Visual Narration in Primary Healthcare for COMPREHENSIV: A Multimodal Approach Using LLaVA and Whisper

S. Ramakrishnan, S. Munuswamy, S. Anand

Contact: ramakrishnanshreya@gmail.com

INTRODUCTION

Universal Health Coverage (UHC) requires innovative solutions to improve healthcare accessibility and patient education. This study explores the potential of AI models in enhancing primary healthcare through multimodal interactions for images collected through COMPREHENSIV head-to-toe app.

AIMS AND OBJECTIVES

Aim- To evaluate the performance of multimodal AI framework in primary healthcare applications.

1. Evaluate AI models' performance in Visual Question Answering (VQA)
2. Assess multimodal translation capabilities in local Indian languages
3. Improve patient education for diseases like leprosy and lymphatic filariasis

METHODOLOGY

Nine AI models were assessed:

- VQA Models: Gemini, Microsoft Phi-3, Llava v-1.6, CogVLM2, Qwen-VL-Plus, MiniGPT-4, Claude, Idefics2, Custom Llava 1.5 + Whisper
- Translation Tools: IndicTrans2, Indic-TTS

Two primary tasks were designed: Image-based VQA and Local Language Translation

KEY RESULTS

Top-performing VQA Models:

- Llava-v1.5 + Whisper
- Llava-v1.6-34b
- Claude
- Microsoft Phi-3

Translation Capabilities:

- 23 Indian languages for text translation
- 13 Indian languages for audio translation

CONCLUSION

Multimodal Large Language Models (LLMs) demonstrate significant potential in generating patient education materials and translations.

Future research should focus on:

- Optimizing model selection
- Exploring frameworks like RouteLLM for efficient task routing
- Improving healthcare accessibility through AI

POLICY RECOMMENDATION

The government should establish a dedicated think tank or research group to create open-source datasets of prompt-output pairs for fine-tuning base models to cater to various healthcare needs. They should also hire local language experts to expand patient education materials to non-English speakers.