# BIMEDIX2 : BIO-MEDICAL EXPERT LMM FOR DIVERSE MEDICAL MODALITIES

Sahal Shaji Mullappilly<sup>1</sup>\*, Mohammed Irfan Kurpath<sup>1</sup>\*, Sara Pieri<sup>1</sup>, Saeed Yahya Alseiari<sup>5</sup> Shanavas Cholakkal<sup>6</sup>, Khaled Aldahmani<sup>3,4</sup>, Fahad Khan<sup>1,2</sup>, Rao Anwer<sup>1</sup>, Salman Khan<sup>1</sup> Timothy Baldwin<sup>1</sup>, Hisham Cholakkal<sup>1</sup>

## **ABSTRACT**

This paper introduces BiMediX2, a bilingual (Arabic-English) Bio-Medical EXpert Large Multimodal Model (LMM) with a unified architecture that integrates text and visual modalities, enabling advanced image understanding and medical applications. BiMediX2 leverages the Llama3.1 architecture and integrates text and visual capabilities to facilitate seamless interactions in both English and Arabic, supporting text-based inputs and multi-turn conversations involving medical images. The model is trained on an extensive bilingual healthcare dataset consisting of 1.6M samples of diverse medical interactions for both text and image modalities, mixed in Arabic and English. We also propose the first bilingual GPT-40 based medical LMM benchmark named BiMed-MBench. BiMediX2 is benchmarked on both text-based and image-based tasks, achieving state-of-the-art performance across several medical benchmarks. It outperforms recent state-ofthe-art models in medical LLM evaluation benchmarks. Our model also sets a new benchmark in multimodal medical evaluations with over 9% improvement in English and over 20% in Arabic evaluations. Additionally, it surpasses GPT-4 by around 9% in UPHILL factual accuracy evaluations and excels in various medical Visual Question Answering, Report Generation, and Report Summarization tasks. The project page, including source code and the trained model, is available at https://github.com/mbzuai-oryx/BiMediX2.

# 1 Introduction

Recent advancements in healthcare AI, including medical large language models (LLMs) and large multimodal models (LMMs), have shown great potential in improving access to quality medical advice. However, these advancements remain predominantly English-centric, leaving significant gaps in addressing the healthcare needs of non-English-speaking populations. This language bias limits the accessibility and utility of AI-driven medical solutions in regions where languages like Arabic are widely spoken, creating a pressing need for more inclusive and multilingual models. Furthermore, conventional medical LMMs (Zhang et al. (2024); Li et al. (2023); Chen et al. (2024) see Fig 3) generally compromise their advanced medical text-based understanding when integrating multimodal capabilities.

Our project, BiMediX2, strives to address the aforementioned limitations of existing medical LMMs by introducing a Bio-Medical EXpert Large Multimodal Model (LMM) built on the Llama3.1 architecture. It achieves state-of-the-art performance in medical image understanding while also excelling in text-based medical LLM evaluations. BiMediX2 facilitates seamless user interactions in both English and Arabic, leveraging our extensive, purpose-built bilingual multimodal healthcare dataset BiMed-V.

BiMediX2 facilitates medical image understanding along with textual-queries, enabling multi-turn conversations about user-provided medical images. It supports a wide range of imaging modalities,

<sup>&</sup>lt;sup>1</sup>Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), <sup>2</sup>Linköping University

<sup>&</sup>lt;sup>3</sup>Shaikh Tahnoon bin Mohammed Medical City (STMC), <sup>4</sup>Tawam Hospital

<sup>&</sup>lt;sup>5</sup>Sheikh Shakhbout Medical City (SSMC), <sup>6</sup>Govt Medical College Kozhikode

<sup>\*</sup>Equal Contribution

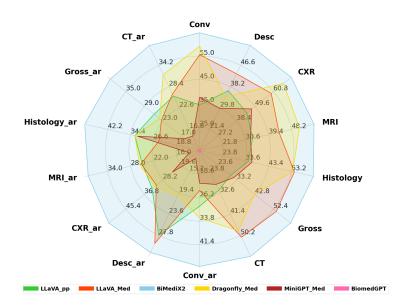


Figure 1: Model Performance Comparison on BiMed-MBench: These comparisons are made across different categories, including CT, MRI, CXR, Histology, Gross, and their Arabic counterparts (CT\_ar, MRI\_ar, CXR\_ar, Histology\_ar, Gross\_ar). The models compared are LLaVA-pp, LLaVA-Med, BiMediX2, Dragonfly-Med, MiniGPT-Med, and BiomedGPT. Each axis represents the performance score in a specific category, allowing for a visual comparison of how each model performs in bilingual medical contexts.

including chest X-rays, CT scans, MRIs, histology slides, and gross pathology images. This multi-modal approach ensures comprehensive and accessible medical assistance in a bilingual setting.

Recent advancements in medical LLMs and vision-language models (VLMs) have shown progress in tasks such as multi-turn conversations (MTC), report generation (RG), and specialized domains like radiology (Rad) and pathology (Path) as shown in Tab 1. While models like LLaVa-pp (Rasheed et al. (2024)) and LLaVa-Med (Li et al. (2023)) enable seamless multi-turn conversations, others, such as MiniGPT-Med (Alkhaldi et al. (2024)), are limited to specific modalities like chest X-rays, and BiomedGPT (Zhang et al. (2024)) requires fine-tuned checkpoints for each downstream task. BiMediXv1 (Pieri et al. (2024)) enables multi-turn conversational capabilities and report summarization within a bilingual framework, setting it apart from other medical LLMs. A comprehensive comparison of state-of-the-art models in Clinical LLM evaluations is shown in Fig 3.

Building on this, BiMediX2 expands its functionality to cover a broader spectrum of medical tasks, including Multi-turn Conversations (MTC), Report Summarization (RS), Report Generation (RG), Radiology (Rad), Ophthalmology (Oph), Pathology (Path), Microscopic Analysis (Micro), in a Unified Model (UM) with LLM and VLM capabilities (LLM+VLM). Fig 1 illustrates the performance of state-of-the-art medical LMMs on our BiMed-MBench evaluation in a bilingual context. Its robust bilingual support for Arabic addresses the healthcare needs of Arabic-speaking regions, providing a comprehensive and inclusive solution for diverse, multilingual, and multimodal healthcare applications.

#### **CONTRIBUTIONS**

In this work, we present BiMediX2, a bilingual multimodal medical LMM with the following key contributions:

- (i) We introduce the *first bilingual medical LMM* that achieves state-of-the-art results on VLM evaluation benchmarks across various medical image modalities, while also excelling on medical LLM evaluation benchmarks.
- (ii) We curated a comprehensive *Arabic-English multimodal bilingual instruction set* named *BiMed-V* comprising over *1.6M* instructions.

Model	MTC	RS	RG	Rad	Oph	n Pat	h Mic	ero 1	LLM+VLM	Bil (Ar)
Meditron (Chen et al. (2023))	Х	Х	Х	Х	Х	Х	Х		Х	Х
Med42 (Christophe et al. (2024))	1	1	X	X	X	X	X		X	X
OpenBioLLM (Ankit Pal (2024))	1	1	X	X	X	X	X		X	X
Llama3.1 (Meta (2024))	1	1	×	X	X	X	X		×	X
BiMediXv1 (Pieri et al. (2024))	1	1	X	X	X	X	Х		X	✓
Model	MTC	RS	RG	Rad	Oph	Path	Micro	UM	LLM+VLM	Bil (Ar)
LLaVA-pp (Rasheed et al. (2024))	1	1	Х	Х	X	Х	X	1	×	Х
MiniGPT-Med (Alkhaldi et al. (2024))	X	1	1	1	X	Х	X	1	×	X
BioMedGPT (Zhang et al. (2024))	X	1	1	1	1	1	1	X	×	X
LLaVA-Med (Li et al. (2023))	1	1	1	1	1			1	X	X
Dragonfly VLM (Chen et al. (2024))	X	1	1	1	1	1	1	1	X	X

Table 1: Comparison of Recent Medical LLMs and VLMs. Abbreviations: MTC (Multiturn conversation), RS (Report Summarization), RG (Report Generation), Rad (Radiology), Oph (Ophthalmology), Path (Pathology), Micro (Microscopic), UM (Unified Model: Single model checkpoint for all downstream tasks), LLM+VLM (Unified LLM + VLM), Bil (Ar) (Bilingual Arabic capabilities).

- (iii) We introduce the *first bilingual* GPT-4o-based *medical LMM benchmark* named *BiMed-MBench*, consisting of 286 medical queries in English and Arabic across various medical image modalities, fully verified by medical experts.
- (iv) Our BiMediX2 LLM outperforms GPT-4 by *more than 8%* on the USMLE benchmark and by *more than 9%* in UPHILL factual accuracy evaluations.
- (v) Our BiMediX2 LMM achieves state-of-the-art results on BiMed-MBench, with over a 9% improvement in English evaluations and more than a 20% improvement in Arabic evaluations. Furthermore, it excels in medical Visual Question Answering, Report Generation, and Report Summarization tasks.

# 2 BIMEDIX2

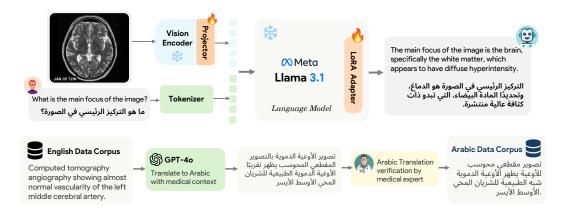


Figure 2: **BiMediX2: Overall Architecture** Our model is designed for medical image analysis and bilingual multi-turn conversations. Medical images are processed through a Vision Encoder and aligned with a Projector, while the text inputs are tokenized using the default tokenizer. The resulting tokens are then passed into the language model (Meta Llama 3.1) to generate responses in the prompted language. We only train the language model using LoRA adapters, while the projector is finetuned for medical image-text alignment. A robust data generation framework translates an English data corpus into Arabic using GPT-40, with verification by a medical expert to ensure accurate and contextually appropriate translations. This approach supports effective training and benchmarking in a bilingual context.

The architecture of BiMediX2 is meticulously designed to facilitate seamless integration of medical image analysis and bilingual multi-turn conversations. At its core, (see Fig 2) the model employs a Vision Encoder (Radford et al. (2021)) to process a diverse array of medical imaging modalities, including chest X-rays, CT scans, MRIs, histology slides, and gross pathology images. This visual data is aligned with textual inputs through a dedicated Projector, ensuring accurate and contextually rich medical image-text mapping. Text inputs are processed using a standard tokenizer, transforming them into the language embedding space of Llama 3.1 (Meta (2024)). This design enables BiMediX2 to generate precise and context-aware responses in either English or Arabic, depending on the user prompt, ensuring flexibility and inclusivity.

Key to BiMediX2's performance is its modular and efficient training approach. LoRA adapters (Hu et al. (2021)) are utilized to fine-tune the language model while maintaining computational efficiency and minimizing resource demands. The projector is simultaneously fine-tuned to optimize image-text alignment in a medical context. Furthermore, the system is supported by a robust data generation framework, where a comprehensive English data corpus is translated into Arabic using GPT-40. A random subset of this translation is meticulously verified by bilingual medical experts to ensure clinical relevance and linguistic accuracy. This pipeline enables BiMediX2 to excel in a wide range of medical tasks, including report generation, radiology analysis, pathology insights, and ophthalmological assessments, all within a unified, bilingual, and multimodal framework.

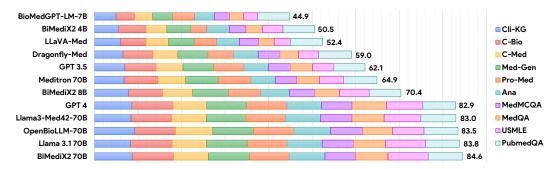


Figure 3: State of the art comparison of models in Clinical LLM Benchmarks

## BIMED-V BILINGUAL AND MULTIMODAL DATASET CREATION

The BiMed-V dataset is a comprehensive bilingual and multimodal instruction set comprising of 1.6M samples, developed to enhance medical image-text alignment and multimodal understanding. It incorporates a diverse range of publicly available datasets, such as PMC-OA (Lin et al. (2023)), Rad-VQA (Lau et al. (2018)), Path-VQA (He et al. (2020)) and SLAKE (Liu et al. (2021)) complemented by custom-curated data. We also curated 163k VQA samples by repurposing the LLaVA-Med (Li et al. (2023)) 60K-IM dataset, which were tailored to align with real-world medical queries. Furthermore, over 10k samples from the LLaVA-Med pretraining dataset were reformatted into interactive conversations using the Llama 3.1 70B model. A subset of the PMC-OA dataset with short question-answer pairs and multiple-choice questions were added to enhance the dataset's diversity. Training splits of Rad-VQA, Path-VQA and SLAKE, which typically feature concise answers, were restructured into more detailed responses using the same Llama 3.1 70B model, enhancing the dataset's depth and usability for complex tasks.

A unique feature of BiMed-V is its bilingual support, facilitated by a multimodal instruction set comprising 326k samples across various medical imaging modalities. This includes 163k Arabic-language samples generated via a comprehensive translation framework (see Fig. 2). English datasets were translated into Arabic using GPT-40, with rigorous verification of a random subset by bilingual medical experts to ensure clinical relevance and linguistic precision. This hybrid approach balances automation and expert validation, significantly reducing reliance on human medical domain experts while maintaining data quality. Additionally, the inclusion of text-based clinical data from BiMediXv1 (Pieri et al. (2024)) ensures the dataset retains robust language understanding capabilities while expanding its multimodal medical proficiency. This extensive dataset forms the foundation for advanced medical image-text alignment and conversational multimodal applications.

## BILINGUAL AND MULTIMODAL MEDICAL INSTRUCTION TUNING

To enable BiMediX2's robust capabilities in both bilingual and multimodal medical tasks, we employ a two-stage training process that ensures precise alignment of visual and textual representations while adapting the language model for complex medical instruction tasks.

# **Stage 1 Training: Medical Concept Alignment**

In the first stage, we finetune the Projector alone to align visual embeddings to the language embedding space. The training utilizes a comprehensive dataset of 467k image-caption pairs sourced from the LLaVA-Med (Li et al. (2023)) pretraining dataset. These pairs span diverse medical imaging modalities and captions that describe clinically relevant features.

## Stage 2 Training: Multimodal Medical Instruction Alignment

The second stage finetunes the LoRA(Hu et al. (2021)) adapters within the language model, enhancing its ability to process and generate multimodal medical instructions. For this, we utilize our BiMed-V 1.6M bilingual multimodal instruction set, which comprises meticulously crafted English and Arabic prompts paired with corresponding visual and textual responses. This dataset enables the model to learn nuanced instructions across a wide array of medical domains, from radiology to pathology, in a bilingual context.

Through these two stages, BiMediX2 achieves seamless integration of bilingual and multimodal capabilities, enabling it to deliver accurate and context-aware medical insights in both English and Arabic, tailored to a variety of clinical scenarios.

#### 3 EXPERIMENTS

In the literature, evaluating medical language models predominantly involves multiple-choice question-answering tasks, with accuracy as the performance metric. We employed the EleutherAI (Gao et al. (2024)) evaluation framework for evaluating text-only medical tasks following (Pieri et al. (2024)). For our bilingual and multimodal evaluations, we utilized several benchmarks and tools. We leveraged our BiMed-MBench for assessing the correctness, helpfulness, relevance and accuracy of the model in a bilingual medical multimodal context. Additionally, we evaluated the model on VQA using datasets such as Path-VQA, SLAKE, and Rad-VQA. For report generation, we used the MIMIC-CXR (Johnson et al. (2019)) dataset, and for report summarization, we used the MIMIC-III (Johnson et al. (2016)) dataset. These evaluations were conducted using the MultiMedE-val (Royer et al. (2024)) toolkit, ensuring a comprehensive assessment of the model's capabilities in both textual and visual medical tasks.

MEDICAL	L	L	M	BEN	CHM.	ARKS
---------	---	---	---	-----	------	------

Model		MMLU						MedQA	USMLE	PubmedOA	Average
	Cli-KG	C-Bio	C-Med	Med-Gen	Pro-Med	Ana	MedMCQA		COMPLE	1 4511104 Q.1	eruge
BioMedGPT-LM-7B	49.4	43.1	41.4	45.0	51.0	45.2	34.8	33.2	31.7	74.0	44.9
BiMediX2 4B	55.1	63.9	47.4	55.0	36.0	52.6	38.1	37.9	47.1	72.2	50.5
LLaVA-Med	59.6	59.7	50.9	59.0	51.5	51.9	44.5	35.7	36.9	74.0	52.4
Dragonfly-Med	65.6	69.4	56.6	69.0	58.4	57.0	49.9	42.8	46.1	75.4	59.0
GPT 3.5	69.8	72.2	61.3	70.0	70.2	56.3	50.1	50.8	49.1	71.6	62.1
Meditron 70B	68.3	77.8	63.6	75.0	74.6	56.3	48.4	53.1	55.4	76.2	64.9
BiMediX2 8B	77.7	79.2	68.8	82.0	74.3	65.9	58.0	57.0	68.6	72.4	70.4
GPT 4	86.0	95.1	76.9	91.0	93.0	80.0	69.5	78.9	83.8	75.2	82.9
Llama3-Med42-70B	84.2	93.1	79.8	91.0	90.1	80.7	72.5	73.8	84.3	80.6	83.0
OpenBioLLM-70B	92.5	93.8	85.6	93.0	93.4	83.7	74.1	68.9	72.0	78.0	83.5
Llama 3.1 70B	83.4	95.1	79.2	93.0	91.5	80.7	71.7	73.8	92.0	77.6	83.8
BiMediX2 70B	86.8	95.1	79.8	94.0	91.5	82.2	70.5	74.3	92.3	79.0	84.6

Table 2: Clinical LLM Evaluation Benchmark

**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 PQAL 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.

**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.

**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.

**USMLE** (Han et al. (2023)) USMLE Self-assessment Test, derived from the United States Medical Licensing Examination (USMLE) self-assessment for Step 1, Step 2, and Step 3, is composed of multiple-choice questions. We employ the version provided by the authors of MedAlpaca, which specifically omits image-based questions and comprises 325 test samples.

The Medical MMLU (Hendrycks et al. (2020)) (Massive Multitask Language Understanding) 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).

**UPHILL OpenQA** (Kaur et al. (2023)) The UPHILL dataset 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.

#### MEDICAL VLM BENCHMARKS

**BiMed-MBench:** We introduced the first bilingual GPT-40-based medical LMM benchmark, consisting of 286 medical queries spanning various medical imaging modalities. The test dataset includes conversational interactions and detailed descriptions for modalities such as chest X-rays (CXR), MRI, histology, gross and CT scans derived from LLaVA-Med (Li et al. (2023)). GPT-40 evaluates the correctness of model responses based on the provided image context and caption. The reference prediction, serving as the upper-bound answer for the teacher model, is generated using GPT-4 (Achiam et al. (2023)). For Arabic evaluations, the ground truth is derived from the upper-bound, translated by GPT-40, and all test samples are comprehensively verified by medical experts for quality assurance. GPT-40 evaluates the responses by comparing the candidate LMM against the upper-bound for English queries and against the Arabic ground truth for Arabic queries. The evaluation assesses the helpfulness, relevance, accuracy, and level of detail of the responses while maintaining the user language. The scoring is conducted on a scale of 1 to 10, with higher scores indicating better performance. Additionally, GPT-40 provides detailed explanations of the evaluation to offer deeper insights into the models' performance. Finally, we compute and compare the relative scores of the two candidates.

**Visual Question Answering (VQA):** To evaluate a model's performance in Visual Question Answering, three datasets are used: Path-VQA, SLAKE, and Rad-VQA. These datasets include a mix of open-ended and close-ended questions, designed to test the model's ability to interpret and respond to queries based on medical images. This evaluation ensures the model's effectiveness in handling a variety of visual and textual inputs.

**Report Generation:** The MIMIC-CXR (Johnson et al. (2019)) dataset, which includes de-identified radiology reports associated with chest X-rays, is utilized to assess the model's performance in generating the findings section of medical reports. For this task, the input consists of one or more radiology images related to a case, followed by a prompt such as, "<i mage> < image> Please caption this scan with findings and impressions." This setup evaluates the model's ability to generate accurate and coherent medical reports based on visual inputs.

Model	Conversation	Description	CXR	MRI	Histology	Gross	CT	Overall
BiomedGPT	15.3	13.3	16.4	13.0	14.1	14.9	15.8	14.8
LLaVA-pp	34.3	36.6	44.7	33.3	34.7	30.2	31.5	34.9
MiniGPT-Med	37.5	29.6	47.6	32.5	36.3	31.8	29.1	35.4
LLaVA-Med	55.6	43.3	59.5	43.4	54.4	53.9	51.0	52.4
Dragonfly-Med	59.2	34.2	67.0	51.2	53.7	42.6	48.3	52.7
BiMediX2 8B	64.9	54.5	71.7	56.8	62.5	61.4	58.9	62.2

Table 3: BiMed-MBench English Evaluation

Model	Conversation	Description	CXR	MRI	Histology	Gross	CT	Overall
BiomedGPT	11.1	11.2	11.4	10.8	11.5	11.3	11.1	11.2
MiniGPT-Med	21.6	12.6	23.7	12.7	32.0	15.8	14.9	20.2
LLaVA-Med	23.9	29.4	31.2	25.3	24.8	23.4	26.4	26.2
LLaVA-pp	29.0	27.8	33.2	25.0	33.0	25.8	25.8	28.7
Dragonfly-Med	32.8	19.9	31.9	25.7	33.0	24.0	31.7	29.5
BiMediX2 8B	54.3	36.2	61.4	44.6	51.5	43.5	50.8	50.5

Table 4: BiMed-MBench Arabic Evaluation

**Report Summarization:** We conduct an evaluation of report summarization using the MIMIC-III (Johnson et al. (2016)) dataset. The task involves generating the impressions section of a radiology report based on the findings section. To achieve this, the model is provided with the free-text findings along with a task prompt, "Summarize the findings". This assessment focuses on the model's ability to distill comprehensive medical information into a concise and precise summary, which is essential for clear and effective medical communication.

#### 4 RESULTS

#### LLM EVALUATION

The performance of BiMediX2 and other models on various medical benchmarks is presented in Fig 3 and Tab 2. Our BiMediX2 70B achieved the highest average score of 84.6%, outperforming other models such as GPT-4 (82.9%) and Llama-3-Med42-70B (83.0%). BiMediX2 70B exhibited superior average performance across multiple datasets, including Medical MMLU, MedMCQA, MedQA, USMLE, and PubMedQA, demonstrating its strong understanding of medical contexts.

## UPHILL OPENQA EVALUATION

Fig 4 illustrates the performance comparison on the UPHILL OpenQA benchmark. BiMediX2 70B achieved the highest overall factual accuracy of 60.6%, and the second highest being BiMediX2 8B (56.1%), surpassing other models such as GPT-4 (51.5%), Meditron 70B (49.6%), and Med42 (53.5%). This highlights BiMediX2's effectiveness in discerning and correcting misinformation in medical contexts.

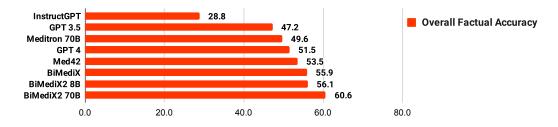


Figure 4: Performance comparison on **UPHILL OpenQA** (Kaur et al. (2023)), assessing the model's ability to address false medical claims at different presupposition levels.

Dataset	Metric	RadFM	LLaVA Med	BioMedGPT	MiniGPT-Med	Phi-3.5 V	BiMediX2 4B	BiMediX2 8B
	BLEU-1↑	0.475	0.033	0.044	0.662	0.377	0.501	0.552
	closed Q accuracy↑	0.577	0.545	0.203	0.829	0.618	0.685	0.725
Rad-VOA	open Q recall↑	0.407	0.246	0.199	0.546	0.295	0.292	0.363
Rau-vQA	recall <sup>↑</sup>	0.438	0.372	0.199	0.703	0.475	0.511	0.565
	open Q accuracy↑	0.335	0.140	0.150	0.490	0.200	0.225	0.305
	F1 ↑	0.442	0.069	0.064	0.675	0.391	0.516	0.569
	BLEU-1↑	0.746	0.036	0.175	0.337	0.089	0.625	0.778
	closed Q accuracy↑	0.752	0.512	0.248	0.572	0.535	0.744	0.831
Slake-VOA	open Q recall↑	0.758	0.429	0.293	0.308	0.377	0.624	0.763
Slake-VQA	recall <sup>↑</sup>	0.695	0.443	0.260	0.396	0.404	0.664	0.786
	open Q accuracy↑	0.725	0.362	0.259	0.278	0.329	0.567	0.729
	F1 ↑	0.714	0.075	0.192	0.349	0.129	0.641	0.787
	BLEU-1↑	0.257	0.021	0.145	0.296	0.283	0.469	0.587
	closed Q accuracy↑	0.505	0.512	0.260	0.581	0.553	0.708	0.872
Path-VQA	open Q recall↑	0.020	0.116	0.093	0.040	0.063	0.239	0.314
rain-vQA	recall <sup>↑</sup>	0.221	0.287	0.176	0.311	0.308	0.474	0.593
	open Q accuracy↑	0.005	0.053	0.077	0.019	0.027	0.210	0.282
	F1 ↑	0.232	0.052	0.154	0.299	0.287	0.475	0.595
	Average	0.461	0.239	0.177	0.427	0.319	0.509	0.611

Table 5: Medical VQA Benchmark (MultiMedEval Royer et al. (2024))

# BIMED-MBENCH EVALUATION

Tables 3 and 4 present the evaluation results of BiMediX2 and other medical LMMs on the English and Arabic BiMed-MBench benchmark, respectively. BiMediX2 8B demonstrated superior performance in both English (overall score of 62.2%) and Arabic (overall score of 50.5%) evaluations, outperforming other models. Fig 1 compares the performance of state-of-the-art medical LMMs on our BiMed-MBench evaluation in a bilingual context. This indicates BiMediX2's strong bilingual capabilities and its effectiveness in handling medical conversations and descriptions across different imaging modalities.

## MEDICAL VQA BENCHMARK

Tab 5 shows the performance of BiMediX2 and other models on the Medical VQA benchmark using the MultiMedEval toolkit. Our BiMediX2 8B achieves the highest average score of 0.611, outperforming other models across datasets such as Rad-VQA, Slake-VQA, and Path-VQA. This demonstrates BiMediX2's proficiency in visual question answering, a critical task in medical diagnostics.

#### REPORT SUMMARIZATION

Tab 6 presents the report summarization performance on the MIMIC-III dataset. BiMediX2 8B achieved the highest average score of 0.416, surpassing other models like LLaVA-Med (0.398) and Dragonfly-Med (0.110). This average score is derived as a unified metric by re-scaling BLUE-4\* and RadCliQ\* metrics. This indicates BiMediX2's effectiveness in generating concise and accurate summaries of medical reports, a vital task for efficient healthcare communication.

#### REPORT GENERATION

Tab 7 shows the report generation performance on the MIMIC-CXR dataset. BiMediX2 8B achieved the highest average score of 0.235, outperforming other models like LLaVA-Med (0.192) and BioMedGPT (0.145). The average score is derived as a unified metric by re-scaling BLUE-4\* and RadCliQ\* metrics. This highlights BiMediX2's capability in generating detailed and accurate medical reports from radiology images, a crucial task for diagnostic purposes.

#### QUALITATIVE EXAMPLES

**Medical Image Understanding in a Conversational Context**: Fig 5 illustrates the capabilities of the BiMediX2 framework in analyzing medical images and providing detailed, context-aware responses. The top section showcases a conversation involving a sagittal CT scan of the lumbar spine. The model identifies the scan type and explains that it is a vertical slice of the body focusing on the lumbar spine. When asked about abnormalities, the model accurately identifies a fracture in

Dataset	Metric	LLaVA Med	Dragonfly-Med	BiMediX2 4B	BiMediX2 8B
	ROUGE-L↑	0.185	0.072	0.209	0.205
	BLEU-1↑	0.192	0.062	0.153	0.178
	BLEU-4↑*	0.520	0.000	0.410	0.449
MIMIC-III	F1-RadGraph↑	0.232	0.000	0.222	0.230
	RadCliQ↑*	0.753	0.247	0.923	0.918
	CheXbert vector↑	0.600	0.326	0.633	0.593
	METEOR↑	0.303	0.060	0.264	0.339
Average		0.398	0.110	0.402	0.416

Table 6: Report Summarization (MultiMedEval Royer et al. (2024))

Dataset	Metric	RadFM	LLaVA Med	BioMedGPT	BiMediX2 4B	BiMediX2 8B
	F1-RadGraph↑	0.042	0.048	0.000	0.083	0.098
	BLEU-1↑	0.006	0.163	0.003	0.046	0.155
MIMIC-CXR	BLEU-4↑*	0.000	0.060	0.000	0.042	0.016
Report	ROUGE-L↑	0.065	0.125	0.012	0.131	0.153
Generation	RadCliQ↑*	0.655	0.660	0.827	0.865	0.860
	CheXbert vector↑	0.197	0.150	0.153	0.205	0.189
	METEOR↑	0.053	0.137	0.016	0.107	0.174
Average		0.145	0.192	0.145	0.211	0.235

Table 7: **Report Generation** (MultiMedEval Royer et al. (2024))

the L4 vertebra, providing an explanation of the condition and its potential causes, such as trauma or stress. The bottom section features a color Doppler ultrasound scan of the left ovary, identifying a potential abnormality *a left ovarian cyst with a solid component*. Our model explains the imaging technique, names the organ, and discusses the detected abnormality, emphasizing the need for further evaluation. These examples demonstrate BiMediX2's ability to interpret complex medical images and provide informative responses to assist in clinical decision-making.

Medical Image Understanding in a Conversational and Bilingual (Arabic) Setting: Fig 6 illustrates BiMediX2's ability to understand medical imagery and converse in Arabic, showcasing its bilingual capabilities. In the top section, our model accurately identifies the organ and the type of scan, providing clear and precise information that can assist medical professionals in diagnosing and treating liver-related conditions. The middle section shows that our model identifies the body part and the type of scan. And in the bottom section, BiMediX2 accurately identifies the organ, explains the imaging technique, and detects abnormalities, providing valuable insights that can aid in the diagnosis and treatment of conditions related to the female reproductive system. This particular example showcases BiMediX2's capability to converse in both English and Arabic simultaneously.

Medical Image Understanding of our BiMediX2 in Multidomain: Fig 7 showcases BiMediX2's versatility and accuracy across various medical imaging modalities. In the top section, it identifies key structures in a scanning electron micrograph of a mosquito head. In subsequent sections, it correctly interprets an MRI of a parotid tumor, a histology slide of adipose tissue (including stain type), a chest X-ray with pneumothorax, and a CT scan of the abdomen, pinpointing adrenal abnormalities. These examples demonstrate BiMediX2's capability to analyze and diagnose diverse imaging types, making it a valuable tool across medical specialties.

#### 5 CONCLUSION

BiMediX2 represents a leap forward in bilingual, multimodal medical AI, addressing the global need for accessible and inclusive healthcare solutions in both English and Arabic. By integrating text and visual modalities within a unified architecture, it enables seamless multi-turn interactions for diverse medical tasks, including medical image analysis and complex medical conversations. Key contributions include the comprehensive bilingual dataset, BiMed-V, which provides diverse multi-modal medical instructions tailored for both languages, and the introduction of BiMed-MBench, the first bilingual GPT-40-based medical benchmark, which showcases the model's ability to excel in a wide range of expert-verified medical scenarios. BiMediX2 paves the way for inclusive, multilingual, and multimodal healthcare applications, significantly enhancing the accessibility and quality of medical assistance worldwide.

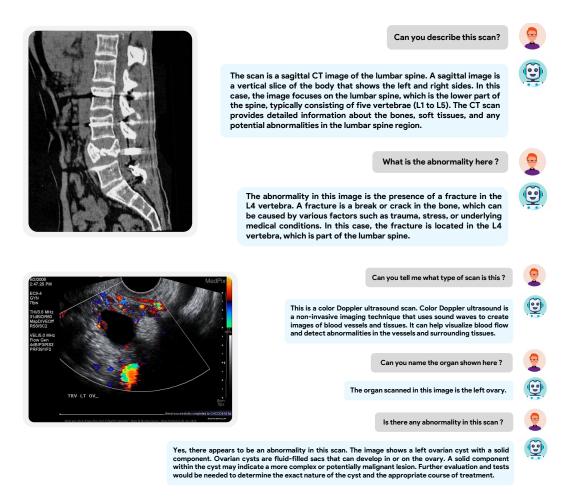


Figure 5: Qualitative Examples of our BiMediX2 for Medical Image Understanding in a Conversational Context.

## 6 LIMITATIONS

Despite its overall improvement, BiMediX2, like other generative language models, may experience issues such as hallucinations, toxicity, and stereotypes. These issues stem from both the inherited limitations of the base models and the nature of the pretraining data. While we have conducted automatic and qualitative evaluations with medical professionals, we acknowledge that our model's medical diagnoses and recommendations may not always be accurate. Extensive human evaluation is more reliable but expensive and time-consuming. The exploration of alternative solutions remains an important focus for ongoing research.

Currently, our models lack explicit mechanisms to curb undesirable behaviors. Future research will focus on enhancing security and alignment strategies. On a brighter note, we believe that releasing our model weights could contribute to investigating and mitigating risks tied to application uses.

## 7 SAFETY AND ETHICAL IMPLICATIONS

We recognize the significant societal impact of BiMediX2 and emphasize the importance of ethical considerations and transparency. This release is intended for research purposes only and is not ready for clinical or commercial use. Ensuring the model's accuracy and reliability is crucial, as incorrect medical advice could have serious health consequences. Robust validation and quality control measures are essential to minimize errors.



Figure 6: Qualitative Examples of our BiMediX2 for Medical Image Understanding in a Conversational and Bilingual (Arabic) Setting: The top section translates to; User: What is the organ shown in the image? BiMediX2: The image shows the liver, which is a large and vital organ in the human body. User: What type of scan is this? BiMediX2: This is a computed tomography (CT) scan of the abdomen. The middle section translates to; User: What type of imaging technique is used in this image? BiMediX2: The imaging technique used in this image is magnetic resonance imaging (MRI). User: What part of the body is shown in the image? BiMediX2: The image shows the head and neck area, with a particular focus on the pituitary gland. And the bottom section translates to; User: What is the organ shown in the image? BiMediX2: The organ shown in the image is the ovary, which is part of the female reproductive system. User: What type of scan is this? BiMediX2: This is an ultrasound image, which is a non-invasive imaging technique that uses high-frequency sound waves to create detailed images of the body's internal structures. User: What is the abnormality in this ultrasound image? BiMediX2: The abnormality in this ultrasound image is the presence of a multilocular ovarian cyst.

Ethical considerations include protecting patient privacy and ensuring the confidentiality of medical data. The model must comply with relevant data protection regulations and ethical guidelines.

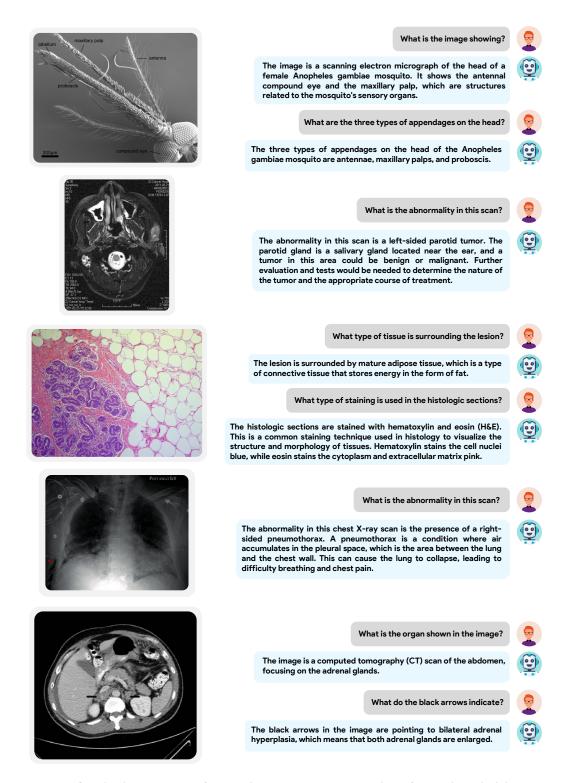


Figure 7: Qualitative Examples for Medical Image Understanding of our BiMediX2 in Multidomain: Capability of BiMediX2 to understand and interpret various medical imaging modalities and provide accurate responses. The examples include describing the anatomy of a mosquito's head in a scanning electron micrograph, detecting a left-sided parotid tumor in an MRI scan, recognizing mature adipose tissue in a histology slide, identifying a right-sided pneumothorax in a chest X-ray, and identifying bilateral adrenal hyperplasia in a CT scan of the abdomen. These examples highlight BiMediX2's versatility and effectiveness in medical image analysis and diagnosis.

Addressing potential biases in the model's outputs is also critical for ensuring fair and equitable healthcare outcomes.

Collaboration with patients, medical professionals, and ethicists is essential for ethical oversight and further research to ensure safety and accuracy in clinical settings. By acknowledging and addressing these considerations, we can continue to refine BiMediX2 for safe and effective use in healthcare.

## 8 ACKNOWLEDGEMENT

We would like to thank Dr. Omair Mohammed, Dr. Mohammed Zidan and Dr. Vishal Thomas Oommen for their contribution in 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.

We are grateful to the *Meta Llama Impact Innovation Awards* for recognizing BiMediX2, as one of the winners awarded in October 2024. This recognition highlights our commitment to advancing AI-driven healthcare solutions.

## REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- Asma Alkhaldi, Raneem Alnajim, Layan Alabdullatef, Rawan Alyahya, Jun Chen, Deyao Zhu, Ahmed Alsinan, and Mohamed Elhoseiny. Minigpt-med: Large language model as a general interface for radiology diagnosis. *arXiv preprint arXiv:2407.04106*, 2024.
- Malaikannan Sankarasubbu Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.
- Kezhen Chen, Rahul Thapa, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Dragonfly: Multi-resolution zoom supercharges large visual-language model, 2024. URL https://arxiv.org/abs/2406.00977.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv* preprint arXiv:2311.16079, 2023.
- Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco AF Pimentel, and Shadab Khan. Med42 evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches, 2024. URL https://arxiv.org/abs/2404.14779.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.
- Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathological visual question answering. *arXiv preprint arXiv:2010.12435*, 2020.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv* preprint arXiv:1909.06146, 2019.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- Navreet Kaur, Monojit Choudhury, and Danish Pruthi. Evaluating large language models for health-related queries with presuppositions. *arXiv preprint arXiv:2312.08800*, 2023.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.
- Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1650–1654. IEEE, 2021.
- Meta. Meta llama 3.1. https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct, 2024.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pp. 248–260. PMLR, 2022.
- Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. Bimedix: Bilingual medical mixture of experts llm, 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.
- Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad S. Khan. Llava++: Extending visual capabilities with llama-3 and phi-3, 2024. URL https://github.com/mbzuai-oryx/LLaVA-pp.
- Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models, 2024.
- Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pp. 1–13, 2024.