# BiMediX: Bilingual Medical Mixture of Experts LLM

**Sara Pieri**[1*], **Sahal Shaji Mullappilly**[1*], **Fahad Shahbaz Khan**[1,2], **Rao Muhammad Anwer**[1], **Salman Khan**[1], **Timothy Baldwin**[1], **Hisham Cholakkal**[1]

[1]Mohamed bin Zayed University of Artificial Intelligence [2]Linköping University

## Abstract

In this paper, we introduce BiMediX, the first bilingual medical mixture of experts LLM designed for seamless interaction in both English and Arabic. Our model facilitates a wide range of medical interactions in English and Arabic, including multi-turn chats to inquire about additional details such as patient symptoms and medical history, multiple-choice question answering, and open-ended question answering. We propose a semi-automated English-to-Arabic translation pipeline with human refinement to ensure high-quality translations. We also introduce a comprehensive evaluation benchmark for Arabic medical LLMs. Furthermore, we introduce BiMed1.3M, an extensive Arabic-English bilingual instruction set that covers 1.3 Million diverse medical interactions, including 200k synthesized multi-turn doctor-patient chats, in a 1:2 Arabic-to-English ratio. Our model outperforms state-of-the-art Med42 and Meditron by average absolute gains of 2.5% and 4.1%, respectively, computed across multiple medical evaluation benchmarks in English, while operating at 8-times faster inference. Moreover, our BiMediX outperforms the generic Arabic-English bilingual LLM, Jais-30B, by average absolute gains of 10% on our Arabic and 15% on our bilingual evaluations across multiple datasets. Additionally, BiMediX exceeds the accuracy of GPT4 by 4.4% in open-ended question UPHILL evaluation and largely outperforms state-of-the-art open source medical LLMs in human evaluations of multi-turn conversations. Our trained models, instruction set, and source code are available at https://github.com/mbzuai-oryx/BiMediX.


Figure 1: BiMediX is adept at conducting precise medical conversations seamlessly in both English and Arabic. It demonstrates advanced interactive capabilities, such as soliciting further details from the patient and seamlessly switch languages based on patient input while maintaining the continuity of the conversation. The Arabic text translates as follows: **User:** *The pain occurs mostly on both sides of the head, sometimes on only one side. It is a throbbing pain that can last for several hours. Light and noise seem to make it worse.* **BiMediX:** *Thank you for the detailed information. Based on your symptoms, it sounds like you may have a migraine. Migraine headaches are usually characterized by a throbbing pain on one or both sides of the head, and can be triggered or increased in intensity by light and noise. It is important to consider that there are other possible causes for your symptoms, so we may need to do further evaluation. Have you noticed any specific factors that trigger these headaches?*

## 1 Introduction

Large Language Models (LLMs) demonstrate excellent ability to understand instructions, generate human-like text, and tackle unseen or complex tasks, rendering them valuable for a broad range of applications. Specifically, LLMs have the potential to impact the medical field by offering virtual chat support across diverse medical departments to enhance diagnostic accuracy. However, despite the promise of LLMs in healthcare, the medical

*Equally contributing first authors.



*Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002
November 12-16, 2024 ©2024 Association for Computational Linguistics

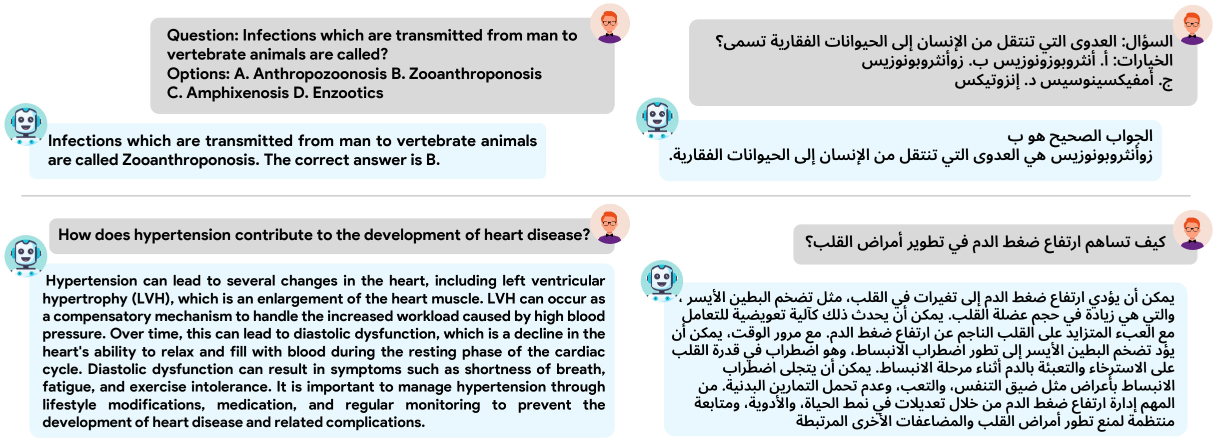


Figure 2: English, Arabic language capabilities of our BiMediX with multiple-choice questions and answers (MCQA) on the top row, and open-ended questions and responses (QA) on the bottom row.

domain poses distinct challenges, including the necessity for domain-specific data and expertise to ensure accurate interpretations, as well as concerns regarding transparency and bias.

Recent methods have emerged to bridge the gap in medical domain knowledge by developing open-source medical LLM alternatives (Chen et al., 2023; Toma et al., 2023; Wu et al., 2023) to proprietary models like ChatGPT (Achiam et al., 2023). However, among the leading medical LLMs in English, Med42-70B (Christophe et al., 2023) has not made its training data and resources available. Meditron-70B (Chen et al., 2023), relies on separate fine-tuning for each evaluation dataset. This approach is likely to compromise the model's ability to serve as a unified solution with diverse interaction capabilities. Moreover, many of the leading open-source medical LLMs (Christophe et al., 2023; Chen et al., 2023) are limited to the English language, posing limitations in their ability to understand and interact in resource-constrained languages such as Arabic. Despite Arabic's potential to cater to a population of more than 400 million people, it remains underrepresented in the medical LLM literature. The unique features of Arabic, such as its distinctive script and less conventional right-to-left writing format, the unavailability of large-scale medical training data in Arabic, and the lack of a comprehensive benchmark to evaluate Arabic medical LLMs, present challenges in the development of Arabic or English-Arabic bilingual medical LLMs.

In this work, we introduce BiMediX, a bilingual medical mixture of experts LLM with seamless conversational capabilities in both English and Arabic. Our BiMediX is a chat agent with state-of-the-art performances on non-chat, Open QA and multi-turn chat evaluations.

### 1.1 Contributions

Our key contributions can be summarized as:
**(i)** We introduce the ***first bilingual medical mixture of experts LLM***, named ***BiMediX***, having seamless interaction capabilities in both English and Arabic languages (see Fig.1). Our model facilitates various medical interactions, including multi-turn chats essential for follow-up inquiries with human patients (Fig.1), multiple-choice question answering (Fig.2 top row) and open-ended question answering (Fig.2 bottom row).
**(ii)** We develop a ***semi-automated iterative translation pipeline***, incorporating a human verification step to ensure high-quality translation of English medical text into resource-constrained Arabic. This pipeline facilitates the compilation of instruction-tuning dataset and a comprehensive benchmark for evaluating Arabic healthcare LLMs and Arabic-English Bilingual LLMs.
**(iii)** We curate a comprehensive ***Arabic-English bilingual instruction set*** named ***BiMed1.3M*** comprising over 1.3 million instructions, resulting in a total of over 632 million healthcare specialized tokens. Our dataset comprises open-ended question-and-answer, multiple-choice question answering, and over 200k synthesized multi-turn chats rooted in authentic medical content enabling follow-up inquiries with human patients.
**(iv)** We introduce a ***parameter-efficient fine-tuning of routing and expert layers*** in Mixtral (Jiang et al., 2024) using BiMed1.3M, requiring fewer training resources compared to Med42 and Chen et al. (Christophe et al., 2023; Chen et al., 2023).

Our BiMediX achieves ***state-of-the-art performance*** on multiple medical exam question datasets ***in both English and Arabic***. Our model outperforms Med42 and Meditron across multiple MCQA, Open-ended QA and multi-turn conversation medical evalautions in English, while operating 8 times faster. Additionally, BiMediX surpasses Jais-30B, by average absolute gains of 10% and 15% on Arabic and bilingual evaluations across multiple medical datasets. Moreover, bilingual instruction tuning with our BiMed1.3M leads to an average absolute 10% gain over the base Mixtral (Jiang et al., 2024) on bilingual evaluations.

## 2 Related Works

Early BERT-style open medical language models leveraged the PubMed corpus for either continuous pre-training (Lee et al., 2020; Huang et al., 2019) or specific domain training from scratch (Beltagy et al., 2019; Gu et al., 2021; Shin et al., 2020). Subsequent research enhanced bidirectional systems to incorporate additional knowledge (Yasunaga et al., 2022b,a), while later studies have tailored GPT-2 based models to medical and scientific literature (Bolton et al.; Luo et al., 2022; Taylor et al., 2022). Recent works such as MedAlpaca (Han et al., 2023) focus on finetuning large scale open-source LLMs such as LLaMA (Touvron et al., 2023a,b) on a set of medical instructions. ChatDoctor (Yunxiang et al., 2023) enhanced a LLaMA model on 100K clinical Q&As, incorporating knowledge retrieval capabilities. ClinicalCamel (Toma et al., 2023) introduced question answering data, converting PubMed articles and MedQA into questions and descriptive answers. Meditron (Chen et al., 2023) performed pre-training on PubMed content and medical texts, with further refinements on individual MCQA datasets. Med42 (Christophe et al., 2023) instruction-tuned LLaMA model for medical tasks, though the details of its training remain undisclosed. Additional details on related English medical LLMs are provided in Appendix B.

Recent years have witnessed progress in Arabic language processing models (Nagoudi et al., 2021; Eddine et al., 2022). In addition to these monolingual models, Arabic has been integrated into multilingual frameworks (Scao et al., 2022; Muennighoff et al., 2022; Qiu et al., 2024; Wang et al., 2024). The recent launch of the Jais model (Sengupta et al., 2023) marked a substantial advancement as a general-purpose bilingual LLM for English and Arabic. Additionally, many of the latest foundational models (Touvron et al., 2023a,b; Jiang et al., 2024) demonstrate the ability to reply in Arabic, despite their limited exposure to the language during pre-training. However, these generic models lack sufficient medical domain knowledge, making them ill-equipped for medical applications in both Arabic and English languages.

## 3 Method

**Overview**: To develop a bilingual healthcare chat assistant, we first introduce a comprehensive bilingual dataset, BiMed1.3M, that encompasses diverse medical interactions in both English and Arabic (Sec. 3.1). This bilingual dataset is created by first compiling English instruction, as outlined in Sec. 3.1.1. Subsequently, in Sec. 3.1.2, we describe our semi-automated English-to-Arabic translation pipeline, which enables us to obtain high-quality Arabic medical benchmarks and instruction sets, as detailed in Sec. 3.1.3. The generated Arabic instructions are then combined with our English instructions to obtain the BiMed1.3M bilingual dataset. Finally, we present the model and instruction-tuning strategy in Sec. 3.2.

### 3.1 BiMed1.3M: Bilingual Dataset with Diverse Medical Interactions

The growing need for AI-driven medical assistants, proficient across a spectrum of NLP tasks, highlights the necessity for comprehensive datasets. In particular, the ability to deliver concise answers is critical for evaluating model performance, yet the functionality for engaging in chat is essential for practical deployment. This holds particular significance in the healthcare domain, where multiple turns of interaction with further inquiries about symptoms, examinations, and pre-existing conditions are often necessary. To this end, we compile our English instruction set as detailed below.

#### 3.1.1 Compiling English Instruction Set

At first, we compile a dataset in English encompassing three types of medical interactions: multiple-choice question answering (MCQA), which focuses on specialized medical knowledge and reasoning for definitive answers; open question answering (QA), that includes real-world consumer questions; and multi-turn chat conversations for dynamic exchanges. For the first two categories, we combined various existing sources into a unified collection of question-answer pairs as

| Data | Samples | Avg. Turns | #Tokens |
|---|---|---|---|
| QA | 423.8 K | 1.00 | 131.8 K |
| MCQA | 638.1 K | 1.00 | 342.5 M |
| Chat | 249.7 K | 4.72 | 158.0 M |
| Total | 1311.6 K | 1.71 | 632.3 M |

Table 1: Statistics of the BiMed1.3M across QA, MCQA, and Chat (with more than one turn of exchanges) totaling 623M tokens and 1.3M samples.

| Length Range | Count |
|---|---|
| L0 - 1024 | 1,167,791 |
| L1024 - 2048 | 135,677 |
| L2048 - 8192 | 8,113 |
| **Metric** | **Value** |
| Average Prompt Length | 147.71 |
| Average Response Length | 134.56 |

Table 2: BiMed1.3M Dataset Overview: Length Distribution and Length Counts.

detailed in Table 8 in Appendix. Specifically, for Multiple-Choice QA, we included data from PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), and MedQA (Jin et al., 2021). For Open QA, we utilized sources such as HealthCareMagic (Yunxiang et al., 2023), iCliniq (Yunxiang et al., 2023), Medical Meadow (Han et al., 2023), UMLS (Wu et al., 2023), LiveQA (Abacha et al., 2017), and MedicationQA (Abacha et al., 2019). Note that these datasets are publicly available and have been previously anonymized. For the multi-turn chat component, we generated realistic conversations grounded on MCQA scenarios as described below.

**MCQA-grounded Multi-turn Chat Generation:** State-of-the-art open-source medical LLMs such as (Chen et al., 2023) generally struggle to support in-depth, multi-turn conversations that include symptom inquiries and further information requests. To address this, we combine the conversational capabilities of ChatGPT with medically accurate MCQAs (multiple-choice question answers) sourced from publicly available datasets. We use ChatGPT to craft realistic doctor-patient dialogues based on MCQAs. The chat creation process involves using ChatGPT to draft messages for both the user and the AI assistant in a dialogue format. Inputs from the PubmedQA, MedQA, and MedMCQA datasets are provided with specific directives that ensure discussions are relevant to the medical question-answer pairs provided. A prompt template is employed to define the format and criteria alongside the input question, generating conversation transcripts for both parties (doctor and patient) until a logical conclusion is reached. To assess the dialogues' relevance and quality, we conducted manual reviews with medical professionals, refining our prompts based on their feedback (see Appendix C). Based on this approach, we produced more than 200,000 high-quality multi-turn medical dialogues, each linked to a specific MCQA, collectively comprising over 74 million tokens.

Following the generation of multi-turn dialogues, our dataset comprehensively encompasses multiple-choice question answering (MCQA), open question answering (QA), and dynamic chat conversations (Chats)—totaling over 860,000 instruction-tuning data in English. Moreover, we leverage various publicly available medical benchmark datasets in English to evaluate the performance of our model, as elaborated in Sec. 3.1.3.

Next, we detail our semi-automated iterative translation pipeline to generate medical evaluation benchmarks and instruction sets tailored for the resource-constrained Arabic language.

#### 3.1.2 Semi-Automated Iterative Translation

Traditional methods like Google Translate or direct usage of large language models often fail to meet the high accuracy required for medical texts, while hiring professional translators is costly and slow, limiting scalability. We have developed a semi-automated translation pipeline that combines machine translation with human oversight, optimizing both scale and quality, as illustrated in Fig. 3.

At first, we perform English to Arabic translation using generic LLM (eg. ChatGPT) by considering the entire input English text as context. Subsequently, we task LLM with evaluating the quality of the translated Arabic text relative to its original English version, with scores from 0 to 100 reflecting translation fidelity and are verified to align to human preferences by native Arabic speakers.

**Realistic translations with selective review:** Scores below a predefined threshold trigger a refinement process where LLM iteratively adjusts the translations based on the original English text, the current translations, and their scores as feedback, enhancing accuracy and consistency. We set a higher threshold (eg. 85/100) for the test set and a relatively lower threshold for the training set. We observe that only 8% of the translated benchmark text falls below the threshold in the initial translation, indicating the effectiveness of GPT-based translation. Translations below the thresholds are refined iteratively, with the majority reaching high

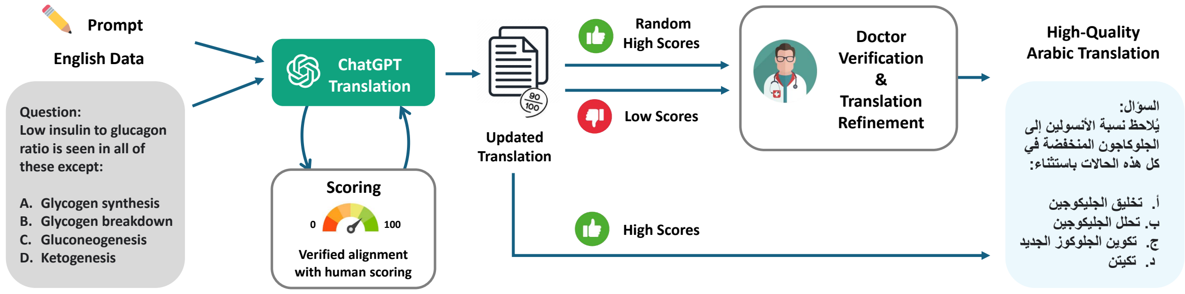


Figure 3: Overview of the proposed semi-automated, iterative translation pipeline featuring human alignment. The process involves iterative translation and score optimization utilizing LLMs (such as ChatGPT) and manual refinement by professionals for samples with low scores along with a random selection of high-scoring samples.

scores within 1 to 2 cycles of iterative refinement. On average, translations required 1.4 iterations, with more than 98% of initially low-scoring data surpassing the threshold after refinement, substantially minimizing manual effort.

Translations persistently scoring below threshold go thorough *manual verification process performed by a medical professional fluent in Arabic* to ensure they meet the academic and clinical standards required for real medical Q&A scenarios. Approximately 90% of such low-scoring translations were subjected to significant modifications upon expert review, underscoring the indispensability of human expertise in these cases. In contrast, the translations that attained high initial scores were generally found to have superior translation quality during human verification of random samples, demonstrating the pipeline's effectiveness in reducing unnecessary expert reviews and highlighting the cost-saving benefits of our approach. Further details are provided in Appendix C.

#### 3.1.3 Bilingual Benchmark & Instruction Set

**Creation of Medical Benchmark:** To evaluate the accuracy and applicability of Arabic medical AI models, we translated the English medical evaluation benchmarks (detailed in Sec. D) into Arabic using the aforementioned iterative translation procedure. By providing a high-quality Arabic medical benchmark aligned with its English counterpart, we aim to establish a fundamental step in bridging the linguistic divide in model evaluation and comparison, thereby offering a valuable asset for further research.

These Arabic benchmarks are combined with the original English evaluation benchmarks to create an English-Arabic bilingual benchmark. This allows us to assess the linguistic capabilities of our bilingual model as well as its medical domain knowledge.

**Bilingual Instruction Set:** We translated 444,995 English samples into Arabic, covering all three types of medical interactions. Following (Sengupta et al., 2023), we adopted a bilingual approach, mixing Arabic and English in a 1:2 ratio. This approach led to the creation of an extensive bilingual instruction tuning dataset that integrates both languages. Consequently, we compiled an English-Arabic bilingual dataset named BiMed1.3M, 1.5 times larger than our English counterpart, comprising more than 1,311,000 samples. Further details are illustrated in Tab. 1 and 2.

Next we introduce our model and its bilingual medical instruction tuning.

### 3.2 Bilingual Medical Instruction Tuning of Mixture of Experts LLM

Recently, Mixtral (Jiang et al., 2024) introduced a Mixture of Experts (MoE) based architecture that achieve impressive performance on generic NLP benchmarks in English. It enables substantial scaling of model size within the same computational budget as traditional dense models. Unlike traditional dense feed-forward network layers, Mixtral employs a set number of "experts" in a sparse manner, replacing the FFN layers in dense networks. Additionally, a gate network or router is employed to direct input tokens to appropriate experts based on learned parameters.

While Mixtral (Jiang et al., 2024) offer advantages such as improved inference times over dense models, they face challenges in Arabic-English bilingual medical applications due to limited medical domain knowledge and Arabic language capabilities. To address these challenges, we perform Arabic-English bilingual medical instruction tuning of Mixtral MoE architecture using our BiMed1.3M dataset.

Note that performing conventional supervised fine-tuning of Mixtral (Jiang et al., 2024) requires substantial computational resources and often encounters challenges in carefully tuning the expert layers and the routing network leading to generalization issues. To this end, we employ Parameter-Efficient Finetuning (PEFT) techniques (Hu et al., 2021; Li and Liang, 2021; Dettmers et al., 2023) to adapt the pre-trained Mixtral model (Jiang et al., 2024) with minimal computational resources while ensuring high performance and adaptability. In our approach, we use QLoRA-based PEFT training. Specifically, we attach the QLoRA adapters to the decoder layers of (Jiang et al., 2024), including the experts and the routing network, to repurpose the Mixture of Experts architecture as a bilingual medical chat assistant by utilizing the proposed BiMed1.3M dataset for instruction tuning. Unlike alternative approaches such as (Chen et al., 2023), we refrained from conducting additional medical pre-training before instruction-tuning it on our medical dataset.

In summary, through careful instruction tuning of the routing network and experts in the Mixtral decoder using our BiMed1.3M dataset, we achieve enhanced medical domain capabilities in both Arabic and English languages. This includes proficiency in multiple-choice question answering, open-ended question answering, and realistic multi-turn question answering between our BiMediX model and the user (e.g., patient), making it suitable for deployment as a bilingual medical chatbot.

# 4 Experiments

## 4.1 Experimental Settings

Our BiMediX is built upon mixtral (Mixtral-8x7B) (Jiang et al., 2024) base network, a sparse mixture of experts language model. This model extends the Mistral-7B architecture (Jiang et al., 2023), featuring a unique design with each layer hosting eight "expert" feedforward blocks. A router network determines which two experts will process each token, merging their outputs. Consequently, it handles 47 billion parameters, with fewer than 13 billion active during inference. This model features a hidden state dimension of 14,336, a context window capable of accommodating 32,768 tokens, 32 layers, 32 attention heads, and a vocabulary size of 32,000.

For fine-tuning Mixtral, we utilized QLoRA, a low-rank adaptation technique, by incorporating a set of learnable low-rank adapter weights into both the experts and the routing network. This adaptation applies to $q$, $k$, and $v$, setting the rank to 128 and the $\alpha$ parameter to 64. Consequently, this approach results in the training of approximately *4% of the original parameters*. Our bilingual model underwent training on roughly 632 million tokens sourced from the Arabic-English corpus and there are 288 million English tokens to train English model. BiMediX's performance was stable across English benchmarks during the first and second epochs, with only a minor enhancement observed in the Arabic benchmark after the second epoch, possibly due to adaptation to Arabic medical data. Consequently, BiMediX was trained for two epochs using eight A100 (80GB) GPUs, completing the process in just 35 hours. Further details are provided in Appendix G.

## 4.2 Evaluation on Medical MCQ Benchmarks

In the literature, evaluating medical language models predominantly involves multiple-choice question-answering tasks, with accuracy as the performance metric. Our analysis incorporates multiple prominent benchmarks in medical multiple-choice question-answering, including PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022) MedQA (Jin et al., 2021) and medical MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2020). All questions from the above English datasets are translated into Arabic using our semi-automated pipeline to create our Arabic and Bilingual evaluation benchmarks. We employed the EleutherAI evaluation framework[1].

### 4.2.1 Quantitative Comparison

**Bilingual Evaluation:** Here, we evaluate our BiMediX on the Arabic-English bilingual evaluation benchmark, derived from evaluating results in both languages. Tab. 3 shows our BiMediX results against the base model, multilingual medical models (Qiu et al., 2024; Wang et al., 2024), Mixtral-8x7B (not fine-tuned) and Jais-30B. The latter is a larger model designed specifically for the Arabic language and capable of functioning in both English and Arabic. BiMediX demonstrates superior performance across all benchmarks, achieving

[1]https://github.com/EleutherAI/lm-evaluation-harness, is the backend for Hugging Face Open LLM Leaderboard.

| Model | MMLU | | | | | | MedMCQA | MedQA | PubmedQA | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| | Cli-KG | C-Bio | C-Med | Med-Gen | Pro-Med | Ana | | | | |
| MMedLM2 (Qiu et al., 2024) | 54.0 | 46.9 | 48.3 | 47.0 | 50.0 | 49.6 | 40.1 | 35.4 | 60.9 | 48.0 |
| Apollo-7B (Wang et al., 2024) | 33.2 | 33.4 | 30.7 | 28.5 | 43.8 | 27.4 | 26.8 | 26.5 | 57.7 | 34.2 |
| Jais-30B (Sengupta et al., 2023) | 57.4 | 55.2 | 46.2 | 55.0 | 46.0 | 48.9 | 40.2 | 31.0 | 75.5 | 50.6 |
| Mixtral-8x7B (Jiang et al., 2024) | 59.1 | 57.6 | 52.6 | 59.5 | 53.3 | 54.4 | 43.2 | 40.6 | 74.7 | 55.0 |
| GPT-3.5 (OpenAI, 2023b) | 58.5 | 55.8 | 50.3 | 53.0 | 55.1 | 47.6 | 41.0 | 40.4 | 56.8 | 51.0 |
| GPT-4 (Achiam et al., 2023) | 82.4 | 88.8 | 74.6 | 88.0 | 86.8 | 71.0 | 64.0 | 71.9 | 71.1 | 77.6 |
| **BiMediX (Bilingual)** | **70.6** | **72.2** | **59.3** | **74.0** | **64.2** | **59.6** | **55.8** | **54.0** | **78.6** | **65.4** |

Table 3: Performance of BiMediX on the **Bilingual benchmark**.

| Model | MMLU | | | | | | MedMCQA | MedQA | PubmedQA | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| | Cli-KG | C-Bio | C-Med | Med-Gen | Pro-Med | Ana | | | | |
| MMedLM2 (Qiu et al., 2024) | 36.6 | 27.1 | 35.3 | 24.0 | 26.5 | 31.8 | 31.8 | 21.0 | 47.0 | 31.2 |
| Apollo-7B (Wang et al., 2024) | 30.9 | 28.5 | 27.8 | 23.0 | 41.9 | 16.3 | 21.5 | 21.0 | 43.2 | 28.2 |
| Jais-30B (Sengupta et al., 2023) | 52.1 | 50.7 | 40.5 | 49.0 | 39.3 | 43.0 | 37.0 | 28.8 | 74.6 | 46.1 |
| Ours (Arabic) | 60.0 | 54.9 | **55.5** | 58.0 | **58.1** | 49.6 | 46.0 | 40.2 | 76.6 | 55.4 |
| Mixtral-8x7B (Jiang et al., 2024) | 37.4 | 32.6 | 37.6 | 42.0 | 26.8 | 36.3 | 30.2 | 23.6 | 72.2 | 37.6 |
| GPT-3.5 (OpenAI, 2023b) | 47.1 | 39.4 | 39.3 | 36.0 | 40.0 | 38.8 | 32.0 | 30.1 | 53.5 | 39.6 |
| GPT-4 (Achiam et al., 2023) | 78.8 | 82.5 | 72.2 | 85.0 | 80.5 | 61.9 | 58.6 | 64.0 | 67.9 | 72.4 |
| **BiMediX (Bilingual)** | **63.8** | **57.6** | 52.6 | **64.0** | 52.9 | **50.4** | **49.1** | **47.3** | **78.4** | **57.3** |

Table 4: Performance of BiMediX on the **Arabic benchmark**.

| Model | MMLU | | | | | | MedMCQA | MedQA | PubmedQA | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| | Cli-KG | C-Bio | C-Med | Med-Gen | Pro-Med | Ana | | | | |
| MMedLM2 (Qiu et al., 2024) | 71.3 | 66.7 | 61.3 | 70.0 | 73.5 | 67.4 | 48.5 | 49.7 | 74.8 | 64.8 |
| Apollo-7B (Wang et al., 2024) | 35.5 | 38.2 | 33.5 | 34.0 | 45.6 | 38.5 | 32.1 | 31.9 | 72.2 | 40.2 |
| Med42-70B (Christophe et al., 2023) | 75.9 | 84.0 | 69.9 | 83.0 | 78.7 | 64.4 | 61.9 | 61.3 | 77.2 | 72.9 |
| Clinical Camel-70B (Toma et al., 2023) | 69.8 | 79.2 | 67.0 | 69.0 | 71.3 | 62.2 | 47.0 | 53.4 | 74.3 | 65.9 |
| Meditron-70B (Chen et al., 2023) | 72.3 | 82.5 | 62.8 | 77.8 | 77.9 | 62.7 | 65.1 | 60.7 | 80.0 | 71.3 |
| Mixtral-8x7B (Jiang et al., 2024) | 80.8 | 82.6 | 67.6 | 77.0 | 79.8 | 72.6 | 56.2 | 57.7 | 77.2 | 72.4 |
| Vicuna-13B-v1.5 (Chiang et al., 2023) | 57.4 | 54.2 | 50.9 | 55.0 | 54.0 | 50.4 | 37.4 | 36.1 | 73.8 | 52.1 |
| Vicuna-13B-v1.5 (Finetuned with Our dataset) | 63.0 | 60.4 | 54.9 | 64.0 | 65.4 | 58.5 | 50.5 | 43.9 | 76.2 | 59.7 |
| GPT-3.5 (OpenAI, 2023b) | 69.8 | 72.2 | 61.3 | 70.0 | 70.2 | 56.3 | 50.1 | 50.8 | 60.2 | 62.3 |
| GPT-4 (Achiam et al., 2023) | **86.0** | **95.1** | **76.9** | **91.0** | **93.0** | **80.0** | **69.5** | **78.9** | 74.4 | **82.8** |
| **BiMediX** | 78.9 | 86.1 | 68.2 | 85.0 | 80.5 | 74.1 | 62.7 | 62.8 | **80.2** | 75.4 |

Table 5: Performance of BiMediX on the **English benchmark**.

accuracy that is, on average, more than 10 and 15 points higher, respectively, when compared to the baseline model and Jais-30B. This achievement underscores the substantial value of our proposed BiMed1.3M dataset and its unmatched effectiveness and adaptability in addressing medical queries within a bilingual framework among the open-source models. The close-source models GPT-3.5 and GPT-4 are highlighted in Tables [3,4,5].

**Arabic Benchmark:** We evaluate BiMediX using our Arabic benchmark, comparing its performance with that of Jais-30B. In Table 4, we present the findings for Jais and our BiMediX in two configurations: one pre-trained exclusively on Arabic content (Arabic) and the other with bilingual data (Bilingual). Our bilingual model outperforms in all categories within the Arabic context, underscoring that the integration of both types of training data significantly enhances the understanding and processing of medical information in an Arabic setting.

**English Benchmark:** We assess the quantitative performance of our English model against previous state-of-the-art English medical models. As shown in Table 5, BiMediX demonstrates remarkable performance across all subtasks, achieving the highest average scores compared to the evaluated models, with the notable exception of GPT-4. Nonetheless, it consistently surpasses the performance of GPT-3.5 and GPT-4 on PubMedQA.

When compared to Clinical Camel-70B model, our BiMediX exhibits around 10% average performance gain. Furthermore, it also outperforms Meditron-70B in almost every subset except for MedMCQA. Notably, our model's success comes without the need for separate fine-tuning on the training set for each evaluation benchmark, in contrast to Meditron, which employs individual fine-tuning for each evaluation benchmark to achieve favorable outcomes. This demonstrates our model's versatility in handling various medical interactions simultaneously. Moreover, our method outperforms Med42 by an average gain of 2.5%

Furthermore, our model is more efficient in terms of prediction speed (latency and tokens per second) than all other models considered, as detailed in Table 6.

| Model | Active Param | Latency | Tokens/sec |
|---|---|---|---|
| Med42-70B (Christophe et al., 2023) | 70B | 24.5 s | 20.9 |
| Meditron-70B (Chen et al., 2023) | 70B | 24.5 s | 20.9 |
| Jais-30B (Sengupta et al., 2023) | 30B | 14.0 s | 36.5 |
| **BiMediX** | 13B | **2.8 s** | **180.6** |

Table 6: Model statistics comparing memory size and inference speed, computed with (LLm-analysis, 2023) on GPU A100-80GB.

| Metric | InstructGPT | ChatGPT | GPT-4 | Med42 | BiMediX |
|---|---|---|---|---|---|
| Accuracy | 28.8 | 47.2 | 51.5 | 53.5 | 55.9 |

Table 7: Performance comparison on UPHILL OpenQA (Kaur et al., 2023), assessing the model's ability to address false medical claims at different presupposition levels.

#### 4.2.2 Baseline Comparisons

Our BiMedix model, using the bilingual BiMed1.3M instruction set, surpasses the Mixtral 8x7B model by 10.4 percentage points on the bilingual evaluation benchmark (see Table 3). For English benchmarks (Table 5), integrating BiMed1.3M with Mixtral-8x7B significantly boosts performance, with improvements exceeding 6% on the MedMCQA dataset and an average gain of 3% across multiple datasets. Employing Vicuna-13B base architecture and fine-tuning with our dataset further enhances the gain by over 13% on MedMCQA, averaging a 7% increase. These results demonstrate BiMed1.3M's broad utility and impact across various base models. Moreover, the addition of MCQ-grounded multi-turn chat data in BiMedix improves average English medical benchmark scores from 74.7 to 75.4, illustrating the value of chat data in boosting overall model performance. In Arabic benchmarks (Table 4), our fine-tuned model outperforms baseline Mixtral-8x7B by over 19 percentage points, underscoring BiMed1.3M's effectiveness.

### 4.3 Open-ended QA Performance Assessment

We evaluate our model on the UPHILL dataset (Kaur et al., 2023), which is designed to evaluate the factual accuracy of LLMs in handling health-related queries in English embedded with varying degrees of presuppositions. Our analysis specifically targeted the zero-shot models' abilities to process and respond to false claims (factually incorrect or debunked by reliable sources) across a spectrum of presupposition levels, a crucial test given the potential real-world implications of misinformation and inaccuracies in the health domain. In this context, the accuracy refers to the model's effectiveness in accurately refuting false health-related claims at different presupposition levels. As shown in Table 7, our model achieves an accuracy of 55.85%, surpassing InstructGPT (Ouyang et al., 2022) (28.8%), Med42 (Christophe et al., 2023) (53.5), ChatGPT (OpenAI, 2023a) (47.2%), and GPT-4 (Achiam et al., 2023) (51.5%), highlighting its superior ability to refute false health-related claims and in combating misinformation in OpenQA settings.

### 4.4 Evaluation of Multi-turn Medical Conversations

We compare the responses generated by BiMedix with Med42, Meditron-70B, GPT-4 and baseline Mixtral 8x7B. All models were prompted with the same initial query and similar follow-up user responses, based on the model's response in different medical specialties. The entire multi-turn conversation between the model and user was documented with anonymized model identifiers and given to LLMs and human Doctors for independent evaluations. They were tasked to select the best model response based on medical knowledge and facts, accuracy of diagnosis, appropriate leading questions, and quality. The results from these experiments in English quantitatively show that BiMediX was preferred in more than 75% of the samples in both GPT-4 and Gemini evaluations. Similarly, during human evaluation by three certified medical doctors, the multi-turn conversation with our model was preferred in more than 80% of samples. Additional qualitative examples are provided in the Appendix.

## 5 Conclusion

Our work introduces the first bilingual medical mixture of experts (LLM) BiMediX designed to facilitate comprehensive medical interactions, including conversations, MCQA, and QA in both English and Arabic languages. The implementation of a semi-automated translation pipeline, coupled with human refinement, ensures the high quality of English-to-Arabic translations. BiMediX demonstrates superior performance over existing models such as Med42 and Meditron in English-based medical evaluation benchmarks, open-qa, and conversations and significantly outperforms the generic bilingual LLM, Jais-30B, in Arabic medical and bilingual evaluations. The introduction of a novel evaluation benchmark and the extensive BiMed1.3M instruction set, encompassing 1.3 million diverse medical interactions, further highlights the significance of our contributions.

## 6 Limitations

Despite its overall improvement, BiMediX, like other language models, may experience issues such as hallucinations, toxicity, and stereotypes due to inherited limitations from base models and pre-training data. Our paper serves as a first step toward developing a medical LLM tailored to the resource-constrained Arabic language. We understand the importance of ensuring both cultural sensitivity and accuracy in medical data, which can sometimes be challenging to capture fully in translations. While our translation pipeline works to address these concerns, there remains considerable room for further refinement. We have rigorously tested BiMediX for factual accuracy through both automated assessments and qualitative evaluations conducted with medical professionals. Nonetheless, our model's medical diagnoses and recommendations may not always be accurate. Extensive human evaluation is more reliable but costlier and time-consuming. The exploration of alternative solutions remains an important focus for ongoing research. Our models lack explicit mechanisms to curb undesirable behaviors. Future research will focus on enhancing security and alignment strategies (Christiano et al., 2017; Bai et al., 2022; Rafailov et al., 2023). On a brighter note, we believe that releasing our weight could contribute to investigating and mitigating risks tied to application uses.

## 7 Safety and Ethical Implications

In alignment with the ACL Ethics Policy, we address the ethical dimensions of our work. We have conscientiously credited the data sources and other open source works on which BiMediX is built upon. We recognize the significant societal impact of this technology, emphasizing ethical considerations and transparency. This release, intended for research, is not ready for clinical or commercial use. Further analysis is needed to ensure safety and accuracy in clinical settings and to prevent patient harm. Collaboration with patients, medical professionals, and ethicists is crucial for ethical oversight. We adhere to the License Agreement of the base models.

## 8 Acknowledgement

We would like to thank Dr. Shanavas Cholakkal, Dr. Amjad Ramzi Kamal Ebdelkari, Dr. Kartik Bhagat, Dr Omair Mohammed and Dr Mohammed Zidan for their contribution in translation and verification of medical responses.

The computations were enabled by resources provided by LUMI hosted by CSC (Finland) and LUMI consortium, and by Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the NSC.

This work is partially supported with Google Research Award titled "A Climate Change and Sustainability Tailored Arabic LLM", awarded in November 2023.

## References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MedInfo*, pages 25–29.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. BioMedLM: a domain-specific large language model for biomedical text. https://crfm.stanford.edu/2022/12/15/biomedlm.html.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Clément Christophe, Avani Gupta, Nasir Hayat, Praveen Kanithi, Ahmed Al-Mahrooqi, Prateek Munjal, Marco Pimentel, Tathagata Raha, Ronnie Rajan, and Shadab Khan. 2023. Med42 - a clinical large language model.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. Arabart: a pretrained arabic sequence-to-sequence model for abstractive summarization. *arXiv preprint arXiv:2203.10945*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Navreet Kaur, Monojit Choudhury, and Danish Pruthi. 2023. Evaluating large language models for health-related queries with presuppositions. *arXiv preprint arXiv:2312.08800*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

LLm-analysis. 2023. Llm-analysis - latency and memory analysis of transformer models for training and inference. Availabe at https://github.com/cli99/llm-analysis.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Arat5: Text-to-text transformers for arabic language generation. *arXiv preprint arXiv:2109.12068*.

OpenAI. 2023a. Chatgpt. Availabe at https://www.openai.com/chatgpt.

OpenAI. 2023b. Openai gpt-3 api [gpt-3-5-turbo]. Availabe at https://platform.openai.com/docs/models/gpt-3-5-turbo.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *arXiv preprint arXiv:2402.13963*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: Larger biomedical domain language model. *arXiv preprint arXiv:2010.06060*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further fine-tuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022a. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022b. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

# Appendix

## A Developing Comprehensive Arabic Healthcare Language Models and Datasets: BiMed1.3M

To the best of our knowledge, unlike English, there are no comprehensive and accurate evaluation benchmarks for Arabic healthcare Language Models (LLMs), nor are there large-scale Arabic healthcare datasets available for instruction-tuning LLMs. This has resulted in a lack of large-scale and accurate medical LLMs in Arabic, a language spoken by more than 400 million people. Motivated by this research gap and its practical implications for a significant population, we aim to address both of these limitations by introducing an Arabic medical evaluation benchmark consisting of four datasets as well as a large-scale instruction-tuning dataset consisting of more than 444k Arabic instructions (344.19 M tokens).

Different from other domains, the factual correctness of instructions is much more important in the healthcare domain. However, most reliable medical textbooks and online materials are in English. The scarcity of reliable and medically accurate public resources in Arabic prompted us to utilize English-to-Arabic translation to create the benchmark and instruction-tuning dataset. This study represents an initial effort to bridge this gap, undertaken within the constraints of the current dataset availability. We hope that this work serves as a step toward fulfilling the need for more contextually accurate medical datasets in Arabic.

Furthermore, we acknowledge the limitations of naive translation of medical content into Arabic, especially when using naive translation services like Google Translate. This motivated us to propose a semi-automated translation pipeline that incorporates a Language Model-based translation. By considering the entire medical conversation (questions and answers) in context, this approach achieves higher translation quality. In our semi-automated translation pipeline, we also involved native Arabic-speaking doctors to validate the quality and relevance of the translated text within the Arabic medical context. Moreover, to minimize human effort, we introduced a novel scoring strategy based on Language Models. This strategy iteratively refines the translated text by scoring its quality on a scale of one to 100, with low-scored text being passed back to the Language Model for improvement. Through this iterative process, we observed an improvement in the quality of translated text (see Fig. 4), thereby reducing human involvement while ensuring the quality and context of translated Arabic medical Q&A. We believe that our proposed translation pipeline will aid in the translation of medical content into Arabic, and the involvement of native Arabic speakers in the translation pipeline can help to reduce the gap with real medical Q&A scenarios. Note that our source code, instruction sets, and evaluation benchmarks are made publicly available to support future research.

## B Further Details on Related Works

**Comparison to Previous Work:** Similar to Clinical Camel, our models are fine-tuned on conversational data, with more than one turn of interactions, to significantly improve understanding and response capabilities in medical contexts. Notably, we are the first to integrate all conceivable interaction types, including Q&A, MCQA, and Chat, into a large-scale instruction tuning dataset. Unlike models that undergo continual pre-training on the base model weights, such as PMC-LLaMA and Meditron, our methodology steers clear of this approach due to the considerable demands for additional data collection, extended training periods, and potential to impair the base model's capabilities. Furthermore, our models adopt Parameter-Efficient Fine-Tuning (PEFT) (Hu et al., 2021; Li and Liang, 2021; Dettmers et al., 2023) techniques to boost model performance efficiently, circumventing the need for substantial resources. A summary of close works is provided in Table 9.

## C Additional Details on Data Creation

In this section, we provide comprehensive details about the data generation strategies implemented in our method.

**Dataset Construction.** To build BiMed1.3M, we first constructed an English dataset comprising three distinct categories of medical interactions: Multiple-Choice Question Answering (MCQA), which emphasizes specialized medical knowledge and reasoning to provide definitive answers; Open Question Answering (QA), incorporating real-world consumer inquiries; and Multi-Turn Chat Conversations, designed for dynamic dialogue. For the first two categories, we sourced data as detailed in Table 8. For the multi-turn chat component, we generated realistic conversations grounded on

| Dataset | Description |
|---|---|
| **Multiple-Choice Question Answering (MCQA)** | |
| PubMedQA (Jin et al., 2019) | Closed-answer questions from medical abstracts in PubMed research papers. |
| MedMCQA (Pal et al., 2022) | Multiple-choice questions from the Indian AIIMS and NEET PG medical entrance exams. |
| MedQA (Jin et al., 2021) | Multiple-choice questions from the USMLE for U.S. medical licensing. |
| **Question Answering (QA)** | |
| HealthCareMagic (Yunxiang et al., 2023) | Specialist-patient Q&As crawled from HealthCareMagic. |
| iCliniq (Yunxiang et al., 2023) | Physician-patient Q&As from the iCliniq consultation site. |
| Medical Meadow (Han et al., 2023) | Q&A pairs from Anki flashcards and Wikidoc's medical textbooks and Patient Information. |
| UMLS (Wu et al., 2023) | Q&As on medical term meanings and entity relationships from knowledge graphs. |
| LiveQA (Abacha et al., 2017) | Annotated consumer health questions from the National Library of Medicine (NLM). |
| MedicationQA (Abacha et al., 2019) | Annotated consumer medications inquiries from MedlinePlus. |

Table 8: Summary of collected data sources for multiple-choice (MCQA) and open question answering (QA).

| Model | Language | Continual Pre-training | PEFT | Instruction Tuning Data Type | Data Size |
|---|---|---|---|---|---|
| ChatDoctor (Yunxiang et al., 2023) | EN | ✗ | ✗ | QA | 100 K |
| MedAlpaca (Han et al., 2023) | EN | ✗ | ✓ | QA | 160 K |
| PMC-LLaMA (Wu et al., 2023) | EN | ✓ | ✗ | QA + MCQA | 514 K |
| Clinical Camel (Toma et al., 2023) | EN | ✗ | ✓ | MCQA + Chat | 174 K |
| Med42 (Christophe et al., 2023) | EN | - | - | - | - |
| Meditron (Chen et al., 2023) | EN | ✓ | ✗ | MCQA | 370 K |
| Ours-BiMediX | AR & EN | ✗ | ✓ | QA + MCQA + Chat | 1312 K |

Table 9: Comprehensive comparison with current open-source medical models, highlighting language, training strategy, and data statistics. PEFT indicates whether models have been adapted for specific tasks with minimal parameter updates. Chat means multi-round QA.

MCQA scenarios. The example below illustrates a sample of our data generation for chats grounded on medical QAs, highlighting how the model leverages deep medical knowledge to produce accurate and pertinent content. Indeed the prompt (see Table 10) is specifically designed to engage in complex interactions that include asking about symptoms, posing follow-up questions, and offering simulated advice from a doctor.

To ensure the content's relevance and accuracy, we undertook a rigorous quality assurance process. This involved a manual review conducted by medical professionals. We selected a random subset of the generated chats and presented these to doctors for verification, ensuring that the responses were not only accurate but also practically applicable in a medical context.

### Generated Chat Example

**User** Hi Doctor.

**Model** Hello, welcome. How can I assist you today?

**User** I have been diagnosed with tuberculosis and I'm currently undergoing treatment. I wanted to ask about vitamin supplementation.

**Model** Sure, I'll be happy to help. What specific vitamin are you interested in?

**User** I would like to know which vitamin supplementation is advised when a person is on treatment for tuberculosis.

**Model** When someone is undergoing treatment for tuberculosis, it is recommended to supplement with pyridoxine, also known as Vitamin B6.

**User** Why is pyridoxine supplementation necessary during TB treatment?

**Model** Isoniazid, which is used in TB regimens, inhibits the enzyme pyridoxine phosphokinase. This enzyme is necessary to activate pyridoxine to its active form. Thus, pyridoxine supplementation is required during TB treatment to prevent the development of peripheral neuropathy.

**User** Thank you for explaining. Should I take pyridoxine supplements on my own or do I need a prescription?

**Model** You will need a prescription for pyridoxine supplementation during your TB treatment. It's important to consult with your healthcare provider who will determine the appropriate dosage for you.

### Corresponding Original MCQA

**Question:** Which vitamin supplementation is advised when a person is on treatment for tuberculosis. Options:

A. Thiamine

B. Biotin

C. Niacin

D. Pyridoxine

**Answer:** Isoniazid used in TB regimens inhibits the enzyme

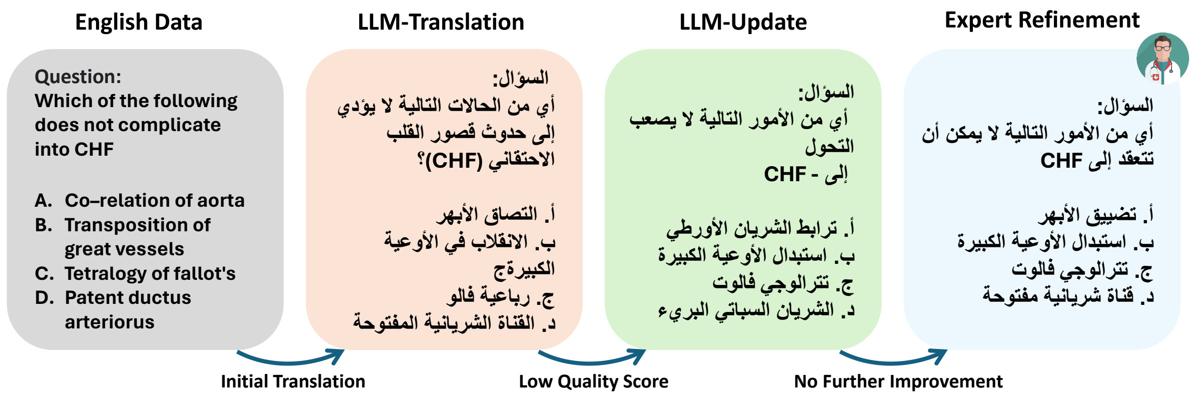


Figure 4: Update of sample in our semi-automated translation pipeline.

pyridoxine phosphokinase. This enzyme is necessary to activate pyridoxine to its active form. Thus pyridoxine (Vitamin B6) is supplemented during TB treatment to prevent the development of peripheral neuropathy. The correct answer is D.

**Iterative Translation Pipeline.** Table 11 describes the scoring system used to assess the quality of the model's translations. In Table 12, we detail the prompt used to improve the Arabic translations, given the Arabic predicted translation, the original English text, and the assigned score. To validate the effectiveness of our semi-automated translation pipeline, we conducted an experiment involving a fixed number of both high-scoring and low-scoring translations, which were evaluated by native Arabic-speaking medical doctors. These experts assessed the translation quality and were tasked to refine the translations if needed. Our findings revealed that approximately 90% of the low-scoring translations required significant modifications upon expert review, such as major adjustments in the sentence construction. In contrast, translations with initially high scores were generally of superior quality, requiring mostly no changes and, in a few cases, minimal intervention— just single-word adjustments to improve the specificity of Arabic medical terminology (e.g., more specific medical terms instead of more general/colloquial ones). Furthermore, we randomly selected 100 question-answer pairs that were scored below the threshold initially and improved to above the threshold through our iterative translation. We asked native Arabic speakers to compare translation quality before and after our iterative translation and rate the change as *'deterioration'*, *'no improvement'*, *'moderate'*, or *'significant'*. In all samples, evaluators indicated moderate or significant improvement. Notably, in our approach, samples that initially scored below the threshold were improved through our iterative approach, significantly reducing human effort. Our iterative pipeline ensures the production of high-quality, medically insightful content that is both reliable for healthcare applications and cost-efficient.

# D Quantitative Medical Benchmarks

Our analysis incorporates multiple prominent benchmarks in medical multiple-choice question-answering.

**(i) PubMedQA** (Jin et al., 2019) is a question-answering dataset derived from biomedical research papers on PubMed. Given a question sourced from the title of a research paper and a context from the abstract, the task is to provide an answer in the form of 'yes', 'no', or 'maybe', mirroring the conclusion of the paper. Building on prior research, our examination focuses on the PQA-L subset of PubMedQA, consisting of 500 manually annotated QA pairs sourced from PubMed. Answering these questions with accuracy requires thorough reasoning over the biomedical contexts and quantitative data analysis.

**(ii) MedMCQA** (Pal et al., 2022) is a multiple-choice dataset constructed from questions featured in the Indian AIIMS and NEET PG medical entrance exams. It covers a broad spectrum of medical subjects, evaluating both professional domain knowledge and language comprehension. The dataset includes a test set of 4,183 questions, each with four options.

**(iii) MedQA** (Jin et al., 2021) is a dataset comprising multiple-choice questions from medical board examinations in the US, Mainland China,

and Taiwan. It features two types of questions: concise sentences to assess specific knowledge and extended paragraphs that detail a patient's condition. Our analysis concentrates on the English portion (USMLE), sourced directly from the National Board of Medical Examiners (NBME), responsible for the assessments required for medical licensure in the U.S., containing 1,273 samples for testing purposes. Engaging with this benchmark involves multiple stages of reasoning and evidence retrieval.
**(iv) The Medical MMLU** (Massive Multitask Language Understanding) (Hendrycks et al., 2020) is a group of six datasets that compiles 1,089 test questions with four options related to different medical subjects. The six MMLU datasets are Clinical Knowledge (Cli-KG), College Biology (C-Bio), College Medicine (C-Med), Medical Genetics (Med-Gen), Professional Medicine (Pro-Med) & Anatomy (Ana).

All questions from the above English datasets are translated into Arabic using our semi-automated pipeline to create our Arabic and Bilingual evaluation benchmarks.

## E Qualitative Experiments

We evaluate BiMediX on complex queries that reflect real-world patient interaction beyond assessing the model's medical knowledge on multiple-choice question answering.
**(i) UPHILL** (Kaur et al., 2023) is a dataset designed to evaluate models' ability to process and respond to false medical claims across different levels of user presupposition, which refers to the assumptions users hold when asking a question. This is a crucial test due to the real-world implications of health-related misinformation. The model's performance is assessed through zero-shot accuracy by measuring using GPT how often it refutes false claims at each presupposition level. A response is considered factually accurate if it disagrees with a false claim. Accuracy for a set of queries $Q_l$ at presupposition level $l$ is calculated as the proportion of responses that correctly refute the false claims in the queries. The reported overall accuracy is the average across all presupposition levels.
**(ii) Multi-Turn Conversations** evaluate the quality of various models' output in multi-turn conversations, with three certified medical doctors serving as evaluators. All compared models are given the same initial query, followed by user responses adapted to each model's replies across different medical specialties. The responses are anonymized, and the experts assess and select the preferred conversation flow based on factors such as medical knowledge and facts, accuracy of diagnosis, appropriate follow-up questions, and natural conversational flow. This process aims to evaluate both the correctness and the naturalness of the conversations, while considering the model's effectiveness in real-world scenarios.

## F Additional Examples

The following example shows the comparative analysis between BiMediX and other models: Med42, Meditron-70B, Mixtral 8x7B and GPT-4. As confirmed by medical experts BiMedix responses often asks the right questions leading to an accurate diagnosis compared to other models.

### Med42

**User** Hi Doctor. I've been experiencing some discomfort lately. I've had this persistent pain in my lower back and abdomen.

**Model** I'm sorry to hear that. Based on your symptoms, it's possible that you might have a condition called hydronephrosis. Have you ever been diagnosed with it before?

**User** No I have not been diagnosed with it, what should I do ?

**Model** It's crucial to seek medical attention promptly to identify the cause of your symptoms accurately. Here are the steps you should follow: 1. Consult a healthcare professional, such as a primary care physician or a nephrologist, to discuss your symptoms and medical history. 2. Undergo a thorough physical examination and provide details about the onset, duration, and progression of your symptoms. 3. Your doctor may recommend imaging tests like ultrasound, CT scan or MRI to evaluate the kidneys and urinary tract. 4. Based on the test results and clinical evaluation, your doctor will determine the diagnosis and recommend a suitable treatment plan, which may include medications, surgery or other interventions. 5. Follow up with your healthcare provider regularly to monitor your condition, adjust medications, and manage symptoms effectively. It's important to remember that early diagnosis and treatment can help prevent complications and improve the quality of life.

### Meditron-70B

**User** Hi Doctor. I've been experiencing some discomfort lately. I've had this persistent pain in my lower back and abdomen.

**Model** I'm sorry to hear that you're experiencing discomfort. As a virtual assistant, I can provide you with some general information and suggest some possible causes and solutions. However, it's important to note that I cannot provide medical advice or diagnose any medical conditions. If

| System | Prompt |
|---|---|
| You are a medical expert,<br>use the following set of question and answer<br>to create a realistic medical conversation.<br>Carefully generate a multistep medical conversation.<br>Return the conversation into a valid JSON format. | Generate a simulated conversation in JSON format that mimics a natural exchange between a patient and a doctor.<br>The conversation should reflect common dialogue that might occur, including greetings,<br>symptom explanations, questions, and the doctor's advice. Ensure the conversation is realistic, exclude<br>any direct answer options like 'Option A'. Format the output according to the following requirements:<br><br>- The dialogue should consist of a sequence of messages between a 'human' (a patient or a person asking<br>medical questions) and 'gpt' (representing a medical professional or a medical AI).<br>- The message values for 'human' should be related to given medical inquiries or descriptions, and the 'gpt'<br>responses should offer professional, correct, and informative answers.<br>- It has to sound like a natural conversation between a patient and a doctor. The conversation has to be based<br>on the topic and content extracted from the following medical QA:<br><br>Question: {question}<br><br>Answer: {answer}<br><br>- The JSON has to be in valid JSON format.<br><br>Here is an example of how the JSON format should look. Without newline characters, backslashes, or extraneous<br>quotes, to be in JSON format:<br><br>{"id":"1","conversations":[<br><br>{"from":"human","value":"I am a 40-year-old male. I have a sharp chest pain that<br>started suddenly two hours ago. Can you tell me what might be the issue?"},<br>{"from":"gpt","value":"The symptoms you're describing could be indicative of several conditions, including<br>a possible myocardial infarction, commonly known as a heart attack.<br>It is critical to seek immediate medical attention in such cases."},<br>{"from":"human","value":"What should I do while I wait for the ambulance?"},<br>{"from":"gpt","value":"It is important to remain calm. If you have aspirin available and there are no allergies<br>or contraindications, chewing an aspirin can be beneficial while you wait for emergency services."}<br><br>More conversation entries as needed...<br>]} |

Table 10: Prompt for generating conversations anchored in Multiple-Choice Questions and Answers.

| System | Prompt |
|---|---|
| You are an English-Arabic medical language expert,<br>score the following translations from English to<br>Arabic Language. Return the score as an integer<br>into a valid JSON format with key "score". | You are an English-Arabic language expert, I will give you a pair of English text and its corresponding Arabic<br>Translation. You need to provide me a score between 1-100 on the quality of the translation. You have to evaluate<br>based on consistency/alignment with the original English text, preservation of technical terms and details.<br><br>English Text: {english text}<br><br>Arabic Translation: {arabic text}<br><br>You only need to give the score for the translation quality as output.<br>The output should be in a valid JSON format with only key "score".<br><br>Example output: {"score": 67} |

Table 11: Prompt for scoring the accuracy of Arabic translations compared to their English originals.

you're concerned about your symptoms, I strongly recommend consulting with a licensed healthcare professional for proper evaluation and treatment.
**User** It's more of a sharp, stabbing pain that comes in waves.
**Model** Lower back and abdominal pain can have a variety of causes, including muscle strain, kidney stones, urinary tract infections, and gastrointestinal issues. It's also possible that the pain could be related to an underlying medical condition, such as appendicitis or hernias.

## Mixtral 8x7B

**User** Hi Doctor. I've been experiencing some discomfort lately. I've had this persistent pain in my lower back and abdomen.
**Model** I'm sorry to hear you're not feeling well. I'm here to help. Can you tell me more about your symptoms?
**User** It's more of a sharp, stabbing pain that comes in waves.
**Model** Hmm, those symptoms could indicate a few different things. It's possible you have a kidney infection, or it could be appendicitis or even pneumonia. Either way, it's important to see a doctor right away.
**User** What can I do to reduce the pain ?
**Model** Taking over-the-counter pain medication like ibuprofen or acetaminophen can help reduce the pain. It's also important to rest and avoid any activities that make the pain worse.

## GPT-4

**User** Hi Doctor. I've been experiencing some discomfort lately. I've had this persistent pain in my lower back and abdomen.
**Model** I'm sorry to hear that you're experiencing discomfort. To better understand your symptoms and provide you with appropriate advice, could you please provide more details on the following: 1. How long have you been experiencing this

| System | Prompt |
|---|---|
| You are an English-Arabic medical language expert, correct the following translations from English to Arabic Language. | As an English-Arabic medical language expert, you are tasked with refining an Arabic translation to ensure optimal consistency and alignment with the original English text. The focus should be on preserving technical terms and details accurately. Below is a pair of English text and its corresponding Arabic translation, accompanied by a quality score between 1-100 indicating the initial translation's quality.<br><br>English Text: {english text}<br><br>Arabic Translation: {arabic text}<br><br>Quality Score: {score}<br><br>Please refine the Arabic translation based on the quality score and the need for improved consistency with the English original. The output should be a valid JSON object with the same format as the input text.<br><br>Note: Ensure technical terms and details from the original text are accurately reflected in the refined translation. |

Table 12: Prompt for improving the Arabic translation given the computed score.

pain? 2. Is the pain constant, or does it come and go? 3. Can you describe the nature of the pain (sharp, dull, throbbing, etc.)? 4. Is there anything that seems to make the pain better or worse? 5. Are you experiencing any other symptoms, such as fever, nausea, changes in bowel movements, or urinary issues? 6. Have you engaged in any recent physical activity or had any injuries that might have triggered this pain? These details will help me assess your situation more accurately.

**User** 1. Since last 2 days 2. Comes in waves 3. Sharp stabbing pain 4. No 5. Not sure 6. No

**Model** Thank you for providing more details. Based on your symptoms, here are a few potential causes for your lower back and abdominal pain: 1. Kidney Stones: Sharp, stabbing pain that comes in waves can be a sign of kidney stones. This pain is often felt in the lower back and can radiate to the abdomen. 2. Urinary Tract Infection (UTI): Though typically associated with a burning sensation during urination, a UTI can also cause lower abdominal and back pain. 3. Muscle Strain: Even if you haven't engaged in any recent physical activity, everyday movements can sometimes result in muscle strains. 4. Gastrointestinal Issues: Conditions such as constipation, gas, or irritable bowel syndrome (IBS) can cause lower abdominal pain that sometimes radiates to the back.

### BiMediX

**User** Hi Doctor. I've been experiencing some discomfort lately. I've had this persistent pain in my lower back and abdomen.

**Model** I'm sorry to hear that. Can you please describe the pain in more detail? Is it sharp, dull, or cramp-like?

**User** The pain is definitely not dull. It's more of a sharp, stabbing pain that comes in waves.

**Model** I see. How long have you been experiencing this pain?

**User** It's been rough the past two days, but I think I might have felt some twinges a few days before that. It's hard to say for sure.

**Model** Based on your description, it's possible that you may be experiencing some kidney-related issues. Have you noticed any changes in your urine?

**User** Actually, come to think of it, I have been noticing I haven't been urinating as much as usual. And when I do go, it feels kind of uncomfortable.

**Model** Okay, that's an important observation. It's possible that there may be an obstruction in your urinary system. I would recommend conducting a thorough medical examination to determine the exact cause. We may need to perform some tests, such as a urine analysis or an ultrasound.

## G Hyperarameters and Training

The training hyper-parameters included an effective batch size of 16 with two gradient accumulation steps and utilized the AdamW optimizer with a learning rate of 0.0002, cosine learning rate schedule, and ten warmup steps. The model implementation leveraged PyTorch (Paszke et al., 2019), along with the Deepspeed and ZeRO (Rasley et al., 2020; Rajbhandari et al., 2020) packages, with gradient checkpointing (Chen et al., 2016). Moreover, to align with the conversational style in our BiMed1.3M dataset, for both English and Arabic instructions, we adopted the Vicuna (Chiang et al., 2023) data format "conversations": ["from": "human/AI," "value": "..."], calculating loss exclusively for the AI's responses within the dialogue.


Figure 5: **Qualitative Examples of our BiMediX (English).**

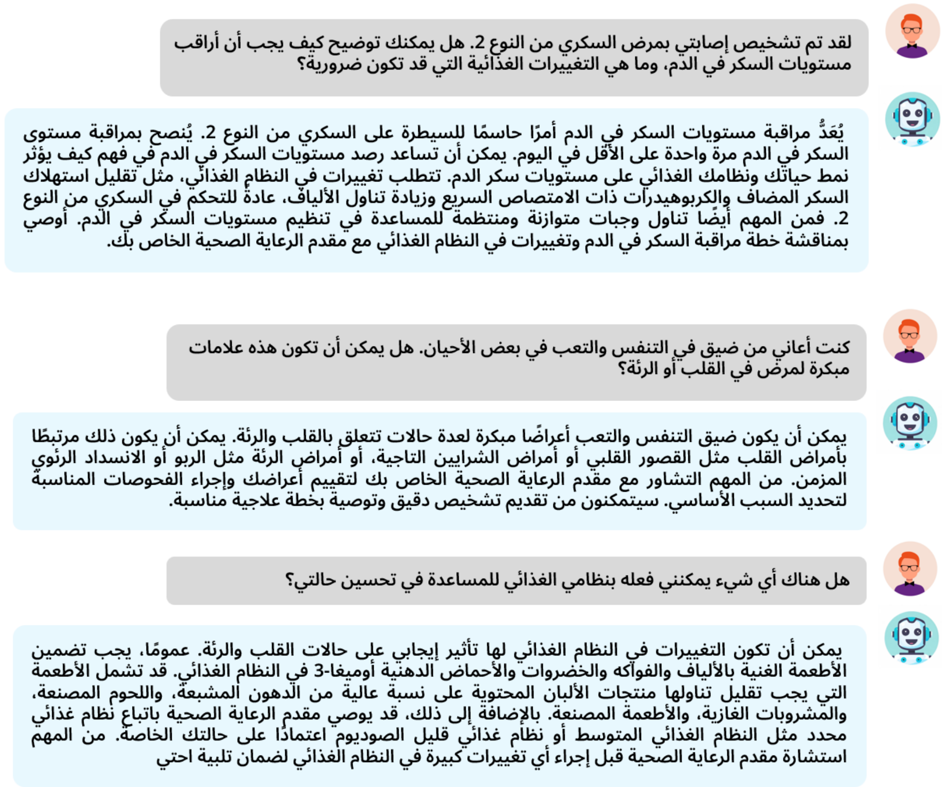


Figure 6: **Qualitative Examples of our BiMediX (Bilingual) which is translated to English as follows:**
**User:** *I have been diagnosed with type 2 diabetes. Can you explain how I should monitor my blood sugar levels, and what dietary changes might be necessary?* **BiMediX:** *Monitoring blood sugar levels is crucial to controlling type 2 diabetes. It is recommended to monitor your blood sugar level at least once a day. Monitoring your blood sugar levels can help you understand how your lifestyle and diet affect your blood sugar levels. It usually requires dietary changes, such as reducing consumption of added sugar and fast-absorbing carbohydrates and increasing fiber intake, to control type 2 diabetes. It is also important to eat regular, balanced meals to help regulate blood sugar levels. I recommend discussing a blood sugar monitoring plan and diet changes with your healthcare provider.*
**User:** *I've been experiencing occasional shortness of breath and fatigue. Could these be early signs of a heart or lung condition?* **BiMediX:** *Shortness of breath and fatigue can be early symptoms of several heart and lung conditions. This can be related to heart disease such as heart failure or coronary artery disease, or lung disease such as asthma or chronic obstructive pulmonary disease. It is important to consult with your healthcare provider to evaluate your symptoms and perform appropriate tests to determine the underlying cause. They will be able to provide an accurate diagnosis and recommend an appropriate treatment plan.*
**User:** *Is there anything I can do with my diet to help improve my situation ?* **BiMediX:** *Dietary changes can have a positive impact on heart and lung conditions. Generally, foods rich in fiber, fruits, vegetables and omega-3 fatty acids should be included in the diet. Foods to reduce your intake may include dairy products high in saturated fat, processed meats, soft drinks, and processed foods. In addition, your healthcare provider may recommend a specific diet such as a moderate-moderate diet or a low-sodium diet depending on your specific condition. It is important to consult your health care provider before making any major dietary changes to ensure your needs are met.*