



21st

WIMC

WARSAW INTERNATIONAL MEDICAL CONGRESS

ABSTRACT

BOOK

Hybrid Session

Session Coordinators:

Piotr Gronowski

Gosia Michalska

Julia Krawczyk

Mikołaj Bańkowski

Oliwia Dudek

Weronika Łosiewicz

CheXGPT: A multimodal reasoning language model for chest X-ray labelling and report generation.

Authors

Sai Keertana Devarapalli, Mohammed Belal, Abdulla Hourani

Trustee of the paper

Abdulla Hourani

Affiliations

Medical University of Warsaw

Khalifa University

Klinika Nefrologii, Dializoterapii i Chorób Wewnętrznych Uniwersyteckie Centrum Kliniczne
WUM

Introduction

Automated chest X-ray (CXR) report labeling enables large-scale imaging research, yet widely used baselines such as CheXbert provide fixed behavior, are limited to the 14 CheXpert labels, and do not support multimodal or multilingual report generation.

Aim of the study

To develop a single compact multimodal model that (1) achieves state-of-the-art performance and surpasses CheXbert on CheXpert-style 14-label report classification, (2) extends supervision to an expanded multi-label taxonomy beyond 14 labels, and (3) generates image-grounded, label-faithful reports across languages.

Materials and methods

We benchmarked many open-source LLMs (Qwen, Llama, Gemma, Phi, GPT-OSS) on CheXpert, MIMIC-CXR, ReXGradient-160K, PadChest (Spanish), and CASIA-CXR (French) using a unified prompt and CheXpert-compatible output schema. We ablated zero-shot, few-shot, and dynamic few-shot retrieval (pool=50), and evaluated quantized inference (INT4/INT8). We then fine-tuned and post-trained Qwen3-V-4B on ~1.3M radiology reports and ~450k CXR images with mixed objectives: supervised multi-label classification (CheXpert-14 plus additional labels) and instruction-style image-to-report generation, including multilingual training. Performance was evaluated on held-out test sets; 95% confidence intervals were obtained by bootstrapping.

Results

Qwen3-V-4B achieved macro-F1 0.93 (0.91–0.95) for 14-label classification, exceeding the CheXbert baseline in our benchmark. For generated reports, label similarity to gold labels reached macro-F1 0.86 (0.84–0.88) with BLEU 0.24, indicating improved label faithfulness. Dynamic few-shot retrieval improved recall for rare and ambiguous labels—most notably Lung Lesion and Pleural Other—and multilingual prompting improved robustness on non-English reports. Quantization preserved performance with reduced latency.

Conclusions

A post-trained 4B multimodal LLM can unify CheXpert-style labeling and multilingual, image-grounded report generation, providing a practical, efficient alternative to CheXbert with tunable prompt- and hardware-level trade-offs.