Med-Gemini and Single-Expert Model Trade-offs in Edge Medical Inference

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What are the trade-offs between inference latency and accuracy when deploying Med-Gemini for multimodal medical tasks compared to specialized single-expert models on edge devices. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Towards Better Dental AI: A Multimodal Benchmark and Instruction Dataset for Panoramic X-ray Analysis. Research question: What are the trade-offs between inference latency and accuracy when deploying Med-Gemini for multimodal medical tasks compared to specialized single-expert models on edge devices?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

16 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.4/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
MMOral-Bench consists of 500 closed-ended and 600 open-ended QA pairs with 100 images.	×	0.04
Images in MMOral-Bench are selected from the dataset proposed by Hoang Viet Do [19] due to clearer and more reliable acq	×	0.03
MMOral-Bench covers five different clinically grounded dimensions: Teeth, Patho, HisT, Jaw, SumRec.	×	0.04
For closed-ended questions in MMOral-Bench, accuracy is used as the evaluation metric.	×	0.06
For open-ended questions in MMOral-Bench, GPT-4-turbo is used to assign a score ranging from 0 to 1 based on each sample	×	0.04
MMOral-Bench is integrated into the standard VLMEvalKit [22] framework.	×	0.03
Zero-shot evaluations are conducted across 64 LVLMs on MMOral-Bench.	×	0.10
GPT-4o achieves 41.45% overall performance on MMOral-Bench.	✓	0.15
MMOral-Attribute dataset contains 904k entries describing the category, position, and correlation of anatomical structur	×	0.09
MMOral-Report dataset contains 41k entries with two types of textual descriptions for each panoramic X-ray image: ground	×	0.08
MMOral-VQA dataset contains 965k entries with two types of visual question answering: closed-ended QA and open-ended QA.	×	0.07
MMOral-Chat dataset contains 296k entries with multi-turn conversations between the assistant and a person asking questi	×	0.05

References

- http://arxiv.org/abs/2506.17337v4
- http://arxiv.org/abs/2503.04521v2
- http://arxiv.org/abs/2509.09254v1