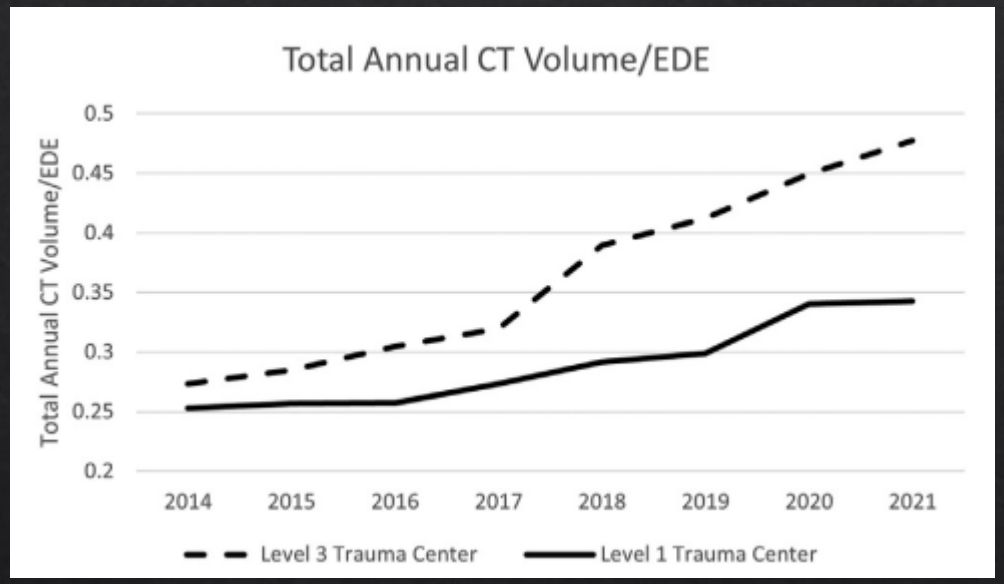
Massachusetts General Hospital
Founding Member, Mass General Brigham

Radiology-Specific Vision-Language Models - Your Future Digital Colleague?

Rehab N. Khalid,
Andreas Schicho,
Christian Stroszczynski,
Quirin D. Strotzer

HARVARD MEDICAL SCHOOL
TEACHING HOSPITAL

Background and Clinical Need



- Imaging **volume rising** sharply: **CT +35.5%, MRI +56.3%** at Level-I trauma centers (2014 → 2021)
- **Workforce strain:** 1-yr separation 13.8% → 19.2% (2014–15 → 2017–18); reports of understaffing, job migration, more part-time roles
- **Residency** positions **not** keeping pace with imaging growth → **staffing shortfalls**
- **Clinical consequences**: longer turnaround times, increased burnout, less time for complex, context-heavy cases (surgical planning, detailed histories)

Why VLMs for Radiology?

- **AI rapid evolution**: **narrow** task models → **multimodal** VLMs (image + language) capable of **interactive** interpretation
- **Radiology-specific VLMs**: designed/tuned for radiology language and imaging features — potential performance gains vs general models
- Clinical promise: act as a "**digital colleague**" - triage, draft findings, routine QA, trainee feedback, workflow triage

Study Objectives

01

Compare diagnostic accuracy (radiology specific models vs human readers) on chest + MSK radiographs

02

Accuracy/sensitivity/ specificity, per-task performance, and error-mode analysis

03

Clinical intent: evaluate readiness as decision-support /triage tools and identify gaps for safe deployment

Methods - Dataset

Single-center, retrospective, IRB-approved

N = **72** radiographs

Pathologic: **39 (54%)**

Normal: **33 (46%)**

Single, de-identified image per case (AP/PA or single view)

Reference standard: clinical/radiologic confirmation (chart + imaging)

Target pathologies: Lung cancer, pneumonia, pneumothorax, fractures

Methods – Models, Prompts, Human readers

- **Harrison.rad.1 (agent + small), GPT-4o, GPT-4V**
- Radiologist-persona prompts; binary output format
- Consistent prompting across models

- 4 board-certified radiologists + 1 trainee
- Blinded, randomized reads
- Majority-vote used for pooled human-reader reference
- Readers independent from model development team

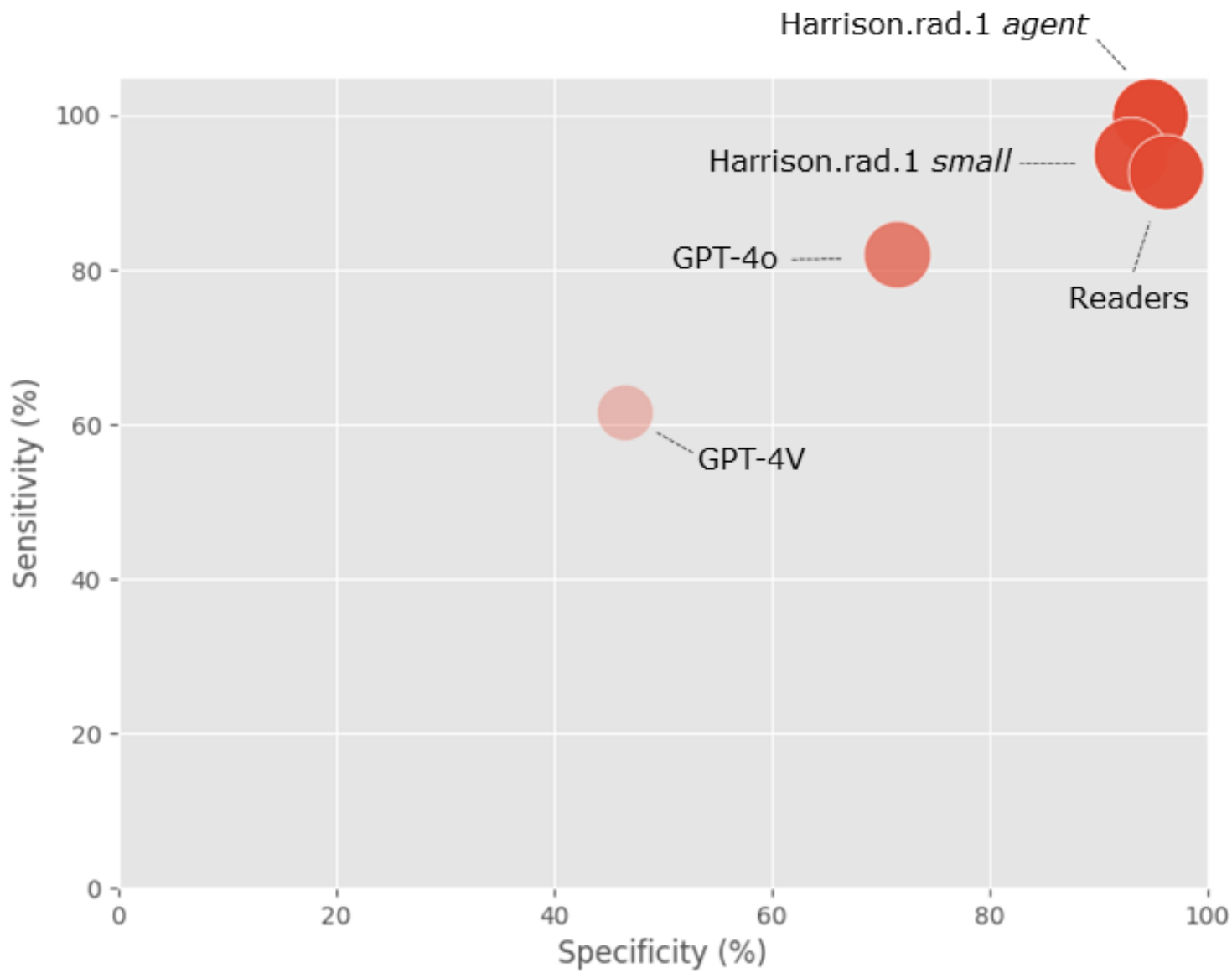
Results – Headline Performance

Key tests: McNemar (pairwise)

Exact binomial CIs for sensitivity/ specificity

Harrison vs humans: no significant difference (reader-level parity)

GPT models: statistically lower performance ($p < 0.001$)


Harrison.rad.1 agent
Harrison.rad.1 small
GPT-4o
Readers
GPT-4V
Sensitivity (%)
Specificity (%)
0
20
40
60
80
100

Task-Specific Results – Lung Cancer

Harrison.agent & small: **100%** accuracy

Readers: **98.6%**

GPT-4o: **92.9%**

GPT-4V: **64.3%**

Task - Pneumonia

Agent & Small: **93.3%**

Readers: **96.0%**

GPT-4o: 100%

GPT-4V: **46.7%**

Task - Pneumothorax

Agent: **100**% - perfect on this set

Small: **96.6**%

Readers: **93.8**%

GPT-4o: **79.3**%

GPT-4V: **55.2**%

Task - Fracture

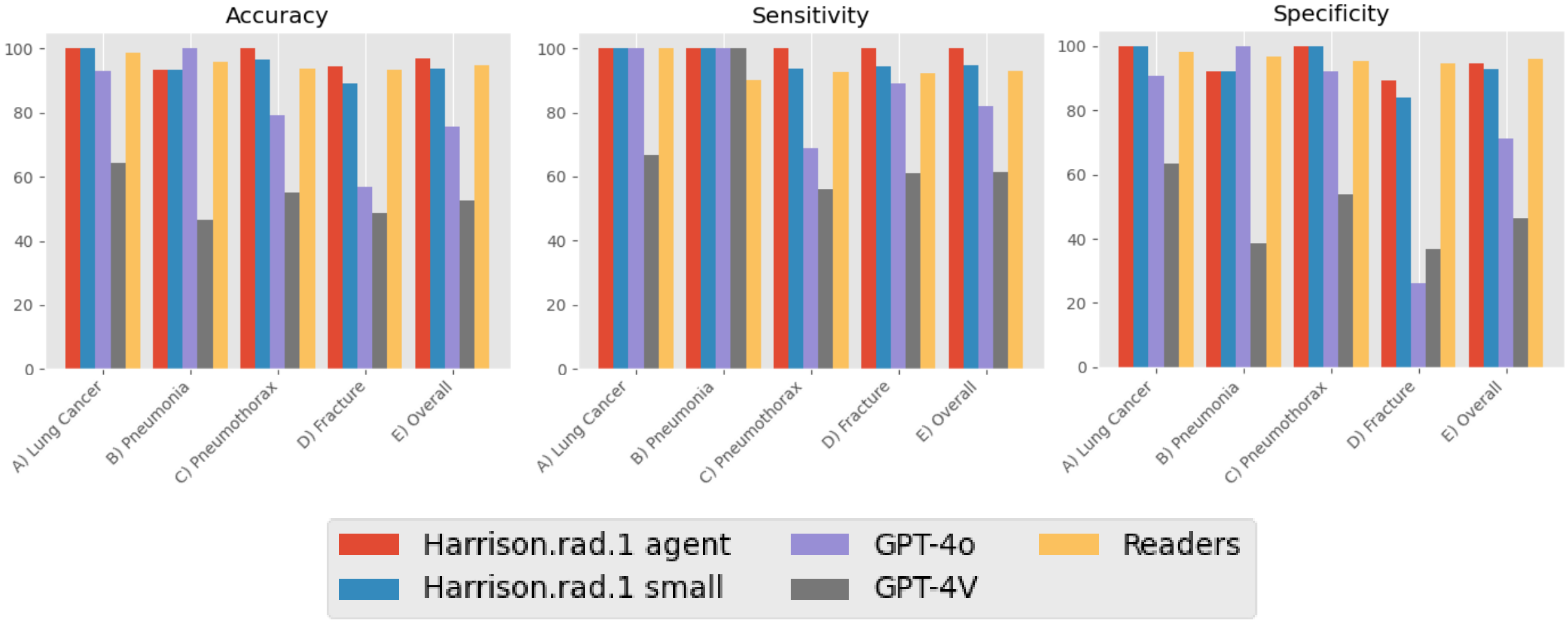
Agent: **94.6**%

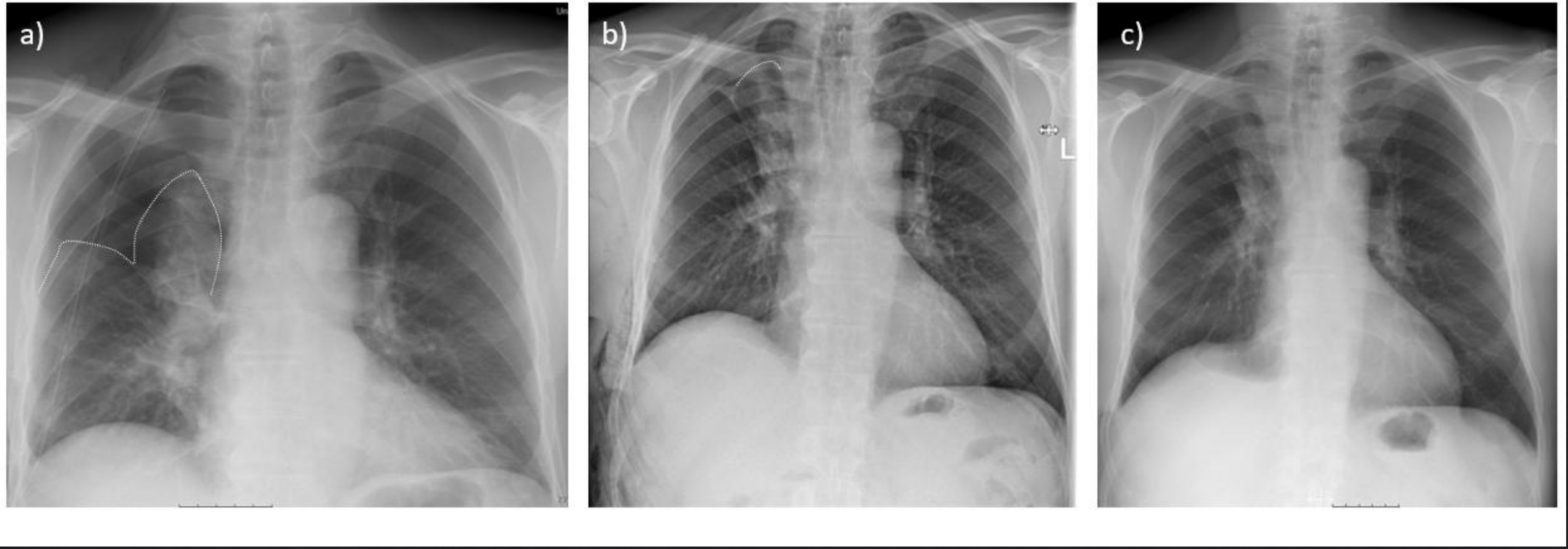
Small: **89.2**%

Readers: **93.5**%

GPT-4o: **56.8**%

GPT-4V: **48.6**%


Accuracy
Sensitivity
Specificity
100
80
60
40
20
0
A) Lung Cancer
B) Pneumonia
C) Pneumothorax
D) Fracture
E) Overall
Harrison.rad.1 agent
Harrison.rad.1 small
GPT-4o
GPT-4V
Readers


a)
b)
L
c)

Error analysis: What went wrong and why?

- **Example case**: **atypical** post-op anatomy → missed apical PTX (Harrison.small + non-radiologist)
- Possible errors: false negatives (**subtle/atypical**), false positives (**artifacts/overlap**)
- Pattern: fractures → **wide specificity variability** (GPTs low); Harrison.agent = few misses
- Cause & action: **single-view images** + uncommon **anatomy** → need **multi-view**/CT validation and **clinician oversight** with clinical correlation

Limitations

- Pre-existing dataset → **small**, sometimes imbalanced sample
- Slightly **different** prompt used for GPT-4V evaluation (GPT-4V discontinued)
- No in-context learning applied (to mimic real-world prompt use)
- Only **single-image** inputs used despite multi-image capability to **ensure uniformity**
- **Excluded multi-pathology** or ambiguous cases to **reduce confounding**
- Design favored **internal validity** but **limited real-world generalizability**
- Clinical radiographs often feature **overlapping abnormalities**

Conclusion

- Radiology-specific VLMs **matched radiologist accuracy** and **outperformed general GPT** models.
- Show promise as reliable **digital colleagues** to **ease workload** and **enhance efficiency**.
- Need **broader validation** and **regulatory approval** before clinical use.
- Future work: extend to **CT/MRI**, add reasoning workflows, and test on **larger**, real-world datasets.
- Emphasis on **privacy**, **safety**, and **clinician oversight** in deployment.

Thank you!