SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Rapid, Low-Cost Structuring of 200 MIMIC-CXR Reports with an Open 4-Billion-Parameter MedGemma Language Model

Mathue Duhaney, MD, SUNY Downstate Health Sciences; Mina Al-Ani, MBBS

Introduction/Background

Rapid structuring of legacy radiology reports can unlock quality dashboards and downstream automation. Open-weights language models such as MedGemma-4B-IT promise vendor-free solutions, but their real-world labeling speed and accuracy remain unreported.

Methods/Intervention

We sampled 200 de-identified chest-radiograph impressions from the public MIMIC-CXR v 2.1.0 corpus. Three workflow labels were targeted—follow-up recommendation, technique quality, and urgency flag—using a zero-shot JSON prompt. Inference ran on a RunPod A100-80 GB PCIe GPU (spot \$1.64 / h). Total wall-clock time and GPU cost were recorded. Two fourth-year (PGY-5) radiology residents independently graded a 20-report subset to estimate label accuracy.

Results/Outcome

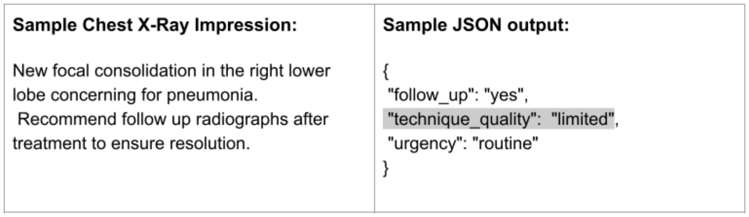
MedGemma labeled all 200 reports in 641 s (3.2 s/report) at a cloud cost of \$0.29. Manual audit showed accuracies of 35 % for follow-up recommendation, 25 % for technique quality, and 65 % for urgency flag. Disagreements were chiefly due to ambiguous language or absence of explicit follow-up phrasing.

Conclusion

A single open 4-billion-parameter model structured 200 public reports in under 11 minutes for < \$0.30. Accuracy was above average for urgency detection but below average for technique quality and follow-up, indicating room for prompt tuning or light task-specific training. The workflow demonstrates that institutions can pilot foundation LLM labeling with no protected-data egress and minimal cloud spend.

Statement of Impact

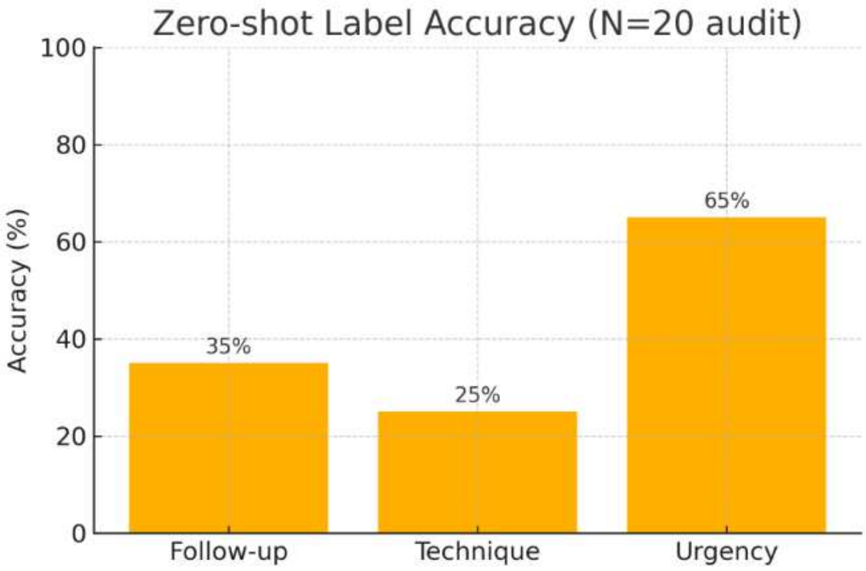
Low-cost, browser-scriptable large language model (LLM) labeling enables rapid QA dashboards and "AI-readiness" scoring across legacy archives—even at community sites without dedicated data-science /information technology teams. By quantifying runtime, dollar cost, and resident-validated accuracy on a public dataset, this study provides a practical benchmark for integrating open LLMs into real-world radiology workflows.



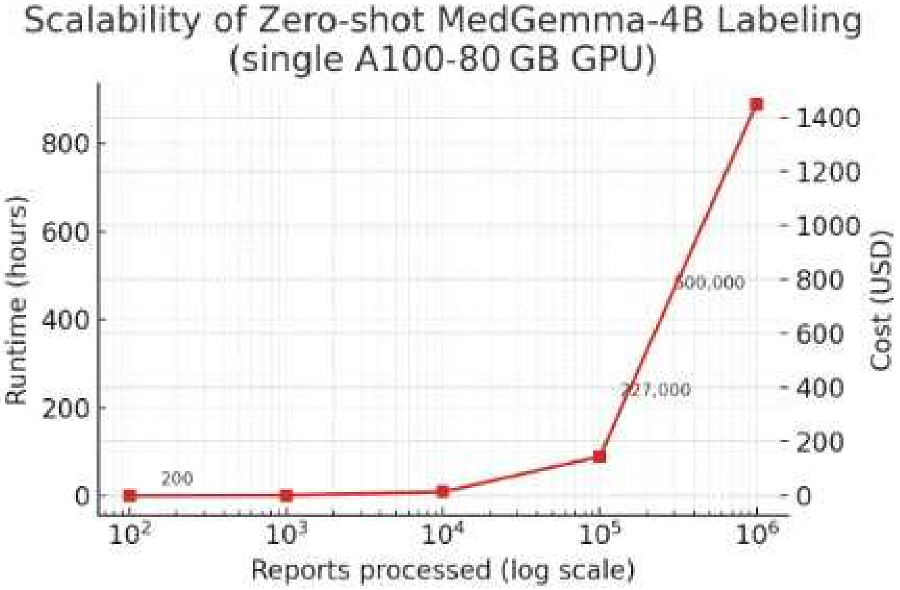
Sample Chest X-Ray Impression:	Sample JSON output:

New focal consolidation in the right lower lobe concerning for pneumonia. Recommend follow up radiographs after treatment to ensure resolution.	{
	"follow_up": "yes",
	"technique_quality": "limited",
	"urgency": "routine"
	}

Qualitative example of model output. Left — excerpt from a de-identified MIMIC-CXR impression. Right — zero-shot MedGemma-4B-IT JSON response produced by the prompt. The example highlights correct identification of follow-up and urgency, but a missed technique-quality limitation (shaded in gray).



Zero-shot label accuracy (N = 20 audit). Bar heights show percent agreement between MedGemma-4B-IT and two PGY-5 radiology residents for each workflow label. Urgency was highest (65 %), technique quality lowest (25 %), and follow-up 35 %. Error bars omitted for clarity (single 20-case sample).



Projected runtime and cloud cost for zero-shot MedGemma-4B-IT report labeling as a function of dataset size, assuming a single RunPod A100-80 GB spot instance ($1.64 / h) and the per-report metrics measured in this study (3.2 s, $0.00145). Linear scaling is valid because the 4-B model fits in GPU memory and reports are processed sequentially without batching. Alternative hardware or batch inference would shift both curves downward proportionally.

Keywords

large language models; MedGemma; natural language processing; radiology report structuring; quality assurance; workflow automation