This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

Coherence Is Not a Green Light: The Limits of Self-Consistency as a Label-Free Reliability Proxy for Medical Vision–Language Models

This supplement provides the complete frozen elicitation prompts, the ground-truth-blind judge rubric and output schema, and the response-extraction rule, for full reproducibility of the results in the main paper.

S1 Report-Elicitation Prompts (frozen set v1.1)

Every (image, finding) pair received the same four-step template; only the step-2 focus line varied by finding and arm. The token ⟨ present⟩ is the finding phrase (cardiomegaly; a pleural effusion; atelectasis; consolidation; pulmonary edema).

```
You are a radiologist examining a frontal chest X-ray.
1. Briefly describe what you see in the image.
2. <step-2 focus line>
3. Reason step by step about whether <present> is present.
4. On the last line write exactly 'FINAL: YES' if the patient
   has <present>, or 'FINAL: NO' if not.
```

Table S1. Step-2 focus lines for every finding and arm (neutral; forced-CTR, Sec. 5.3; categorical; directed sign-check, Sec. 5.4).

Finding	Arm	Step-2 focus line
Cardiomegaly	neutral	Focus on the heart: describe the cardiac silhouette and its size relative to the width of the chest.
Cardiomegaly	forced-CTR (5.3)	Focus on the heart: describe the cardiac silhouette and estimate the cardiothoracic ratio (CTR) as a decimal.
Cardiomegaly	categorical	Focus on the heart: state whether the cardiac silhouette is enlarged or normal.
Pleural Effusion	neutral	Focus on the pleural spaces and the lung bases: describe what you see.
Pleural Effusion	sign-check (5.4)	Focus on the pleural spaces and the lung bases: check specifically for signs of a pleural effusion such as blunting of the costophrenic angles or a meniscus sign.
Atelectasis	neutral	Focus on the lung fields: describe any areas of opacity and the degree of lung expansion.
Atelectasis	sign-check (5.4)	Focus on the lung fields: check specifically for volume-loss signs such as fissure displacement, mediastinal shift toward the opacity, an elevated hemidiaphragm, rib crowding, or hilar displacement.
Consolidation	neutral	Focus on the lung fields: describe any areas of increased opacity and whether lung volume appears preserved.
Consolidation	sign-check (5.4)	Focus on the lung fields: check specifically for airspace-consolidation signs such as air bronchograms or the silhouette sign.
Edema	neutral	Focus on the cardiac silhouette, pulmonary vascularity, lung fields and interstitial area and describe what you see.
Edema	sign-check (5.4)	Focus on the cardiac silhouette, pulmonary vascularity, and the interstitial and airspace regions: check specifically for interstitial or alveolar edema signs such as Kerley B lines, septal lines, or perihilar airspace shadowing.

S2 Ground-Truth-Blind Judge (Claude Opus 4.8)

The judge received only the model's free-text report. The full prompt is the per-finding rubric (below), followed by the report and the output schema:

```
Return ONLY a JSON object, no prose:
{"evidence_points_to": "PRESENT|ABSENT|AMBIGUOUS",
 "final_answer": "YES|NO|UNCLEAR",
 "coherent": true|false,
 "positioning_or_caveat_noted": true|false,
 "rationale": "<=20 words"}
```

A report maps to unclear when final_answer = UNCLEAR or evidence_points_to = AMBIGUOUS.

S2.1 Per-finding judge rubric (frozen source of truth)

Cardiomegaly. You assess whether a radiology model's report is INTERNALLY COHERENT: does its FINAL yes/no follow from the cardiac evidence IT described? Do NOT use outside knowledge of the real diagnosis. ENLARGED-supporting evidence: CTR at/above ~0.50, an explicitly enlarged/prominent cardiac silhouette, chamber enlargement. NORMAL-supporting evidence: CTR below ~0.50, a normal/unremarkable cardiac silhouette, or an explicit 'no cardiomegaly'. COHERENT = (FINAL YES with enlarged-supporting evidence) OR (FINAL NO with normal-supporting evidence). INCOHERENT = the FINAL contradicts the model's own stated evidence (e.g. describes a normal silhouette / sub-threshold CTR yet says YES).

Pleural Effusion. You assess whether a radiology model's report is INTERNALLY COHERENT: does its FINAL yes/no follow from the pleural evidence IT described? Do NOT use outside knowledge. EFFUSION-supporting: blunted/obscured costophrenic angle, meniscus sign, dependent layering fluid, basal opacity consistent with fluid. NO-EFFUSION-supporting: sharp costophrenic angles, no meniscus, clear bases. COHERENT = FINAL matches the model's own stated evidence. INCOHERENT = FINAL contradicts it (e.g. describes sharp angles and clear bases yet says YES).

Atelectasis. You assess whether a radiology model's report is INTERNALLY COHERENT: does its FINAL yes/no follow from the pulmonary evidence IT described? Do NOT use outside knowledge. ATELECTASIS-supporting: opacity WITH volume loss — displaced fissure, raised hemidiaphragm, mediastinal/tracheal shift TOWARD the opacity, crowded vessels or ribs, linear/band-like or lobar collapse. ABSENT-supporting: clear, well-expanded lungs with no volume loss. COHERENT = FINAL matches the model's own stated evidence. INCOHERENT = FINAL contradicts it (e.g. describes clear, fully expanded lungs yet says YES).

Consolidation. You assess whether a radiology model's report is INTERNALLY COHERENT: does its FINAL yes/no follow from the pulmonary evidence IT described? Do NOT use outside knowledge. CONSOLIDATION-supporting: dense homogeneous airspace opacity, air bronchograms, silhouetting of the heart or diaphragm border, lobar/segmental opacity WITHOUT volume loss. ABSENT-supporting: clear lungs, no airspace opacity. COHERENT = FINAL matches the model's own stated evidence. INCOHERENT = FINAL contradicts it (e.g. describes clear lungs yet says YES).

Edema. You assess whether a radiology model's report is INTERNALLY COHERENT: does its FINAL yes/no follow from the pulmonary/vascular evidence IT described? Do NOT use outside knowledge. EDEMA-supporting: bilateral perihilar ('batwing') opacity, septal (Kerley B) lines, peribronchial cuffing, vascular redistribution or cephalization, often with pleural fluid. ABSENT-supporting: clear lungs, normal vascular pattern, no septal lines. COHERENT = FINAL matches the model's own stated evidence. INCOHERENT = FINAL contradicts it (e.g. describes clear lungs with normal vasculature yet says YES).

S3 Response Extraction

A response is committed if a final yes/no can be parsed from a 'FINAL: YES' / 'FINAL: NO' line by regular expression. Responses that merely echo the instruction template (no genuine conclusion) are mapped to uncertain and excluded from committed responses. Generation was greedy (deterministic) for all three models; ground truth was stored but never sent to the models or the judge.