Supplementary Table 1. REFINE-oriented reporting map

REFINE domain	Location	Revision-specific content
Model specification	Methods 2.2; Table 1; Supplementary Tables 3 and 21	Model identifiers, vendors, access dates, inference-only use, snapshot differences, and observed cost context
Prompt design	Methods 2.2; Supplementary Materials 2-3; Supplementary Table 3	Locked prompt, JSON schema, version history, and provenance
Stochasticity control	Methods 2.5; Results 3.4; Supplementary Table 16; Supplementary Figure 4	Temperature 0, three independent runs, prompt variants, and pairwise/all-run consistency
Dataset integrity	Methods 2.1, 2.3, 2.4; Figure 1; Supplementary Tables 2, 4-6; Supplementary Material 4	Corpus, exclusions, sampling, reference standard, lexicon audit, and source separation
Output evaluation	Methods 2.6; Tables 2-6; Supplementary Tables 7-20	QWK, paired bootstrap, per-class behavior, error direction, sensitivity, discovery caveat, and token accounting
Implementation and non-use	Discussion; Conclusion; Supplementary Tables 18-21	Research-stage scope, cost/token accounting, and no autonomous clinical triage

The completed 44-item REFINE checklist is provided as Supplementary Material 1. QWK, quadratic weighted kappa.

Supplementary Table 2. Original executable 43-term lexicon

Category	Executable terms
Epistemic (n = 12)	possible; possibly; probable; probably; likely; unlikely; may; might; could; perhaps; apparent; apparently
Lexical (n = 16)	suspicious; suspicious for; suggestive; suggestive of; concerning; concerning for; compatible with; consistent with; questionable; equivocal; indeterminate; nonspecific; non-specific; ill-defined; ill defined; borderline
Evidential/exclusion (n = 15)	appears; appear to; seems; seem to; cannot be excluded; cannot exclude; cannot be ruled out; cannot rule out; may represent; might represent; could represent; could be; difficult to; unclear; not clearly

Original implementation: lowercased raw substring search; each distinct term counted once; score = min (distinct term count, 3). Corrected implementation: alphanumeric boundaries plus longest non-overlapping match selection. Corrected occurrence counts all accepted matches; corrected distinct counts unique accepted terms.

Supplementary Table 3. Reproducibility and model-provenance parameters

Parameter	Value
Random seed	42
Robustness subset sampling	50-report simple random sample without replacement from 3,927 eligible reports using NumPy seed 42; 30-report repeated-run set = first 30 identifiers of the archived 50-report sample
Bootstrap iterations	1,000
Primary endpoint	Cohen quadratic weighted kappa against the multi-reader consensus
Claude validation/discovery identifier	claude-opus-4-5-20251101
Claude validation access	April 2026; temperature 0
Claude source separation/sensitivity	Six validation-ID predictions differed between the locked and later full-corpus files despite the same recorded API identifier. The cause was not established; primary validation scores were not overwritten (Supplementary Material 4).
Gemini validation identifier	gemini-2.5-pro-preview-05-06
Gemini validation access	April 2026; temperature 0; structured JSON output; no directly comparable thinking- budget value established in archived metadata
Gemini discovery/robustness identifier	gemini-2.5-pro
Gemini discovery access	July 18, 2026; Batch API; temperature 0; documented thinking budget 1,024
Gemini discovery requests	3,827 successful; 0 failed
Archived metadata prompt_sha256 value	34dd1c7eacce2a259224323d88dc10707588397845aa8e4603ce4e6fa52dc00d
Validation/archive provenance	Input-archive and validation-ID SHA-256 values preserved in the reviewer reproducibility package
Software	Python 3.12; pandas; numpy; scikit-learn; scipy; statsmodels; irrCAC-compatible formulas; google-genai; anthropic

API, Application Programming Interface.

Supplementary Table 4. Validation lexicon implementations

Validation implementation	n	Mean count	Score 0/1/2/3	QWK (95% CI)
Legacy distinct substring	100	1.61	25/25/25/25	0.622 (0.518-0.712)
Corrected longest-match occurrence	100	1.37	25/37/22/16	0.572 (0.440-0.677)
Corrected longest-match distinct	100	1.20	25/41/24/10	0.585 (0.446-0.692)

The legacy validation distribution is fixed at 25/25/25/25 by the sampling design. QWK, quadratic weighted kappa; CI, confidence interval.

Supplementary Table 5. Discovery lexicon audit (n = 3,827)

Discovery implementation	Score 0/1/2/3	Count changed	Ordinal score changed
Legacy distinct substring	3009/540/204/74	0	0
Corrected longest-match occurrence	3010/571/170/76	231	213
Corrected longest-match distinct	3010/645/136/36	161	145

Corrected occurrence changed 213/3,827 ordinal scores (5.6%). Across all 3,927 eligible reports, 245 scores changed (6.2%).

Supplementary Table 6. Four-level ordinal linguistic-uncertainty rubric

Score	Operational definition	Interpretive guardrail
0-None	Definitive language; no meaningful diagnostic qualification	Negative findings alone are not uncertainty.
1-Low	Localized mild qualifier or limited uncertainty that does not dominate the report	Examples include isolated borderline or apparent wording.
2-Moderate	Meaningful unresolved characterization, hedged differential, or multiple plausible explanations	Context and diagnostic framing are required.
3-High	Extensive ambiguity, multiple competing explanations, or additional evaluation because an important finding cannot be resolved	Not an automatic risk or urgency score; recommendations are interpreted in context.

Primary mapping: none = 0, low = 1, moderate = 2, high = 3. The secondary 0-10 model field was not used analytically.

Supplementary Table 7. Pairwise reader and reader-to-consensus agreement

Comparison	QWK	95% bootstrap CI	Exact agreement
PGY-2 vs. senior resident	0.421	0.246-0.572	38/100 (38.0%)
PGY-2 vs. attending	0.659	0.506-0.782	61/100 (61.0%)
Senior resident vs. attending	0.497	0.322-0.639	48/100 (48.0%)
PGY-2 vs. consensus	0.722	0.557-0.849	72/100 (72.0%)
Senior resident vs. consensus	0.576	0.403-0.719	59/100 (59.0%)
Attending vs. consensus	0.902	0.822-0.961	89/100 (89.0%)

Reader-to-consensus comparisons are structurally dependent because each reader contributes to the consensus. Attending-to-consensus agreement is further increased because the attending supplied the adjudicated score in all complete three-way disagreements. PGY-2, Postgraduate Year 2; QWK, quadratic weighted kappa; CI, confidence interval.

Supplementary Table 8. Overall validation metrics

Method	QWK	95% CI	Exact	Over/under	1/2/3-level errors	MAE	Macro F1
Claude Opus 4.5	0.807	0.722-0.876	68 (68.0%)	25/7	30/2/0	0.34	0.640
Gemini 2.5 Pro	0.814	0.735-0.877	71 (71.0%)	17/12	28/1/0	0.30	0.584
Legacy 43-term lexicon	0.622	0.518-0.712	46 (46.0%)	48/6	41/13/0	0.67	0.440
Corrected longest- match occurrence	0.572	0.440-0.677	45 (45.0%)	41/14	44/11/0	0.66	0.422
Corrected longest- match distinct	0.585	0.446-0.692	51 (51.0%)	34/15	40/9/0	0.58	0.475

QWK, quadratic weighted kappa; CI, confidence interval; MAE, mean absolute error.

Supplementary Table 9. Per-category validation metrics

Model	Score	Support	Precision	Recall	F1
Claude Opus 4.5	0	40	0.969	0.775	0.861
Claude Opus 4.5	1	31	0.600	0.581	0.590
Claude Opus 4.5	2	21	0.519	0.667	0.583
Claude Opus 4.5	3	8	0.455	0.625	0.526
Gemini 2.5 Pro	0	40	0.972	0.875	0.921
Gemini 2.5 Pro	1	31	0.667	0.645	0.656
Gemini 2.5 Pro	2	21	0.484	0.714	0.577
Gemini 2.5 Pro	3	8	0.333	0.125	0.182

Supplementary Table 10. Claude Opus 4.5 confusion matrix

Consensus score	Pred 0	Pred 1	Pred 2	Pred 3
True 0	31	9	0	0
True 1	1	18	10	2
True 2	0	3	14	4
True 3	0	0	3	5

Rows represent consensus/reference scores and columns represent predicted scores.

Supplementary Table 11. Gemini 2.5 Pro confusion matrix

Consensus score	Pred 0	Pred 1	Pred 2	Pred 3
True 0	35	5	0	0
True 1	1	20	10	0
True 2	0	4	15	2
True 3	0	1	6	1

Rows represent consensus/reference scores and columns represent predicted scores.

Supplementary Table 12. Legacy 43-term lexicon confusion matrix

Consensus score	Pred 0	Pred 1	Pred 2	Pred 3
True 0	22	16	2	0
True 1	2	8	11	10
True 2	1	1	10	9
True 3	0	0	2	6

Rows represent consensus/reference scores and columns represent predicted scores.

Supplementary Table 13. Corrected longest-match occurrence lexicon confusion matrix

Consensus score	Pred 0	Pred 1	Pred 2	Pred 3
True 0	22	16	2	0
True 1	2	12	11	6
True 2	1	7	7	6
True 3	0	2	2	4

Rows represent consensus/reference scores and columns represent predicted scores.

Supplementary Table 14. Reference-standard and prompt-overlap sensitivity analyses

Analysis	n	Model	QWK	95% CI
Primary 3-reader consensus	100	Claude Opus 4.5	0.807	0.722-0.876
Primary 3-reader consensus	100	Gemini 2.5 Pro	0.814	0.735-0.877
Attending-reference sensitivity	100	Claude Opus 4.5	0.831	0.758-0.886
Attending-reference sensitivity	100	Gemini 2.5 Pro	0.832	0.761-0.885
Complete-disagreement cases excluded	93	Claude Opus 4.5	0.816	0.736-0.882
Complete-disagreement cases excluded	93	Gemini 2.5 Pro	0.810	0.738-0.874
Attending reference + complete- disagreement excluded	93	Claude Opus 4.5	0.841	0.769-0.894
Attending reference + complete- disagreement excluded	93	Gemini 2.5 Pro	0.829	0.766-0.882
Prompt-development overlap excluded, ID 1000	99	Claude Opus 4.5	0.799	0.714-0.866
Prompt-development overlap excluded, ID 1000	99	Gemini 2.5 Pro	0.814	0.731-0.873

Report ID 1000 was the only validation report identified among the early prompt-development examples. It had consensus Score 3, Claude Score 3, and Gemini Score 2. Excluding it did not materially alter aggregate agreement. QWK, quadratic weighted kappa; CI, confidence interval.

Supplementary Table 15. Prompt-overlap exclusion details

Model	Exact agreement	≥ 2-level discrepancies	Score-3 recall
Claude Opus 4.5	67/99 (67.7%)	2	4/7 (57.1%)
Gemini 2.5 Pro	71/99 (71.7%)	1	1/7 (14.3%)

Paired Gemini-minus-Claude Delta QWK: 0.015 (95% CI, -0.061 to 0.086). QWK, quadratic weighted kappa; CI, confidence interval.

Supplementary Table 16. Prompt sensitivity and repeated-run consistency

Model	Analysis	Comparison	n	QWK	Exact
Claude Opus 4.5	Prompt sensitivity	Original vs. short	50	0.899	44/50 (88.0%)
Claude Opus 4.5	Prompt sensitivity	Original vs. chain-of-thought	50	0.881	43/50 (86.0%)
Claude Opus 4.5	Prompt sensitivity	Short vs. chain-of-thought	50	0.818	41/50 (82.0%)
Gemini 2.5 Pro	Prompt sensitivity	Original vs. short	50	0.913	44/50 (88.0%)
Gemini 2.5 Pro	Prompt sensitivity	Original vs. chain-of-thought	50	0.895	43/50 (86.0%)
Gemini 2.5 Pro	Prompt sensitivity	Short vs. chain-of-thought	50	0.861	42/50 (84.0%)
Claude Opus 4.5	Three-run stochasticity	All three runs identical	30		29/30 (96.7%)
Claude Opus 4.5	Stochasticity pairwise	Run 1 vs. run 2	30	0.962	29/30 (96.7%)
Claude Opus 4.5	Stochasticity pairwise	Run 1 vs. run 3	30	0.962	29/30 (96.7%)
Claude Opus 4.5	Stochasticity pairwise	Run 2 vs. run 3	30	1.000	30/30 (100.0%)
Gemini 2.5 Pro	Three-run stochasticity	All three runs identical	30		28/30 (93.3%)
Gemini 2.5 Pro	Stochasticity pairwise	Run 1 vs. run 2	30	1.000	30/30 (100.0%)
Gemini 2.5 Pro	Stochasticity pairwise	Run 1 vs. run 3	30	0.919	28/30 (93.3%)
Gemini 2.5 Pro	Stochasticity pairwise	Run 2 vs. run 3	30	0.919	28/30 (93.3%)

The 50-report prompt-sensitivity subset was a simple random sample drawn without replacement from the 3,927-report eligible frame with NumPy seed 42. Report ID 569 was the sole overlap with validation. The 30-report repeated-run set corresponded to the first 30 identifiers of the archived 50-report sample and had no validation overlap. Exact identifiers are included in the reproducibility package. QWK, quadratic weighted kappa.

Supplementary Table 17. Model–reference discrepancies of at least two ordinal levels

Report ID	Model	Consensus	Model score	Difference	Source impression
1557	Claude Opus 4.5	1	3	+2	1. Enlarged cardiac silhouette, possibly consistent with mild cardiomegaly versus a small pericardial effusion. 2. Mild bibasilar atelectasis and/or infiltrates with probable small bilateral pleural effusions, on the right greater than left, and pleural thickening versus loculated pleural fluid along the peripheral aspect of the right upper lobe.
2002	Claude Opus 4.5	1	3	+2	Possible lower posterior lateral left rib fractures as described above. If further concern for rib fractures dedicated rib films would better evaluate. Otherwise, Otherwise no acute cardiopulmonary disease.
397	Gemini 2.5 Pro	3	1	-2	XXXX 7 XXXX nodular density at the left costophrenic XXXX. Recommend CT scan for further evaluation.

Report 1557 is presented as a contested rubric-boundary case rather than an unambiguous model error because its multiple qualifiers and competing alternatives make the distinction between low and higher linguistic uncertainty debatable under the study rubric. The locked consensus label was retained, and no quantitative result was altered. Post-hoc attending re-review maintained the locked consensus scores in all three cases; because the attending contributed to the original reference standard, this review represents confirmation rather than independent adjudication. Extracted phrases were present in the source reports, but source-text presence alone does not establish correct contextual weighting. Figure 3 and this table use the locked Claude validation source; in the later full-corpus source, the identity of one ≥ 2-level discrepancy changed from report 2002 to report 2188 while aggregate performance was essentially unchanged (Supplementary Material 4). Phrase-source audit summary: no fully fabricated evidence phrase was identified for either model. Partial string normalization or punctuation differences required manual review in a small number of Gemini outputs; the audit does not establish correct interpretation of every phrase.

Supplementary Table 18. Model-derived discovery score distributions (n = 3,827)

Model	Score	n	Percent	95% Wilson CI (%)
Claude Opus 4.5	0	2649	69.22%	67.74-70.66
Claude Opus 4.5	1	802	20.96%	19.70-22.27
Claude Opus 4.5	2	316	8.26%	7.43-9.17
Claude Opus 4.5	3	60	1.57%	1.22-2.01
Gemini 2.5 Pro	0	2871	75.02%	73.62-76.37
Gemini 2.5 Pro	1	635	16.59%	15.45-17.80
Gemini 2.5 Pro	2	299	7.81%	7.00-8.71
Gemini 2.5 Pro	3	22	0.57%	0.38-0.87

No independent human reference standard was available; these distributions describe model outputs, not true uncertainty prevalence. CI, confidence interval.

Supplementary Table 19. Descriptive discovery inter-model agreement

Metric	Discovery result
Quadratic weighted kappa	0.858
Exact agreement	3,353/3,827 (87.61%)
Within-one-category agreement	3,818/3,827 (99.76%)
Two-category differences	9
Three-category differences	0

Supplementary Table 20. Gemini discovery token accounting

Token type	n	Mean	SD	Median	IQR	Total
Prompt tokens	3,827	331.6	28.5	326	313-343	1,269,101
Output tokens	3,827	72.7	13.3	70	62-80	278,103
Thinking tokens	3,827	563.1	151.5	507	441-686.5	2,155,168
Total tokens	3,827	967.4	177.3	903	824.5-1,121	3,702,372

Model: gemini-2.5-pro; temperature 0; documented discovery Batch thinking budget 1,024; 3,827 successful requests and 0 failures. This value should not be assumed to match the April validation or robustness configurations. SD, standard deviation; IQR, interquartile range.

Supplementary Table 21. Cost and deployment record

Cost/deployment item	Observed value	Interpretation
Claude validation-period billing record	USD 27.17	Gross author billing associated with the April 2026 validation period; token-level attribution was insufficient to isolate a cohort-pure 100-report inference cost
Gemini validation-period billing record	Approximately USD 1.80	Author billing associated with the April 2026 validation run; not used to derive a cross-model fold-cost claim
Gemini discovery/robustness expenditure	TRY 904 (approximately USD 19 at the transaction-date rate)	Entire July 2026 workload: 3,827-report Batch API discovery plus prompt-variant and repeated-run robustness analyses; components were not separately billed
Recommended reporting	Model identifier, date, billing mode, token counts, failures/ retries, and invoice-defined workload cost	Model snapshot, request mode, and vendor pricing differ over time

These figures are run-period billing records, not normalized vendor price estimates. The discovery token accounting is provided in Supplementary Table 20. The July 2026 invoice covered the complete Gemini discovery/robustness workload and was not separable into each component; the USD equivalent depends on the transaction-date exchange rate. Because the Claude validation-period record cannot be cleanly isolated to the 100 locked inference calls, no per-report or fold-difference comparison is reported. API, Application Programming Interface.