Developing a Medical Question-Answering Chatbot for Egyptian Arabic: A Comparative Study

Esraa Ismail[1], Marryam Yahya[2], Mariam Nabil[3], Yomna Ashraf[4], Ziad Elshaer[5], Ghada Khoriba[6]
School of Information Technology and Computer Science
Nile University, Giza, Egypt
Emails: {E.Ismail2165, M.Yahya2163, M.Nabil2184, Y.Ashraf2278, ZElshaer, ghadakhoriba }@nu.edu.eg

***Abstract*—The Egyptian Arabic medical chatbot's development portrays a notable development in the medical field in natural language processing (NLP). This research addresses the Egyptian speech data limitation for question-answering (QA) systems by applying a QA model competent in processing speech and text as input. The model generates answers by implementing the Large Language Model (LLM), creating a dataset for future training models. We implemented Prompt Engineering using the Chain of Thoughts (CoT) approach to improve the model's capability to generate precise Arabic medical answers. Applying the CoT enabled the model to think logically before responding to the question based on the given instructions. We implemented two LLM's: Silma and Phi. The models were evaluated based on BertScore, achieving accuracies of 65.2% and 59.8%, respectively. This paper aims to focus on advancing the NLP in the medical Arabic language and contribute a novel dataset for training speech QA models.**

***Index Terms*—Egyptian Arabic, Medical Chatbots, NLP, LLM, Prompt Engineering, Silma, Phi.**

I. Introduction

The discipline of NLP is growing, and it has changed the medical field and enabled the creation of various innovative medical applications that aim to provide a real-life solution to improve patient outcomes. Despite that, the process of creating Arabic Medical Chatbots has slowed due to the shortage of Arabic speech data for QA systems. Implementing the Arabic Medical bots faces several challenges related to the morphological structure, the variety of Arabic dialects, and the need for a suitable comprehensive and detailed medical database [1]. To overcome this limitation, we aim to develop a comprehensive QA model that analyzes the user's question as a text or speech. Aiming to improve the healthcare services for all Egyptian Arabic speakers, our project intends to create a speech dataset from the QA chatbot using two LLMs, Phi and Silma. By closing this gap, we aim to implement a chatbot that will enable the user to ask the question as a Speech or Text, and the model will generate a text response, forming a complete system for Medical Services. The ability to transform the capabilities of the NLP has revolutionized communication, logical decisions, and effective retrieval processes [2]. We aim to create an Arabic-Egyptian dataset for future training models. The used medical QA dataset consists of 87,930 of 12 different medical fields, all sourced from a creditable medical website, divided into training, validation, and test sets [3]. We used the SILMA and Phi LLMs to generate relevant answers based on the instructions given in the prompt. The models were evaluated using BERTScore [4] to compare the ground truth answer with the generated answer based on semantic meaning. The models achieved an accuracy of 65.2% and 59.8%, respectively.

Section II presents a comprehensive review of the related work, focusing on advancements in natural language processing (NLP), particularly in Arabic, and previous efforts in developing medical question-answering (QA) systems. Section III outlines the methodology employed in this research, detailing the application of the Chain of Thought (CoT) approach, Large Language Models (LLMs), and Prompt Engineering to develop the Egyptian Arabic medical chatbot pipeline. The implementation and evaluation process of two LLMs, Silma and Phi, are also discussed. Section IV presents the results and discussions, where the performance of the models is assessed using the BertScore metric to compare the generated answers with the ground truth. Finally, Section V concludes the paper by summarizing key findings and contributions and outlining potential future directions for further improving NLP in the medical Arabic domain.

II. Related Work

Recently, healthcare services, most extensively medical bots, have grown popular. The instant access they provide for medical services is highly needed for doctors and patients. One of the most challenging bots to implement is an Arabic one. Yet, Abdelhay et al. [5] developed one of the largest Q&A datasets, MAQA [6], with over 430,000 questions spread throughout 20 medical specialties. Utilizing deep learning models, an average cosine similarity of 80.81% and a BLeU [7] score of 58% were achieved.

Similarly, DZchatbot, a Medical Assistant Chatbot in the Algerian Arabic Dialect, was implemented using Long Short Term Memory (LSTM) [8], Bi-directional LSTM (Bi-LSTM) [9], and Gated Recurrent Unit (GRU) [10]. The chatbot assisted with understanding the user's inquiry and providing the most reasonable answer, addressing the complexity of the Arabic language and the insufficiency of the data resources. The highest achieved accuracy was the GRU model, with an overall accuracy of 90% during the model training [11].

979-8-3503-6777-5/24/$31.00 ©2024 IEEE

Another medical visual question-answering system, Med-VQA [12], responds to NLP questions based on medical images as input. This work presents a domain-specific pre-training strategy that involves a novel contrastive learning pre-training method to mitigate the issue of the small datasets for the Med-VQA task. The model was evaluated based on the model's visual reasoning, which used the evidence verification techniques, resulting in an accuracy of 60% on the VQA-Med 2019 test dataset similar to the other leading Med-VQA models [13].

The BiQA corpus [14] is a large-scale, question-article pair dataset, obtaining 7,453 questions and 14,239 question-article pairs. The BiQA corpus has some limitations when compared to expert annotations. Corpora used in community challenges, such as BioASQ and MEDIQA, require questions and answers obtained from user experts, while PubMedQA [15] and emrQA [16] use synthetically generated questions [14].

Another study by Soufyane, Abdelhakim, and Ahmed presents a new strategy in favor of improving accessibility to healthcare through the use of artificial intelligence. The authors propose a medical chatbot to assist patients in diagnosing diseases and giving essential information about various health conditions. The chatbot uses NLP supported by TF-IDF to enhance the engagement of patients and reduce healthcare costs. The proposed hybrid architecture will enable personalized diagnosis based on symptoms as a medical reference tool and improve the overall patient experience in seeking medical advice [17].

The comprehensive explorations in LLMs have significantly influenced the effectiveness of AI in healthcare. These models, for instance, OpenAI's GPT-3.5 and GPT-4, have demonstrated exceptional performance across a broad range of natural language comprehension tasks as illustrated by Al Nazi and Peng [26]. Wang et al. [20] have implemented Apollo Models to enhance the capabilities of the medical multilingual LLMs, including Apollo 7B [27], a State-Of-The-Art (SOTA) model, that achieved an accuracy of 58.78%. Exploring more LLM advancements in text and speech research, a comparative study was conducted by Abdelali et al. to illustrate the performance of several LLMs and a State-of-the-art (SOTA) model on a variety of 21 Arabic NLP tasks. SOTA models mostly outperformed LLMs in zero-shot learning, resulting in a benchmark that joins the open and close source LLMs on Arabic NLP tasks, namely LAraBench [28].

Similarly, another paper examined the use of AI chatbots for patient-specific QA from clinical notes using certain LLMs: ChatGPT3.5 and ChatGPT 4, Google Bard, and Claude. The evaluation was done based on the accuracy, importance, thoroughness, and consistency of the generated responses by each model using a 5-point Likert scale on a group of patient-certain questions. Based on this, the results for ChatGPT 3.5 and Claude showed that they could provide accurate, meaningful, and thorough responses to diverse questions [22].

The MedMCQA [29] dataset is a large-scale, multiple-choice question-answering dataset designed to test real-world medical entrance MCQS. This dataset was collected from open websites and books that put together several mock tests and online test series created by medical professionals. This research also mentions using pre-trained language models such as BERT, SciBERT [30], BioBERT [31], and PubMedBERT [32], all types of LLMs commonly used as baselines for a MedMCQA dataset. The baseline experiments on this dataset with the most current state-of-the-art methods answer only 47% of the question correctly, which is far behind the human performance of 90%, indicating possibilities for improvement in models' reasoning ability & constitutes a challenging benchmark for future research [18].

There is also another research paper [24] by Ming Zhu et al., which proposes a new dataset called MASH-QA [24] is a multiple-answer spans Healthcare question-answering dataset about the consumer health domain. The MASH-QA dataset is challenging because the answers can consist of multiple sentences from non-consecutive document parts. The MASH-QA dataset comprises 35,000 questions and 696,000 sentences from the consumer health domain. This research also mentions the use of pre-trained language models such as TANDA [33], BERT [34], RoBERTa [35], XLNet [36], and MultiCo [37], which are all types of LLMs used as baselines, and the MultiCo achieving the highest performance, indicating possibilities for improvement in models' ability to capture the relevance among multiple answer spans and constitutes a challenging benchmark for future research.

Another research paper discusses the challenges of question answering in the medical domain; they face two main challenges related to question analysis and answer extraction. Several QA approaches were proposed in the literature for the open and medical domains. They propose a new QA approach based on question entailment using information retrieval (IR) models to retrieve question candidates and the Recognizing Question Entailment (RQE) model to identify entailed questions and return their answers. They evaluate RQE methods (a deep learning model and logistic regression). They used five different datasets for training and testing, including the Stanford Natural Language Inference corpus (SNLI), the multiNLI corpus, the Quora dataset of similar questions, the Clinical-QE [38] dataset, and a new test dataset of consumer health questions. Logistic regression achieved the best accuracy of 98.60% on Clinical-RQE [39].

Another work uses patient-derived questions and LLMs to approximate human judgment, considering time consumption by medical professionals. The interactions of patients form a basis for a dataset that covers a wide range of specialties for human-LLM comparison analysis. This study considered only 94 assessments, and data from only one database were used. Moreover, knowledge gaps exist in LLM on infrequent conditions. There is also single human reviewer bias. Increasing the diversity of data and having multiple reviewers will improve accuracy and reliability [40].

Another research paper uses large language models to address the question-answering task in natural language over clinical notes. The authors apply the Medical Information Mart for Intensive Care-MIMIC-IV [41] dataset. They do an

TABLE I
COMPARISON OF RESEARCH, MODELS, OWNERS, PARAMETERS, AND SIZE

RESEARCH	MODELS	OWNER	PARAMETERS	SIZE
[13]	VGG-16 + LSTM + Concatenation	Visual Geometry Group	138 million	528 MB
[13]	ResNet-50 + LSTM + Concatenation	Microsoft Research	25.6 million	98 MB
[13]	ResNet-152 + LSTM + Concatenation	Microsoft Research	60.2 million	232 MB
[1]	Bi-LSTM	N/A	526k	N/A
[1]	Transformers	Google AI	Varies (e.g., BERT has 110 million)	420 MB
[13]	VGG-16 + BERT + Concatenation	Google AI	110 million	420 MB
[13]	VGG-16 + BERT + SAN	Google AI	110 million	420 MB
[13], [18], [19]	VGG-16 + BioBERT + Concatenation	Korea University	110 million	420 MB
[18]	SciBERT	Allen Institute for AI	110 million	420 MB
[18]	PubMedBERT	National Library of Medicine	110 million	420 MB
[20]	Apollo (1.8B to 7B)	Xidong Wang et al.	1.8 billion to 7 billion	7 GB to 28 GB
[21]	GPT-4	OpenAI	Estimated 1.7 trillion	350 GB
[21]	GPT-4 Turbo	OpenAI	Estimated 1.3 trillion	280 GB
[21]	Claude 3.5 Sonnet	Anthropic	Estimated 100 billion	200 GB
[21], [22]	Claude 3 Opus	Anthropic	Estimated 100 billion	200 GB
[21], [22], [23]	GPT-3.5 Turbo	OpenAI	175 billion	350 GB
[21]	Llama3 (70B)	Meta	70 billion	140 GB
[21], [23]	Llama3 (8B)	Meta	8 billion	16 GB
[21]	Llama400B	Meta	400 billion	800 GB
[21]	Gemini 1.5 Pro	Google DeepMind	Estimated 500 billion	1 TB
[21]	Gemini Ultra	Google DeepMind	Estimated 1 trillion	2 TB
[21]	Gemini 1.5 Flash	Google DeepMind	Estimated 250 billion	500 GB
[24]	DrQA Reader	Facebook AI Research	N/A	N/A
[24]	BiDAF	Allen Institute for AI	N/A	N/A
[18], [24]	BERT	Google AI	110 million	420 MB
[24]	SpanBERT	Facebook AI Research	110 million	420 MB
[24]	XLNet	Google Brain	340 million	800 MB
[24]	MultiCo	University of Washington	N/A	N/A
[24]	TANDA	University of Maryland	N/A	N/A
[24]	RoBERTa	Facebook AI Research	125 million	500 MB
[11], [1]	LSTM	Sepp Hochreiter and Jürgen Schmidhuber	1 million	4 MB
[11]	GRU	Kyunghyun Cho et al.	1 million	4 MB
[19], [11]	BiLSTM	Schuster and Paliwal	2 million	8 MB
[25]	Vicuna	LMSYS	13 billion	26 GB
[25]	Wizard Vicuna	Open Assistant team	13 billion	26 GB
[25]	RedPajama-Chat	Together Computer	7 billion	14 GB
[25]	Aplaca v1	Stanford University	7 billion	14 GB
[25]	Aplaca v2	Stanford University	7 billion	14 GB
[25]	Med Alpaca	Stanford University	13 billion	26 GB
[25]	GPT 4 x Alpaca	OpenAI + Stanford University	7 billion	14 GB
[25]	FastChat - T5	Salesforce Research team	3 billion	6 GB
[25]	Flan T5 xl	Google Research	3 billion	6 GB
[25]	LexPodLM (64bits)	unknown	13 billion	26 GB

intensive assessment of various embedding model-LLM pairs and find that Wizard Vicuna, a 13 billion parameter LLM, gives the highest result combined with Sentence-Transformers. It investigates RAG, which was better than fine-tuning domain-specific data, suffering from model hallucinations. The evaluation was challenging, for such a small dataset and used GPT-4 as probably flawed ground truth. However, this work underlines the potential of LLMs to improve clinical information retrieval [25].

To sum up the previous work, a comparison of all the related research was made based on the used models & their owners, the model sizes and parameters are shown in Table I Additionally, a table that provides a comprehensive summary of the papers, their respective datasets, descriptions, and licenses, detailing key resources and their associated licensing information is available in Table II.

III. METHODOLOGY

In this section, we will cover the details of the dataset, data pre-processing, the use of Large Language Models in speech-to-text and text generation, and an overview of the system's architecture.

A. Dataset Description

To implement the model, we used an Arabic Medical QA dataset available on Kaggle [3]; the dataset contains 87,930 Arabic divided into training, testing, and validation. Figure 2 shows that the dataset is divided into 12 medical labels. The dataset was collected from various trusted medical websites for research and development in Arabic Medical NLP.

B. Data Pre-processing

Multiple reprocessing steps are done for cleaning the dataset and preparing it for testing. We removed the rows that

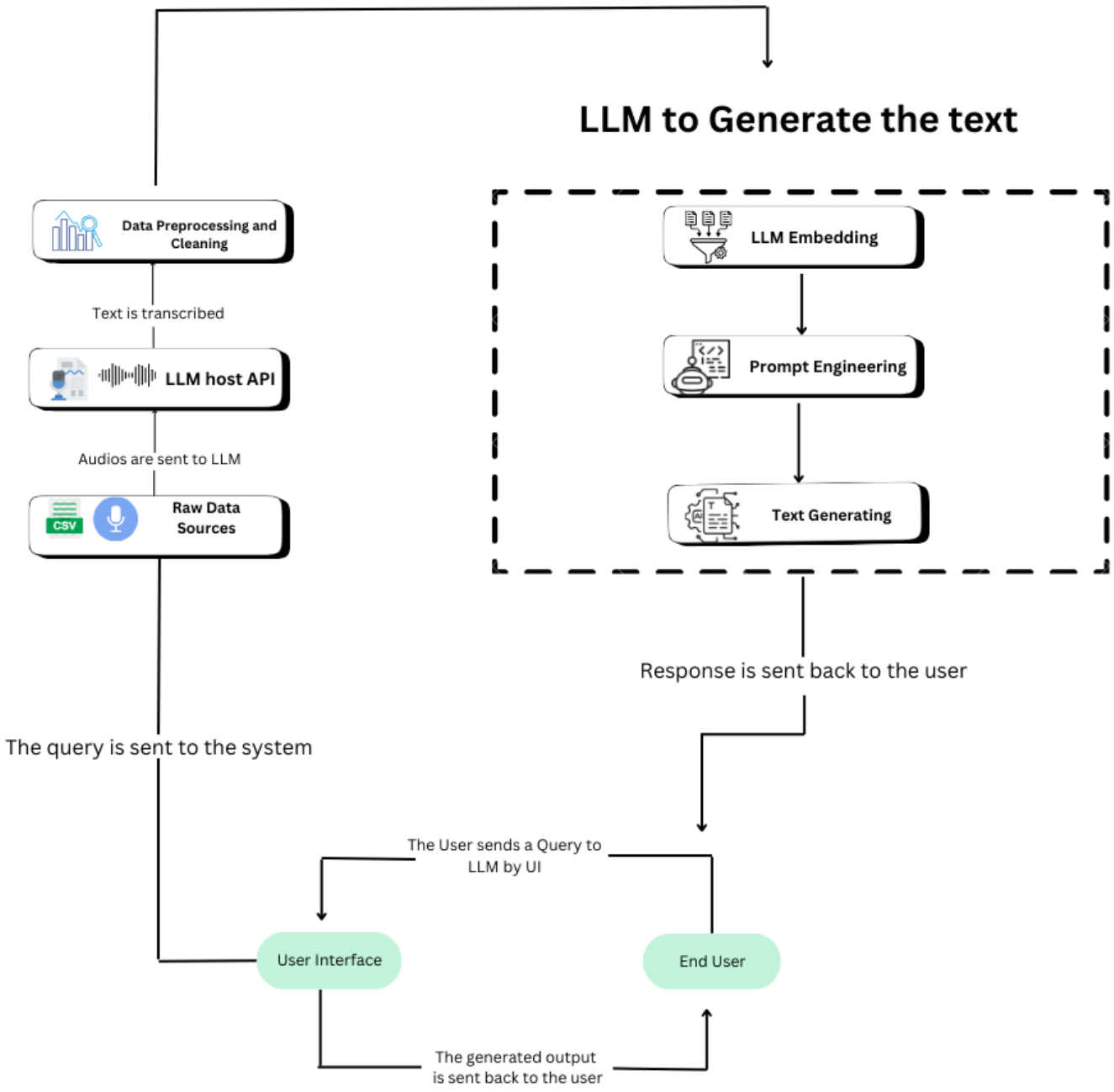

Fig. 1. The full architecture of the proposed model pipeline

contained missing answers from the training, testing, and validation datasets. Normalizing labels is done by replacing underscores and hyphens with spaces, then we standardized specific Arabic letters and removed extra whitespace. Leading/trailing whitespace is stripped in questions and answers, replacing newline characters with a space. We also removed dates and trailing digits from answers. There are several ways to pre-process the text, such as eliminating punctuation, inserting space around punctuation, removing stop-words, lemmatizing words, and normalizing Arabic text to standardize the variations of letters and remove diacritics. These systematic cleaning processes ensure the text data's homogeneity and prepare it for testing.

C. *Speech-to-text translation*

The system's design to process audio (input to the model) by converting the speech into the corresponding text, allowing users to submit their audio, which is then transcribed into written form. To achieve this, a large language model was integrated into the system by its API, enabling the system

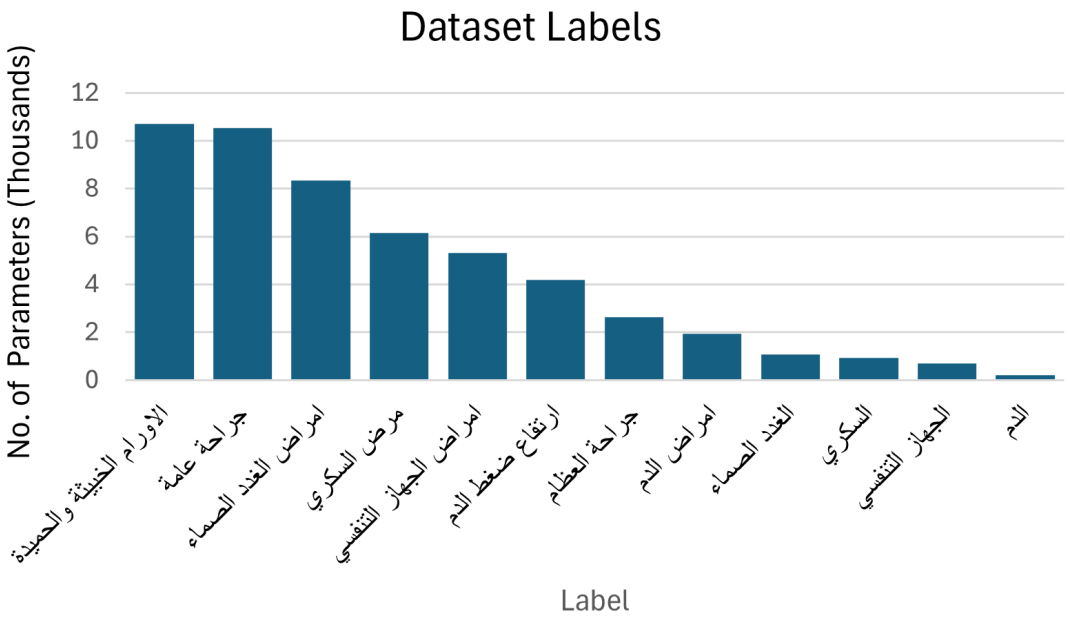

Fig. 2. This figure represents the labels of the dataset with its number of occurrences.

TABLE II
SUMMARY OF PAPERS, DATASETS, AND THEIR DESCRIPTIONS WITH LICENSES.

Research	Dataset	Description	License
[20]	ApolloCorpora multilingual medical dataset	The Apollo dataset includes medical books, encyclopedias, clinical guidelines, papers, exam questions, doctor-patient dialogues, and online medical content, supplemented by mathematical reasoning and coding tasks	Public
[11]	DZMedicaldata	2150 pairs of questions and answers	Public
[13]	VQA-Med 2019	4,200 images from the MedPix database with 15,992 corresponding question and answer pairs	Public
[22]	Uses admission notes from the MIMIC-III dataset, specifically a subset released as part of the TREC 2016 Clinical Decision Support track	Patient-specific EHR data	Restricted Access
[18]	MedMCQA	194k high-quality medical domain MCQs covering 2.4k healthcare topics and 21 medical subjects	Public
[14]	BiQA	7,453 questions and 14,239 question-article pairs from biomedical Q&A forums	Restricted Access
[5]	MAQA	430,000 questions across 20 medical specializations	Public
[1]	MedQuAD	47,457 question-answer pairs from reliable medical sources	Public
[26]	Not mentioned	Review paper on LLMs in the medical domain	—
[28]	Multiple datasets	61 publicly available datasets and 98 testing setups to evaluate models	—
[24]	MASH-QA	5,574 contexts and 34,808 QA pairs (single-span and multi-span subsets)	Public
[40]	No Name	Questions and answers based on patient interactions in healthcare institutions	—
[17]	Not mentioned	Dataset details not provided	—
[25]	MIMIC-IV	Information from over 40,000 ICU patients, specifically clinical notes (MIMIC-IV-Note)	Public
[23]	MedQA	Medical questions sourced from USMLE	Public

to perform speech-to-text conversion by processing the user's audio.

SeamlessM4T (Massively Multilingual and Multi-Model Machine Translation) [42], developed by Meta, is a multilingual, multitask model for major natural language tasks, such as translation and speech recognition. It supports various tasks, including speech-to-text, speech-to-speech, text-to-speech, text-to-text, and Automatic Speech Recognition (ASR) in nearly 100 languages. Available in v2-large version, the model aims to enable smooth communication across languages by integrating transcription, translation, and voice synthesis within a unified framework. With its ability to handle written and spoken language processing tasks, seamless v2 was integrated into our system to cover the first part of the question-answering process.

D. Answer Generating

Upon the transcription of audio into text, the data is subsequently transferred to the next model for the generation of accurate and reliable answers. This model employs its natural language understanding (NLU) mechanisms to interpret the responses, which involves a thorough analysis of the text and the extraction of contextual meaning. In this study, multiple large language models were implemented to assess and compare their performance on the dataset.

1) SILMA: The first model is SILMA AI [43], a large language model that outperformed 72-billion-parameter models in most Arabic language tasks, rendering it highly suitable for business applications. It is built upon the robust foundational models of Google Gemma [44] to achieve exceptional performance. Additionally, SILMA operates as an open-weight model, available as an open-source model, enhancing flexibility for diverse projects. The quantized 4-bit version was used on our data.

2) Phi: The last LLM applied was Phi [45]. Phi-3.5-mini is a lightweight, state-of-the-art model developed using datasets from the Phi-3 framework, which includes synthetic data and curated publicly available websites. It emphasizes high-quality, reasoning-rich content and supports a long context as part of the Phi-3 family. Phi-3.5-mini underwent a comprehensive optimization process to enhance its performance, including supervised fine-tuning, proximal policy optimization, and direct preference optimizations.

E. Prompt Engineering

Various approaches are needed to enhance Large Language Models' performance and extend their capabilities; one of them is prompt engineering. Prompt engineering is a task-specific method to improve results without adjusting the model or modifying any layers. This instruction improves the LLM's performance by focusing exclusively on the given task [46]. There are several techniques to apply prompt engineering to the model, such as the Chain of Thoughts (CoT). Chain of Thoughts was first introduced by Wei et al. [47], which is a technique to provide a detailed process, in step-by-step instructions, to overcome the issue of complex reasoning. For both large language models applied in this paper, SILMA AI and Phi-3, specific, detailed prompt engineering was performed to enhance the model's capabilities and improve its score.

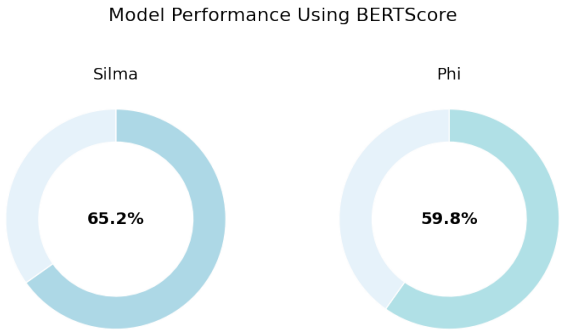

Fig. 3. This illustrates the Model Evaluation Results Based on BertScore

F. Model Architecture

The system's architecture starts from the user interface, where the audios are sent to the model. The text output is also returned to the user after it is generated in Figure 1.

IV. Results and Discussions

This section will cover the details of the evaluation metric used and the model's performance on the LLMs used.

1) Evaluation Metric: BERTScore is an automatic evaluation metric used for text generation. It computes the similarity score for each token in the generated sentence with each token in the reference. Rather than the exact match, the token similarity is calculated using the contextual embeddings [4]. BERTScore uses the pre-trained contextual embeddings from BERT, which will match the generated response and the reference response words using cosine similarity. Additionally, BERTScore calculates the precision, recall, and F1-measure, making it very useful when evaluating different language generation tasks [48].

2) Performance Analysis: The quality of the generated responses for the models is measured using BERTScore by comparing generated responses against reference responses in the Arabic medical Q&A dataset. Results for each question, using SILMA and Phi to generate predictions, are saved. The analysis reveals an average BERTScore of 65.2% and 59.8% for SILMA and Phi, respectively, showing that there is fairly good agreement between outputs from the model with the reference answers. This underlines the efficiency of SILMA and Phi approaches in generating responses that have coherence and relevance to the context. The average BERTScores for the models are as follows: SILMA at 65.2% and Phi at 59.8% as shown in Figure 3.

V. Conclusion

To summarize, there are limitations in the Question-answer systems regarding the Arabic language and Arabic datasets. Natural language's main tasks, including understanding natural language, such as speech-to-text and question-answering

systems, have many challenges. The challenges addressed are mainly due to the lack of available resources in Arabic, especially Egyptian text and speech. This study aims to fill this gap by providing a question-answering system for the medical field in the Egyptian dialectic. The architecture starts by sending the questions as audio to the user interface, transcribed into text by the SeamlessM4T v2 LLM. The next step is to apply several text-generating Large Language Models to send the output back to the user. Applying SILMA and Phi-3, the models achieved a BERTScore of 65.2% and 59.8%, respectively. Future work will optimize our generative LLMs further to send more accurate responses and provide the system for use in the medical field.

REFERENCES

[1] F. Gobet and P. Lane, "Chunking mechanisms and learning," *Encyclopedia of the sciences of learning*, pp. 541–544, 01 2012.

[2] A. Rayhan, R. Kinzler, and R. Rayhan, "Natural language processing: transforming how machines understand human language (2023)," *DOI: https://doi. org/10.13140/RG*, vol. 2, no. 34900.99200.

[3] Y. Abdulmahdi, "Arabic medical qa dataset," Dec 2023.

[4] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

[5] M. Abdelhay, A. Mohammed, and H. A. Hefny, "Deep learning for arabic healthcare: MedicalBot," *Soc. Netw. Anal. Min.*, vol. 13, p. 71, Apr. 2023.

[6] M. Abdelhay, "MAQA: Medical Arabic QA Dataset," 2022.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (P. Isabelle, E. Charniak, and D. Lin, eds.), (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.

[8] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," 2014.

[9] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," 2015.

[10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014.

[11] A. Boulesnane, Y. Saidi, O. Kamel, M. M. Bouhamed, and R. Mennour, "Dzchatbot: A medical assistant chatbot in the algerian arabic dialect using seq2seq model," in *2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pp. 1–8, 2022.

[12] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, "Medical visual question answering via conditional reasoning," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2345–2354, 2020.

[13] L. Canepa, S. Singh, and A. Sowmya, "Visual question answering in the medical domain," 2023.

[14] A. Lamurias, D. Sousa, and F. M. Couto, "Generating biomedical question answering corpora from qa forums," *IEEE Access*, vol. 8, pp. 161042–161051, 2020.

[15] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," 2019.

[16] A. Pampari, P. Raghavan, J. Liang, and J. Peng, "emrqa: A large corpus for question answering on electronic medical records," *arXiv preprint arXiv:1809.00732*, 2018.

[17] A. Soufyane, B. A. Abdelhakim, and M. B. Ahmed, "An intelligent chatbot using nlp and tf-idf algorithm for text understanding applied to the medical field," in *Emerging Trends in ICT for Sustainable Development* (M. Ben Ahmed, S. Mellouli, L. Braganca, B. Anouar Abdelhakim, and K. A. Bernadetta, eds.), (Cham), pp. 3–10, Springer International Publishing, 2021.

[18] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering," 2022.

[19] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "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)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 2567–2577, Association for Computational Linguistics, Nov. 2019.

[20] X. Wang, N. Chen, J. Chen, Y. Wang, G. Zhen, C. Zhang, X. Wu, Y. Hu, A. Gao, X. Wan, H. Li, and B. Wang, "Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people," 2024.

[21] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," 2024.

[22] A. Hamidi and K. Roberts, "Evaluation of ai chatbots for patient-specific ehr questions," 2023.

[23] G. Xiong, Q. Jin, X. Wang, M. Zhang, Z. Lu, and A. Zhang, "Improving retrieval-augmented generation in medicine with iterative follow-up questions," 2024.

[24] M. Zhu, A. Ahuja, D.-C. Juan, W. Wei, and C. K. Reddy, "Question answering with long multiple-span answers," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 3840–3849, Association for Computational Linguistics, Nov. 2020.

[25] R. Elgedawy, I. Danciu, M. Mahbub, and S. Srinivasan, "Dynamic qa of clinical documents with large language models," 2024.

[26] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," *Informatics*, vol. 11, no. 3, 2024.

[27] X. Wang, N. Chen, J. Chen, Y. Hu, Y. Wang, X. Wu, A. Gao, X. Wan, H. Li, and B. Wang, "Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people," 2024.

[28] A. Abdelali, H. Mubarak, S. Chowdhury, M. Hasanain, B. Mousi, S. Boughorbel, S. Abdaljalil, Y. El Kheir, D. Izham, F. Dalvi, M. Hawasly, N. Nazar, Y. Elshahawy, A. Ali, N. Durrani, N. Milic-Frayling, and F. Alam, "LAraBench: Benchmarking Arabic AI with large language models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (Y. Graham and M. Purver, eds.), (St. Julian's, Malta), pp. 487–520, Association for Computational Linguistics, Mar. 2024.

[29] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Proceedings of the Conference on Health, Inference, and Learning* (G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, eds.), vol. 174 of *Proceedings of Machine Learning Research*, pp. 248–260, PMLR, 07–08 Apr 2022.

[30] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," 2019.

[31] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[32] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," 2020.

[33] S. Garg, T. Vu, and A. Moschitti, "TANDA: transfer and adapt pre-trained transformer models for answer sentence selection," *CoRR*, vol. abs/1911.04118, 2019.

[34] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[36] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2020.

[37] M. Zhu, A. Ahuja, D.-C. Juan, W. Wei, and C. Reddy, "Question answering with long multiple-span answers," pp. 3840–3849, 01 2020.

[38] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinform.*, vol. 20, no. 1, pp. 511:1–511:23, 2019.

[39] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinformatics*, vol. 20, Oct. 2019.

[40] J. Krolik, H. Mahal, F. Ahmad, G. Trivedi, and B. Saket, "Towards leveraging large language models for automated medical qa evaluation," 2024.
[41] A. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. Pollard, S. Hao, B. Moody, B. Gow, L.-w. Lehman, L. Celi, and R. Mark, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, p. 1, 01 2023.
[42] S. Communication, L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim, J. Hoffman, M.-J. Hwang, H. Inaguma, C. Klaiber, I. Kulikov, P. Li, D. Licht, J. Maillard, R. Mavlyutov, A. Rakotoarison, K. R. Sadagopan, A. Ramakrishnan, T. Tran, G. Wenzek, Y. Yang, E. Ye, I. Evtimov, P. Fernandez, C. Gao, P. Hansanti, E. Kalbassi, A. Kallet, A. Kozhevnikov, G. M. Gonzalez, R. S. Roman, C. Touret, C. Wong, C. Wood, B. Yu, P. Andrews, C. Balioglu, P.-J. Chen, M. R. Costa-jussà, M. Elbayad, H. Gong, F. Guzmán, K. Heffernan, S. Jain, J. Kao, A. Lee, X. Ma, A. Mourachko, B. Peloquin, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, A. Sun, P. Tomasello, C. Wang, J. Wang, S. Wang, and M. Williamson, "Seamless: Multilingual expressive and streaming speech translation," 2023.
[43] S. Team, "Silma," 2024.
[44] J. Banks and T. Warkentin, "Gemma: Introducing new state-of-the-art open models," 2024.
[45] Microsoft, "Phi-3 mini-128k-instruct," 2024.
[46] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," 2024.
[47] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.
[48] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.