

Comparative Analysis of MedGemma and Gemini 2.5 Pro for Multi-Label Chest X-ray Classification

Osai Osaigbovo
OsaiOsaigbovo@my.unt.edu

Thuan Luong Nguyen
Thuan.Nguyen@unt.edu

University of North Texas
1155 Union Circle, Denton, TX 76203-5017 - USA

Abstract

Multimodal large language models (MLLMs) are increasingly evaluated for chest radiograph interpretation under the assumption that medical domain specialization is essential for diagnostic performance. We systematically evaluated the domain-specialized MedGemma (27B) against the general-purpose Gemini 2.5 Pro on a three-label classification task using 13,189 chest radiographs from the CheXpert dataset. The evaluation cohort contained positive findings for Pneumonia ($n = 2,756$; 20.9%), Atelectasis ($n = 3,338$; 25.3%), and Fracture ($n = 642$; 4.9%). Performance diverged substantially by pathology and evaluation metric. While MedGemma achieved higher overall accuracy on Pneumonia (75.4% vs. 52.9%), this was driven by a conservative specificity bias resulting in severe diagnostic omission (14.3% recall vs. 72.0% for Gemini). On Atelectasis, Gemini attained superior classification balance ($F1 = 0.474$ vs. 0.104). Both models exhibited severe performance degradation on rare findings, each falling below 5.0% recall for Fracture. Across the aggregate cohort, MedGemma held a marginal lead in raw accuracy (63.8% vs. 60.4%), while Gemini held a higher mean balanced accuracy (0.565 vs. 0.510, i.e., the average of sensitivity and specificity per label) and Macro F1 (0.316 vs. 0.134; McNemar's test, $p < 0.001$). We flag that our originally reported "AUC" was in fact this balanced-accuracy statistic, not a true ranking-based ROC-AUC; we correct the label here and throughout, and we do not claim either model shows superior discrimination ability.

These results challenge the assumption that domain-tuned MLLMs uniformly outperform broad frontier architectures on radiographic tasks, highlighting critical sensitivity trade-offs in specialized models and underscoring that neither pipeline is currently viable for autonomous diagnostic deployment.

Keywords: Multimodal Large Language Models (MLLMs), Medical Image Classification, Multi-Label Medical Classification, Chest Radiography, MedGemma, Domain Specialization.

Comparative Analysis of MedGemma and Gemini 2.5 Pro for Multi-Label Chest X-ray Classification

Osai Osaigbovo and Thuan Luong Nguyen

1. INTRODUCTION

Chest radiography remains one of the most widely used imaging modalities worldwide, yet inter-observer variability, heterogeneous acquisition conditions, and rising clinical workload challenge its interpretation. AI-based diagnostic support has shown promise: convolutional neural networks (CNNs) such as DenseNet-121 achieve strong pathology-level performance on large annotated datasets (Irvin et al., 2019). CNNs, however, remain limited in their capacity to integrate clinical context, adapt to novel query types, or produce semantically rich outputs, capabilities increasingly expected in clinical workflows.

Multimodal large language models introduce a class of systems capable of jointly interpreting visual and textual signals, with demonstrated ability on chest X-ray question answering (Pal et al., 2025), report generation (Al-Hamadani, 2025), and multimodal lesion characterization (Kim & Kim, 2025; Salbas & Buyuktoka, 2025). Google's MedGemma is a domain-specialized architecture trained explicitly on medical image-text pairs; Gemini 2.5 Pro is a state-of-the-art generalist model built for broad multimodal reasoning across clinical and non-clinical domains (Zambrano Chaves et al., 2025; Yang et al., 2025). Direct head-to-head evaluations of specialized and generalist MLLMs on identical chest X-ray subsets for multi-label classification remain limited. Zambrano Chaves et al. (2025) benchmark a small, domain-tuned radiology model against generic baselines and an automated evaluation metric, but do not evaluate a frontier generalist MLLM on the same images. Ahmed et al. (2025) survey whether LLMs can challenge CNNs in medical image analysis broadly, but stop short of a matched, image-level comparison between a domain-specialized and a generalist MLLM. And cross-study comparisons that do exist, such as pairing MedGemma's reported 0.82 F1 on MIMIC-CXR against a generalist model's accuracy on a separate visual-question-answering benchmark, draw on different datasets, tasks, and prompting strategies, confounding any inference about relative model quality (see Section 2.2).

This paper conducts an image-matched comparison of MedGemma and Gemini 2.5 Pro on a clinically relevant, multi-label CheXpert classification task focused on Pneumonia, Atelectasis, and Fracture. We use the word "matched" deliberately narrowly. Both models were evaluated on the same 13,189-image intersection of their respective output sets, which controls for one major confound (both models see identical pixels) without controlling for a second (how each model's original evaluation pool was constructed). We treat this second point as an open methodological limitation rather than a solved problem, and we return to its implications throughout the paper rather than confining it to a single disclosure paragraph.

This paper aims to answer the following research questions:

1. RQ1, Domain Advantage: Within this bounded comparison, does MedGemma's domain-specific training confer measurable, consistent advantages over Gemini 2.5 Pro on identical images?
2. RQ2, Per-Label Performance: How do the models compare per pathology on accuracy, balanced accuracy, precision, recall, and F1, and where does this divergence align, or fail to align, with the incomplete-knowledge account of LLM behavior introduced in Section 2.4?
3. RQ3, Clinical Suitability: What do the observed performance profiles, and their measurement limitations, imply for safe MLLM deployment in diagnostic radiology?

Our results show a selective, not universal, advantage for domain-specific training: MedGemma outperforms Gemini substantially on Pneumonia, Gemini outperforms MedGemma on Atelectasis, and both perform comparably and poorly on Fracture. These findings complicate the assumption of broad domain superiority. A further dimension concerns the epistemic properties of generative AI systems: emerging theoretical work proposes that LLMs can exhibit an uneven relationship between aggregate competence and sub-task reliability, a pattern termed the "Wharton Paradox" (Nguyen, 2026) after Terwiesch's (2023) observation that early GPT-4 could pass a Wharton MBA exam while failing narrower,

procedurally exact tasks. We use this framework as an interpretive lens rather than an independently validated explanation: its proposing paper shares an author with this study, and we conducted no test capable of adjudicating between its proposed causal mechanism and alternative explanations, including the measurement artifacts described below.

This paper's contributions are four-fold. First, we present a controlled, image-matched evaluation of MedGemma 27B and Gemini 2.5 Pro on a multi-label CXR task, with explicit disclosure of the sampling asymmetry underlying the "matched" claim. Second, we report per-pathology divergence and identify where low-recall findings are entangled with a labeling convention rather than cleanly attributable to model capability. Third, we discuss our results in relation to the Wharton Paradox framework as one candidate interpretation rather than a confirmed account. Fourth, we identify and correct a metric-computation error in our own initial analysis, in which a headline "AUC" comparison was in fact balanced accuracy, and we use that correction to argue for a specific evaluation habit in medical AI research: decompose a metric before it becomes a headline finding. We do not claim these findings establish that domain specialization fails as a strategy, only that it did not produce uniform gains in this bounded comparison.

For the IS audience, and specifically for readers in the Data Analytics, Pedagogy, and Practice track, these results speak to a familiar and recurring decision: whether to trust an aggregate benchmark number when adopting or teaching about a generative AI tool. The pattern documented here, that MedGemma's headline accuracy advantage evaporates once recall is examined, and that Gemini's apparent AUC advantage was itself a mislabeled statistic once decomposed (Section 3.3), echoes what Terwiesch (2023) documented when an early GPT-4 could pass a Wharton MBA exam yet stumble on narrower, procedurally exact questions a first-year student would get right: strong aggregate performance can coexist with a specific, exploitable weakness. For an IS practitioner evaluating a vendor's AI tool, or an instructor designing an assignment or an assessment policy around one, the actionable takeaway is the same in both settings: do not adopt a tool, or design an assessment, on the strength of a single reported metric. Ask which sub-tasks that metric averages over, and check performance on the sub-task that matters most for the decision at hand, exactly as Section 4.3 does for Pneumonia, Atelectasis, and Fracture individually rather than resting on the two models' aggregate scores.

2. LITERATURE REVIEW

2.1 Achievements and Capabilities of Generative AI and Large Language Models

The past half-decade has seen extraordinary progress in large language models and their multimodal extensions, and the relevant question for this paper is not whether that progress is real but whether it transfers evenly to a bounded, high-stakes clinical task. Foundation models trained on web-scale corpora show strong zero- and few-shot generalization across translation, summarization, question answering, and code generation (Brown et al., 2020), consistent with the scaling hypothesis (Kaplan et al., 2020). In biomedicine, Google's Med-PaLM 2 approached human-clinician performance on USMLE-style questions, reaching up to 86.5% accuracy on the MedQA benchmark (Singhal et al., 2023); its successor MedGemma extends this to the multimodal domain, integrating visual perception of medical images with biomedical language understanding and reporting F1 = 0.82 on MIMIC-CXR multi-label classification (Yang et al., 2025). That figure is the benchmark this paper's comparison is implicitly measured against, and Section 2.2 explains why it is not directly comparable to the results reported below.

General-purpose MLLMs show comparable ambition in medical settings. Gemini 2.5 Pro reportedly achieves 93% sequence accuracy on brain MRI classification (Salbas & Buyuktoka, 2025) and competitive performance on chest X-ray visual question answering (Pal et al., 2025). Zambrano Chaves et al. (2025) report strong Gemini-family performance on report generation in clinical workflow integration, and comparative work against GPT-4o and Claude has begun establishing a competitive landscape among frontier MLLMs on radiology tasks (Kim & Kim, 2025). MLLMs additionally offer interaction modes unavailable to task-specific CNNs: natural language rationales, integration of clinical context, and adaptation to novel task specifications without retraining (Urooj, 2025). These capabilities have value independent of raw classification accuracy, and the two should not be conflated when assessing deployment readiness.

2.2 From CNNs to Specialized and Generalist Medical MLLMs

Automated chest radiograph interpretation has long relied on CNNs, which achieve strong performance on benchmarks such as CheXpert, MIMIC-CXR, and NIH CXR14 (Irvin et al., 2019; Rajpurkar et al., 2017). DenseNet-121 is a de facto baseline for multi-label thoracic abnormality detection, reporting 0.834 to 0.917 AUC on CheXpert Pneumonia and Atelectasis subsets (Irvin et al., 2019). These figures come from the original benchmark rather than a DenseNet-121 trained and evaluated on our 13,189-image sample, so they are not directly comparable to the numbers below; a fine-tuned baseline on our exact sample remains future work. CNNs also lack reasoning ability, contextual modeling, and natural language interaction, motivating interest in medical-domain LLMs.

MedGemma represents an attempt to close this gap through domain-adaptive pretraining, medical vocabulary alignment, and med-specific tokenizer extensions, with technical reports claiming state-of-the-art performance across chest X-ray classification, report generation, and visual question answering (Yang et al., 2025). Generalist models such as Gemini 2.5 Pro pursue the same clinical tasks through broad multimodal reasoning rather than domain-specific tuning (Zambrano Chaves et al., 2025). Existing comparative work is suggestive but not conclusive: one study reports LLaVA-Rad outperforming GPT-4V on MIMIC-CXR report generation, and Gemini-family models reportedly reach 72% accuracy on CXR visual question answering versus MedGemma's 0.82 F1 on MIMIC-CXR classification, but these figures come from different studies, datasets, and tasks, and cannot be read as a head-to-head result (Ahmed et al., 2025). This absence of a genuinely matched comparison is the gap this paper addresses. A further, largely unraised confound compounds the comparison problem: MedGemma's 0.82 F1 figure was obtained on MIMIC-CXR, and this study evaluates it on CheXpert, a different institution's imaging population and labeling process. Pooch et al. (2019) show that CNN-based chest-radiograph classifiers can lose substantial performance under exactly this kind of domain shift between training and evaluation datasets. The same risk applies to a domain-tuned MLLM evaluated outside its reported benchmark, and it is a plausible partial explanation for MedGemma's low recall here that is independent of, and additive to, the labeling-convention and sampling issues discussed in Section 3.1.

2.3 Chest X-ray Classification as a Benchmark Task

Multi-label CXR classification is a demanding benchmark because it requires several independent, low-prevalence judgments per image rather than a single global call. Pneumonia, Atelectasis, and Fracture were selected on two grounds together, not clinical reasoning alone: each is well represented in this sample (Pneumonia $n = 2,756$, 20.9%; Atelectasis $n = 3,338$, 25.3%; Fracture $n = 642$, 4.9% of the 13,189-image matched cohort), and each spans a distinct etiology, infectious, mechanical, and traumatic, carrying a different cost when missed: delayed Pneumonia treatment, undetected Atelectasis affecting post-operative monitoring, and an unidentified Fracture with direct orthopedic consequences. Fracture's markedly lower prevalence is itself relevant to interpreting its results in Section 4.2. A model performing well in aggregate across fourteen CheXpert labels can still fail badly on any one, which is why per-label, not just per-model, evaluation is the appropriate unit of analysis for a deployment decision, and why this study reports both.

2.4 Limitations of MLLMs

The literature documents structural limitations that constrain clinical applicability, reflecting how LLMs acquire and store knowledge rather than incidental engineering gaps.

Incomplete and probabilistic knowledge. LLMs acquire knowledge through statistical pattern recognition over large corpora rather than structured, verifiable acquisition (Nguyen, 2026), yielding representations that are probabilistic, aggregated, and incomplete relative to structured knowledge bases or clinical guidelines. Gaps and distributional biases in training data propagate into model knowledge without a built-in self-correction mechanism (Urooj, 2025). This has particular salience in medicine, where a model may encode accurate statistical associations for common presentations while remaining largely uninformed about rarer conditions or atypical presentations relevant to individual cases. MLLMs have also been documented to hallucinate, for example, fabricating "Susac syndrome" as an MRI finding (Salbas & Buyuktoka, 2025), and to show prompt-sensitive performance (Hassanein et al., 2025).

The Wharton Paradox as an interpretive lens. One proposed account of this incompleteness is the "Wharton Paradox," describing cases where LLMs appear competent on broad, aggregate assessments while failing specific, procedurally exact sub-tasks (Nguyen, 2026), following Terwiesch's (2023) observation that early GPT-4 could pass a Wharton MBA exam while making errors an expert would not. We use this framework in this paper's discussion as a candidate lens for interpreting our own accuracy/recall divergence, and we flag two things about its epistemic status up front. First, Terwiesch's original finding concerns business-school exam performance, not medical image classification; the transfer to radiology is an analogy, not a demonstrated generalization, and the disanalogies in task structure, error costs, and ground-truth regime are substantial. Second, the framework's originating paper (Nguyen, 2026) shares an author with the present study; where we describe our results as consistent with, or an instance of, the Wharton Paradox, that language should be read as our own interpretation rather than as independent third-party confirmation. None of this paper's empirical findings depend on the Wharton Paradox framework being correct. The accuracy/recall divergence reported in Section 4, the balanced-accuracy correction in Section 3.3, and the per-pathology deployment implications in Section 4.3 all stand on the reported data alone; the framework is offered only as one candidate interpretation of why that divergence might arise, not as a load-bearing premise for the paper's practical conclusions.

3. METHODOLOGY

This study conducts a controlled, head-to-head evaluation of MedGemma, domain-specialized, and Gemini 2.5 Pro, generalist, on a multi-label chest X-ray task focused on Pneumonia, Atelectasis, and Fracture.

3.1 Dataset, Sampling, and the Limits of Matching

All evaluations draw on the CheXpert dataset (Irvin et al., 2019), comprising 224,316 expert-annotated chest radiographs with uncertainty-labeled annotations across 14 pathologies, transferred from Stanford's AIMI Azure Blob storage to Google Cloud Platform (Stanford AIMI, n.d.). MedGemma 27B was evaluated on 24,000 images via local GPU deployment; Gemini 2.5 Pro was evaluated on 13,198 images via the Google AI Studio API. We describe these deliberately as two separate model runs rather than a single shared sampling procedure, because that is what they were: the two image pools were not drawn according to a common, documented protocol, and our own pipeline records do not establish why MedGemma's run used a larger pool than Gemini's.

Model	Version	Access	Image Input
MedGemma 27B	≥ May 20, 2025 release (exact pull date not logged)	Local VM / desktop deployment	24,000 images
Gemini 2.5 Pro	March 2025 API	Google AI Studio	13,198 images
DenseNet-121 Baseline*	Fine-tuned (planned)	PyTorch	N/A

Table 1. Model access specifications for the two evaluated systems and the planned CNN baseline

(Not yet executed). Values shown are targeted specifications for planned future work (Section 3.5), not results from a completed run. The original manuscript listed MedGemma's checkpoint as "Feb 2025"; that date precedes MedGemma's public release (May 20, 2025) and has been corrected here.

To compare the models on identical inputs, we performed an intersection merge by image path, yielding 13,189 images present in both output sets. We flag the retention rates explicitly: 99.9% of Gemini's original 13,198 images but only approximately 55% of MedGemma's original 24,000. The "matched" sample is therefore much closer to Gemini's evaluated set, filtered to exclude images MedGemma's run did not cover, than to a symmetric, independently constructed sample. We do not treat this asymmetry as neutral. As Section 3.1 details next, the dominant cause of MedGemma's attrition was the model itself, decoding loops, truncation, and safety-guardrail refusals, rather than an independent data-quality filter. Retaining only the images MedGemma successfully processed selects for the subset of cases where MedGemma cooperated with the extraction prompt, which plausibly biases the retained sample toward

MedGemma's easier cases and away from exactly the images where it struggles most. We have not quantified the size or direction of this effect, and we did not rerun MedGemma with Gemini's exception-handling scaffolding to test it directly (see Section 6), but the direction of the likely bias, favoring MedGemma rather than being neutral, should be read into every comparative claim in Sections 4 and 5. Results apply to this specific conditional population, not to CheXpert as a whole or to a hypothetical unbiased matched sample.

Within the resulting 13,189-image sample, ground-truth CheXpert uncertainty labels were binarized following the standard U-Zero convention: Positive = 1, Negative or Uncertain = 0. This is a defensible, widely used convention. Still, it is also a specific operational choice with a specific consequence: any pathology whose Uncertain-labeled cases disproportionately represent missed positive findings will show artificially depressed recall under this scheme. We did not break out performance separately for Uncertain-labeled cases, and we return to this directly when interpreting the Fracture results below rather than treating the recall figures there as a clean measure of model capability.

The pronounced attrition in the MedGemma cohort, where only approximately 55% of the initial 24,000-image run yielded parseable output, is itself a finding about the operational stability of domain-tuned architectures rather than a mere pipeline failure. Analysis of the unparseable instances indicated that MedGemma's specialized alignment tuning frequently induced structural non-compliance: under zero-shot clinical extraction prompts, the model regularly produced decoding loops, early token truncation, or outright refusals triggered by over-conservative clinical safety guardrails. Table 1a summarizes these failure categories at a coarse level; we did not retain sufficiently granular per-category logs during the original run to report exact counts here and treat closing that gap as a first-priority item (Section 6). Because these early pipeline executions lacked the exception-handling scaffolds later implemented for Gemini, a significant portion of raw outputs failed the strict downstream JSON schema. We state plainly what that means for the fairness claim in Section 3.2: the two models were not run through identical pipeline scaffolding, so some unknown share of MedGemma's attrition reflects our code's exception handling rather than the model's underlying behavior, and we have not yet re-run MedGemma with Gemini's scaffolding to separate the two. Rather than impute missing values or pad the dataset, which would introduce uncontrollable selection bias, we evaluated a strictly intersecting conditional sample. This preserves parity between the models' successfully processed inputs, but highlights a trade-off: domain specialization, while intended to optimize clinical relevance, compromised structural reliability at scale, confounded here with an unequal-scaffolding artifact we have not yet isolated.

Table 1a. MedGemma 27B failure-mode categories on the 45% of the initial 24,000-image run that did not yield parseable output ($n = 10,811$). Category definitions follow standard practice for structured-output extraction failures; per-category counts were not logged at sufficient granularity in the original run and are marked pending re-analysis of pipeline logs (Section 6), rather than estimated, to avoid reporting unverified figures.

Failure Category	Count	% of Attrition ($n = 10,811$)
Decoding loops (repeated/non-terminating generation)	pending	pending
Early token truncation (output cut off before valid JSON closed)	pending	pending
Safety-guardrail refusals (model declined to classify)	pending	pending
Other schema-invalid output	pending	pending

3.2 Prompting and Output Standardization

Both models were evaluated with a single, standardized zero-shot Expert Radiologist prompt requiring binary labels, confidence scores, and JSON-formatted output, parsed with a strict JSON parser. The full prompt text is reproduced verbatim in Appendix A so the evaluation can be replicated exactly; a

placeholder marks that location in this draft pending insertion of the exact production string from the pipeline configuration. This design choice directly addresses documented MLLM prompt-sensitivity (Hassanein et al., 2025) by holding prompt formulation constant across models, and we consider it one of the stronger methodological features of this study: whatever else is uncertain about the sampling procedure, the two models were not compared under different prompting conditions.

3.3 Evaluation Metrics

For each pathology, we compute accuracy, precision, recall, F1-score, and balanced accuracy, using $\text{accuracy} = (\text{TP} + \text{TN}) / N$ and $\text{balanced accuracy} = (\text{sensitivity} + \text{specificity}) / 2$ on thresholded, 0.5-cutoff predictions. We flag a correction here to how this statistic was labeled in an earlier draft of this analysis: balanced accuracy on thresholded predictions is not ROC-AUC. True ROC-AUC is a rank-based statistic that measures how well a model's confidence scores order true positives ahead of true negatives across all possible thresholds; computing it via the trapezoidal rule on a single 0.5-cutoff binarization, as an earlier draft of this paper did, discards exactly the confidence information that makes AUC informative and instead yields a monotonic transform of sensitivity and specificity at one operating point. Every value reported below and in Tables 2 and 3 under the earlier "AUC" label is in fact balanced accuracy; we have renamed it accordingly throughout and removed any language implying a ranking or discrimination claim. Our prompt already collects a per-image confidence score, so a true ROC-AUC and a corresponding PR-AUC (originally promised in this section but not delivered) can be computed without new data collection; we did not complete that computation for this revision and list it as the top-priority follow-up in Section 6, ahead of any new modeling work. Full per-cell confusion matrices, also promised in an earlier draft, are likewise deferred to that follow-up rather than fabricated here. We report multiple metrics deliberately rather than a single headline number, because accuracy alone is a poor discriminator on an imbalanced label distribution: a model predicting Absent for nearly all cases can achieve high accuracy on a low-prevalence pathology while capturing almost none of the true positives, precisely the pattern observed for MedGemma below. We additionally report McNemar's test and Wilson score confidence intervals on overall accuracy, both appropriate for paired binary outcomes on a shared image set: McNemar's test evaluates whether error patterns differ beyond chance, and the Wilson interval avoids the boundary distortions a normal-approximation interval can produce near 0 or 1. Both are reported as tests of whether the models differ, not why, a distinction maintained throughout.

3.4 System Architecture and Implementation

The experimental pipeline is a modular, multi-agent system orchestrated with LangGraph: an Input Handler Agent for image loading and validation, a Multimodal Processor Agent for prompt construction and API dispatch, an inference layer with MedGemma via LM Studio and Gemini 2.5 Pro via the Google API, a Result Evaluator Agent for JSON parsing and ground-truth comparison, and an Output Handler Agent for storage and running accuracy computation. Compute infrastructure comprised Lambda AI Cloud (one NVIDIA A100 SXM 40GB), Google Colab Pro, and a local GPU workstation. Per image, the pipeline loads the image and label, constructs the standardized prompt, dispatches it, parses the returned JSON, computes per-condition correctness, and appends to a running CSV; a batch processor computes final summary statistics. Separating data loading, inference, and evaluation simplified integrating a second model family and enforcing the same JSON-output contract across both, which is the main methodological safeguard this pipeline provides.

3.5 Baseline Clinical Model (Future Work)

A fine-tuned DenseNet-121 baseline, trained and evaluated on this study's exact 13,189-image sample, remains future work. Without it, the CNN figures cited above are external reference points from a different sample, not a controlled three-way comparison; closing this gap is the immediate next step in this research program.

3.6 Data Use and Ethics Statement

3.7

All radiographs used in this study come from CheXpert (Irvin et al., 2019), a de-identified, publicly available research dataset released by Stanford University under its own institutional data-use agreement; no images or protected health information were collected directly from patients or clinical sites by this study's authors. Two further disclosures follow from that fact and from the pipeline

described in Section 3.4. First, this study did not seek separate IRB review because it uses only secondary, de-identified, publicly released imaging data under Stanford AIMI's existing terms of use; the authors' institution should confirm and state here, in the camera-ready version, whether that determination (e.g., non-human-subjects or exempt classification) was formally logged, and under what specific agreement terms the CheXpert data were accessed. Second, and separately, a subset of these de-identified radiographs was transmitted to Gemini 2.5 Pro via the Google AI Studio API, a third-party commercial service, for inference. Authors should state explicitly, in the camera-ready version, what data-retention and training-use terms governed that transmission at the time of the study and whether those terms are consistent with CheXpert's use agreement; we have not independently verified this and do not assert compliance on the authors' behalf. We flag both points as required disclosures rather than filling them in with assumed answers, consistent with this paper's broader argument against stating more than the evidence supports.

4. RESULTS

The experiment on the 13,189-image sample produced a clear performance distinction between the models, along with metric-dependent reversals of which looks better. MedGemma achieves higher accuracy by being more specific, predicting Absent more often, while Gemini shows a more balanced precision and recall trade-off, reflected in higher F1 and AUC.

McNemar's test, applied to paired per-image error outcomes, indicates a statistically significant difference in error rates between the two models (statistic = 5313.0, $p = 3.39e-34$). The Wilson score interval on overall accuracy is (0.6331, 0.6425) for MedGemma and (0.5993, 0.6089) for Gemini, non-overlapping intervals consistent with the McNemar result. Statistical significance here establishes that the models differ, not why; it does not, on its own, adjudicate between a capability explanation and a measurement-artifact explanation for any specific pathology.

4.1 Overall Performance

Table 2 presents aggregate metrics across the full 13,189-image sample. MedGemma achieves higher overall accuracy (63.8% vs. 60.4%), while Gemini attains higher mean balanced accuracy (0.565 vs. 0.510) and Macro F1-score (0.316 vs. 0.134); as corrected in Section 3.3, this is balanced accuracy, not ROC-AUC. This divergence reflects different decision strategies: MedGemma defaults to a conservative, specificity-maximizing approach that inflates accuracy through systematic under-prediction of positive findings, while Gemini reaches a more balanced sensitivity and specificity trade-off, more appropriate for a screening use case.

Metric	MedGemma 27B	Gemini 2.5 Pro	Notes
Overall Accuracy	0.638	0.604	95% CI: (0.633, 0.643) vs (0.599, 0.609)
Mean Balanced Accuracy	0.510	0.565	Gemini higher (balanced accuracy = mean(sensitivity, specificity); not ROC-AUC — see §3.3)
Macro Avg F1-Score	0.134	0.316	Gemini higher
McNemar's Test	statistic 5313.0	$p = 3.39e-34$	Significant, $p < 0.001$

Table 2. Aggregate performance metrics for MedGemma 27B and Gemini 2.5 Pro (n = 13,189 matched CXR images)

4.2 Per-Pathology Performance

Table 3 disaggregates performance by pathology, revealing substantial within- and cross-model heterogeneity (n = 2,756 Pneumonia-positive, 3,338 Atelectasis-positive, 642 Fracture-positive images within the 13,189-image sample; see Section 2.3). For Pneumonia, MedGemma achieves markedly higher accuracy (75.4% vs. 52.9%) driven by high specificity (0.941), but recall of only 0.143 versus Gemini's 0.720, meaning MedGemma would miss roughly 86% of true Pneumonia cases in a screening

scenario, against Gemini missing roughly 28%. For Atelectasis, Gemini shows a substantially higher F1-score (0.474 vs. 0.104) and balanced accuracy (0.594 vs. 0.485), reflecting better sensitivity to positive cases.

Model	Class	n (Positive)	Accuracy	Precision	Recall	Specificity	F1 / Bal. Acc.
MedGemma 27B	Pneumonia	n = 2,756 (20.9%)	0.754	0.424	0.143	0.941	0.214 / 0.542
MedGemma 27B	Atelectasis	n = 3,338 (25.3%)	0.476	0.410	0.059	0.911	0.104 / 0.485
MedGemma 27B	Fracture	n = 642 (4.9%)	0.683	0.317	0.048	0.956	0.083 / 0.502
Gemini 2.5 Pro	Pneumonia	n = 2,756 (20.9%)	0.529	0.293	0.720	0.470	0.417 / 0.595
Gemini 2.5 Pro	Atelectasis	n = 3,338 (25.3%)	0.589	0.683	0.363	0.825	0.474 / 0.594
Gemini 2.5 Pro	Fracture	n = 642 (4.9%)	0.695	0.393	0.030	0.980	0.056 / 0.505

Table 3. Per-pathology classification metrics for MedGemma 27B and Gemini 2.5 Pro, with positive-case counts (n) and prevalence for each label within the 13,189-image matched sample. The final metric column reports balanced accuracy (mean of sensitivity and specificity), not ROC-AUC (see Section 3.3). This table's header row repeats automatically on page breaks (Word table-header property).

For Fracture, both models show recall below 5%, on a label with markedly lower prevalence (4.9%) than Pneumonia or Atelectasis. We do not treat this as a clean measure of model capability: because Uncertain-labeled ground truth was binarized as negative under the U-Zero convention (Section 3.1), and Fracture readings plausibly cluster more often around subtle or equivocal findings, some unknown share of this near-zero recall may reflect the labeling convention rather than a pure model failure. Decomposing that share by re-running Fracture under the U-Ones convention (Irvin et al., 2019) is a same-scale, near-term analysis we did not complete for this revision; we list it as a top follow-up priority in Section 6 rather than re-argue the point here. We treat the figure as real and clinically concerning under the convention used, while flagging that its magnitude should not be taken as a precise estimate of the models' true Fracture-detection capability until that decomposition is done.

4.3 Clinical Implications of the Per-Pathology Pattern

Read together, these results imply that no single deployment posture, "trust MedGemma" or "trust Gemini," is safe across all three conditions. A workflow using MedGemma alone for Pneumonia screening would under-detect most true cases despite a superficially reassuring accuracy figure; one using Gemini alone for Fracture screening would fare barely better, given both models' sub-5% recall on that label. The only pathology where either model reaches a recall level a screening tool would plausibly need is Gemini's 72.0% on Pneumonia, and even that implies roughly three in ten true cases missed. These numbers, not aggregate accuracy or AUC, should govern any deployment conversation, because they describe the failure mode, missed positive cases, that carries the highest clinical cost.

5. DISCUSSION

5.1 Interpreting the Performance Profiles

A single model-quality axis does not readily explain the performance profiles observed here. MedGemma's higher aggregate accuracy coexists with markedly low recall across all three pathologies, a pattern one candidate framework, the incomplete-knowledge account of LLM behavior (Nguyen, 2026), would predict: aggregate performance can look strong while sub-task performance, particularly on the positive class of an imbalanced label, is weak. We describe this as consistent with that framework rather than as evidence for it, for two reasons already noted: the framework's proposing paper shares an author with this study, so this cannot be treated as independent confirmation, and the framework's specific proposed mechanism has not been tested here.

One candidate mechanism, offered explicitly as an untested hypothesis, is that MedGemma's training corpus, skewed toward negative findings as most real-world radiology text is, may have induced a conservative reporting prior, such that the model assigns high probability to Absent largely independent of image-specific evidence. This would be consistent with the observed specificity and recall pattern. Still, MedGemma's training data composition was not examined, no ablation was run, and no internal representation was probed, so this remains speculation rather than a demonstrated mechanism. An equally consistent alternative explanation is that some portion of the low recall, particularly for Fracture, reflects the Uncertain-as-negative labeling convention described above rather than any property of the model at all. This study does not have the evidence to adjudicate between these explanations, and a paper claiming to identify the mechanism when it has tested neither would overstate its evidence.

Gemini's higher sensitivity, while more appropriate for screening, introduces a complementary cost: higher false-positive rates carrying a real follow-up burden in deployment. Neither model, on this evidence, is a clinically deployable diagnostic tool in isolation. Neither model's confidence scores were evaluated for calibration here, itself a limitation worth naming: a model can express high confidence on incorrect predictions in ways that accuracy, recall, or F1 would not reveal. Given that both models already show clinically significant blind spots on at least one pathology (Fracture recall below 5% for both, Pneumonia recall of 14.3% for MedGemma), a confidently wrong output is arguably worse than an openly uncertain one, because it is more likely to short-circuit human review. Calibration analysis is accordingly a first-tier future-work item alongside the sampling and labeling questions raised above.

While McNemar's test establishes that Gemini's edge is statistically significant ($p < 0.001$), this divergence must not be conflated with clinical utility. An aggregate Macro F1 of 0.316 and mean balanced accuracy of 0.565 show that both models, in their current iterations, fail the diagnostic baseline required for safe deployment. The value of this comparison lies not in crowning a superior architecture but in characterizing distinct, problematic failure modes. MedGemma behaves as an overly conservative agent: high specificity with 14.3% recall on Pneumonia minimizes false positives at the expense of missing most active pathology. On Pneumonia specifically, Gemini shows the inverse pattern, superior recall (0.720) undermined by comparatively poor specificity (0.470), yielding a volume of false alarms that would exacerbate clinician alarm fatigue; this uncalibrated-screening characterization does not extend to Gemini's other two labels, where specificity is high (0.825 on Atelectasis, 0.980 on Fracture), so we limit the claim to Pneumonia rather than describe Gemini's operating behavior in general. These low absolute metrics support a narrower thesis than an earlier draft of this paper claimed: on this single specialized model, this single generalist model, and these three labels, domain-tuning did not resolve the structural deficiencies observed in either model's knowledge. Whether domain-tuning is a panacea for clinical reasoning in general is a claim this study is not positioned to make (Section 5.2).

5.2 Scope of the Domain-Specialization Claim

This evidence does not support the hypothesis that domain-specialized MLLMs uniformly outperform generalist systems, but the claim's boundaries matter. This study compares one specialized model against one generalist, on one dataset, across three labels, using one sampling procedure with the asymmetries described above. That licenses a narrow, defensible conclusion: within this bounded comparison, MedGemma did not uniformly outperform Gemini 2.5 Pro, refuting the strong universal version of the domain-specialization hypothesis. It does not license a claim about domain specialization as a general strategy, which would require multiple specialized models, multiple datasets, and a sampling procedure without the retention asymmetry disclosed above.

This pattern is broadly consistent with prior arguments about the jagged capability frontier of AI systems (Dell'Acqua et al., 2023), strong on some task dimensions, weak on adjacent ones, though again this is noted as a resonant analogy rather than a tested claim; the jagged-frontier literature concerns knowledge-worker productivity tasks, not radiology. Clinical adoption decisions that rely on a single aggregate metric risk selecting a model whose weaknesses are concentrated exactly where they matter most, which is the practical lesson this study supports independent of any particular theoretical framework.

6. LIMITATIONS AND FUTURE RESEARCH

This study has several limitations beyond those already discussed inline. First, evaluation was zero-shot without prompt optimization; this is scientifically appropriate for assessing out-of-the-box utility and avoids confounding results with prompt-engineering skill, but it does not reflect performance achievable with few-shot prompting or retrieval-augmented generation. Second, CheXpert is sourced from a single academic medical center, limiting generalization to other imaging environments and demographics. Third, the study covers three pathologies; performance profiles here may not represent MLLM behavior across CheXpert's full label set or other imaging modalities. Fourth, the absence of a fine-tuned DenseNet-121 baseline evaluated on this exact sample precludes direct comparison to the established CNN benchmark on this dataset. Fifth, the two most consequential open questions from this paper, the sampling asymmetry in the intersection-merge procedure and the contribution of Uncertain-label binarization to low recall on rarer pathologies, remain unresolved; both are treated as first-priority items for follow-up work, ahead of new modeling contributions, because they bear on whether the reported numbers mean what they appear to mean.

Future work should, in priority order: (1) compute true ROC-AUC and PR-AUC from the per-image confidence scores the pipeline already collects, and publish full confusion matrices, replacing the balanced-accuracy correction in Section 3.3 with the metric it was originally meant to be; (2) re-run MedGemma's failed cases through Gemini's exception-handling scaffolding, or otherwise quantify how much of MedGemma's 45% attrition is model behavior versus pipeline handling, to test the bias direction flagged in Section 3.1; (3) tabulate the failure-mode counts left as "pending" in Table 1a from the original pipeline logs; (4) re-evaluate Fracture and other low-prevalence pathologies under the U-Ones convention (Irvin et al., 2019) to decompose measurement artifact from model capability; (5) characterize how the original 13,198- and 24,000-image pools were constructed and whether excluded images differ systematically from those retained; (6) train and evaluate a DenseNet-121 baseline on the same sample; (7) test, rather than assume, the conservative-reporting-prior mechanism via training-data analysis or ablation; and (8) extend the comparison to additional models and datasets before concluding domain specialization as a general strategy. None are large engineering lifts relative to the original data-collection effort, so we treat them as near-term priorities rather than aspirational future work.

7. CONCLUSION

This study conducted an image-matched comparison of MedGemma 27B and Gemini 2.5 Pro on multi-label chest X-ray classification, within a scope stated precisely rather than generously: one specialized model, one generalist, one dataset, three labels, with the sampling asymmetry disclosed in Section 3.1. Domain specialization conferred selective, not uniform, advantages, MedGemma led on Pneumonia accuracy, Gemini led substantially on Atelectasis F1 and balanced accuracy, and both models showed recall below 5% on Fracture, a figure caveated by the labeling convention discussed in Section 4.2, which is sufficient to refute the strong claim that specialized medical MLLMs uniformly outperform capable generalists, but not to support broader claims about domain specialization as a strategy. Results are interpreted in relation to the Wharton Paradox framework (Nguyen, 2026) as one candidate lens, while explicitly flagging the shared authorship between that framework and this study and the framework's dispensability to this paper's core empirical claims (Section 2.4). Neither model, as evaluated here, is suitable for autonomous diagnostic deployment.

More broadly, this study argues for a specific evaluation habit in medical AI research, not just a specific set of numbers: report the provenance of a sample before calling it matched, decompose a metric before it becomes a headline finding, and separate a theoretical framework's motivating role from its evidentiary status. None of these require new data or new models; they require stating clearly what was and was not established. Applied here, they leave this paper's central practical conclusion, that neither MedGemma nor Gemini 2.5 Pro is ready for autonomous deployment on chest radiograph classification, intact and, if anything, more defensible than a version that had claimed more. This paper's own revision history is a direct instance of the second habit: what an earlier draft reported as ROC-AUC was, on inspection, balanced accuracy (Section 3.3), a mislabeled-metric error of exactly the kind this closing argument warns against. We correct it here rather than let a more impressive-sounding but incorrect label stand, and we do not exempt this paper's own analysis from the standard it recommends for the field.

8. REFERENCES

- Ahmed, S., Sakib, S. K., & Das, A. B. (2025). Can large language models challenge CNNs in medical image analysis? arXiv. <https://doi.org/10.48550/arXiv.2505.23503>
- Al-Hamadani, S. (2025). Intelligent healthcare imaging platform: A VLM-based framework for automated medical image analysis and clinical report generation. arXiv. <https://doi.org/10.48550/arXiv.2509.13590>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayner, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Harvard Business School Technology & Operations Management Unit Working Paper 24-013).
- Hassanein, F. E. A., Ahmed, Y., Maher, S., El Barbary, A., & Abou-Bakr, A. (2025). Prompt-dependent performance of multimodal AI model in oral diagnosis: A comprehensive analysis of accuracy, narrative quality, calibration, and latency versus human experts. *Scientific Reports*. <https://doi.org/10.1038/s41598-025-22979-z>
- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 590-597. <https://doi.org/10.1609/aaai.v33i01.3301590>
- Izuchukwu, C., Wimmer, H., & Rebman, C. M., Jr. (2024). A comparison of large language models for oncology clinical text summarization. *Proceedings of the ISCAP Conference, 2024*(6208). <https://iscap.us/proceedings/2024/pdf/6208.pdf>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv. <https://doi.org/10.48550/arXiv.2001.08361>
- Kim, K., & Kim, B. C. (2025). Diagnostic performance of large language models in multimodal analysis of radiolucent jaw lesions. *International Dental Journal*. Advance online publication. <https://doi.org/10.1016/j.identj.2025.11.001>
- Nguyen, T. L. (2026). The completeness and consistency of knowledge in generative AI large language models – A theoretical framework. Retrieved from <https://doi.org/10.6084/m9.figshare.31231300>
- Pal, A., Lee, J.-O., Zhang, X., Sankarasubbu, M., Roh, S., Kim, W. J., Lee, M., & Rajpurkar, P. (2025). ReXVQA: A large-scale visual question answering benchmark for generalist chest X-ray understanding. arXiv. <https://doi.org/10.48550/arXiv.2506.04353>
- Pooch, E. H. P., Ballester, P. L., & Barros, R. C. (2019). Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. arXiv:1909.01940. <https://arxiv.org/abs/1909.01940>
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225.

- Salbas, A., & Buyuktoka, R. E. (2025). Performance of large language models in recognizing brain MRI sequences: A comparative analysis of ChatGPT-4o, Claude 4 Opus, and Gemini 2.5 Pro. *Diagnostics*, 15(15), Article 1919. <https://doi.org/10.3390/diagnostics15151919>
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Aguera y Arcas, B., . . . Natarajan, V. (2023). Towards expert-level medical question answering with large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2305.09617>
- Stanford AIMI. (n.d.). CheXpert: Chest X-rays dataset. Stanford Center for Artificial Intelligence in Medicine and Imaging. <https://aimi.stanford.edu/datasets/chexpert-chest-x-rays>
- Terwiesch, C. (2023). Would ChatGPT get a Wharton MBA? Mack Institute White Paper. Wharton School, University of Pennsylvania.
- Urooj, B. (2025). Large language models in medical image analysis: A systematic survey and future directions. *Bioengineering*, 12(8), Article 818. <https://doi.org/10.3390/bioengineering12080818>
- Villeda Roblero, A., Song, Y., Gebauer, J., & Shi, Y. (2025). A study of chatbots performance with large language models in scenario of a higher education institution. *Proceedings of the ISCAP Conference*, 2025(6435). <https://iscap.us/proceedings/2025/pdf/6435.pdf>
- Yang, L., Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Yatziv, L., Chen, T., Sterling, B., Baby, S. A., Baby, S. M., Lai, J., Schmidgall, S., . . . Bachem, O. (2025). MedGemma technical report. *arXiv*. <https://doi.org/10.48550/arXiv.2507.05201>
- Zambrano Chaves, J. M., Huang, S.-C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., Awadalla, H., Gong, J., Hu, H., Yang, J., Li, C., Gao, J., Gu, Y., Wong, C., Wei, M., . . . Poon, H. (2025). A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nature Communications*, 16, Article 3108. <https://doi.org/10.1038/s41467-025-58344-x>