The supplement is organized into three separate numbered series: Supplementary Materials 1–4, Supplementary Tables 1–21, and Supplementary Figures 1–4. The terminology and Figure 3 source references are synchronized with the revised manuscript.

Supplementary Material 1. Completed 44-item REFINE checklist

Item	REFINE requirement	Status	Location	Clarification
1.1	Model name, version, developer, release/ cutoff	Partial	Methods 2.2; Table 1; Limitations	Model identifiers and inference months are reported; proprietary release/ training-cutoff details are incompletely available.
1.2	Model architecture and scale	Partial	Methods 2.2; Discussion	Commercial transformer-based LLMs; proprietary architecture/parameter details are unavailable.
1.3	Pretraining/fine-tuning description	Partial	Methods 2.2; Discussion; Limitations	No study-specific fine-tuning; vendor pretraining corpora are proprietary and possible OpenI exposure is discussed.
1.4	Task input/output specification	Yes	Methods 2.2 and 2.4; Supplementary Material 2; Supplementary Table 6	Available Findings and/or Impression text is inserted under the corresponding section labels; structured JSON and ordinal mapping are specified.
1.5	Language(s)	Yes	Methods 2.1 and 2.2; Limitations	English-language reports only.
1.6	Access mode/API	Yes	Methods 2.2	Public commercial APIs; Batch API used for Gemini discovery.
1.7	Availability of code/prompts/derived data	Yes	Data and Code Availability; Supplementary Table 3; Supplementary Material 4	Prompts, scripts, audit outputs, and derived data are supplied; raw report text is not redistributed.
1.8	Local hardware/accelerator details	N/A	Methods 2.2 and 2.5	Vendor-hosted APIs; local accelerators were not controlled.
2.1	Prompt development process	Yes	Methods 2.2; Supplementary Material 3	Three prompt versions were reconstructed; final v3 was locked.
2.2	Prompting strategy	Yes	Methods 2.2; Supplementary Material 2	Zero-shot, single-turn, structured- output prompting.
2.3	Full prompt provided	Yes	Methods 2.2; Supplementary Material 2	Full English prompt and placeholders are provided.
2.4	Input context and preprocessing	Yes	Methods 2.1 and 2.2; Supplementary Material 2	Available Findings and/or Impression text only; no image, identifier, or clinical outcome input.
2.5	Conversation memory/context	Yes	Methods 2.2	Independent stateless calls; no conversation history.
2.6	Output parsing/mapping	Yes	Methods 2.2; Supplementary Table 6	JSON parsing, ordinal mapping, and source-text phrase audit are described.
3.1	Inference parameters	Yes	Methods 2.2 and 2.5; Table 1; Supplementary Tables 3 and 20	Temperature 0; Gemini discovery thinking budget 1,024; identifiers reported.
3.2	Human interaction/operator effects	Yes	Methods 2.2, 2.4, 2.5; Supplementary Material 3; Supplementary Table 16	Operator dual role, development examples, and robustness procedures are disclosed.
3.3	Selection/best-of-n/failure handling	Yes	Methods 2.2	One structured output per request; no best-of-n selection; failures reported.
4.1	Dataset source and provenance	Partial	Methods 2.1; Reference 15	Public OpenI source is cited; source demographics are limited.
4.2	Setting/modality/institution	Yes	Methods 2.1; Table 1; Figure 1	Single-institution English-language chest radiograph reports.
4.3	Ethics/consent/de-identification	Yes	Methods 2.1	Public de-identified data; IRB approval not required under local policy.
4.4	Potential pretraining contamination	Yes	Introduction; Methods 2.1; Discussion; Limitations	Public release predates model cutoffs; possible pretraining exposure is explicitly disclosed.
4.5	Synthetic data	N/A	Methods 2.1	No synthetic data were used.
4.6	Population/data composition	Partial	Methods 2.1; Discussion; Limitations	Report composition and single- institution/English/CXR scope are reported; detailed demographics are unavailable.
4.7	Reference standard/annotators	Yes	Methods 2.4; Table 1; Supplementary Tables 6 and 7	Three radiologists, majority consensus, attending rule for complete disagreement.
4.8	Image registration/segmentation	N/A	Methods 2.1 and 2.3	Text-only study; no image registration.
4.9	Missing data/exclusions	Yes	Methods 2.1; Figure 1	28 reports with neither Findings nor Impression text were excluded.
4.10	Train/validation/test separation	Partial	Methods 2.1, 2.2, 2.4; Discussion; Limitations	No fine-tuning; validation/discovery separation is explicit; pretraining overlap cannot be excluded and no external test set was available.
5.1	Primary and secondary metrics	Yes	Methods 2.5-2.6; Tables 2-5	QWK, exact agreement, ordinal error, per-category behavior, sensitivity, and robustness metrics reported.
5.2	Human agreement/reliability	Yes	Methods 2.4 and 2.6; Results 3.1; Table 2	Pairwise QWK plus Fleiss kappa and Gwet AC1/AC2; structural dependence disclosed.
5.3	Uncertainty / CI / resampling	Yes	Methods 2.6; Supplementary Tables 7-20	Bootstrap CIs and seed 42 reported; paired bootstrap used for comparative QWK.
5.4	Subgroup performance	Partial	Results 3.2 and 3.6; Tables 3 and 5	Score-specific behavior and model outputs reported; demographic subgroup data unavailable.
5.5	Error analysis	Yes	Methods 2.5; Results 3.2 and 3.5; Figure 3; Supplementary Table 17	Ordinal error distance and all three >=2-level model-reference discrepancies are described, including the contested rubric-boundary interpretation of report 1557; attending re-review is not independent adjudication.
5.6	Repeatability/stochasticity	Yes	Methods 2.5; Results 3.4; Table 5; Supplementary Table 16	Three temperature-0 runs evaluated.
5.7	Prompt sensitivity	Yes	Methods 2.5; Results 3.4; Supplementary Table 16; Supplementary Figure 4	Three prompt variants compared.
5.8	Model-version sensitivity	Partial	Methods 2.2; Table 1; Discussion; Limitations	Gemini validation and discovery used different dated snapshots. Six Claude validation predictions also differed between the locked and later source despite the same recorded identifier; their origin was not established.
5.9	Interpretability/rationale analysis	Partial	Methods 2.2; Results 3.5; Supplementary Table 17	Key phrases and selected semantic errors are assessed; no mechanistic interpretability claim is made.
5.10	Baseline comparator	Yes	Methods 2.3 and 2.6; Results 3.3; Table 4; Supplementary Tables 4, 5, 12, 13	Legacy and corrected lexicons are reported as transparent design- linked floors.
6.1	Intended use	Yes	Clinical Significance; Discussion; Conclusion	Research-stage corpus curation and reporting-language audit are proposed.
6.2	Clinical integration	No	Discussion; Conclusion	No clinical integration or triage was evaluated.
6.3	Clinical/downstream utility	No	Discussion; Limitations	Downstream utility was not measured.
6.4	Limitations and non-use cases	Yes	Discussion; Limitations; Conclusion	No risk triage, urgency assignment, autonomous decisions, or broad generalization claims.
6.5	Safety testing/monitoring	N/A	Discussion; Limitations	No prospective deployment; therefore no live safety-monitoring protocol.
6.6	Privacy/data governance	Yes	Methods 2.1; Data and Code Availability	Public de-identified text; no identifiers, images, or outcomes; raw text not redistributed.
6.7	Reproducibility/governance package	Partial	Supplementary Table 3; Supplementary Material 4; Data and Code Availability	Prompts, scripts, hashes, and audit outputs are supplied; institutional governance is outside study scope.

Status reflects the information available in this study and vendor documentation. N/A denotes items not applicable to a text-only, non-deployed API study.
Checklist reference: Mese I, Akinci D'Antonoli T, Bluethgen C, et al. Reporting checklist for foundation and large language models in medical research (REFINE): an international consensus guideline. *Diagn Interv Radiol.* 2026. [Crossref]
LLM, large language model; API, Application Programming Interface; IRB, Institutional Review Board; CXR, chest X-ray; QWK, quadratic weighted kappa; CI, confidence interval; N/A, not applicable.

Supplementary Material 2. Final locked prompt and JSON output schema

You are a radiologist analyzing the linguistic certainty in a chest X-ray report.

Report:

FINDINGS: <REPORT FINDINGS>

IMPRESSION: <REPORT IMPRESSION>

Analyze the level of diagnostic uncertainty expressed in this report. Consider:

- Hedging language (e.g., "may represent", "possibly", "suspicious for")

- But do NOT count negations like "no evidence of pneumothorax" as uncertainty

- Consider context: "consistent with X" is more certain than "may represent X"

Primary categorical output:

- uncertainty_level must be one of: none, low, moderate, high.

- The primary analysis maps uncertainty_level to the 0-3 ordinal scale: none=0, low=1, moderate=2, high=3.

Secondary descriptive output:

- uncertainty_score should be an integer from 0 to 10.

- uncertainty_score is collected only as a secondary descriptive output and is not used in the primary 0-3 analysis.

Respond ONLY with valid JSON in this exact format:

```
{
"uncertainty_level": "none|low|moderate|high",
"uncertainty_score": 0,
"key_phrases": ["list of actual hedging phrases found"],
"reasoning": "brief 1-sentence explanation"
}
```

Supplementary Material 3. Prompt version history and blinding provenance

Version	Key changes	Primary limitation/use
v1	Minimal 0-3 uncertainty instruction	No explicit negation handling or structured output; prompt-development only
v2	Added negation handling and examples	No fixed JSON schema; limited semantic weighting; prompt-development only
v3-final	Added contextual weighting, categorical none/low/moderate/high output, secondary 0-10 score, key phrases, one-sentence explanation, and JSON-only response	Locked before formal validation inference; available Findings and/or Impression text inserted under corresponding labels

Early prompt development used five preliminary OpenI reports (IDs 1, 1016, 1019, 1682, and 1000). Report ID 1000 was the only overlap with the final 100-report validation cohort. The final v3 prompt was locked before formal validation inference. Report ID 1000 had consensus Score 3, Claude Score 3, and Gemini Score 2. Exclusion of this report in the secondary n=99 analysis did not materially alter aggregate performance.

Supplementary Material 4. Reproducibility package and validation/discovery source separation

The accompanying reproducibility package contains the recovered 43-term lexicon, longest-match audit scripts and outputs, paired-bootstrap analyses, Gemini discovery metadata without redistributed report text, prompt-overlap sensitivity files, and revised manuscript-facing outputs. Raw OpenI report text is not redistributed.

Validation/discovery source separation: formal Claude validation predictions were taken from the locked analysis-ready validation file. A later full-corpus Claude output file was used only to select discovery-cohort scores after excluding all validation IDs. Six validation IDs differed between these inference sources despite the same recorded application programming interface identifier; the origin of these differences could not be established from the archived metadata and is described as unexplained provenance variability. The later file was not used to revise the primary validation results. In a provenance sensitivity audit using the later source for those six validation IDs, all six changes were one category, quadratic weighted kappa changed from 0.807 to 0.803, exact agreement remained 68/100, over-/under-grading remained 25/7, the number of discrepancies of at least two levels remained 2, and Score-3 recall remained 5/8. The identity of one Claude ≥2-level discrepancy changed from report 2002 to report 2188. Figure 3 and Supplementary Table 17 therefore retain the locked validation source and are not intended to represent the later full-corpus run.