IndicMedQA: Multimodal Medical Query Analysis in Indian Languages

AKASH GHOSH, Indian Institute of Technology Patna, India
ARKADEEP ACHARYA, IBM Research India, India
MUHSIN MUHSIN, Indira Gandhi Institute of Medical Sciences, India
AMAN CHADHA, Apple, United States
VINIJA JAIN, Google, United States
SETU SINHA, Indira Gandhi Institute of Medical Sciences, India
SRIPARNA SAHA, Indian Institute of Technology, Patna, India

In the rapidly advancing field of AI-driven telehealth services, effective medical communication in India remains a significant challenge due to the language barrier, as most of the population is not proficient in English. Additionally, many patients struggle to accurately describe their medical conditions using text alone. As a result, an essential feature of any telehealth service is the ability to supplement textual queries with medical images, enabling doctors to conduct a more careful analysis and provide well-informed diagnoses and treatment recommendations. In this work, we introduce ***IndicMedQA***, a novel multimodal AI framework that integrates Indic large language models (LLMs) and visual encoders to analyze patient inquiries using both textual and visual cues. To support this, we create a multilingual multimodal medical corpus spanning seven major Indian languages, translated using a semi-automated approach. This dataset facilitates medical understanding for every input query and its associated medical image—the output is a detailed patient summary, symptom analysis, probable conditions, additional findings, and severity assessment with medical precision. Our framework significantly enhances personalized healthcare experiences, ensuring context-aware multimodal understanding of patient needs in Indic languages. Extensive experiments demonstrate that ***IndicMedQA*** surpasses all baselines, establishing a new benchmark for Indic AI in healthcare.

Disclaimer: This work contains medical images that depict the subject matter of the study, which may be disturbing to some readers.

CCS Concepts: • **Computing methodologies** → *Artificial Intelligence*; *Natural Language Processing*.

Additional Key Words and Phrases: Radiology Impression Generation, Language Models, Trustworthy NLP, and Healthcare AI.

1 Introduction

Artificial intelligence, particularly through advances in frontier models such as large language models (LLMs) and vision-language models (VLMs), is transforming healthcare by improving diagnosis, treatment planning, patient engagement, and clinical documentation, making these processes more efficient and effective[11, 17, 19]. In India, the unequal distribution of doctors across states and healthcare sectors worsens existing medical challenges

Authors' Contact Information: Akash Ghosh, Indian Institute of Technology Patna, Patna, India, akashghosh.ag90@gmail.com; Arkadeep Acharya, IBM Research India, Patna, India, acharyarka17@gmail.com; Muhsin Muhsin, Indira Gandhi Institute of Medical Sciences, Patna, India, contactmuhsin@gmail.com; Aman Chadha, Apple, California, United States, aman@amanchadha.com; Vinija Jain, Google, California, United States, vinija@gmail.com; Setu Sinha, Indira Gandhi Institute of Medical Sciences, Patna, India, drsinhasetu@gmail.com; Sriparna Saha, Indian Institute of Technology, Patna, Patna, India, sriparna@iitp.ac.in.

ACM 2637-8051/2026/7-ART
https://doi.org/10.1145/3833083

[13]. Additionally, language barriers not only hinder doctor–patient communication but also affect both the performance and trustworthiness of frontier models[16, 22–24, 42].

Despite growing progress in multilingual and multimodal NLP for Indian languages [27–29, 37] and cross-domain multimodal reasoning [9, 10], the application of these capabilities to Indian clinical settings remains largely unexplored. In this situation, extracting important details from multilingual clinical queries helps speed up patient responses and saves time for healthcare professionals, allowing for faster and more effective communication between doctors and patients. Moreover, since many patients struggle with medical terminology, accurately describing their condition becomes challenging, making it difficult for doctors to fully understand and provide the best possible guidance or medical advice. In this regard, the ability to incorporate medical cues through images significantly enhances and personalizes the doctor-patient interaction, leading to more effective communication and better-informed medical decisions[6].

Research Gap: In the evolving AI landscape, Large Language Models (LLMs) and Multimodal LLMs (MLLMs) are reshaping the field. However, the dominance of English and Chinese in LLMs has led to the underrepresentation of Indic languages[47], making it difficult to develop models that effectively understand them. Moreover, most existing closed-source LLMs are difficult to customize and come with high operational costs. These limit the progress of Indic LLMs, restricting their ability to serve diverse linguistic and cultural needs. To address this, researchers have developed Indic Language Models specifically trained on Indic languages, such as Airavata [15] for Hindi and Tamil Llama [4] for Tamil. These models incorporate language-specific tokens, improving their linguistic adaptability and effectiveness. The medical domain, however, presents unique challenges that require specialized data, domain expertise, and careful handling of transparency and bias concerns [45]. Recent advancements have introduced open-source medical LLMs, including Meditron [7], PMC-Llama [57], and Med-Flamingo [39], which are tailored for healthcare applications. However, their English-centric training limits their effectiveness in low-resource languages. To bridge this gap, recent works such as Apollo [54] and Multilingual Medical LLMs [47] have emerged, but they primarily focus on English, Chinese, European languages, and Hindi, leaving a significant gap in support for other Indic languages. Moreover, these multilingual medical models are limited to text-based processing, and, to the best of our knowledge, no existing work on Indic medical models specifically addresses multimodal medical queries incorporating both text and visual inputs.

Motivation: This study aims to address the under-representation of speakers of major Indic languages in the medical LLM and Multimodal Large Language Model (MLLM) literature. Focusing on Bengali, Gujarati, Tamil, Marathi, Hindi, Telugu, and Malayalam widely spoken languages across India we strive to expand the scope of Indic LLMs and MLLMs in the medical domain.[1] Motivated by the need for more inclusive healthcare solutions, this paper introduces a novel multimodal, multilingual patient query analyzer, ***IndicMedQA***, integrating Indic language decoders with vision encoders. The model is trained using a multi-objective loss function to ensure high-quality generation from both text and image modalities. To train this model, we curate a large-scale dataset, **IndicMed**, consisting of 21,105 meticulously annotated and translated medical data points. *Each data point includes a multilingual patient query paired with a corresponding medical image, while the output provides a condition focused analysis, encompassing patient summaries, symptom assessments, probable conditions, additional findings, and severity evaluations in English.*[2] The dataset was developed in close collaboration with medical practitioners to uphold quality and ethical standards throughout its creation.

Contributions: The main contributions of the work can be summarized as follows:

(1) We introduce a novel **task** of multimodal medical query analysis in the context of seven major Indian languages. In this task, patients articulate their health concerns using native language text accompanied by visual

[1]These languages cover a total of 80% of the Indian population.

[2]Since the model will be utilized by doctors ,so we have kept the output generations in English as for doctors English is the language that they are comfortable[36]

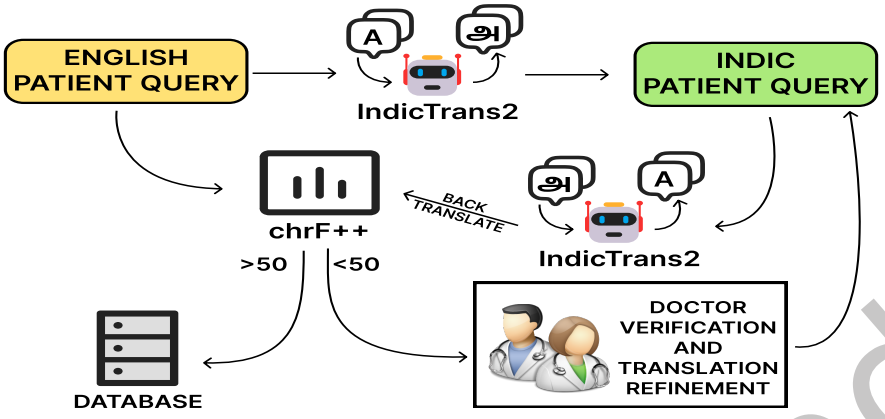


Fig. 1. Overview of our semi-automated, iterative pipeline engineered specifically for maintaining the quality of our translation task. This innovative approach integrates human expertise to refine translations while leveraging automated ChrF++ metrics to drive decision-making.

cues. This method aims to provide medical condition-focused analysis, enabling doctors to understand patient cases quickly and effectively.

(2) We assemble a **dataset** IndicMed for multimodal, multilingual medical query analysis covering Hindi, Bengali, Gujarati, Marathi, Tamil, Telugu, and Malayalam. Our approach employs a **semi-automated** iterative translation pipeline, complemented by human validation to guarantee precise translation of English medical text into these resource-constrained languages.

(3) We introduce ***IndicMedQA***, a novel **framework** designed as a lightweight multimodal transformer. Combining a static image encoder with an Indic LLM, ***IndicMedQA*** excels in understanding Indic Languages. Bridging the modality gap, it employs a Querying Transformer (Q-Former) [33] in a two-stage training process, beginning with representation learning using the image encoder and then progressing to generative learning with the Indic LLM.

(4) We conducted thorough **human and automated evaluations**, supplemented by detailed qualitative analysis and risk assessments. This examination was crucial to ensure the safety and dependability of the model's results, confirming its suitability for its potential decision-support use in real-world healthcare settings with a high degree of certainty.

2 Related Works

Medical Diagnosis System and Automation : Agent-derived Multi-Specialist Consultation was used by [52] to simulate the diagnosis process in real-world settings. A hierarchical policy structure for dialogue systems, designed to effectively handle diagnoses involving a multitude of diseases and symptoms, was proposed by [34]. To improve medical conversational systems, [50] trained a multi-turn dialogue model using single-turn question-and-answer pairs from healthcare websites. An online medical forum dataset was compiled by [55], who also developed a task-oriented dialogue system for automated patient diagnosis, capable of gathering additional symptoms through conversation. A hierarchical reinforcement learning model incorporating medical knowledge for disease diagnosis was introduced by [59], enhancing the overall success rate of dialogue-based diagnostic systems. For a more complete approach, [51] developed the Multi-modal Disease Diagnosis Virtual Assistant (MDD-VA), an end-to-end system leveraging reinforcement learning. Their work included a Context-aware

Symptom Image Identification module and the curation of a multi-modal conversational medical dialogue corpus in English. Complementary efforts in clinical safety have addressed adverse drug event summarization through frameworks that align language model outputs with human preferences for pharmacovigilance in oncology [26].More recently, advances in reasoning-capable frontier models have enabled long-horizon clinical automation [21].

Applications of Multimodality and Multilinguality in Healthcare: Recent strides in medical AI have focused on multilingual and multimodal applications. Text-based multilingual models like MMedLM2 [48], BioMistral [30], and MedExpQA[3] have expanded language coverage and benchmarking. In parallel, bilingual models like BiMediX [45] have emerged for English and Arabic. Multimodal approaches are also advancing. EHR-KnowGen [41] improves diagnosis generation from health records, while other BERT-based models [38] enhance medical image-report retrieval and generation. More generalist models like MedViLaM [58] now interpret diverse data types, including imaging and video, with MedPLIB [25] advancing pixel-level understanding. Despite this, many multilingual efforts focus on European languages, and prominent vision-language models like Med-flamingo[39] and LLava-med[32] are English-centric. While BiMediX2 has recently added Arabic-English multimodal capabilities, a significant gap remains for Indic languages. The closest works for multimodal Indic languages for healthcare are MedSumm [18] and HealthAlignSumm[20] but both of them targets Hindi-English codemixed text and no other Indian languages .

3 IndicMed Dataset Construction

For this study, we expanded upon the MMQS dataset introduced in [17], comprising 3015 multimodal medical queries alongside corresponding annotated summaries. The text is in English, while visual cues are categorized broadly into Ear, Nose, Throat (ENT), EYE, LIMB, and SKIN. We chose this dataset for its open accessibility and annotated summaries, which are vital for patient query analysis. The 18 medical symptoms were selected from the most common ENT-related conditions, focusing on those where patients can easily click and share photos. A few samples of our dataset are shown in Figure 2. However, one limitation of this dataset is the lack of fine-grained analysis of multimodal queries. For medical evaluation, doctors require detailed descriptions of visual cues along with the severity assessment, which are currently missing in this dataset.

3.1 Data Annotation

The data annotation comprises two subtasks: (A) Annotating medical images with their corresponding textual descriptions and severity assessments and (B) Translation of English patient queries into Indic Languages.

(1) Annotation of Medical Images with their textual descriptions: Through collaboration with medical professionals, we identified that medical analysis necessitates detailed descriptions of visual cues alongside patient queries. To address this, we incorporated precise medical terminology for symptoms observed in images, supplementary medical findings inferred from the visuals, and severity assessments to enhance the depth and accuracy of the analysis. The guidelines for image annotation provided by the doctor are as follows:

i) *Identify Primary Symptom/Sign:* Determine the main symptom or sign based on the provided description.

ii) *Additional Findings:* Note any additional signs or symptoms in the images that are not included under the primary symptom/sign.

iii) *Severity Grading Basis:* The severity grading on the primary symptom/sign into one of the three categories mild, moderate, and severe. For findings with established classification systems (e.g., tonsillitis, edema), follow those systems for severity grading. For findings without established classification systems (e.g., lip swelling), grade severity is based on the clinical picture.

(2) Semi-Automated Translation of English Patient Queries into Indic Languages: Initially, we evaluated ChatGPT [43], NLLB [8], and IndicTrans2 [14] for translation by randomly selecting 40 samples and having

annotators rate their translations on a five-point scale. Our analysis showed that IndicTrans2 consistently outperformed the other models in translating the English language to the corresponding Indic language. Based on these findings, we selected IndicTrans2 as our primary translation model. Next, we translated the English text into the respective Indic languages and validated translation quality through back-translation into English. To ensure accuracy, we retained only translations with a chrF++ [46] score above 50.[3] chrF++ is a character-level evaluation metric that has been shown to correlate well with human judgment for morphologically rich and low-resource languages, including Indic languages[46]. In a pilot analysis conducted with domain experts, translations with chrF++ scores below 50 frequently exhibited semantic distortions, incomplete medical terminology, or meaning shifts, whereas samples above this threshold were generally fluent and faithful to the source. Therefore, we adopted chrF++ > 50 as a conservative filtering criterion to balance translation fidelity and dataset coverage. For translations scoring below this threshold, we manually reviewed and refined them. After modification, only samples that achieved a chrF++ score above 50 were included in the final dataset, while lower-scoring samples underwent further refinement. This semi-automated iterative process required three rounds to achieve high-quality translations, ensuring the complexity and precision of medical text were effectively preserved. The full pipeline is illustrated in Figure 1.[4]

3.2 Annotation and Quality Validation

To guarantee the quality and adherence to ethical guidelines of the annotations, we enlisted the support of four capable medical interns who were fluent in the Indic languages and supervised by medical experts who are co-authors of this work. The medical interns are in the third year of their MD program, while the medical expert is a specialist physician with over ten years of clinical experience. Initially, the medical expert annotated 50 samples, provided guidance, and conducted sessions to address any queries. The dataset was divided into four segments, with annotators assessing data quality based on criteria such as fluency, adequacy, informativeness, and persuasiveness. Following each validation iteration, annotators collaboratively reviewed and discussed inaccuracies in each other's annotations. This manual review was critical for correcting translation errors, specifically targeting terminology mismatches, the loss of clinical nuance in symptom descriptions, and grammatical inaccuracies that could alter medical meaning. The quality of annotations significantly improved from the initial phase (fluency = 3.2, adequacy = 3.25, informativeness = 3.91, persuasiveness = 3.35) to the final phase (fluency = 4.23, adequacy = 4.17, informativeness = 4.56, persuasiveness = 3.91) with an interannotator agreement of 0.82. To maintain top-notch quality, we scheduled breaks every 30 minutes. It took approximately 45 sessions to complete the annotation process in a span of 2 months .[5]

3.3 Quality of IndicMed Dataset

We have tried to make *IndicMed* dataset a high-quality multimodal multilingual medical corpus in the following ways:

(1) **Reliability of sources:** IndicMed is based on MMQS [17] dataset which has recently been accepted in AAAI. We further verified the quality of that dataset by randomly selecting 30 samples, and we found that the dataset is of good quality.

(2) **Manual annotations and correction:** Given the specialized clinical nature of the dataset, we have partnered with a medical team to provide expert annotation of the images. To maintain the highest quality in translation, we use the chrF++ scoring system to evaluate accuracy. Any data point that falls below our

[3]ChrF++ [46] measures character n-gram matches using the F-score statistic, incorporating word n-grams for better correlation with human assessments. It has been shown to outperform BLEU in aligning with human judgment [31].

[4]An equal number of data points is maintained across all languages to ensure a balanced evaluation.

[5]The medical students were compensated with gift vouchers and honorarium per minimum wage requirements reported here: https://www.minimum-wage.org/international/india

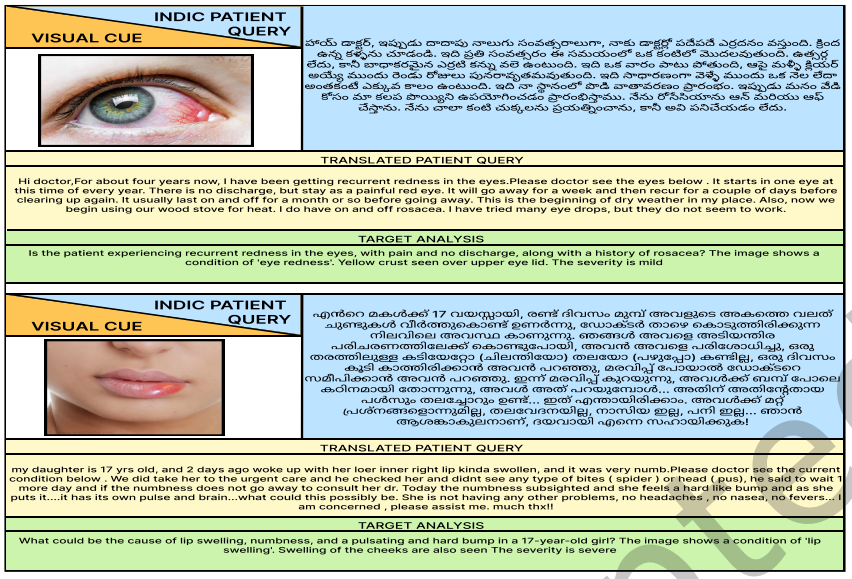


Fig. 2. Samples from our IndicMed dataset. The first sample is in Telugu, while the second one is in Malayalam.

predetermined threshold is meticulously reviewed and corrected manually, ensuring precision and reliability in the final dataset.

(3) **Evaluation of the final translated dataset:** To ensure the quality of translations in our dataset, we conducted a thorough human evaluation. We randomly sampled 40 examples from each of the the languages in the IndicMed dataset, resulting in a total of 120 translation pairs. These samples were reviewed by three annotators fluent in the respective Indic languages, who assessed each example based on fluency and meaning [6]. Our evaluation revealed average fluency scores of 4.4 for Tamil, 4.25 for Telugu, and 4.12 for Malayalam. For meaning, the average scores were 4.31 for Tamil, 4.15 for Telugu, and 4.10 for Malayalam.

3.4 Statistics Related to IndicMed Dataset

The different notable statistics for the IndicMed dataset are mentioned in Table 1 below:

Entries	Value
Number of question-analysis pairs	21,105
Number of questions-per language	3015
Median length of patient query	97
Median length of query analysis	44
Total number of unique images	587
Number of medical conditions	18

Table 1. *IndicMed* dataset statistics.

The dataset encompasses medical disorders categorized into four groups. In the ENT group, conditions include lip swelling, mouth ulcers, and swollen tonsils. The EYE group covers conditions like eye redness, swollen eyes, itchy eyelids, and eye inflammation. Conditions under LIMB comprise edema, hand lump, neck swelling, and foot

[6]definitions provided in the Appendix section

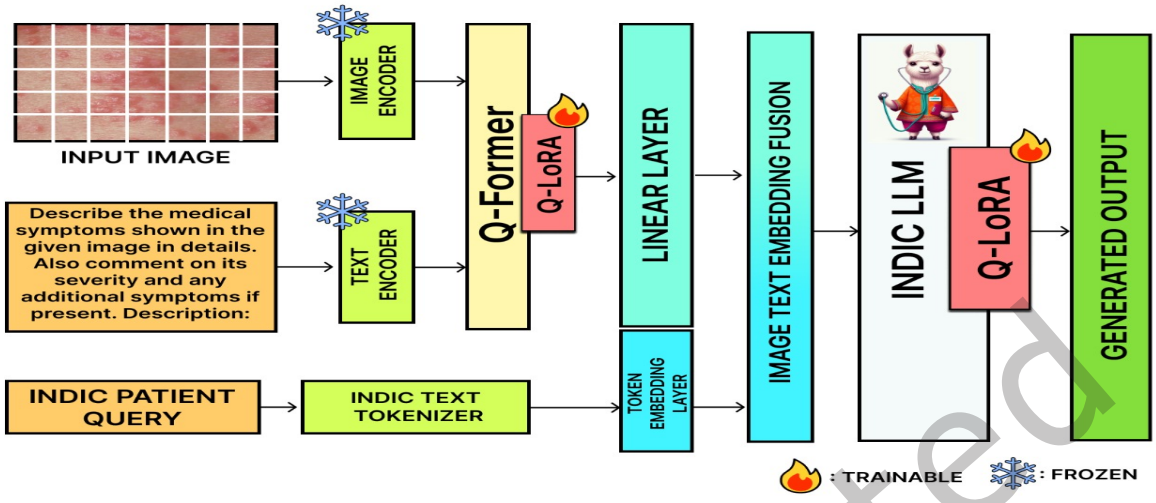


Fig. 3. The diagram illustrates the architectural layout of ***IndicMedQA***, delineating various stages of our proposed model. Additionally, the frozen and trainable layers are shown with appropriate symbols.

swelling. Lastly, the SKIN group includes conditions such as skin rash, cyanosis, dry scalp, skin growth, skin inflammation, and skin dryness.

4 Methodology

This section outlines the problem statement and different components of our proposed model.

4.1 Task Formulation

Our task involves analyzing patient queries in their native language with accompanying medical images provided by them. The analysis comprises two parts: firstly, generating a concise English summary incorporating insights from both modalities and secondly, conducting a medical image analysis to identify the condition depicted in the image along with additional findings and severity assessments. Formally, given an input Indic textual query (T) represented by $T = \{t_0, t_1, \ldots, t_n\}$, where n is the sequence length of the query, and a visual image (V) corresponding to the medical image of the patient, the model's task is to generate a clinically nuanced analysis, represented as

$$M(I, V) = \{S, A\} \tag{1}$$

Where M represents our novel multimodal model, S represents the succinct summary incorporating insights from both modalities, and A represents the medical image analysis encompassing the identification of the medical condition, additional findings, and severity assessments based on the visual cue of the patient. Figure 3 shows a diagrammatic representation of our proposed architecture ***IndicMedQA***. The complete architecture of our proposed model ***IndicMedQA*** has four main components, namely: (i) Visual and Indic text representation, (ii) Bootstrapping and fusion of vision-language representation, (iii) Parameter efficient finetuning of Indic LLM, and (iv) Multitask loss function.

4.2 Visual and Indic Text Representation

We utilize a pre-trained frozen Vision Transformer (ViT-L/14) [49] to extract image embeddings from the visual cues supplied by the patient. The frozen image encoder processes raw images, generating corresponding

image features. These features are then flattened, which is subsequently fed into the Q-former. Patient queries, articulated in their native languages, undergo preprocessing wherein they are tokenized using language-specific Indic tokenizers of Indic LLMs. The tokenized text is then fed into the Indic LLM to generate embeddings.

4.3 Bootstrapping and Fusion of Vision-Language Representation

We leverage the Q-Former model, as introduced in [33], to generate task-specific image embeddings and address the modality gap. We fine-tuned the Q-Former to enhance the mutual information between visual and prompt representations by adding LoRA adapters. This decision was motivated by the aim of enhancing visual language understanding. We use the Indic LLM instead of Q-Former for patient queries because the Q-Former (initialized from English BERT) cannot effectively handle Indic tokens due to mismatched token distributions and limited multilingual training data. The prompt aligns seamlessly with the objective of our task, offering a reliable foundation for producing more task-aware visual embeddings. The prompt used for image-prompt alignment is presented below:

Prompt used for image prompt alignment

Describe the medical symptoms shown in the given image in detail. Also, comment on its severity and any additional symptoms if present.

4.4 Parameter Efficient Fine-tuning of Indic LLM

The multimodal embedding representation is fed into an Indic LLM, a pre-trained language model that has been specifically trained on the Indic languages selected for this task. For this, we chose Sarvam-2B [7], a lightweight yet effective LLM, as it has been extensively trained on Indic languages [8] during both its pretraining and fine-tuning stages, making it well-suited for our multilingual multimodal framework. The preference for Indic LLMs over-generalized ones stems from a significant drawback: the insufficient representation of Indic languages in contemporary model training data, resulting in suboptimal performance across various linguistic contexts. Additionally, Indic LLMs undergo enhancement with language-specific token augmentations to refine text generation and improve language comprehension. In pursuit of a more efficient fine-tuning process for our Indic LLM, we have adopted QLoRA [12], a leading parameter-efficient fine-tuning (PEFT) technique. QLoRA utilizes 4-bit quantization, enabling the fine-tuning process within our existing resource constraints. The prompt used to generate the final analysis is given below:

Prompt used to generate the final analysis

Summarize the given question and comment on the medical symptom along with its severity and additional symptoms that are inferred from the image provided. Question: <patient query>

4.5 Loss Function for Multitask Learning

In our medical query analysis framework, we jointly optimize two generation tasks: (i) patient query summarization and (ii) multimodal clinical analysis based on the medical image. Each task has its own supervision signal in terms of target output sequence. For the summarization task, we compute the token-level cross-entropy loss between the generated summary and the reference summary. Similarly, for the clinical analysis task, we compute the cross-entropy loss between the generated analysis and the corresponding ground truth description.

[7] https://huggingface.co/rachittshah/sarvam-2b-v0.5-GGUF
[8] It is trained on ten Indian languages—Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia (Oriya), Punjabi, Tamil, and Telugu

The overall multitasking objective is defined as:

$$L = \alpha L_1 + \beta L_2, \tag{2}$$

where L_1 denotes the cross-entropy loss for the summarization task and L_2 denotes the cross-entropy loss for the clinical analysis task. We set $\alpha = 0.6$ and $\beta = 0.4$. Since clinically coherent summary generation is the primary objective, a slightly higher weight is assigned to L_1. Empirically, multitask training consistently improved performance across languages compared to single-task training (see Fig. 4), demonstrating the effectiveness of jointly optimizing both objectives.

5 Experiment Results with Analysis

This section details the experimental setup and evaluation of the proposed model ***IndicMedQA***'s performance using automatic, human, and qualitative evaluation.

Model	Hindi		Bengali		Marathi		Gujarati		Tamil		Telugu		Malayalam	
	R1	RL	R1	RL	R1	RL	R1	RL	R1	RL	R1	RL	R1	RL
GPT-4.o-mini	0.53	0.44	0.51	0.42	0.52	0.43	0.50	0.42	0.50	0.42	0.51	0.42	0.51	0.43
GPT-4.o	0.55	0.45	0.54	0.44	0.54	0.44	0.52	0.43	0.54	0.44	0.53	0.43	0.53	0.43
LLAMA-VL	0.40	0.36	0.44	0.37	0.43	0.39	0.38	0.36	0.39	0.36	0.38	0.34	0.41	0.37
PaliGemma	0.32	0.25	0.33	0.25	0.34	0.26	0.35	0.26	0.34	0.26	0.33	0.25	0.34	0.25
Qwen 2 VL	0.53	0.47	0.48	0.43	0.49	0.45	0.52	0.47	0.47	0.43	0.47	0.43	0.44	0.39
Med LLaVA	0.51	0.43	0.50	0.42	0.50	0.41	0.40	0.33	0.48	0.41	0.45	0.38	0.33	0.27
BimediX2	0.55	0.47	0.53	0.45	0.53	0.46	0.53	0.45	0.43	0.37	0.49	0.42	0.51	0.43
IndicMedQA	**0.59**	**0.54**	**0.57**	**0.53**	**0.56**	**0.51**	**0.59**	**0.54**	**0.59**	**0.54**	**0.57**	**0.52**	**0.57**	**0.52**
IndicMedQA(GB R-1)	0.54	0.50	0.52	0.49	0.52	0.46	0.50	0.47	0.54	0.50	0.52	0.48	0.52	0.48
IndicMedQA(GB R-4)	0.50	0.46	0.48	0.45	0.49	0.43	0.47	0.43	0.50	0.46	0.48	0.44	0.48	0.44
IndicMedQA(without Q-Former)	0.53	0.48	0.53	0.46	0.51	0.43	0.55	0.49	0.52	0.49	0.54	0.50	0.52	0.48
IndicMedQA(without Indic LLM)	0.50	0.42	0.48	0.40	0.42	0.36	0.38	0.33	0.28	0.23	0.26	0.20	0.22	0.18

Table 2. Performance comparison between **IndicMedQA** and other state-of-the-art multimodal language models across multiple Indian languages, using ROUGE-1 (R1) and ROUGE-L (RL) as evaluation metrics.The white rows indicate closed source models , the blue rows indicate open weight models . The pink rows are various ablation studies with different gaussian blurs with radius 1 and radius 4 . The red rows shows ablations without QFormer and Indic LLM.

Experimental Setup

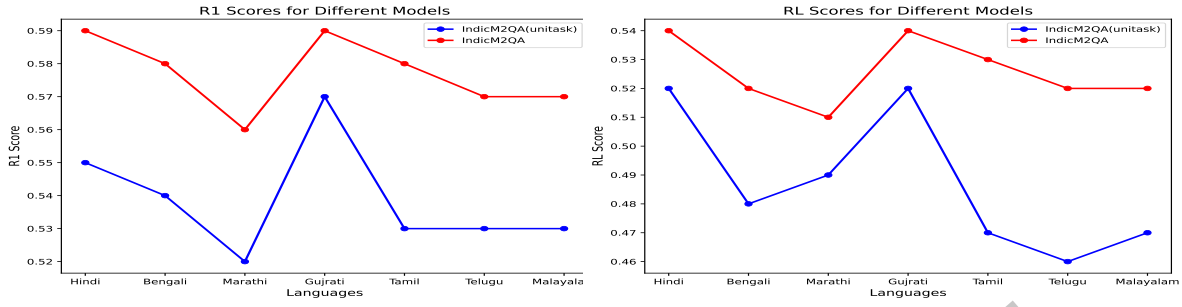
Our experiments were conducted on an NVIDIA RTX 3090 (24GB) GPU, with each model requiring approximately 6 to 7 hours to complete training. We implemented both the baseline models and our proposed IndicMedQA architecture using PyTorch [44], Hugging Face Transformers [56], and Unsloth [9]. The dataset was divided into training (80%), validation (5%), and test (15%) splits [10].

To optimize performance, we set a maximum of 30 epochs[11], a sequence length of 512, and trained the model using the Adam optimizer with a learning rate of 2e-4. For QLoRA fine-tuning, we applied a dropout rate of 0.1, with the LoRA rank set to 64 and the LoRA alpha set to 128. The LoRA adapters were specifically applied to QKVO and MLP layers in Indic LLMs, while for Q-Former, they were used in the Query, Key, and Value layers.

[9]https://unsloth.ai/

[10]This split enables efficient hyperparameter tuning with minimal validation overhead, ensure a robust and low-variance final evaluation using a substantial test set, and maximize training data (especially important in our low-resource Indic medical domain with 21105 samples).

[11]Early stopping was applied with a patience value of 10. It means if the loss does not decrease for 10 consecutive iterations the process will be halted and the weights will get saved.



(a) Comparisons of ROUGE-1 scores

(b) Comparison of ROUGE-L scores

Fig. 4. Comparison of ROUGE-1 and ROUGE-L scores for IndicMedQA variants, with the red curve showing multitask loss training and the blue curve without it.

Additionally, we employed a multilingual decoder from Sarvam, as it has been trained on all Indian languages included in our study, ensuring robust multilingual generation.

Baselines Setup

For baselines, we used both general and medical-specific Vision language models (VLMs). Among general VLMs, we use LLAMA-VL [2], PaliGemma [5], and Qwen 2 VL 8B [53], GPT-4.o-mini and GPT-4.o for each Indic language. Among medical-specific VLMs, we use MedLLAVA [32] and BiMediX2 [40] .

Evaluation Metrics

We evaluated the performance of generated analysis using both automated metrics like ROUGE [35] scores, specifically ROUGE-1 and ROUGE-L scores, and human evaluation metrics like clinical evaluation score, factual recall [1], omission rate [1] and MMFCM Score .[17].

Results and Findings

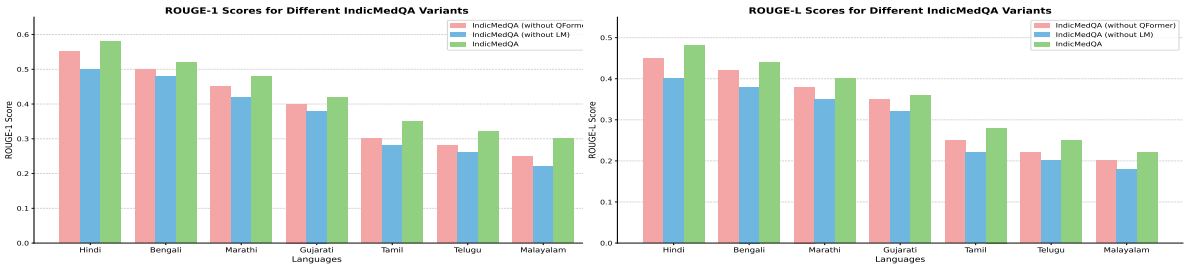
Table-2 shows the comparison of ***IndicMedQA*** with respect to various baselines. We have tried to answer the below research questions based on the results from the automated evaluation in Table-2.

Research Questions: We have addressed the following research questions in this work.

R1.How is the performance of IndicMedQA with respect to the baselines? ***IndicMedQA*** outperforms all baseline models in our evaluation. Among the baselines, QwenVL performed the best among openweight models and GPT 4.0 is the best performing model amongst closed source models , while PaliGemma had the lowest performance. **R2. How is the performance of medical VLMs compared to general-purpose VLMs?** Medical vision-language models (VLMs) generally perform better than general-purpose VLMs in most cases and is comparable to GPT-4.o. Among them, BiMediX2 showed the best results, performing better than MedLLAVA. However, we noticed a common trend: medical VLMs struggled with Dravidian languages such as Tamil, Telugu, and Malayalam. Their performance in these languages was weaker compared to other widely spoken languages, indicating a gap in multilingual medical understanding, especially for these language groups. **R3.How is the performance of VLMs across the languages?** ***IndicMedQA*** achieved state-of-the-art performance across all the baselines considered in this study. One key observation was that vision-language models (VLMs) faced more challenges when processing Dravidian languages such as Tamil, Telugu, and Malayalam. In contrast, their performance was significantly better in other Indian languages, highlighting a disparity in multilingual medical understanding.

Ablation Studies

Impact of freezing various components of IndicMedQA : In our experiments, we compared the performance of ***IndicMedQA*** with and without freezing the Q-Former and the language model decoder, as shown in



(a) ROUGE-1 scores (b) ROUGE-L scores

Fig. 5. Comparison of ***IndicMedQA*** variants with different frozen components: Green represents fine-tuning both LLM and Q-Former, blue represents fine-tuning only Q-Former, and pink represents fine-tuning only LLM.

Figure 5. In most cases, freezing the decoder resulted in a slight drop in performance. However, freezing the Q-Former had minimal impact on performance, and interestingly, in the case of Marathi, freezing the Q-Former led to a slight improvement in performance compared to the standard ***IndicMedQA*** configuration.

Impact of multitask loss: The impact of multitask loss is illustrated in Figure 4, showing a clear trend of performance improvement across all languages in our experiments. Enforcing multitask loss has proven beneficial, as explicitly applying loss to the outputs from both modalities (text and image) significantly boosts overall performance.

Impact of blurring: To measure how much performance drops with blurring, we applied Gaussian blur with radius 1 and radius 4.The pink rows in Table 2 show the results after blurring. We can see that the scores decrease with radius 1 and drop further with radius 4. This confirms that stronger blurring leads to lower performance. However, the results are still higher than those of open-weight models such as LLAMA-VL and PaliGemma.

Impact of Q-Former and IndicLLM: To evaluate the contribution of the fusion module, we removed the Q-Former and replaced it with a simple projection layer that directly maps and combines the visual and textual embeddings, following a fusion strategy similar to that used in the LLaVA architecture. In addition, to assess the impact of the language backbone, we replaced the Indic LLM (Sarvam 2.0) with Mistral-7B, a comparable open-weight model, to isolate the effect of language-specific modeling. The results as shown in red in Table 2 indicate that the choice of the Indic LLM backbone has a substantially larger impact on overall performance than the Q-Former, highlighting the importance of strong multilingual and domain-aligned language modeling for this task.

5.1 Human Evaluation

A team of medical interns, supervised by a medical professional, evaluated 25% of the test data across all languages. They assessed the top-performing models from automated evaluations: QwenVL, IndicMedQA (without Q-Former), and the full IndicMedQA. The evaluation used two main criteria: a 1-to-5 Clinical Evaluation Score (CEval) for relevance, consistency, and fluency, and Medical Fact-Based Metrics like Factual Recall and Omission Rate to measure accuracy against gold-standard annotations. The findings in Figure-6 show that IndicMedQA consistently outperformed the other models in all languages. Hindi and Gujarati yielded the strongest results, while Dravidian languages proved more challenging. The full IndicMedQA model showed the most notable performance improvements over its ablated version in Tamil and Telugu.

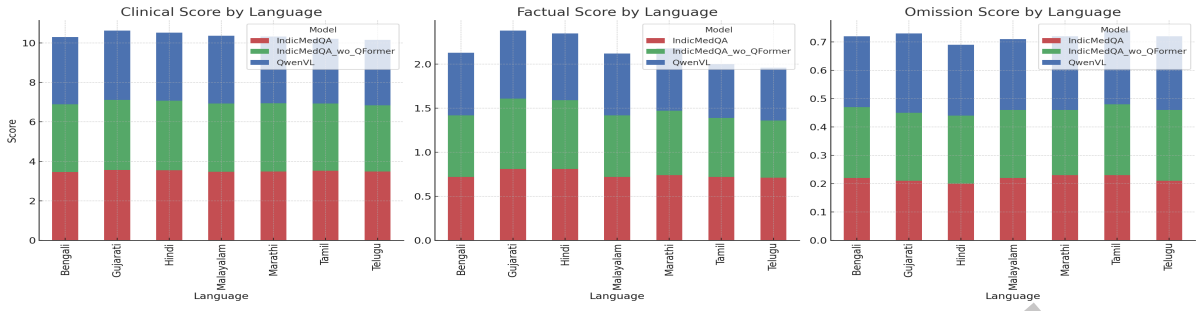


Fig. 6. Human evaluation scores of our proposed model with respect to the baseline. Our model scored higher across all metrics in all languages across all metrics like clinical eval scores , omission rate and factual recall.

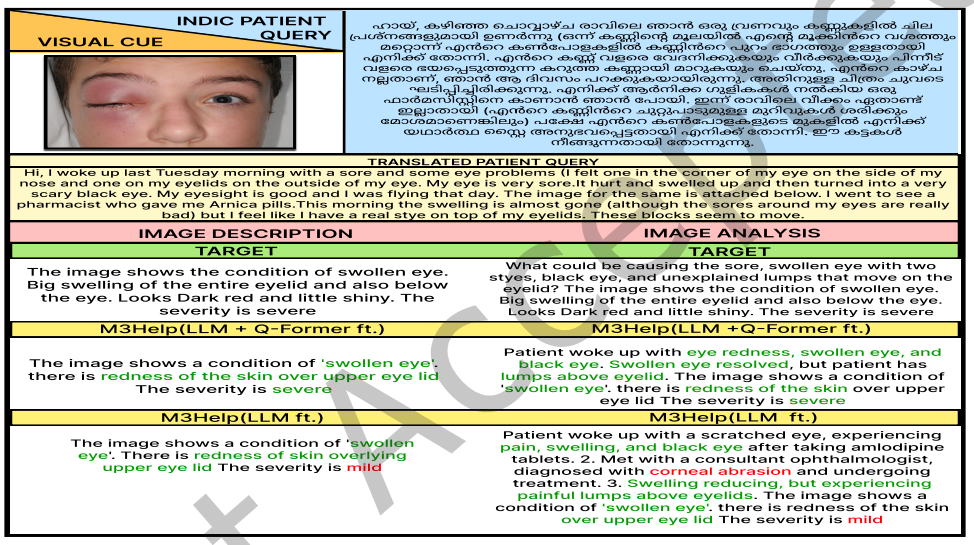


Fig. 7. Qualitative Analysis of the generated analysis in both configurations of ***IndicMedQA***. Green denotes the correct medical facts, and red means the hallucinated text.

6 Qualitative Analysis

Figure 7 highlights a clear pattern in the effectiveness of different configurations of ***IndicMedQA***—one with only LLM fine-tuning, referred to as M3Help(LLMft), and the other with both Q-Former and LLM fine-tuning, denoted as M3Help(LLM + Q-Former ft). The evaluation of annotated samples, as shown in Figure 7, was conducted with valuable input from our collaborating medical professionals. In the figure, green-highlighted phrases indicate correct analyses of patient queries and images, while red-highlighted phrases denote incorrect interpretations. While in most of the samples there were very minute changes in generation but in very rare cases we identified the following key insights: Both configurations of ***IndicMedQA*** successfully captured most medical facts, but the choice of model configuration significantly influenced the quality of generated outputs. ***IndicMedQA*** (with Q-Former fine-tuning) exhibited fewer hallucinations and better formatting, whereas ***IndicMedQA*** (LLM only)

tended to produce longer outputs, which, in some cases, introduced hallucinated or incorrect tokens, as highlighted in red.

7 Risk Analysis

While visual cues are vital for clinical analysis, over-reliance on them poses risks like overlooking textual details or being misled by poor-quality images. Our findings show that ***IndicMedQA*** effectively balances visual and textual information, which reduces hallucinations. Nevertheless, in high-stakes medical situations, human expert validation remains indispensable to ensure analyses are clinically meaningful and trustworthy

8 Conclusion

In this study, we introduce the task of multimodal medical query analysis for seven Indian languages, allowing patients to use native language text and medical images to express health concerns. To support this, we created a high-quality, multilingual, and multimodal dataset using a semi-automated, human-validated translation pipeline. We propose ***IndicMedQA***, a lightweight multimodal transformer that combines a static image encoder with an Indic LLM for better language comprehension. The model uses a multitask loss function to improve generation quality from both text and image inputs. Through extensive evaluations, including human and automated analysis, we ensure the model's reliability and safety. The overall goal of ***IndicMedQA*** is to assist healthcare professionals by enhancing doctor-patient communication for the Indian population.

9 Limitations

Our work exhibits several notable limitations, outlined as follows:

(1) The dataset's current scope, with 21,105 samples covering only 18 common medical conditions, restricts the model's generalizability. A significant expansion with a wider variety of conditions is essential before the model can be safely deployed in real-world clinical scenarios.

(2) By focusing on only seven Indian languages, the model's applicability is limited and does not cover the full linguistic diversity of the country. This presents a significant challenge for broader, nationwide deployment.

(3) The model is tailored for image data, but the medical field is increasingly using videos and voice recordings. To remain relevant and effective, future work must expand the model to handle these additional data modalities.

10 Resources

The code and the dataset are made available in this github page: IndicMedQA.

11 Ethics Statement

Ensuring ethical rigor is very important when working with multimodal medical AI systems, especially in terms of privacy, safety, and responsible use. In this study, we worked closely with medical professionals throughout the process—from selecting the dataset source to validating the results. To maintain high ethical standards and protect patient privacy, all data were carefully reviewed, and any images that could reveal private information were removed during preprocessing. We used the publicly available MMQS dataset in our work. Since we did not collect new images or add new labels—and only analyzed the existing data, which had already been annotated by medical experts—no additional data handling or annotation was involved apart from the severity analysis which was only a minor addition to the already existing dataset . Because of these reasons, our work was exempt from the Institutional Review Board (IRB) oversight.

12 Acknowledgements

This work was funded by the POWER scheme of the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India. Akash Ghosh and Sriparna Saha gratefully acknowledge this

support. This work was also supported by computational resources provided by the Aryabhatta Supercomputing Centre (ASC) at the Indian Institute of Technology Patna, established under the National Supercomputing Mission (NSM), Government of India.

References

[1] Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An Investigation of Evaluation Metrics for Automated Medical Note Generation. *arXiv preprint arXiv:2305.17364* (2023).

[2] Meta AI. 2024. Llama 3.2-11B-Vision. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision Accessed: March 18, 2025.

[3] Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine* 155 (2024), 102938.

[4] Abhinand Balachandran. 2023. Tamil-Llama: A New Tamil Language Model Based on Llama 2. *arXiv preprint arXiv:2311.05845* (2023).

[5] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726* (2024).

[6] Beech Burns, James Heilman, Shana Kusin, Laura Chess, and Mary Elizabeth Tanski. 2022. Turn that frown upside down: implementation of a visual cue improves communication during emergency department to inpatient hand-offs. *BMJ Open Quality* 11, 4 (2022), e002078.

[7] 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* (2023).

[8] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

[9] Sarmistha Das, Akash Ghosh, Sriparna Saha, Koustava Goswami, and Joseph KJ. 2025. Reasoning and Planning for Multimodal Large Language Models: A Multilingual and Cross-Domain Exploration. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 14342–14343.

[10] Sarmistha Das, Vaibhav Vishal, Syed Ibrahim Ahmad, Manish Gupta, and Sriparna Saha. 2026. FIND: Toward Multimodal Financial Reasoning and Question Answering for Indic Languages. *arXiv preprint arXiv:2605.13330* (2026).

[11] Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal* 6, 2 (2019), 94.

[12] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems* 36 (2024).

[13] Katherine Douglass, Lalit Narayan, Rebecca Allen, Jay Pandya, and Zohray Talib. 2021. Language diversity and challenges to communication in Indian emergency departments. *International Journal of Emergency Medicine* 14 (2021), 1–8.

[14] Jay Gala, Pranjal A Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. *arXiv preprint arXiv:2305.16307* (2023).

[15] Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M Khapra, Raj Dabre, Rudra Murthy, Anoop Kunchukuttan, et al. 2024. Airavata: Introducing Hindi Instruction-tuned LLM. *arXiv preprint arXiv:2401.15006* (2024).

[16] Soumya Suvra Ghosal, Vaibhav Singh, Akash Ghosh, Soumyabrata Pal, Subhadip Baidya, Sriparna Saha, and Dinesh Manocha. 2025. Relic: Enhancing reward model generalization for low-resource indic languages with few-shot examples. *arXiv preprint arXiv:2506.16502* (2025).

[17] Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2023. Clipsyntel: Clip and llm synergy for multimodal question summarization in healthcare. *arXiv preprint arXiv:2312.11541* (2023).

[18] Akash Ghosh, Arkadeep Acharya, Prince Jha, Aniket Gaudgaul, Rajdeep Majumdar, Sriparna Saha, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024. MedSumm: A Multimodal Approach to Summarizing Code-Mixed Hindi-English Clinical Queries. *arXiv preprint arXiv:2401.01596* (2024).

[19] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214* (2024).

[20] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, and Setu Sinha. 2024. Healthalignsumm: Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 11546–11560.

[21] Akash Ghosh, Tajamul Ashraf, Rishu Kumar Singh, Numan Saeed, Sriparna Saha, Xiuying Chen, and Salman Khan. 2026. CarePilot: A Multi-Agent Framework for Long-Horizon Computer Task Automation in Healthcare. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition. 9695–9705.
[22] Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457* (2025).
[23] Akash Ghosh, Raghav Jain, Anubhav Jhangra, Sriparna Saha, and Adam Jatowt. 2025. A survey on medical document summarization: From machine learning techniques to large language models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 15, 4 (2025), e70045.
[24] Akash Ghosh, Srivarshinee Sridhar, Raghav Kaushik Ravi, Muhsin Muhsin, Sriparna Saha, and Chirag Agarwal. 2025. CLINIC: Evaluating Multilingual Trustworthiness in Language Models for Healthcare. *arXiv preprint arXiv:2512.11437* (2025).
[25] Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. 2025. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 3779–3787.
[26] Sofia Jamil, Sriparna Saha, and Rajiv Misra. 2026. Enhancing adverse drug event extraction and summarization for cancer drugs through large language models. *Journal of Biomedical Informatics* (2026), 105009.
[27] Raghvendra Kumar, Deepak Prakash, Sriparna Saha, and Shubham Sharma. 2024. Indicbart alongside visual element: multimodal summarization in diverse indian languages. In *International Conference on Document Analysis and Recognition*. Springer, 264–280.
[28] Raghvendra Kumar, Devankar Raj, and Sriparna Saha. 2026. BhashaSutra: A Task-Centric Unified Survey of Indian NLP Datasets, Corpora, and Resources. *arXiv preprint arXiv:2604.18423* (2026).
[29] Raghvendra Kumar, Mohammed Salman SA, Aryan Sahu, Tridib Nandi, Pragathi YP, Sriparna Saha, and Jose G Moreno. 2025. COSMMIC: Comment-Sensitive Multimodal Multilingual Indian Corpus for Summarization and Headline Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8728–8748.
[30] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373* (2024).
[31] Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics* 11, 4 (2023), 1006.
[32] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* 36 (2023), 28541–28564.
[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
[34] Kangenbei Liao, Qianlong Liu, Zhongyu Wei, Baolin Peng, Qin Chen, Weijian Sun, and Xuanjing Huang. 2020. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning. *arXiv preprint arXiv:2004.14254* (2020).
[35] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
[36] John Maher. 1987. English as an international language of medicine. 283–284 pages.
[37] Arijit Maji, Raghvendra Kumar, Akash Ghosh, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, Sriparna Saha, et al. 2025. Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 1289–1313.
[38] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics* 26, 12 (2022), 6070–6080.
[39] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*. PMLR, 353–367.
[40] Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, et al. 2024. BiMediX2: Bio-Medical EXpert LMM for Diverse Medical Modalities. *arXiv preprint arXiv:2412.07769* (2024).
[41] Shuai Niu, Jing Ma, Liang Bai, Zhihua Wang, Li Guo, and Xian Yang. 2024. EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Information Fusion* 102 (2024), 102069.
[42] Eric Onyame, Akash Ghosh, Subhadip Baidya, Sriparna Saha, Xiuying Chen, and Chirag Agarwal. 2026. CURE-Med: Curriculum-Informed Reinforcement Learning for Multilingual Medical Reasoning. *arXiv preprint arXiv:2601.13262* (2026).
[43] OpenAI. 2023. GPT4. (2023). https://openai.com/research/gpt-4.
[44] 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 (2019).
[45] Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMediX: Bilingual Medical Mixture of Experts LLM. *arXiv preprint arXiv:2402.13253* (2024).
[46] Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*. 612–618.

[47] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925* (2024).
[48] 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* (2024).
[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[50] Nazib Sorathiya, Chuan-An Lin, Daniel Chen Daniel Xiong, Scott Zin, Yi Zhang, He Sarina Yang, and Sharon Xiaolei Huang. 2021. Multi-turn dialog system on single-turn data in medical domain. *arXiv preprint arXiv:2105.12887* (2021).
[51] Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari. 2022. Dr. can see: towards a multi-modal disease diagnosis virtual assistant. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 1935–1944.
[52] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. Beyond Direct Diagnosis: LLM-based Multi-Specialist Agent Consultation for Automatic Diagnosis. *arXiv preprint arXiv:2401.16107* (2024).
[53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
[54] Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, et al. 2024. Apollo: A Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. *arXiv preprint arXiv:2403.03640* (2024).
[55] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 201–207.
[56] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
[57] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454* (2023).
[58] Lijian Xu, Hao Sun, Ziyu Ni, Hongsheng Li, and Shaoting Zhang. 2024. MedViLaM: A multimodal large language model with advanced generalizability and explainability for medical data understanding and generation. *arXiv preprint arXiv:2409.19684* (2024).
[59] Ying Zhu, Yameng Li, Yuan Cui, Tianbao Zhang, Daling Wang, Yifei Zhang, and Shi Feng. 2023. A Knowledge-Enhanced Hierarchical Reinforcement Learning-Based Dialogue System for Automatic Disease Diagnosis. *Electronics* 12, 24 (2023), 4896.