and Mathematics, Manchester Metropolitan University, Chester Street, Manchester M15GD, UK

* Correspondence: rasheed.mohammad@bcu.ac.uk (R.M.); omer.alkhnabshi@dubaihealth.ae (O.S.A.)

Abstract: Recent studies have highlighted the growing integration of Natural Language Processing (NLP) techniques and Large Language Models (LLMs) in healthcare. These technologies have shown promising outcomes across various healthcare tasks, especially in widely studied languages like English and Chinese. While NLP methods have been extensively researched, LLM applications in healthcare represent a developing area with significant potential. However, the successful implementation of LLMs in healthcare requires careful review and guidance from human experts to ensure accuracy and reliability. Despite their emerging value, research on NLP and LLM applications for Arabic remains limited particularly when compared to other languages. This gap is largely due to challenges like the lack of suitable training datasets, the diversity of Arabic dialects, and the language's structural complexity. In this study, a panel of medical experts evaluated responses generated by LLMs, including ChatGPT, for Arabic healthcare inquiries, rating their accuracy between 85% and 90%. After fine tuning ChatGPT with data from the Altibbi platform, accuracy improved to a range of 87% to 92%. This study demonstrates the potential of LLMs in addressing Arabic healthcare queries especially in interpreting questions across dialects. It highlights the value of LLMs in enhancing healthcare communication within the Arabic-speaking world and points to a promising area for further research. This work establishes a foundation for optimizing NLP and LLM technologies to achieve greater linguistic and cultural adaptability in global healthcare settings.

Keywords: Large Language Model; Natural Language Processing; artificial intelligence in Arabic; patient medical query



Citation: Mohammad, R.; Alkhnabshi, O.S.; Hammoudeh, M. Optimizing Large Language Models for Arabic Healthcare Communication: A Focus on Patient-Centered NLP Applications. *Big Data Cogn. Comput.* **2024**, *8*, 157. <https://doi.org/10.3390/bdcc8110157>

Academic Editors: Min Chen, Moulay A. Akhloufi and Carson K. Leung

Received: 20 August 2024

Revised: 9 November 2024

Accepted: 13 November 2024

Published: 14 November 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

Healthcare remains a primary concern for nations worldwide [1,2]. The increase in health-related inquiries highlights the need for automated systems to manage these demands especially given the shortage of medically qualified personnel and the importance of maintaining expert-level precision [2,3]. These inquiries can encompass a range of topics, including patient medical history, potential drug allergies, disease-related concerns, and treatment protocols [1,4,5]. The essential role of medical staff in responding to patient queries underscores the need for highly efficient automated systems capable of matching human performance levels [6–8]. At the same time, the development of Medical Dialogue Systems is expected to improve patient access to healthcare services, enhance overall care quality, and reduce costs [4,5,9,10].

Conversational agents (CAs) currently show high accuracy in applications such as symptom checking, medical triage, health monitoring, and diagnosis [4,11]. Achieving this level of precision, however, requires advanced techniques to understand human language, identify intentions, and accurately automate data entry tasks [2]. A significant barrier to the widespread adoption of these technologies is the lack of accessible labeled datasets in many languages [2,12,13]. This limitation is especially relevant for the Arabic language in medical contexts, where usable labeled datasets remain scarce [2,14,15]. Intent classification, a core component of Natural Language Understanding, faces particular challenges when training data are limited, resulting in reduced generalization ability, especially for less frequently encountered terms [16].

The study by Marie-Sainte et al. [17] established foundational research for applying machine learning and deep learning to Arabic Natural Language Processing (NLP). Their work revealed several key insights, particularly regarding the unique challenges posed by the Arabic language for developers and researchers. Key characteristics, such as extensive inflection and the lack of standardized rules for capitalization and punctuation, add to these challenges. In response, researchers in Arabic Natural Language Processing (ANLP) have developed fundamental tools, including sentence splitters, tokenizers, and lightweight stemmers, which are essential for preparing text data for higher-level processing. Marie-Sainte et al. [17] concluded that due to the presence of heterogeneity and interoperability issues, researchers often need to adapt or reconstruct these tools to meet the specific requirements of their projects. This highlights the urgent need for standardized tools with consistent, interoperable components that can be seamlessly integrated within a unified framework. Al-Ayyoub et al. [18] reached a similar conclusion, emphasizing the need for standardization in NLP applications within social networks, which reinforces the call for tools that improve efficiency and consistency across different fields.

A clear example of the need for NLP and LLM techniques in Arabic is the Altibbi platform, which is the largest Arabic-language medical consultation website. Altibbi provides free consultations for Arabic-speaking users, which presents significant linguistic challenges, including handling multiple dialects and the complex morphology of Arabic. To address these challenges, researchers applied LSTM and Bi-LSTM models to classify over 1.5 million patient questions into 15 distinct medical categories. This approach achieved a classification accuracy of up to 87%, highlighting the effectiveness of NLP models in overcoming linguistic barriers in Arabic healthcare communication [2]. Additionally, when combined with fine-tuned LLMs such as ChatGPT, accuracy and contextual understanding can be further enhanced, especially in responding to questions posed in various Arabic dialects.

This research examines the current landscape of Arabic-based Natural Language Processing (NLP) within the medical field. The primary aim is to assess the potential and challenges of using Large Language Models (LLMs) in Arabic to automate healthcare services. The study provides an in-depth analysis of challenges highlighted in existing research on Arabic NLP and LLMs. The article is structured as follows: a literature review introduces the current state of NLP in the medical domain with a focus on patient-intent discovery. This review also highlights methodologies and their effectiveness in interpreting patient needs and intentions. The second part of the literature review discusses existing applications of LLMs in healthcare, outlining successes and challenges encountered in integrating these advanced models into medical processes. The third section explores the use of LLMs specifically in Arabic medical domains, examining the unique challenges and opportunities for healthcare automation.

The study also includes an experimental application of LLMs to answer medical queries in Arabic. The methodology, results, and implications of this experiment demonstrate the practicality and effectiveness of LLMs in this context. This comprehensive analysis provides insights into the current state, challenges, and future potential of Arabic-based NLP research in healthcare, setting the stage for advancements in automated healthcare services tailored to the Arabic-speaking population.

This study presents a novel application of fine-tuning Large Language Models (LLMs) for Arabic healthcare communication with a specific focus on patient-centered Natural Language Processing (NLP) tasks. Unlike previous studies that primarily survey or review Arabic NLP applications, this work provides a practical evaluation of a fine-tuned model's capability to handle diverse Arabic dialects and complex medical queries in real-world scenarios. The contribution of this study lies in both the technical implementation of the fine-tuning process and the empirical evaluation of the model's performance, as validated by medical experts.

2. Literature Review

Natural Language Understanding (NLU) is a critical component that determines the effectiveness of goal-oriented spoken dialogue systems [16,19]. Traditionally, NLU consists of intent classification and slot filling, which is aimed at constructing a semantic representation of users' utterances [2,16]. Intent classification predicts the user's intent in a query, while slot filling extracts relevant semantic concepts. A major challenge in NLU is the lack of human-labeled data, which limits the generalization capacity of proposed models [16]. To address this limitation, various methods have been introduced to train general-purpose language representation models using large amounts of unannotated text. Techniques such as ELMo [20] and Generative Pre-trained Transformer (GPT) [21] have demonstrated high accuracy especially when combined with task-specific annotated data. Additionally, Bidirectional Encoder Representations from Transformers (BERT) [22] has become a widely used model, contributing to the success of multiple NLP tasks, including question answering (SQuAD v1.1) and natural language inference [16].

2.1. AI in the Medical Domain

In recent years, significant progress has been made in applying AI across various medical domains, including medical imaging, clinical diagnosis, drug discovery, computer-aided decision support systems, and online health services [23]. Kwak and Hui [24] provided a comprehensive overview of deep learning applications in health informatics, highlighting the primary challenges in digitizing and transitioning traditional healthcare into data-driven services. These challenges include managing high-dimensional data, addressing data sparsity and heterogeneity, handling non-stationary and temporal data, and ensuring model scalability and security [2]. Mulani et al. [25] investigated the use of deep reinforcement learning to provide personalized health recommendations, including guidance on healthcare providers, nutrition, medications, and exercise. They emphasized the importance of incorporating advanced deep learning algorithms and accounting for patients' medical histories. Kumar et al. [26] applied Convolutional Neural Networks (CNNs) to detect and classify malaria, achieving a high accuracy rate of 96.8%.

Akselrod-Ballin et al. [27] introduced a deep learning model for breast cancer prediction that combines electronic health records (EHRs) with mammograms. This approach demonstrated satisfactory accuracy and was recommended as a supplementary tool for radiologists with further improvements suggested through additional data. Shah et al. [28] proposed another deep learning approach to assess the quality of a healthcare platform using a fusion of visual and textual patient feedback. Their results indicated that a fusion model significantly improved classification accuracy compared to relying on textual data alone.

Vidhya and Shanmugalakshmi [29] applied deep belief networks to predict type-2 diabetes complications, achieving an accuracy of 81%. Lauritsen et al. [30] developed a deep learning model for sepsis prediction using electronic health records from Danish hospitals over seven years. However, they acknowledged challenges with interpretability, bias, and reproducibility. Faes et al. [31] introduced an automated deep learning approach for classifying and diagnosing medical images using Google Cloud AutoML, achieving promising accuracy and discussing both the advantages and limitations of this tool. Estrada et al. [32] developed FatSegNet, an automated model for identifying adipose tissue in

abdominal MRI images. This tool improved detection accuracy by 7% over previous algorithms and was further adapted to handle 3D MRI images in near real-time. Edara et al. [33] proposed a deep learning approach using Long Short-Term Memory (LSTM) networks to analyze and predict the mood of cancer patients through the sentiment analysis of tweets, revealing a tendency toward positive outlooks. Liu et al. [34] outlined criteria for managing narrative electronic health records (EHRs) to advance clinical Natural Language Processing and decision support systems, discussing key issues in extracting valuable information from digitized healthcare data, although clinical applications are still under exploration.

The integration of deep learning, machine learning, and transfer learning methods has significantly transformed traditional healthcare frameworks. Zhang et al. [35] conducted a comprehensive survey examining the benefits and challenges of using advanced learning approaches for medical prognosis within health management systems. They emphasized that the effective implementation of deep learning algorithms depends on their specific applications and the data characteristics they are designed to process.

2.2. Conversational Agents in the Medical Domain

Conversational agents have been developed to augment medical personnel, handling specific tasks and providing primary clinical advice to patients, serving as a complementary tool for healthcare professionals [4]. However, a significant obstacle for medical conversational agents is the limited availability of medical dialogue corpora, which restricts their ability to engage effectively with users [36]. Medical datasets in languages other than English remain sparse, as demonstrated by Zeng et al. [9], Liu et al. [36], and Young-Min et al. [37]. Nonetheless, Gu et al. [38] suggest that the creation of biomedical datasets is feasible, opening avenues to overcome the limitations posed by the scarcity of non-English data and providing opportunities to challenge and refine pretrained models like BERT within the medical conversational agent domain. Zhou et al. [39] proposed several task frameworks for healthcare, while Bao et al. [40] developed a chatbot to identify patient intents, and Bai et al. [41] proposed incremental intent detection for medical contexts. Intent-context relationships for healthcare applications were also explored by Razzaq et al. [42], Amato et al. [43], and Zhang et al. [44].

Mondal et al. [45] reported gaps in the performance of state-of-the-art commercial products for Natural Language Understanding (NLU) tasks in low-resource languages in Asia and Africa. Similarly, Bai et al. [41] noted that pre-trained intent categories in medical intent detection do not fully address emerging intents, which evolve frequently in the medical field. Meanwhile, Hijjawi and Elsheikh [46] identified a lack of Arabic-based conversational agents, which is a finding echoed in recent research by Wael et al. Among the few studies applying NLP to Arabic in the medical domain, Habib et al. [5] developed a medical recommendation system for five specializations, and Mounsef et al. [14] created a dataset to assist users in finding answers to medical questions through a healthcare assistant [15].

Several efforts have focused on developing intent recognition systems in the medical domain across languages. For instance, Liu et al. [36] created a Chinese-language dataset for gastrointestinal medical dialogues, while Zeng et al. [9] proposed MedDialog, which is an English and Chinese dataset for medical dialogue generation. Zhang et al. [47] developed a Chinese medical intent evaluation dataset, and Young-Min et al. [37] created a Korean health intent dataset.

Building new medical datasets is a complex process that involves named entity recognition, information extraction, clinical diagnosis normalization, and establishing an online platform for model evaluation, comparison, and analysis [47]. Despite the challenges, creating new medical datasets is essential for developing conversational agents that approach human-level accuracy [47]. General evaluation datasets and benchmarks play a pivotal role in NLP [38,47,48]. The Arabic Dialect Dataset (ADD), developed by Mounsef et al. [14], supports users in finding answers to medical questions but is limited in scope. Many studies on medical NLP for Arabic reference the Altibbi website, which is one of the largest

Arabic-language medical consultation platforms offering free advice to Arab users. Faris et al. [2] used data from Altibbi to classify patient questions into 15 categories, employing LSTM and BiLSTM models, and reported an accuracy rate of 87%.

2.3. Arabic and Medical Domain

Arabic, a Semitic language, presents unique linguistic and morphological challenges [4,49]. Researchers such as Abdelhay et al. [4] identified issues including dialectal variations, text direction, and script complexities that impact patient intent discovery. Studies applying NLP to Arabic, particularly in Natural Language Understanding (NLU), are limited, and further investigation is recommended [4,50].

Few studies explore the use of Arabic in the medical domain. For example, Mezzi et al. [51] developed an intent detection model for mental health patients using BERT and the International Neuropsychiatric Interview (MINI), achieving an accuracy of approximately 90%. Abdelhay et al. [4] created the Medical Question and Answer database (MAQA), containing 430,000 entries across 20 medical specializations. Their use of deep learning models, including LSTM, Bi-LSTM, and Transformers, showed that the Transformer model outperformed others with an average cosine similarity of 80.81% and a BLEU score of 58%. Habib et al. [5] proposed a recommendation system for medications and treatments based on patient input from the Altibbi database, although model accuracies were below 80%, indicating that using limited input words for predictive recommendations is insufficient. Another study by Wael et al. involved a small dataset with Saudi and Egyptian dialects collected from the children's section of Elconsolto.com, focusing on patient intent. However, this dataset was limited, with only a few intent classes, including drugs, food products, symptoms, and vaccines.

There are numerous gaps in current research on Arabic NLP, especially in the medical domain. These gaps, summarized in Table 1, include the following: Researchers such as Faris et al. [2], Abdelhay et al. [4], Habib et al. [10], Mounsef et al. [14], Alhassan et al. [52], and Boudjellal et al. [53] have highlighted the lack of Arabic-based, task-oriented datasets, particularly for the medical domain.

Table 1. Studies focused on developing Arabic-based NLP approaches for the medical domain.

Author(s)	NLP Solution	Domain	Dataset	Approaches
[54]	Chatbot	General medical Questions	2150 pairs	Three sequence-to-sequence models (LSTM, BiLSTM, GRU)
[55]	Empathetic bot	Human-like conversational model	38K	AraBERT
[5]	Medical recommendations	Telemedicine service	Altibbi databases	(LSTM and BiLSTM)
[14]	Understanding Arabic dialects for better diagnose	Healthcare	Develop Arabic Dialect Dataset (ADD)	(LSTM and BiLSTM)
[15]	Healthcare assistant chatbot	Healthcare	3664 pair of questions and answers from elconsolto.com	BERT, Logistic regression TF-IDF and Doc2vec
[53]	Corpora	Healthcare	49,856 sentences tagged with 13 entity types	
[2]	Medical questions classification	Healthcare	Altibbi	LSTM and BiLSTM
[10]	Specialty-based question classification	Healthcare	1.5 million medical consultations from Altibbi	Word embedding model

A key challenge in Arabic-related NLU studies is the difficulty of providing a general NLU solution, which often results in reduced accuracy [50]. General intent recognition models frequently rely on labeled datasets and yield broad intent categories when they

encounter untrained intents. This supports the need for specialized NLU solutions tailored to specific domains.

Many studies have used LSTM, Bi-LSTM, and embedding models like Word2Vec, which require substantial input to achieve high accuracy, particularly for question–answer contexts. Habib et al. [10] developed AltibbiVec, which is an Arabic language embedding model. However, there is a lack of studies leveraging Large Language Models (LLMs) within Arabic healthcare/medical tasks, which could improve performance.

The literature review found only a few studies, e.g., [2,10], that classify patient questions into categories, which is a classification that could help direct inquiries to the appropriate specialists.

Intent recognition studies for Arabic are generally limited with even fewer focused on the medical domain. There is a need for intent recognition systems that can effectively recognize medical terms in Arabic, echoing global trends in countries like China, India, and Korea, where local intent recognition systems have proven valuable.

2.4. LLMs and Medical Domain

The review of Arabic NLP studies in Section 2.3 revealed a notable gap in the use of Large Language Models (LLMs) for NLP tasks within the medical domain. Although LLMs are widely recognized for their effectiveness in producing human-like and contextually appropriate responses, their precision in specific medical domains requires further evaluation [56,57].

LLMs, such as GPT-3, BERT, and their multilingual variants, offer significant advantages over traditional text embeddings like Word2Vec, GloVe, and FastText, especially in various NLP tasks [58,59]. A distinctive feature of LLMs is their ability to capture contextual information by considering surrounding words when representing a word or phrase [60]. This contextual awareness helps resolve word ambiguities and accurately capture meaning across different contexts, an area where traditional embeddings, which provide static, context-independent representations, often fall short [59,61].

A key advantage of LLMs is their multilingual capability, allowing them to handle text in multiple languages without requiring language-specific embeddings [62,63]. Unlike traditional embeddings, which are constrained by language dependencies, LLMs are pre-trained on extensive text corpora, acquiring broad world knowledge, including contemporary information [61]. This attribute is especially useful for NLP tasks that require general knowledge or an understanding of current language usage [63]. Moreover, LLMs exhibit strong adaptability through few-shot learning, enabling them to adapt to new tasks or domains with minimal training examples [64], whereas traditional embeddings typically require substantial retraining for task-specific adaptations.

Some LLMs, such as BERT, enable researchers to delve into data intricacies through masked language modeling, uncovering deeper linguistic patterns [58,65]. This capability is valuable for conducting in-depth linguistic analyses. Additionally, LLMs offer a robust fine-tuning approach, where the model can be tailored for specific NLP tasks by adjusting the top layers with task-specific data, often achieving state-of-the-art performance [61]. In contrast, traditional embeddings require more custom engineering and additional layers to perform well on specialized tasks [64]. The versatility of LLMs is further evident as they excel across a broad range of NLP tasks, including sentiment analysis, question answering, text generation, translation, and medical inquiries [64]. This contrasts with traditional embeddings, which may necessitate task-specific feature engineering. Furthermore, LLMs can be fine-tuned for domain-specific tasks, making them adaptable across diverse fields, from medical and legal to financial applications [64].

However, it is important to recognize the relevance of traditional embeddings, such as GloVe, in NLP, particularly in scenarios with limited computational resources or when simpler tasks do not require the full capabilities of LLMs. In practice, the choice between LLMs and traditional embeddings depends on the specifics of the NLP task, data availability, and the balance between model complexity and computational resources. Researchers

and practitioners often test both approaches to determine the most suitable one for their specific problem. This pragmatic approach allows for the consideration of each method's advantages and limitations in light of the task's unique requirements.

The global accessibility of LLMs can sometimes pose challenges for researchers applying them in medical contexts [59]. For example, an evaluation of ChatGPT-3.5, ChatGPT-4.0, and Google Bard using 37 common medical queries showed that ChatGPT-4.0 performed best [59], although this assessment was limited to ocular symptom-related queries. In a similar study, Sengupta et al. evaluated LLMs on gynecological inquiries with ChatGPT-4.0 outperforming its predecessor. Lee et al. [66] reported inconsistencies in expert satisfaction with ChatGPT's responses, though the study involved only eight experts. Reflecting these findings, Reddy [64] advises caution when using LLMs in healthcare particularly with regard to accuracy. Reddy [64] also advocates for implementing a structured evaluation process to assess the performance and practical value of LLMs in medical applications, highlighting the importance of a systematic approach when integrating these powerful models into healthcare settings.

2.5. LLMs for Arabic NLP Tasks in the Medical Domain

The role of Large Language Models (LLMs) in medical research and related tasks is essential, as emphasized by recent studies [61,67]. The application of LLMs for Arabic-based NLP tasks in the medical field shows considerable potential. However, concerns about the accuracy of these models' outputs remain, as noted in recent literature [58,59,64]. Addressing these concerns requires training LLMs with accurate and reliable medical data [64]. Models like GPT-3, BERT, and specialized variants offer several advantages for healthcare and medical research, as summarized in Table 2.

Table 2. Some of studies that investigate employing LLMs within medical domain.

Authors	LLMs	Domain	Result	Limitation
[61]	ChatGPT 3, 3.5, 4.0	Gynecology enquires	ChatGPT-4.0 outperformed ChatGPT-3.5, though ChatGPT-3.5 was excellent in many aspects	Inconsistencies were observed among LLMs, and the assessment dataset was limited in size
[59]	ChatGPT 3.5, 4.0, and Google Bard	Ocular symptom related queries	ChatGPT-4.0 outperformed others	It is limited to the responses of LLMs for only 37 inquiries
[66]	ChatGPT	Eight common questions about colonoscopy	The responses were more organized compared to those provided on professional healthcare websites	It is limited to the satisfaction of a group of experts with the responses provided by ChatGPT
[68]	Bing Chatbot powered by ChatGPT	Three common radiologic examinations and procedures	93% of answers were classified as entirely correct	Limited to radiologic examination
[67]	No specific LLMs	Ophthalmology	Discussions and an extensive literature review recommend the use of LLMs with certain limitations. Exercising caution when introducing these models into clinical practice is crucial, as issues related to safety, efficacy, and ethics remain subjects of debate and ongoing investigation	A practical assessment was not performed

Table 2. Cont.

Authors	LLMs	Domain	Result	Limitation
[69]	OpenAI's Generative Pre-trained Transformer-4 (GPT-4)	For clinicians	They recommended that healthcare professionals seize the opportunity to explore, participate in, and lead the responsible adoption of LLMs, leveraging their potential to enhance patient care and drive progress in the continuously evolving healthcare landscape	No practical assessment was conducted
[70]	Bing Chat and ChatGPT 3.5, 4.0	Ophthalmology: 250 questions from the Basic Science and Clinical Science Self-Assessment Program	Human experts rated the accuracy as follows: ChatGPT-3.5 achieved 58.8%, ChatGPT-4.0 achieved 71.6%, and Bing Chat achieved 71.2%	The authors did not clarify that Bing Chat relies on ChatGPT to generate answers
[58]	ChatGPT	113 questions related to esophagogastroduodenoscopy (EGD), colonoscopy, endoscopic ultrasound (EUS), and endoscopic retrograde cholangiopancreatography (ERCP)	The accuracy for providing comprehensive answers was highest for questions related to EGD, at 52.6%, which is considered low	No information was provided regarding any settings to adjust ChatGPT's output to fit a medical context
[63]	ChatGPT	Literature review in the context of oral and maxillofacial surgery (OMS)	There is a need to scientifically explore the use of LLMs in the core areas of oral and maxillofacial surgery (OMS)	The paper highlights that the current research on LLMs primarily covers secondary tasks such as research assistance and patient information with limited evidence on their core applications directly related to OMS, like surgical procedures and critical medical decision making
[71]	No specific LLMs	Anatomic pathology	LLMs can be used for specific medical tasks, but emphasis should be placed on validation	Absence of empirical results
[56]	ChatGPT-3.5, ChatGPT-4.0, and Google Bard	31 commonly asked inquiries related to myopia care	Accuracy: ChatGPT-4.0 achieved 80.6%, ChatGPT-3.5 achieved 61.3%, and Google Bard achieved 54.8%	It is limited to the responses of LLMs for only 31 inquiries
[72]	ChatGPT-3.5, ChatGPT-4.0, and Google Bard	47 medical questions	ChatGPT-4.0 achieved an accuracy of 91%, as rated by experts	The conclusion was overly general despite the assessment focusing on the performance of LLMs with 47 questions

Enhanced Natural Language Understanding (NLU): LLMs' capacity for understanding and generating human-like text can be harnessed for medical tasks such as information extraction, text summarization, and question answering.

Multilingual Support: Many LLMs support multiple languages, including Arabic. This feature is valuable for reducing language barriers in medical data and research.

Semantic and Contextual Understanding: LLMs excel in disambiguating terms, understanding complex medical terminology, and generating coherent medical reports.

Text Classification: LLMs can assist in categorizing medical texts, such as grouping diseases, symptoms, or medical literature into relevant domains.

Automation of Routine Tasks: LLMs can automate tasks like extracting information from electronic health records, summarizing medical literature, and generating clinical reports.

Synthetic Data Generation: LLMs can create synthetic data for training and validating other machine learning models, which is especially valuable when medical datasets are limited.

Continual Learning: By fine-tuning with domain-specific medical data, LLMs can adapt to the nuances of the medical field.

Beyond typical challenges such as security, privacy, and integration, additional considerations arise, particularly the lack of domain-specific knowledge. While LLMs exhibit broad knowledge, their depth of expertise may be limited in specific medical subfields, necessitating careful application in specialized medical decision-making tasks. Another crucial factor is the need for interpretable AI, especially in healthcare, where medical professionals require transparency in the model's decisions, as these can directly impact patient care.

The outcomes generated by LLMs can present interpretation challenges. Additionally, the use of LLMs, like other AI applications in healthcare, raises ethical concerns, particularly regarding data bias and decision making. Ensuring fairness and transparency in the deployment of LLMs is therefore essential. Furthermore, any AI model used in healthcare, including LLMs, must undergo rigorous validation to ensure safety and efficacy. These considerations highlight the importance of a comprehensive approach when integrating LLMs into the healthcare field.

Numerous applications of LLMs with Arabic texts have emerged. For example, Jackson [73] introduced Jais, which is an LLM designed for Arabic–English text alignment. However, Jais is a general-purpose model and not specifically tailored to the medical domain; its primary focus is on aligning Arabic and English words. This approach may inadvertently overlook many Arabic terms, particularly dialectal words that lack direct English equivalents. The University of California, San Francisco (UCSF) team developed an LLM capable of generating personalized messages to encourage patient participation in clinical trials [74], although its primary purpose remains research focused. Similarly, the King Abdullah University of Science and Technology (KAUST) created an LLM proficient in identifying adverse drug reactions (ADRs) from Arabic social media posts [75]. Additionally, the Johns Hopkins University team developed an LLM capable of generating summaries of Arabic medical records, achieving quality comparable to human-written summaries [76]. However, these cases only partially address NLP tasks associated with patient interactions with healthcare professionals, such as consultations, question-and-answer sessions, and intent detection. While these developments demonstrate the diverse applications of LLMs in Arabic, there remains significant potential for further exploration and specialization, especially within the complex landscape of medical NLP tasks involving patient–provider interactions.

3. Methodology

3.1. Method

Figure 1 illustrates the step-by-step methodology used in our study. The process began with the web scraping of medical questions, answers, and the specialties of the medical team members who responded to the questions on altibbi.com. The collected data then underwent a rigorous pre-processing phase to ensure suitability for further analysis. A major effort was made to add Arabic translations for non-Arabic words (e.g., English, French) found in the queries. After discussion, it was decided to retain these non-Arabic words, as a considerable number of queries include them—either in their original language or transliterated using Arabic script to approximate pronunciation. Keeping these terms reflects the reality of medical queries on Arabic medical websites, which often feature non-Arabic terms.

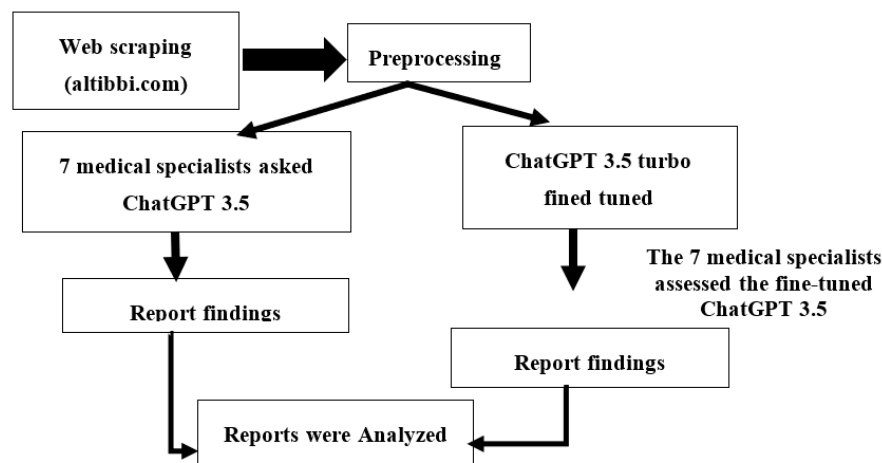


Figure 1. Research steps.

Following this, the methodology split into two parallel actions. The first involved eight medical specialists, each within their field, using ChatGPT-3.5 to answer relevant medical questions and documenting their findings. Concurrently, the second action used the pre-processed data to fine-tune ChatGPT-3.5 Turbo, tailoring its responses more closely to the medical domain. Training begins with a high initial training loss of 1.3859, indicating the model's early stage, where its predictions significantly diverge from the target outputs. As training progresses, the training loss decreases, reflecting the model's improving ability to generalize from the training data. This decrease is marked by fluctuations in loss values, which indicate the dynamic learning process and the challenges presented by certain epochs or data peculiarities. Notably, a significant reduction in training loss is observed at step 1201, reaching 0.6334, suggesting substantial learning or optimization. However, subsequent increases in loss illustrate the non-linear optimization path, where improvements are interspersed with setbacks. The absence of validation loss data suggests a focus on monitoring training loss to adjust model parameters, potentially aiming for rapid iteration with plans to address overfitting or validate performance later.

Upon completing these parallel actions, the methodology proceeds to post-tuning evaluation. The same eight medical specialists reassess ChatGPT-3.5, now fine-tuned with domain-specific data, using the same set of questions to measure improvements in response accuracy and relevance. The final stage is a comprehensive report consolidating findings from the initial and post-tuning evaluations, providing a detailed comparison to demonstrate the effectiveness of the fine-tuning process in enhancing ChatGPT-3.5 Turbo's performance in a medical context.

3.2. Dataset

This study aimed to systematically evaluate the technical proficiency and thematic relevance of ChatGPT-3.5, a freely accessible version, in responding to medical inquiries in Arabic, covering both Modern Standard Arabic and various dialects.

The dataset used in this study was sourced from the Altibbi platform [10], one of the largest online medical forums in the Middle East, offering free medical consultations to Arabic-speaking users. The selection process for questions and answers was conducted in multiple stages to ensure representation across different medical domains, dialects, and question complexities. Initially, questions were filtered for relevance to healthcare, focusing on 15 medical specialties available on Altibbi, including cardiology, neurology, and oncology, using keyword-based filtering. To capture the linguistic diversity of Arabic, the dataset was further expanded to include questions in various Arabic dialects, such as Egyptian, Saudi, Syrian, and Moroccan. Dialect classification was performed using an automatic dialect detection model. The dataset also prioritized a balanced representation of complex, multi-part medical inquiries alongside simpler, symptom-based questions. Finally, a panel

of medical experts reviewed the answers to ensure clinical accuracy and appropriateness for patient interaction. This curation process yielded approximately 100,000 question-and-answer pairs, providing a robust dataset for fine tuning LLMs in the medical domain.

A set of approximately 550 question-and-answer pairs was selected by seven medical experts. This diverse set covered various medical specialties with each expert choosing questions relevant to their area of expertise. Notably, the set included questions in distinct Arabic dialects, such as Egyptian, Syrian, Saudi, Sudanese, and Moroccan. These questions were then input into ChatGPT-3.5, and the generated responses were compiled for evaluation by the medical experts. The experts assessed the accuracy of ChatGPT-3.5's responses and identified any potential warning signs that could compromise patient safety. Additionally, the answers were carefully analyzed by the research team, focusing on NLP-related features individually, which was followed by a group discussion to synthesize a unified conclusion from the findings. This comprehensive approach ensured a thorough evaluation of ChatGPT-3.5's performance across a wide range of medical queries in diverse linguistic contexts.

Regarding the dataset content, the decision to retain non-Arabic words reflects the linguistic realities of medical queries in the Arabic-speaking world, where code switching between Arabic and languages like English or French is common, particularly in medical and technical contexts. Many patients and healthcare professionals use English medical terminology (e.g., "diabetes", "blood pressure") or French terms in North African regions, even when speaking primarily in Arabic. By including these non-Arabic words, we aimed to ensure that the model could effectively handle these real-world communication patterns.

3.3. Participants

The selection of medical experts for evaluating ChatGPT-3.5 was conducted through non-random convenience sampling. Given the high level of expertise required and the limited availability of medical professionals, convenience sampling was used to ensure timely participation from experts across various medical specialties. While this method enabled the practical execution of the evaluation, we acknowledge the potential for bias, as the selected experts may not fully represent the broader population of healthcare professionals. However, considering the time-sensitive nature of the study and the need for domain-specific expertise in evaluating medical-related queries, convenience sampling allowed us to obtain high-quality feedback from experienced professionals within the study's timeline. This approach ensured that the experts possessed the necessary background to provide informed evaluations of the model's responses to medical queries.

In conducting this study, a targeted approach was initially employed to invite fifty medical experts known for their significant contributions and expertise in the field. Selection criteria were designed to ensure comprehensive representation across relevant domains, aiming to enrich the study's findings with diverse, high-quality insights. However, practical challenges in engaging busy professionals led to a lower-than-expected response rate. Although several experts expressed initial interest, many were ultimately unable to participate due to time constraints and other commitments—a challenge commonly encountered in qualitative research, where the specialized knowledge and demanding schedules of participants often limit availability [77,78].

Of the experts contacted, seven responded positively within the timeframe required for meaningful inclusion in the study. These participants were selected through convenience sampling, which, while non-random, was the most pragmatic approach given the circumstances. This method allowed for the efficient inclusion of available and willing participants, enabling the study to proceed without further delay. Convenience sampling is widely supported in the literature as a practical strategy when faced with recruitment challenges, permitting researchers to make effective use of available resources while still capturing valuable expert insights [79].

The participation of these seven medical experts provided a valuable cross-section of knowledge and experience within the specified domain. Each participant contributed

unique insights that significantly enriched the research findings, adding depth to the study's contribution. Despite the smaller-than-intended sample size, the quality of data collected from these experts was not compromised. Their expertise and input were instrumental in achieving the research objectives, offering perspectives that broader, less-focused methodologies might not capture [80].

While acknowledging the limitations of the sample size, this study also emphasizes the rich, qualitative nature of the data obtained. By utilizing the in-depth expertise of its participants, this research contributes detailed insights into the complexities of the field. Future research could expand upon this work by incorporating a larger pool of experts to build on the foundational understanding established here [77].

The medical team comprised seven professionals: a Consultant in Neurosurgery, a Professor of Oncology and Internal Medicine, a Professor of Surgery, a Pediatric Cancer Consultant, a Consultant in Chest Diseases, a Urology Consultant, and a Public Health Expert. These titles reflect the experts' self-identifications. The team was given a one-month period, from 1 October to 30 October 2023, to evaluate the responses provided by ChatGPT-3.5.

3.4. Fine-Tuning

To enhance ChatGPT-3.5's performance in handling Arabic medical queries, a fine-tuning process was implemented. The steps for retraining the Large Language Model (LLM) are outlined as follows:

Dataset Preparation: The dataset was pre-processed by removing irrelevant data, tokenizing the Arabic text, and normalizing dialects to ensure compatibility across dialectal variations. Punctuation normalization and stop-word removal were also applied.

Model Architecture and Fine-Tuning Setup: We utilized OpenAI's GPT-3.5 architecture as the base model, fine-tuning it for healthcare-specific tasks in Arabic using supervised learning with pairs of questions and expert-verified answers. The fine-tuning process employed a supervised learning algorithm with gradient descent optimization, using cross-entropy as the loss function to effectively handle multi-class classification tasks in healthcare.

Training Parameters: The model was fine-tuned with a batch size of 32 to optimize memory usage for large medical datasets. A variable learning rate was applied, starting at 5×10^{-5} (0.00005) and gradually decreasing as training progressed. The model was trained over 10 epochs with checkpoints saved after each epoch to retain the best-performing model for deployment. The fine-tuning was conducted on an NVIDIA A100 GPU cluster, leveraging its computational power to manage the complexity of training a large model like GPT-3.5 on a substantial Arabic dataset.

After fine tuning, the model was evaluated using a separate validation set of 5000 question-answer pairs, which were distinct from the training dataset. This validation set included questions from multiple medical domains and a variety of Arabic dialects. The fine-tuned model's performance was assessed using metrics such as accuracy, precision, recall, and F1-score, among which accuracy showed the most significant improvement (from 86.55% to 90%).

During training, issues with dialect handling emerged, as the model occasionally confused different Arabic dialects. To address this, dialect-specific tokens were introduced to help the model distinguish between dialects in both input questions and expected answers. Additionally, data augmentation techniques were used to broaden the model's exposure to medical vocabulary by generating synthetic data through minor modifications of existing queries. This fine-tuning process not only improved the model's ability to handle specific medical queries in Arabic but also enhanced its understanding of regional dialects and medical terminology.

Focusing on training loss during the fine-tuning process allowed us to monitor the model's learning progress and optimize weight updates efficiently. Training loss provides immediate feedback on the model's fit to the current data, which is particularly useful in

large-scale models like ChatGPT-3.5. This feedback helps adjust learning rates, batch sizes, and other hyperparameters, enabling efficient learning in a complex domain like healthcare. However, focusing solely on training loss without concurrent validation introduces the risk of overfitting, especially in a domain as diverse and complex as medical queries. Without validation to assess generalization, the model might learn patterns specific to the training data, leading to suboptimal performance on real-world queries. In a medical context, overfitting could result in overly specific or inaccurate responses that fail to generalize across different patient cases, medical conditions, or regional dialects, posing risks to both model reliability and patient safety.

To mitigate overfitting risks without concurrent validation, we employed early stopping techniques, monitoring training loss for signs of diminishing returns to avoid overfitting. Additionally, validation runs were conducted at set intervals to assess generalization and adjust hyperparameters accordingly. In future iterations, more frequent validation checks could be incorporated to ensure the model generalizes effectively to unseen data and performs reliably in real-world medical applications.

4. Results

Code switching, or the practice of alternating between languages within a single conversation, presents both challenges and opportunities for NLP models. In the medical domain, this phenomenon is particularly prominent, as patients may use Arabic for general communication but switch to English or French for medical terms. Including non-Arabic words in the dataset allows the model to adapt to these linguistic behaviors, enabling it to interpret mixed-language queries accurately and provide contextually appropriate responses. For instance, a query like “عندي diabetes” (“I have diabetes”) combines Arabic with English terminology, requiring the model to understand this code-switched input to deliver a relevant medical response.

The initial high training loss of 1.3859 observed during the early stages of fine tuning indicated that the model’s predictions were far from the desired outputs, particularly for domain-specific medical queries. This suggested that the model’s baseline knowledge, primarily derived from general language data, was insufficient to accurately process specialized medical terminology in Arabic. In response, we adjusted the learning rate and applied a gradual decay to help the model better adapt to the complexities of medical language without overfitting to initial patterns.

Fluctuations in training loss throughout fine tuning reflected the challenges of optimizing the model for context-specific medical queries. These fluctuations often arose when the model encountered complex or rare medical terminology or processed queries in different Arabic dialects. For example, at certain points during training, the model struggled to generalize across similar medical terms with slight dialectal variations, leading to temporary increases in training loss. These fluctuations indicated the model’s difficulty in balancing overfitting to specific medical terms with generalizing across diverse queries.

To address these challenges, we implemented adaptive strategies during training. First, we introduced additional examples of complex medical queries, especially those involving less common medical conditions and regional dialects, to improve the model’s ability to generalize across a broader range of inputs. Additionally, we employed early stopping and regularization techniques to prevent overfitting when the model’s performance on the validation set plateaued, suggesting that further fine tuning on the training data might compromise its generalizability.

4.1. Statistical Analysis of Expert Evaluations

To ensure a rigorous evaluation of the model, we conducted a statistical analysis of assessments provided by the seven medical experts. The analysis included the following steps:

Descriptive Statistics: The evaluations from the medical experts were compiled, and the accuracy, precision, and recall of the model’s responses were calculated. Each expert

rated the accuracy of 50–100 responses in their field. The average accuracy ratings provided by the experts before and after fine tuning were measured (see Table 3).

Table 3. Statistical analysis of the model.

Statistical Procedure	Pre-Fine Tuning	Post-Fine Tuning	Statistical Analysis
Accuracy	86.55%	90%	$p < 0.01$ (paired t -test)
Cohen's Kappa (Inter-Rater)	0.78	0.85	Indicates substantial to almost perfect agreement
95% Confidence Interval	(85.5%, 87.6%)	(89%, 91%)	

Inter-Rater Reliability: To assess the consistency of evaluations among the medical experts, we calculated Cohen's Kappa coefficient for inter-rater reliability. This statistic measured the level of agreement beyond chance between each pair of experts in rating the model's accuracy. The Kappa value for pre-fine-tuning evaluations was 0.78, indicating substantial agreement, while the post-fine-tuning Kappa value improved to 0.85, reflecting increased consistency among experts (see Table 3).

Paired t -Test: A paired t -test was conducted to statistically confirm the improvement in accuracy following fine tuning. This test compared the experts' evaluations of the model before and after fine tuning. The p -value obtained from the t -test was less than 0.01, indicating that the improvement in the model's performance post-fine tuning was statistically significant (see Table 3).

Confidence Intervals: Confidence intervals were calculated for the accuracy ratings to quantify measurement uncertainty. The 95% confidence interval for pre-fine-tuning accuracy was (85.5%, 87.6%), while for post-fine-tuning accuracy, it was (89%, 91%) (see Table 3).

This statistical analysis demonstrated not only an improvement in model performance after fine tuning but also consistency and reliability in expert evaluations. The substantial agreement among experts and the statistically significant improvement highlights the robustness of the fine-tuning process.

4.2. Model Performance

The evaluation process began with the categorization of 550 medical queries across seven medical specializations. The initial responses generated by ChatGPT-3.5 were rated for accuracy by a panel of medical experts with each expert providing feedback on the relevance, correctness, and clarity of the responses. Following this, the model was fine-tuned using a subset of 100,000 question–answer pairs from the Altibbi platform. The fine-tuned model's performance was then reassessed by the same panel of experts to quantify improvements in accuracy and overall response quality.

Response accuracy was determined by comparing the model's answers to expert evaluations with each response graded on correctness, clinical relevance, and its ability to mimic human-like interaction. Improvements in accuracy were attributed to the model's enhanced understanding of medical terminology and dialects after fine tuning. Additionally, the response length and conciseness were assessed by comparing the average word count before and after fine tuning, revealing a 30% reduction in verbosity.

On average, before fine tuning, the seven experts rated the accuracy of ChatGPT-3.5 at 86.55%. The individual accuracy ratings provided by each medical expert are detailed in Table 4. Overall, the accuracy remained above 85%, which is considered high in many technical domains. After fine tuning, the overall accuracy rating given by the medical experts improved to 90%.

Table 4. Medical experts' evaluations of ChatGPT-3.5's responses to medical inquiries.

Experts' Specializations	No. of Answers Assessed	Accuracy Average (Before Fine Tuning)	Accuracy Average (After Fine Tuning)
Consultant Neurosurgery	50	90%	92%
Professor of Oncology and Internal Medicine	100	85%	87%
Professor in surgery	70	85%	88%
Pediatric cancer consultant	80	86.25%	91%
Consultant Chest Diseases	60	86.60	90%
Urology consultant	70	87%	90%
Public Health	120	86%	92.5%
Average	550	86.55%	90%

4.3. Medical Experts' Feedback—Before Tuning

However, the experts identified several concerns in ChatGPT-3.5's responses prior to fine tuning: (1) A small subset of answers was entirely irrational, which experts suggest may be due to the model's potential misinterpretation of questions posed in different dialects, although such instances were rare. (2) In a few cases, ChatGPT-3.5 specified medication dosages for patients, which is a significant concern given that these medications can have serious side effects and should not be prescribed without thorough medical supervision. (3) The medical experts commended ChatGPT-3.5 for often concluding its responses by advising users to consult their doctors, demonstrating an awareness of its limitations in providing medical advice. (4) Compared to responses from actual doctors on altibbi.com, ChatGPT-3.5's answers were more elaborate but less precise. (5) It was observed that the more detailed the question posed by the user, the more accurate ChatGPT-3.5's response tended to be. (6) The medical experts recognized ChatGPT-3.5 as a potential educational tool for knowledge dissemination and even for training medical students, though they emphasized that users should always seek professional medical consultation. (7) Unlike traditional search engines, ChatGPT-3.5 was noted for reducing the time taken to obtain information; however, the lack of cited sources in its responses raises questions about the reliability of the information provided.

4.4. Medical Experts' Feedback—ChatGPT Fine Tuning

On the other hand, the medical experts' evaluation of the fine-tuned ChatGPT-3.5's responses can be summarized as follows: (a) responses were concise and resembled answers typically provided by medical staff; (b) the fine-tuned ChatGPT-3.5 used dialects rather than standard Arabic in its answers though not necessarily matching the dialect of the original query; (c) no medical prescriptions, such as medication dosages, were included in the responses; (d) a few irrelevant answers were identified; and (e) the fine-tuned ChatGPT-3.5 more frequently advised users to consult their doctors and provided limited general medical information.

4.5. NLP-Based Analysis for Answers

On the other hand, the authors' analysis of the textual content of ChatGPT-3.5's responses (prior to fine tuning) revealed several key observations. (a) Specialist Referral: ChatGPT-3.5 consistently emphasized the need to consult a specialist, reflecting its limitations in providing conclusive answers. For instance, referral statements were the third most frequent element in responses, following initial greetings and expressions of sympathy. (b) Generalized Responses: Answers were generally broad, often concluding with advice to "contact your doctor." (c) Length Discrepancy: ChatGPT-3.5's responses were significantly longer than those of experts on Altibbi.com. Expert responses averaged 45 words, while ChatGPT-3.5's responses averaged 100 words. However, when prompted for specificity,

ChatGPT-3.5's responses shortened to 30 words, often sacrificing meaningful content and defaulting to advice to consult a doctor. (d) Handling Multiple Medical Issues: In cases with multiple symptoms, ChatGPT-3.5 recommended consulting a doctor in 80% of instances, often interpreting such queries as covering broad symptom sets. (e) Jaccard Similarity: The Jaccard similarity between ChatGPT-3.5's responses and expert answers was low (0.05–0.5), indicating that expert responses were more concise and direct, whereas ChatGPT-3.5's responses tended to be verbose and general. It should be noted that Jaccard similarity relies on literal word overlap, which may not fully capture contextual similarity in this study. (f) Dialect Understanding: ChatGPT-3.5 demonstrated a strong understanding of various dialects in queries yet consistently responded using standard Arabic.

These findings highlight both the strengths and limitations of ChatGPT-3.5 in addressing medical queries, particularly its cautious approach with frequent referrals to specialists, albeit with verbosity and a tendency to generalize.

In contrast, the fine-tuned ChatGPT-3.5 displayed less sensitivity to matching the dialect of the query in its responses. For example, a question posed in an Egyptian dialect might receive a response containing terms from another dialect. Generally, most responses were in standard Arabic. This likely reflects a limitation in the fine-tuning process, where emphasis on dialect matching was minimal. Notably, most Altibbi responses used standard Arabic or, less frequently, the same or a different dialect than the query. Additionally, many responses from the fine-tuned model were short and sometimes irrelevant.

4.6. Comparative Analysis of LLMs for Different Languages

While this study focuses on optimizing Arabic language models for healthcare, examining Large Language Models (LLMs) in other languages, particularly English and Chinese, offers valuable insights. English and Chinese LLMs have been extensively developed and optimized due to the availability of large, diverse datasets and advancements in Natural Language Processing (NLP). For instance, models like GPT-3 and ChatGPT have shown substantial effectiveness in healthcare contexts in English, achieving accuracy rates of over 85% in medical query tasks. For instance, Rao et al. [81] consulted the Mass General Brigham Registry and found that ChatGPT achieved approximately 72% accuracy in clinical decision making across various medical specialties, highlighting its potential in healthcare contexts. Regarding the Chinese language, ERNIE 3.0, developed by Baidu Sun et al. [82], scored ~83.77% in terms of natural language inference. Additionally, research on Chinese language model Pangu- α [83] has shown promising results in interpreting medical terms and providing accurate healthcare advice (accuracy around 72%). Studies on these models demonstrate that they benefit from mature NLP frameworks and extensive linguistic resources, allowing them to handle complex medical terminology with high accuracy and contextual understanding.

In contrast, Arabic LLMs face unique challenges due to limited labeled data, the diversity of dialects, and the linguistic complexity of Arabic. This study's approach—fine tuning a model specifically for Arabic healthcare communication—aims to bridge these gaps, but the advancements observed in English and Chinese NLP provide a useful benchmark.

5. Discussion

Understanding patient queries and intents within the medical domain requires the development of models tailored to local languages, enhancing Natural Language Understanding (NLU) capabilities [1,47]. Moreover, there is a growing need for unsupervised approaches that leverage unlabeled datasets. The challenges in maintaining pre-trained models—given the evolving nature of queries and intents and the scarcity of high-quality medical datasets—call for innovative methodologies. The incremental learning approach proposed by Bai et al. [41] aims to construct a robust query/intent detection system. However, in situations with unknown inputs, medical conversational agents risk generating inaccurate responses [11]. Thus, there is a preference for models that can operate unsupervised or recognize both familiar and novel queries/intents, or at the very least, identify

unknown intents. Open intent discovery, as proposed by Vedula et al. [19], emerges as a potential solution, helping address the evolving nature of intents and the limitations in labeled datasets to improve adaptability in intent recognition within dynamic medical conversations.

In the Arabic context, research on patient intent recognition remains limited. Similarly, studies on Arabic chatbots for the medical field are scarce, as noted by Wael et al. [15]. Despite this, various studies emphasize the importance of high-quality, annotated data for training deep learning models—a resource lacking in Arabic. Consequently, there is an increasing demand to prioritize the collection and annotation of extensive Arabic medical conversation datasets. Abdelhay et al. [4] highlighted the significance of their MAQA dataset focused on medical-oriented Arabic content, although it is limited by its modest size (~430 K entries) and coverage of only 20 medical specializations. This limitation is echoed by Boulesnane et al. [54]. Most proposed NLP solutions for Arabic are general rather than domain-specific, as pointed out by Hijjawi and Elsheikh [46], who attribute this to the linguistic complexity of Arabic compared to English. These challenges underscore the need for medical NLP solutions that accommodate the diversity of Arabic dialects, yet a substantial gap persists in developing tools that support both standard Arabic and its dialects.

The integration of Large Language Models (LLMs) into NLP tasks for the medical domain holds significant potential, as demonstrated by the enhanced accuracy of LLMs over time [60]. Some multilingual LLMs, such as mBERT, are pre-trained on a wide range of languages, including Arabic, making them suitable for cross-lingual tasks common in medical discussions on Arabic forums. However, the limited availability of Arabic pre-training data compared to English may affect performance. Additionally, Arabic's numerous dialects introduce an additional layer of complexity, and pre-training solely on Modern Standard Arabic (MSA) may not capture dialectal nuances. Furthermore, many medical queries include terms from other languages, often written phonetically in Arabic, which presents an additional challenge.

In summary, while LLMs have demonstrated substantial capability in understanding Arabic, as shown by ChatGPT-3.5's responses, studies on NLP applications—including LLMs—highlight persistent challenges such as dialectal variation and script complexity. Researchers emphasize the need for LLMs trained on accurate, specialized medical datasets to enhance their efficacy. This focus aligns with findings on the limitations of LLMs in addressing medical questions in English. For Arabic medical NLP tasks, combining LLMs with domain-specific data and fine tuning presents a promising strategy for improving performance. LLMs designed for general responses benefit from active learning when trained on medical contextual data. Importantly, LLMs should support medical teams rather than make decisions independently, as emphasized by Harrer [62] and Hunter et al. [69]. This cautious approach reflects the current capabilities of LLMs and underscores the need for further development before they can assume a decision-making role in healthcare.

Most studies agree that ChatGPT-4.0 outperforms its counterparts. However, a critical review of the literature reveals a notable gap: there is a lack of comprehensive empirical findings specific to Arabic. Existing studies often report accuracy metrics for highly specialized queries or provide general insights on LLM applications in the medical field. To our knowledge, few studies within the Arabic context have empirically pre-trained or fine-tuned LLMs on specialized medical datasets to fully assess their capabilities. This approach has been successfully applied to other models, such as FastText, GloVe, and BERT, achieving high accuracy before the advent of LLMs.

By including non-Arabic words in the dataset, we aimed to enhance the model's adaptability to real-world usage, where multilingualism and code switching are common. This decision improves the model's ability to interpret medical queries as they are often posed by users, enabling more accurate responses. Additionally, it ensures that the model can effectively handle queries with medical terms or technical jargon in English or French within otherwise Arabic text, which is a frequent occurrence in Arabic-speaking regions.

One challenge with including non-Arabic words was ensuring that the model could accurately interpret the mix of languages, particularly when medical terms in English or French appeared within Arabic phrases. This required the model to contextualize non-Arabic medical terms within the broader structure of Arabic sentences. During fine tuning, we utilized dialect-specific tokens and multilingual embeddings to help the model distinguish between Arabic content and non-Arabic terms, allowing it to manage diverse linguistic inputs effectively.

This work contributes to Arabic NLP research by presenting a technically robust approach to fine tuning LLMs for Arabic medical queries. Our results demonstrated measurable improvements in the model's ability to interpret medical inquiries across various dialects with a significant increase in response accuracy validated by a panel of medical experts. The technical contributions of this work include the fine-tuning methodology, hyperparameter optimization, and the novel application of dialect-specific tokens to enhance model performance.

Incorporating diverse Arabic dialects—such as Egyptian, Syrian, and Moroccan—into the dataset enabled ChatGPT-3.5 to handle a broader range of medical queries from different regions, increasing the accessibility of healthcare information. However, this dialect diversity initially led to a slight drop in accuracy when the model encountered dialect-specific terms or expressions. For instance, different ways of describing medical symptoms in Moroccan Arabic versus Egyptian Arabic sometimes caused the model to misinterpret queries or provide less relevant answers. After fine tuning, we observed an overall accuracy improvement from 86.55% to 90%, with increased accuracy in specific dialects, although some challenges persisted due to significant linguistic variations. Unlike Modern Standard Arabic (MSA), which is relatively uniform and commonly used in formal contexts, spoken Arabic varies widely across regions, with distinct vocabulary, grammar, and syntax. This variation complicates NLP modeling, as it requires the model to process both standard language and dialect-specific nuances. For example, the term “headache” might be expressed differently as “صداع” in Egyptian Arabic and “راس كيضرنى” in Moroccan Arabic, adding complexity to model training, as the same medical condition can be described with entirely different expressions depending on the dialect.

One of the medical experts who evaluated ChatGPT's responses likened its output to that of a well-educated individual with broad, though superficial, knowledge across diverse topics. However, this expert noted that the model's depth of understanding in specific domains, particularly in medicine, remains limited. Consequently, they concluded that such language models, despite their extensive knowledge base, are not yet capable of assuming responsibilities in the medical field in the foreseeable future.

The effective utilization of LLMs for Arabic NLP tasks, particularly in patient-intention discovery and healthcare-related queries, requires a systematic and strategic approach. Based on insights from existing LLM literature and findings from this study, the following key recommendations emerged:

(a) **Data Collection:** It is essential to collect an extensive, diverse dataset of Arabic healthcare-related text, including patient inquiries and medical records. This dataset should cover a broad range of Arabic dialects and forms. However, the inclusion of non-Arabic words (either written phonetically in Arabic script or in their original alphabets) remains underexplored. Notably, few studies have collected substantial Arabic datasets, with the exception of Habib et al. [10], who amassed 1.5 million records.

(b) **LLM Selection:** Choosing the right LLM is critical, with preference given to models that explicitly support the Arabic language, such as mBERT or Arabic-specific LLMs.

(c) **Pre-Training and Fine Tuning:** It is important to ensure that the selected LLM has been pre-trained on a diverse Arabic text corpus and subsequently fine-tuned on medical-specific tasks, such as patient-intention recognition and healthcare-related queries. Integrating Named Entity Recognition (NER) models alongside LLMs can further enhance task-specific performance.

(d) Labeled Datasets: The creation of sufficiently labeled datasets for training and validation is essential to enable the LLM to effectively understand and respond to medical queries.

(e) Dialectal Variations: Given the inherent diversity in Arabic dialects, it is necessary to fine-tune the model using data specific to certain regions or dialects. This process enhances the model's performance on localized healthcare-related tasks.

By understanding the aforementioned notes, researchers and practitioners can establish robust foundations for utilizing LLMs in Arabic NLP tasks within the healthcare domain, ensuring both accuracy and adaptability to diverse linguistic nuances. Beyond direct patient communication, the improvements observed in the accuracy of the fine-tuned LLM suggest its potential application in automating medical documentation tasks. With further refinement, LLMs could assist healthcare providers by generating detailed patient records based on doctor-patient interactions, reducing the administrative burden and allowing medical professionals to focus more on patient care. Similarly, in public health surveillance, the model's ability to process vast amounts of textual data could be utilized to monitor emerging health trends, aiding in early detection and rapid response to health crises. For instance, if fine tuning the LLM improves accuracy in answering patient questions, it can be tested on how this same fine-tuning approach might be useful in generating accurate medical documentation or assisting in public health data analysis.

On the other hand, in the medical context, optimizing Large Language Models (LLMs) presents unique challenges due to the high risks involved in producing accurate and contextually relevant responses. The fluctuating training loss we observed was a direct result of the model encountering specialized medical terminology and domain-specific knowledge. This challenge was amplified by the necessity to accurately interpret and respond to queries posed in various Arabic dialects, which often have distinct ways of expressing medical symptoms and conditions. Ensuring that the model accurately captured both the medical content and the regional linguistic variations required a careful balance of fine-tuning strategies.

A key challenge for ChatGPT 3.5 in this multilingual environment was the need to generalize across different dialects while maintaining the semantic meaning of the queries. Due to the lexical and grammatical differences between dialects, the model sometimes overfitted to a specific dialect or struggled to generalize across dialects that shared similarities. For example, while the model performed well in recognizing standard medical terms in both Egyptian and Syrian Arabic, it struggled with regional idiomatic expressions or idioms common in Moroccan Arabic. Testing the model with queries related to common symptoms like 'fever' or 'headache', the accuracy was 92% for the Egyptian dialect but dropped to 85% for the Moroccan dialect. A qualitative analysis of the responses showed that while the model correctly interpreted the medical condition in the Egyptian case, it misinterpreted the context in some Moroccan queries due to less familiar phrasing. To address this, we introduced dialect-specific tokens during training, which improved the model's ability to contextualize and differentiate between regional variations while preserving overall accuracy.

Finally, this decision to include non-Arabic words has broader implications for NLP models in multilingual environments, where code switching is a common feature of everyday communication. The ability to handle queries that combine languages reflects the growing need for NLP models that can operate effectively in multilingual and culturally diverse contexts. In the medical domain, where accuracy is crucial, this adaptability ensures that the model can serve a wider audience and respond effectively to queries that may not adhere to strict monolingual patterns but rather reflect the dynamic linguistic landscape of the Arabic-speaking world.

Additionally, the level of professional experience among the evaluators could influence the evaluation process. More experienced medical professionals might focus on clinical accuracy and the potential for medical harm in a way that differs from less experienced professionals who may be more familiar with recent medical technologies and advancements in digital health. Evaluators at different career stages might also differ in their ability

to recognize the nuances of emerging medical terminology or practices, which could affect their assessments of the model's performance. Including a broader range of experience levels in future evaluations could provide additional insights into how ChatGPT 3.5 performs across different user demographics.

Additionally, our model incorporates a code-switching detection mechanism, which is a capability that is rare in the field of Arabic NLP. Given the frequent use of English or French medical terminologies alongside Arabic in patient queries, our model is designed to seamlessly handle and interpret mixed-language inputs. This is particularly crucial in real-world healthcare settings where code switching is predominant. Our system's ability to process and respond to queries that involve multiple languages simultaneously highlights a significant advancement in multilingual and culturally relevant healthcare communication systems.

Furthermore, we introduce the concept of a culturally aware NLP system, which is designed to adapt to the social and cultural norms that exist in Arabic healthcare interactions. Unlike other models that merely translate text, our approach goes beyond language processing to incorporate distinctions in conversation style, formality, and tone, which are critical in patient–doctor communication in Arabic cultures. This adds an additional layer of personalization and relevance to the user experience, making the system more aligned with the cultural expectations of Arabic-speaking patients.

Lastly, by focusing on localized linguistic and cultural challenges, our work provides a specialized, domain-specific solution that can outperform generic models in Arabic healthcare settings.

In sum, this research presents a novel application of fine-tuned LLMs specifically for Arabic healthcare, incorporating dialect management, code-switching detection, and cultural sensitivity.

6. Conclusions

This study has successfully demonstrated the significant potential of utilizing advanced AI and natural language processing technologies, specifically through the fine tuning of ChatGPT 3.5, to create a more patient-centered and culturally attuned digital healthcare experience. By carefully training the model with data that capture the patient–doctor interactions typical in Arab cultures, we have made strides in ensuring that the AI-generated responses are not only accurate and informative but also resonate with the expectations and social norms of Arab patients.

This study successfully demonstrated the potential of fine-tuning Large Language Models (LLMs) for Arabic healthcare communication. We observed notable improvements in accuracy when responding to patient queries. Before fine tuning, ChatGPT 3.5 achieved an average accuracy of 86.55% across different medical specializations as evaluated by a panel of seven medical experts. After fine tuning, the accuracy increased to 90%, with individual specializations such as neurosurgery and public health reaching accuracy levels of 92% and 92.5%, respectively. This represents a significant improvement of 3.45% overall. In addition to accuracy, we also observed that the fine-tuned model provided shorter, more precise answers, which were better aligned with the responses given by healthcare professionals on platforms such as Altibbi. The fine-tuned model also demonstrated enhanced dialect handling, further improving its ability to serve Arabic-speaking patients across diverse linguistic backgrounds.

In summary, the incorporation of LLMs into Arabic healthcare communication significantly enhances the quality and relevance of automated responses, making it a viable solution for scaling medical services to meet the needs of the Arabic-speaking population. The improvement in performance metrics, including a 3.45% increase in accuracy post-tuning, highlights the importance of domain-specific fine tuning in achieving higher precision in medical dialogue systems.

The main contribution of our work lies in its innovative approach to enhancing AI communication within the healthcare domain, ensuring that it goes beyond the mere

translation of languages to truly understand and replicate the cultural subtleties that define effective and comforting patient–doctor dialogues. This fine-tuning process has resulted in a version of ChatGPT that is significantly more adept at mimicking the empathetic and patient-centric approach characteristic of Arab medical professionals, bridging a crucial gap in the digital healthcare landscape.

Moreover, this research highlights the importance of cultural competence in healthcare innovations, highlighting how AI can be shaped to cater to diverse patient populations. The implications of our findings are far reaching, suggesting that similar methodologies could be applied to adapt AI technologies for various cultural contexts, thus democratizing access to personalized and culturally sensitive healthcare advice.

As we look to the future, it is evident that the intersection of AI and healthcare holds immense promise for improving patient outcomes and satisfaction. The advancements demonstrated in this paper pave the way for further research into the customization of AI tools for healthcare, encouraging a more inclusive approach that considers the cultural and personal needs of patients worldwide. By continuing to refine and expand upon the capabilities of AI like ChatGPT, we can aspire to create a global healthcare environment that is more accessible, understanding, and responsive to the unique needs of every patient.

7. Limitations and Future Recommendation

Limitations are represented in many aspects:

- (a) The use of non-random convenience sampling in selecting medical experts may limit the generalizability of the findings. Since the experts were chosen based on availability and their specific expertise, there may be a degree of selection bias that could influence the evaluation outcomes. For example, the experts included in the study may have had specific experiences or biases toward certain medical technologies or practices that may not be representative of the broader medical community. Consequently, this may affect how well the results can be generalized to a wider population of healthcare professionals. To mitigate this limitation in future research, a more diverse and randomly selected pool of medical professionals could be included in the evaluation. Additionally, including healthcare providers from different regions and medical systems could help improve the generalizability of the findings.
- (b) One potential bias introduced by using convenience sampling is that the selected medical experts may have similar backgrounds or expertise, which could skew the evaluation toward a particular perspective. For instance, experts from certain medical specialties may focus on specific criteria when evaluating the model's responses, which might not be prioritized by experts from other domains. This could limit the breadth of feedback received.
- (c) In real-world medical applications, the implications of overfitting can be particularly severe. A model that performs well on training data but lacks generalization could provide inaccurate or irrelevant medical advice to users, which could lead to misinformation or misdiagnosis. This risk is amplified in multilingual environments where the model needs to handle a wide variety of dialects and linguistic variations. Therefore, while optimizing for training loss may lead to short-term gains in model accuracy on the training set, it may not necessarily translate into improved performance on real-world data unless regular validation is conducted to ensure the model's robustness and generalizability.
- (d) Including a broader range of medical specialties—such as dermatology, cardiology, psychiatry, etc., the evaluation might have revealed different strengths and limitations of ChatGPT 3.5. For instance, specialties dealing with more subjective patient-reported symptoms, such as psychiatry, might place a greater emphasis on the model's ability to recognize distinctions in patient descriptions, which may differ from specialties that focus on more objective metrics, such as radiology or surgery. Including more diverse specialties could provide a more well-rounded assessment of the model's ability to handle a wider variety of medical queries.

Future research could employ a mixed methods approach, combining qualitative feedback from a smaller panel of medical experts with quantitative evaluations from a larger pool of healthcare professionals. For example, a larger group of general practitioners or medical students could rate the accuracy of ChatGPT's responses using predefined metrics, while a smaller expert panel could provide deeper qualitative feedback on the model's medical relevance and nuance. This approach would ensure that the depth of qualitative insights is maintained while allowing for broader generalizability and more representative data.

The Delphi method could be employed in future research to gather expert evaluations from a larger number of medical professionals while maintaining the depth of qualitative insights. In this methodology, experts participate in multiple rounds of structured surveys, where they provide feedback on the model's responses. After each round, a summary of the group's feedback is shared, and the experts are encouraged to revise their assessments, ultimately leading to a consensus. This iterative process ensures rigorous expert feedback while mitigating the challenges posed by a smaller sample size.

Author Contributions: Conceptualization, R.M. and O.S.A.; methodology, R.M. and O.S.A.; software, R.M.; validation, O.S.A. and M.H.; formal analysis, R.M. and O.S.A.; investigation, M.H.; resources, O.S.A.; data curation, R.M.; writing—original draft preparation, R.M. and O.S.A.; writing—review and editing, M.H.; visualization, R.M. and O.S.A.; supervision, M.H.; project administration, O.S.A. and M.H.; funding acquisition, O.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to acknowledge the support received from the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant JRCAL-RG-08.

Data Availability Statement: These data were derived from the following resources available in the public domain <https://shorturl.at/dR46Y> (accessed on 12 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest. AI has been used only for proofreading.

References

- Mullick, A.; Mondal, I.; Ray, S.; Raghav, R.; Chaitanya, G.S.; Goyal, P. Intent Identification and Entity Extraction for Healthcare Queries in Indic Languages. In Proceedings of the EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 1870–1881.
- Faris, H.; Habib, M.; Faris, M.; Alomari, A.; Castillo, P.A.; Alomari, M. Classification of Arabic healthcare questions based on word embeddings learned from massive consultations: A deep learning approach. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *13*, 1811–1827. [[CrossRef](#)]
- Gebbia, V.; Piazza, D.; Valerio, M.R.; Borsellino, N.; Alberto, F. Patients with cancer and COVID-19: A whatsapp messenger-based survey of patients' queries, needs, fears, and actions taken. *JCO Glob. Oncol.* **2020**, *6*, 722–729. [[CrossRef](#)] [[PubMed](#)]
- Abdelhay, M.; Mohammed, A.; Hefny, H.A. Deep learning for Arabic healthcare: MedicalBot. *Soc. Netw. Anal. Min.* **2023**, *13*, 71. [[CrossRef](#)] [[PubMed](#)]
- Habib, M.; Faris, M.; Qaddoura, R.; Alomari, A.; Faris, H. A predictive text system for medical recommendations in telemedicine: A deep learning approach in the Arabic context. *IEEE Access* **2021**, *9*, 85690–85708. [[CrossRef](#)]
- Maniou, T.A.; Veglis, A. Employing a chatbot for news dissemination during crisis: Design, implementation and evaluation. *Future Internet* **2020**, *12*, 109. [[CrossRef](#)]
- Li, Y.; Grandison, T.; Silveyra, P.; Douraghy, A.; Guan, X.; Kieselbach, T.; Li, C.; Zhang, H. Jennifer for COVID-19: An NLP-powered chatbot built for the people and by the people to combat misinformation. In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Online, 5–10 July 2020.
- Li, B.; Shi, J.; Gutman, B.A.; Baxter, L.C.; Thompson, P.M.; Caselli, R.J.; Wang, Y.; Initiative, A.s.D.N. Influence of APOE Genotype on Hippocampal Atrophy over Time—An N = 1925 Surface-Based ADNI Study. *PLoS ONE* **2016**, *11*, e0152901. [[CrossRef](#)] [[PubMed](#)]
- Zeng, G.; Yang, W.; Ju, Y.; Wang, S.; Zhang, R.; Zhou, M.; Zeng, J.; Dong, X.; Zhang, R.; Fang, H.; et al. Meddialog: Large-scale medical dialogue datasets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 9241–9250.
- Habib, M.; Faris, M.; Alomari, A.; Faris, H. AltibbiVec: A Word Embedding Model for Medical and Health Applications in the Arabic Language. *IEEE Access* **2021**, *9*, 133875–133888. [[CrossRef](#)]

11. Aftab, H.; Gautam, V.; Hawkins, R.; Alexander, R.; Habli, I. Robust Intent Classification using Bayesian LSTM for Clinical Conversational Agents (CAs). In Proceedings of the 10th EAI International Conference on Wireless Mobile Communication and Healthcare, Virtual Event, 13–14 November 2021.
12. Mehta, D.; Santy, S.; Mothilal, R.K.; Srivastava, B.M.L.; Sharma, A.; Shukla, A.; Prasad, V.; Venkanna, U.; Sharma, A.; Bali, K. Learnings from technological interventions in a low resource language: A case-study on Gondi. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 2832–2838.
13. Daniel, J.E.; Brink, W.; Eloff, R.; Copley, C. Towards automating healthcare question answering in a noisy multilingual low-resource setting. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 948–953.
14. Mounsef, J.; Hasib, M.; Raza, A. Building an Arabic Dialectal Diagnostic Dataset for Healthcare. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 859–868. [[CrossRef](#)]
15. Wael, T.; Hesham, A.; Youssef, M.; Adel, O.; Hesham, H.; Darweesh, M.S. Intelligent Arabic-Based Healthcare Assistant. In Proceedings of the 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 23–25 October 2021.
16. Chen, Q.; Zhuo, Z.; Wang, W. BERT for Joint Intent Classification and Slot Filling. *arXiv* **2019**, arXiv:1902.10909.
17. Marie-Sainte, S.L.; Alalayani, N.; Alotaibi, S.; Ghouzali, S.; Abunadi, I. Arabic Natural Language Processing and Machine Learning-Based Systems. *IEEE Access* **2018**, *7*, 7011–7020. [[CrossRef](#)]
18. Al-Ayyoub, M.; Nuseir, A.; Alsmearat, K.; Jararweh, Y.; Gupta, B. Deep learning for Arabic NLP: A survey. *J. Comput. Sci.* **2018**, *26*, 522–531. [[CrossRef](#)]
19. Vedula, N.; Lipka, N.; Maneriker, P.; Parthasarathy, S. Towards Open Intent Discovery for Conversational Text. In Proceedings of the Conference’17, Washington, DC, USA, 25–27 July 2017.
20. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
21. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *Comput. Sci. Linguist.* **2018**. Available online: <https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035> (accessed on 12 November 2024).
22. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805v2. [[CrossRef](#)]
23. Dash, S.; Acharya, B.R.; Mittal, M.; Abraham, A.; Kelemen, A. Deep Learning Techniques for Biomedical and Health Informatics. In *Studies in Big data*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 68.
24. Kwak, G.-J.; Hui, P. DeepHealth: Deep Learning for Health Informatics. *arXiv* **2019**, arXiv:1909.00384.
25. Mulani, J.; Heda, S.; Tumdi, K.; Patel, J.; Chhinkaniwala, H.; Patel, J. Deep reinforcement learning based personalized health recommendations. In *Deep Learning Techniques for Biomedical and Health Informatics*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 231–255.
26. Kumar, A.; Sarkar, S.; Pradhan, C. Malaria disease detection using cnn technique with sgd, rmsprop and adam optimizers. In *Deep Learning Techniques for Biomedical and Health Informatics*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 211–230.
27. Akselrod-Ballin, A.; Chorev, M.; Shoshan, Y.; Spiro, A.; Hazan, A.; Melamed, R.; Barkan, E.; Herzel, E.; Naor, S.; Karavani, E. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* **2019**, *292*, 331–342. [[CrossRef](#)]
28. Shah, A.M.; Yan, X.; Shah, S.A.A.; Mamirkulova, G. Mining patient opinion to evaluate the service quality in healthcare: A deep-learning approach. *J. Ambient. Intell. Humaniz. Comput.* **2019**, *11*, 2925–2942. [[CrossRef](#)]
29. Vidhya, K.; Shanmugalakshmi, R. Deep learning based big medical data analytic model for diabetes complication prediction. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 5691–5702. [[CrossRef](#)]
30. Lauritsen, S.M.; Kalør, M.E.; Kongsgaard, E.L.; Lauritsen, K.M.; Jørgensen, M.J.; Lange, J.; Thiesson, B. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif. Intell. Med.* **2020**, *104*, 101820. [[CrossRef](#)]
31. Faes, L.; Wagner, S.K.; Fu, D.J.; Liu, X.; Korot, E.; Ledsam, J.R.; Back, T.; Chopra, R.; Pontikos, N.; Kern, C. Automated deep learning design for medical image classification by health-care professionals with no coding experience: A feasibility study. *Lancet Dig. Health* **2019**, *1*, 232–242. [[CrossRef](#)]
32. Estrada, S.; Lu, R.; Conjeti, S.; Orozco-Ruiz, X.; Panos-Willuhn, J.; Breteler, M.M.; Reuter, M. Fatsegnet: A fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. *Magn. Reson. Med.* **2020**, *83*, 1471–1483. [[CrossRef](#)] [[PubMed](#)]
33. Edara, D.C.; Vanukuri, L.P.; Sistla, V.; Kolli, V.K.K. Sentiment analysis and text categorization of cancer medical records with lstm. *J. Ambient. Intell. Humaniz. Comput.* **2019**, *14*, 5309–5325. [[CrossRef](#)]
34. Liu, F.; Weng, C.; Yu, H. Advancing clinical research through natural language processing on electronic health records: Traditional machine learning meets deep learning. In *Clinical Research Informatics*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 357–378.
35. Zhang, L.; Lin, J.; Liu, B.; Zhang, Z.; Yan, X.; Wei, M. A review on deep learning applications in prognostics and health management. *IEEE Access* **2019**, *7*, 162415–162438. [[CrossRef](#)]
36. Liu, W.; Tang, J.; Qin, J.; Xu, L.; Li, Z.; Liang, X. MedDG: A Large-scale Medical Consultation Dataset for Building Medical Dialogue System. *arXiv* **2020**, arXiv:2010.07497.

37. Young-Min, K.; Lee, T.-H.; Na, S.-O. Constructing novel datasets for intent detection and ner in a korean healthcare advice system: Guidelines and empirical results. *Appl. Intell.* **2022**, *53*, 941–961.
38. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **2020**, *3*, 1–23. [[CrossRef](#)]
39. Zhou, B.; Yang, G.; Shi, Z.; Ma, S. Natural Language Processing for Smart Healthcare. *arXiv* **2021**, arXiv:2110.15803. [[CrossRef](#)] [[PubMed](#)]
40. Bao, Q.; Ni, L.; Liu, J. Hhh: An online medical chatbot system based on knowledge graph and hierarchical bi-directional attention. In Proceedings of the Australasian Computer Science Week Multiconference, Melbourne, VIC, Australia, 4–6 February 2020; pp. 1–10.
41. Bai, G.; He, S.; Liu, K.; Zhao, J. Incremental intent detection for medical domain with contrast replay networks. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022.
42. Razzaq, M.A.; Khan, W.A.; Lee, S. Intent-context fusioning in healthcare dialogue-based systems using jdl model. In Proceedings of the International Conference on Smart Homes and Health Telematics, Singapore, 10–12 July 2018; pp. 61–72.
43. Amato, F.; Marrone, S.; Moscato, V.; Piantadosi, G.; Picariello, A.; Sansone, C. Chatbots meet ehealth: Automating healthcare. In Proceedings of the WAIHA@ AI* IA, Bari, Italy, 14 November 2017; pp. 40–49.
44. Zhang, C.; Du, N.; Fan, W.; Li, Y.; Lu, C.; Philip, S.Y. Bringing semantic structures to user intent detection in online medical queries. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 1019–1026.
45. Mondal, I.; Ahuja, K.; Jain, M.; O’Neill, J.; Bali, K.; Choudhury, M. Global Readiness of Language Technology for Healthcare: What Would It Take to Combat the Next Pandemic? In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 4320–4335.
46. Hijjawi, M.; Elsheikh, Y. Arabic language challenges in text based conversational agents compared to the English language. *Int. J. Comput. Sci. Inf. Technol.* **2015**, *7*, 1–13. [[CrossRef](#)]
47. Zhang, N.; Chen, M.; Bi, Z.; Liang, X.; Li, L.; Shang, X.; Yin, K.; Tan, C.; Xu, J.; Huang, F.; et al. Cblue: A chinese biomedical language understanding evaluation benchmark. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 7888–7915.
48. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. Pubmedqa: A dataset for biomedical research question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, Hong Kong, China, 3–7 November 2019; pp. 2567–2577.
49. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the LREC 2020 Workshop Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020.
50. Alruily, M. ArRASA: Channel Optimization for Deep Learning-Based Arabic NLU Chatbot Framework. *Electronics* **2022**, *11*, 3745. [[CrossRef](#)]
51. Mezzi, R.; Yahyaoui, A.; Krir, M.W.; Boulila, W.; Koubaa, A. Mental Health Intent Recognition for Arabic-Speaking Patients Using the Mini International Neuropsychiatric Interview (MINI) and BERT Model. *Sensors* **2022**, *22*, 846. [[CrossRef](#)]
52. Alhassan, N.A.; Albarrak, A.S.; Bhatia, S.; Agarwal, P. A Novel Framework for Arabic Dialect Chatbot Using Machine Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 1844051. [[CrossRef](#)] [[PubMed](#)]
53. Boudjellal, N.; Zhang, H.; Khan, A.; Ahmad, A.; Naseem, R.; Dai, L. A Silver Standard Biomedical Corpus for Arabic Language. *Complexity* **2020**, *2020*, 8896659. [[CrossRef](#)]
54. Boulesnane, A.; Saidi, Y.; Kamel, O.; Bouhamed, M.M.; Mennour, R. DZchatbot: A Medical Assistant Chatbot in the Algerian Arabic Dialect using Seq2Seq Model. In Proceedings of the 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS), Oum El Bouaghi, Algeria, 12–13 October 2022.
55. Naous, T.; Antoun, W.; Mahmoud, R.; Hajj, H. Empathetic BERT2BERT Conversational Model: Learning Arabic Language Generation with Little Data. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19 April 2021.
56. Lim, Z.W.; Pushpanathan, K.; Er Yew, S.M.; Lai, Y.; Sun, C.-H.; Lam, J.S.H.; Chen, D.Z.; Goh, J.H.L.; Tan, M.C.J.; Sheng, B.; et al. Benchmarking large language models’ performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *eBioMedicine* **2023**, *95*, 104770. [[CrossRef](#)]
57. Varghese, J.; Chapiro, J. ChatGPT: The transformative influence of generative AI on science and healthcare. *J. Hepatol.* **2023**, *80*, 977–980. [[CrossRef](#)]
58. Ali, H.; Patel, P.; Obaitan, I.; Mohan, B.P.; Sohail, A.H.; Smith-Martinez, L.; Lambert, K.; Gangwani, M.K.; Easler, J.J.; Adler, D.G. Evaluating ChatGPT’s Performance in Responding to Questions About Endoscopic Procedures for Patients. *iGIE* **2023**, *2*, 553–559. [[CrossRef](#)]
59. Pushpanathan, K.; Lim, Z.W.; Er Yew, S.M.; Chen, D.Z.; Lin, H.A.H.; Goh, J.H.L.; Wong, W.M.; Wang, X.; Marcus, C.J.T.; Koh, V.T.C.; et al. Popular Large Language Model Chatbots’ Accuracy, Comprehensiveness, and Self-Awareness in Answering Ocular Symptom Queries. *iScience* **2023**, *26*, 108163. [[CrossRef](#)]
60. Vaishya, R.; Misra, A.; Vaish, A. ChatGPT: Is this version good for healthcare and research? *Diabetes Metab. Syndr. Clin. Res. Rev.* **2023**, *17*, 102744. [[CrossRef](#)]

61. Sengupta, P.; Dutta, S.; Chakravarthi, S.; Jegasothy, R.; Jeganathan, R.; Pichumani, A. Comparative efficacy of ChatGPT 3.5, ChatGPT 4, and other large language models (LLMs) in gynecology and infertility research. *Gynecol. Obstet. Clin. Med.* **2023**, *3*, 203–206. [CrossRef]
62. Harrer, S. Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine* **2023**, *90*, 104512. [CrossRef]
63. Puladi, B.; Gsaxner, C.; Kleesiek, J.; Hölzle, F.; Röhrig, R.; Egger, J. The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: A narrative review. *Int. J. Oral Maxillofac. Surg.* **2023**, *53*, 78–88. [CrossRef]
64. Reddy, S. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Inform. Med. Unlocked* **2023**, *41*, 101304. [CrossRef]
65. Comi, D.; Christofidellis, D.; Piazza, P.F.; Manica, M. Z-BERT-A: A zero-shot Pipeline for Unknown Intent detection. *arXiv* **2022**, arXiv:2208.07084.
66. Lee, T.-C.; Staller, K.; Botoman, V.; Pathipati, M.P.; Varma, S.; Kuo, B. ChatGPT Answers Common Patient Questions About Colonoscopy. *Gastroenterology* **2023**, *165*, 509–511. [CrossRef] [PubMed]
67. Tan, T.F.; Thirunavukarasu, A.J.; Campbell, J.P.; Keane, P.A.; Pasquale, L.R.; Abramoff, M.D.; Kalpathy-Cramer, J.; Lum, F.; Kim, J.E.; Baxter, S.L.; et al. Generative Artificial Intelligence Through ChatGPT and Other Large Language Models in Ophthalmology: Clinical Applications and Challenges. *Ophthalmol. Sci.* **2023**, *3*, 100394. [CrossRef] [PubMed]
68. Kuckelman, I.J.; Yi, P.H.; Bui, M.; Onuh, I.; Anderson, J.A.; Ross, A.B. Assessing AI-Powered Patient Education: A Case Study in Radiology. *Acad. Radiol.* **2023**, *31*, 338–342. [CrossRef]
69. Hunter, R.B.; Mehta, S.D.; Limon, A.; Chang, C.A. Decoding ChatGPT: A primer on large language models for clinicians. *Intell.-Based Med.* **2023**, *8*, 100114. [CrossRef]
70. Cai, L.Z.; Shaheen, A.; Jin, A.; Fukui, R.; Yi, J.S.; Yannuzzi, N.; Alabiad, C. Performance of Generative Large Language Models on Ophthalmology Board-Style Questions. *Am. J. Ophthalmol.* **2023**, *254*, 141–149. [CrossRef]
71. Hart, S.N.; Hoffman, N.G.; Gershkovich, P.; Christenson, C.; McClintock, D.S.; Miller, L.J.; Jackups, R.; Azimi, V.; Spies, N.; Brodsky, V. Organizational preparedness for the use of large language models in pathology informatics. *J. Pathol. Inform.* **2023**, *14*, 100338. [CrossRef]
72. Tariq, R.; Malik, S.; Khanna, S. Evolving Landscape of Large Language Models: An Evaluation of ChatGPT and Bard in Answering Patient Queries on Colonoscopy. *Gastroenterology* **2023**, *166*, 220–221. [CrossRef]
73. Jackson, A. Jais: A New Pinnacle in Open Arabic NLPz. Available online: <https://www.cerebras.net/blog/jais-a-new-pinnacle-in-open-arabic-nlp> (accessed on 29 September 2023).
74. Fultinavičiūtė, U. It's a Match! Connecting Patients to Clinical Trials with AI. Available online: <https://www.clinicaltrialsarena.com/features/clinical-trial-matching-ai/?cf-view> (accessed on 1 October 2023).
75. Alammary, A.S. BERT Models for Arabic Text Classification: A Systematic Review. *Appl. Sci.* **2022**, *12*, 5720. [CrossRef]
76. Tang, L.; Sun, Z.; Idnay, B.; Nestor, J.G.; Soroush, A.; Elias, P.A.; Xu, Z.; Ding, Y.; Durrett, G.; Rousseau, J.; et al. Evaluating large language models on medical evidence summarization. *npj Digit. Med.* **2023**, *6*, 158. [CrossRef] [PubMed]
77. Merriam, S.B.; Tisdell, E.J. *Qualitative Research: A Guide to Design and Implementation*, 4th ed.; Jossey-Bass: San Francisco, CA, USA, 2015.
78. Denzin, N.K.; Lincoln, Y.S. *The Sage Handbook of Qualitative Research*, 5th ed.; Sage Publications: Thousand Oaks, CA, USA, 2017.
79. Creswell, J.W.; Poth, C.N. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*; Sage Publications: Thousand Oaks, CA, USA, 2017.
80. Lohr, S.L. *Sampling Design and Analysis*, 3rd ed.; Chapman and Hall/CRC: New York, NY, USA, 2021. [CrossRef]
81. Rao, A.; Pang, M.; Kim, J.; Kamineni, M.; Lie, W.; Prasad, A.K.; Landman, A.; Dreyer, K.; Succi, M.D. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *J. Med. Internet Res.* **2023**, *25*, e48659. [CrossRef] [PubMed]
82. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv* **2021**, arXiv:2107.02137. [CrossRef]
83. Zeng, W.; Ren, X.; Su, T.; Wang, H.; Liao, Y.; Wang, Z.; Jiang, X.; Yang, Z.; Wang, K.; Zhang, X.; et al. PANGU- α : Large-Scale Autoregressive pretrained Chinese language models with auto-parallel computation. *arXiv* **2021**, arXiv:2104.12369.

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.