Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264147

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

©Copyright 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Ordinal uncertainty grading in chest radiograph reports using large language models: a REFINE-aligned feasibility and methodology study

Çevik Rıfkı Bilgin
Muhammed Batuhan Daştan
Gökhan Gökalp

Uludağ University Faculty of Medicine, Department of Radiology, Bursa, Türkiye

PURPOSE

Radiologists differ in how strongly they hedge. We tested whether large language models can grade linguistic uncertainty in chest radiograph reports on a 0–3 scale and whether aggregate agreement conceals category-specific differences.

METHODS

We used 3,927 public chest radiograph reports (Open-I, Indiana University). Three radiologists independently scored 100 reports from 0 (definitive wording) to 3 (extensive ambiguity). Majority vote, with prespecified attending adjudication for seven complete three-way disagreements, formed the reference standard. Claude Opus 4.5 and Gemini 2.5 Pro scored the same reports from a fixed prompt without task-specific training. We measured agreement using Cohen's quadratic weighted kappa (QWK). We also recorded per-score recall, ordinal error, prompt and reference-standard robustness, and comparison with a keyword-counting baseline. We then scored the remaining 3,827 reports with both models.

RESULTS

Human pairwise QWK ranged from 0.421 to 0.659. Claude and Gemini had similar aggregate agreement: QWK: 0.807 [95% confidence interval (CI), 0.722–0.876] and 0.814 (95% CI, 0.735–0.877), respectively; paired difference: 0.007 (95% CI, −0.062 to 0.079). In this validation sample, Claude identified 5/8 reports exactly as Score 3 and Gemini 1/8; Score-3 precision was 5/11 and 1/3, respectively. Of Gemini's seven exact Score-3 misses, six were assigned the adjacent Score 2 category. Reference-standard sensitivities yielded QWK: 0.810–0.841; prompt-variant pairwise agreement ranged from 0.818 to 0.913; repeat runs were identical for 29/30 Claude and 28/30 Gemini reports. All 3,827 Gemini discovery requests completed successfully, with token accounting recorded for the discovery cohort.

CONCLUSION

The approach is workable at the corpus scale, but similar aggregate agreement concealed differences in exact Score-3 classification, largely involving the Score-2/3 boundary. The score reflects wording, not clinical risk, and must not be used for risk triage. Clinical or large-scale use requires external validation enriched for ambiguous cases and prior radiologist calibration.

CLINICAL SIGNIFICANCE

These scores may help radiologists audit reporting language and build balanced research evaluation sets, but downstream benefit must be demonstrated rather than assumed.

KEYWORDS

Large language models, radiology reports, linguistic uncertainty, hedging language, ordinal classification, inter-rater reliability, REFINE

Corresponding author: Çevik Rıfkı Bilgin

E-mail: cevikbilgin@uludag.edu.tr

Received 20 May 2026; revision requested 02 June 2026; last revision requested 02 September 2026; accepted 07 October 2026.

Epub: 03.11.2026

DOI: 10.4274/dir.2026.264147

You may cite this article as: Bilgin ÇR, Daştan MB, Gökalp G. Ordinal uncertainty grading in chest radiograph reports using large language models: a REFINE-aligned feasibility and methodology study. *Diagn Interv Radiol*. 03 November 2026 DOI: 10.4274/dir.2026.264147 [Epub Ahead of Print].

1. Introduction

Radiology reports communicate not only imaging observations but also the radiologist's degree of commitment to an interpretation. Expressions such as "may represent," "cannot exclude," and "suggestive of" appropriately signal ambiguity, and referring clinicians act on them: A finding reported as "cannot exclude a neoplastic process" rather than "consistent with post-inflammatory scarring" can change decisions about additional imaging, follow-up interval, specialist referral, and the urgency of those steps. Uncertainty phrasing is therefore part of the diagnostic message rather than stylistic ornament. Its use, however, is not standardized: The frequency and intensity of hedging vary across radiologists, subspecialties, and institutional cultures,[1-4] as well as with reader experience.[5] Because patients increasingly read their own reports without interpretive support, this variability reaches them directly.[2]

The same variability carries over into the computational use of reports. Report text serves as the label source, retrieval target, and training corpus for imaging and natural language processing systems, so a heavily hedged report and a definitive one may become indistinguishable once both are reduced to a categorical finding label. Computational approaches to report uncertainty have progressed from dictionaries and rules for negation and uncertainty detection[6,7] to contextualized supervised classifiers.[3,8,9] Lexicon systems are transparent and inexpensive but vulnerable to term overlap, substring artifacts, negation, and context. Purpose-built neural systems capture context but require labeled data, task-specific training, and implementation infrastructure. General-purpose large language models (LLMs) offer a third approach: At inference time, they can be instructed to integrate qualifiers, negation, differential framing, and sentence context without task-specific weight updates.[10-12] Structured retrieval studies show that such models can normalize complex radiology prose into predefined fields, but accurately extracting what was reported does not show that they reliably recover how strongly it was asserted.[12]

Main points

- Claude and Gemini reached almost identical aggregate agreement with the multi-reader consensus, yet exact Score-3 recall differed (5/8 vs. 1/8), largely at the adjacent Score-2/3 boundary.
- The rubric grades how uncertainty is worded, not disease severity, diagnostic risk, or urgency, and must not be used as a triage signal.
- Aggregate agreement can conceal operationally relevant category-specific behavior; per-category recall and precision, error direction, and auditable lexicon and model provenance should be reported alongside it.

Most prior computational work has framed the task as term detection, a binary certain/uncertain decision, or assignment to a small set of certainty categories.[3,6-9] An ordinal framework asks a different question. Binary detection asks whether hedging is present; ordinal grading asks how much, which is the quantity relevant to stratifying a corpus. An ordered scale preserves more linguistic information than a binary label while remaining simpler to annotate than an unconstrained continuous score, and it allows a one-category error to be weighted differently from a two- or three-category error. Its validity, however, rests on two conditions: that radiologists can apply the rubric consistently, and that aggregate metrics reveal rather than conceal behavior in the rare, most heavily hedged category.

These two conditions determined the study design. Because variability between readers is part of the phenomenon under examination, a single-annotator reference standard would confound one reader's idiosyncrasy with the construct being measured; we therefore used three readers with deliberately different experience levels and report their agreement as a primary result rather than a preliminary quality check. Because uncertainty categories are severely imbalanced in unselected reports, chance-corrected coefficients are sensitive to marginal prevalence and the assumed chance model;[13] therefore, we report unweighted and ordinally weighted coefficients as a complementary panel rather than selecting a single figure. Because a model can achieve a favorable weighted kappa while rarely predicting the highest category, we report per-category recall and the direction of ordinal error alongside aggregate agreement.

We therefore conducted a REFINE-aligned feasibility and methodology assessment of zero-shot ordinal linguistic-uncertainty grading in chest radiograph reports.[14] The objectives were to (i) quantify inter-reader reliability of a four-level rubric; (ii) compare two commercial LLM families with a transparent, design-linked lexicon floor using paired analyses; (iii) examine category-specific behavior, prompt sensitivity, stochasticity, and semantic errors; (iv) audit the executable lexicon implementation; and (v) describe full-cohort model behavior and computational requirements without treating model-derived discovery labels as clinical ground truth.

2. Methods

2.1 Study design, source corpus, eligibility, and sampling frame

This retrospective feasibility and methodology study used the publicly available, de-identified Open-I Indiana University Chest X-ray Collection.[15] The source archive contained 3,955 reports. Reports were eligible when at least one of the Findings or Impression sections contained parsable text. Of the 3,955 source reports, 3,419 contained both sections, 6 contained Findings only, 502 contained Impression only, and 28 contained neither section. The 28 reports with neither Findings nor Impression text were excluded without imputation, leaving 3,927 eligible reports (Figure 1). The corpus contains English-language free-text reports from routine chest radiograph examinations at a single United States institution. Only report text was analyzed; images, patient identifiers, and clinical outcomes were not used. Under local policy, the analysis of public de-identified data did not require institutional review board approval. Reporting follows the REFINE checklist for foundation and LLMs in medical research;[14] the completed checklist (Supplementary Material 1) and an item-by-item reporting map (Supplementary Table 1) accompany this article. Methods are organized by analytical component: corpus and sampling frame, LLM inference, lexicon comparator and its implementation audit, human reference standard, robustness and error analysis, and statistical analysis.

A validation cohort of 100 reports was selected using seed 42 by balanced random sampling from four strata defined by the original executable lexicon count: 0, 1, 2, and at least 3 distinct matched terms (25 reports per stratum). The remaining 3,827 reports formed a discovery cohort. The validation size was chosen as a pragmatic feasibility sample rather than a definitive powered test of rare-category performance. Specifically, the final reference standard contained only eight Score-3 reports.

The actual sampling and baseline implementation was recovered from the executable notebook and comprised 43 unique terms across epistemic, lexical, and evidential/exclusion categories (Supplementary

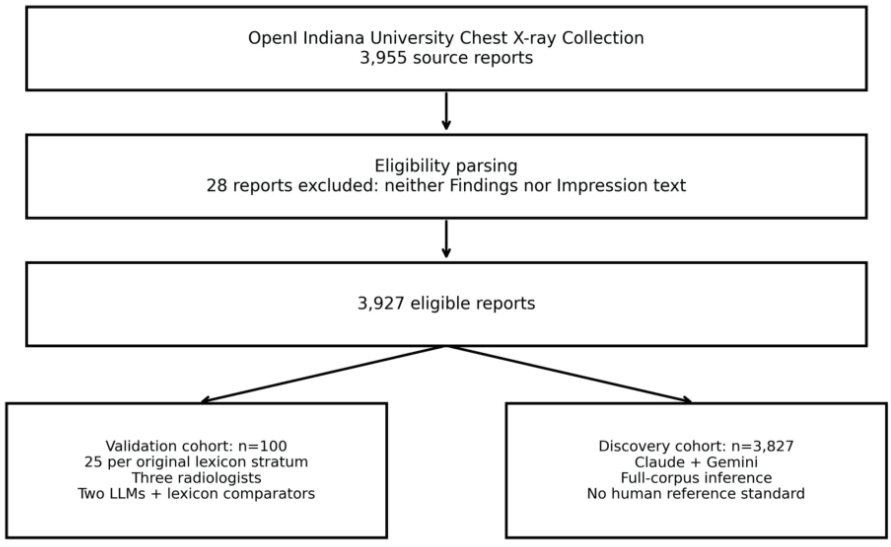


Figure 1. Study workflow. The 3,955-report Open-I source corpus yielded 3,927 eligible reports after exclusion of 28 reports lacking both Findings and Impression text. A 100-report validation cohort was sampled from four original lexicon strata, annotated by three radiologists, and evaluated using two large language models and lexicon comparators. The remaining 3,827 reports underwent descriptive full-cohort inference. Discovery outputs lack a human reference standard. LLM, large language model.

Table 2). For the original implementation, lowercased combined Findings and Impression text was searched by raw substring membership, and each distinct term contributed at most once. The count was mapped to the ordinal score as 0 to Score 0, 1 to Score 1, 2 to Score 2, and at least 3 to Score 3. The 59-term list in the previous supplementary documentation did not match the executable implementation and was not used for cohort stratification or baseline scoring. The three semantic categories were descriptive labels for the term list; the score depended only on the number of matched terms and not on category membership.

2.2 Large language model inference and output mapping

Two commercially available LLMs from independent vendors were assessed in the validation cohort: Claude Opus 4.5 [Anthropic; Application Programming Interface (API) identifier claude-opus-4-5-20251101] and Gemini 2.5 Pro (Google; validation identifier gemini-2.5-pro-preview-05-06). Formal validation inference was performed in April 2026 at temperature 0 using zero-shot prompting without task-specific fine-tuning.

The locked prompt inserted available Findings and/or Impression text under the corresponding section labels, instructed the model to consider hedging and semantic context, excluded negated findings as uncertainty, and required JSON output (Supplementary Material 2). Prompt-development history and blinding provenance are documented in Supplementary Material 3, and model/reproducibility parameters are summarized in Supplementary Table 3.

The primary model field was uncertainty-level, mapped as none = 0, low = 1, moderate = 2, and high = 3 according to the four-level rubric defined in Section 2.4. A 0–10 uncertainty-score was collected as a secondary structured output but was not used analytically because the rubric did not support equal metric distances between values. Extracted key phrases were checked against source report text; source-text presence did not imply correct contextual weighting.

For full discovery deployment, Claude outputs were available for all 3,827 discovery reports using the same API identifier as the validation model; the later full-corpus file was used only for discovery IDs and did not overwrite the locked validation scores. Six predictions for validation identifiers differed between the locked and later Claude sources despite the same recorded API identifier. The cause could not be established from archived metadata; these differences were treated as unexplained provenance variability (Supplementary Material 4). Gemini was applied to all 3,827 discovery reports by Batch API in July 2026 using gemini-2.5-pro, temperature 0, and a documented thinking budget of 1,024 tokens; all requests completed successfully. The archived records document this 1,024-token budget for the July discovery Batch run only; they do not establish a directly comparable reasoning-budget setting for the April Gemini validation run or the robustness runs, and no additional reasoning-budget parameter was documented for the Claude validation run. Because Gemini validation and discovery used different dated API snapshots and incompletely matched inference-configuration metadata, discovery outputs were treated as descriptive model behavior rather than an extension of validation performance.

2.3 Legacy lexicon floor and implementation audit

Beyond its role in defining the sampling strata, the same lexicon count was retained as the comparator against which model outputs were assessed. Because the same count defined the four sampling strata and the baseline prediction, the validation baseline distribution was fixed at 25 reports in each category by design. The legacy lexicon was therefore interpreted as a transparent, context-blind floor rather than a calibrated independent comparator. This design linkage was addressed directly in interpretation and through paired bootstrap comparisons rather than marginal confidence interval (CI) overlap.

An implementation audit replaced raw substring membership with alphanumeric word boundaries and longest, non-overlapping matching. Two corrected summaries were evaluated: occurrence count, where each accepted non-overlapping match contributed; and distinct-term count, where each repeated accepted term contributed once. The original 43-term implementation reproduced the archived validation scores in all 100 reports. Audit results were quantified in both validation and discovery cohorts; the original validation sample was not reselected. The resulting validation score transitions are shown in Supplementary Figure 1, with validation and discovery lexicon-audit summaries presented in Supplementary Tables 4, 5.

2.4 Ordinal rubric, reference standard, and blinding

Three radiologists independently scored all 100 validation reports: a second-year resident, a senior resident with approximately 4 years of radiology training, and an attending radiologist with 23 years of experience.

The structured rubric defined Score 0 as definitive language without meaningful uncertainty; Score 1 as mild or localized qualification; Score 2 as moderate uncertainty with a meaningful differential or unresolved characterization; and Score 3 as extensive ambiguity, multiple competing explanations, or an explicit need for additional evaluation because the report could not resolve a diagnostically important finding. Technical limitations alone were not automatically treated as diagnostic uncertainty (Supplementary Table 6). A four-level ordinal scale was adopted deliberately: A continuous score would imply metric distances the rubric cannot support, and a raw term count is context-blind by construction, whereas four ordered categories were the smallest set that separated definitive language, incidental qualification, an unresolved differential, and irreducible ambiguity.

The rubric grades the linguistic expression of uncertainty in the report. It does not grade disease severity, probability of harm, urgency, or clinical risk. A routine recommendation for additional imaging may reflect ambiguity, defensive reporting, or standard workup; its presence was interpreted in context rather than as an automatic Score 3.

Consensus was determined by majority vote. In seven reports with complete three-way disagreement, the attending score was adopted according to the prespecified adjudication rule. Readers were blinded to each other's scores, lexicon outputs, and final model scores. Because the second-year resident also operated the API workflow, model outputs were generated before manual annotation and were stored without being reviewed during annotation; this dual role remains a structural limitation addressed by attending-reference sensitivity analysis.

Early prompt development used five preliminary Open-I reports (IDs 1, 1016, 1019, 1682, and 1000). Comparison with the final validation identifiers showed that report ID 1000 was the only overlap with the 100-report validation cohort. Complete separation between prompt development and validation therefore cannot be claimed for this single report. The final v3 prompt was locked before formal validation inference, and an additional sensitivity analysis excluded ID 1000 while retaining the prespecified n = 100 cohort as the primary analysis.

2.5 Robustness, error analysis, and computational accounting

Prompt sensitivity was assessed in a simple random sample of 50 reports drawn without replacement from the 3,927-report eligible corpus using NumPy (NumPy Developers; open-source, https://numpy.org) with random seed 42. Repeated-run stochasticity was assessed in a 30-report sample generated from the same ordered eligible-report frame using the same seed; the resulting 30 identifiers corresponded to the first 30 identifiers of the archived 50-report robustness sample. Forty-nine of the 50 prompt-sensitivity reports were outside the final validation cohort, with report ID 569 as the sole overlap; none of the 30 repeated-run reports overlapped the validation cohort. The locked prompt was compared with a shortened direct variant and a chain-of-thought-style variant; only categorical outputs, not hidden reasoning, were analyzed. Exact archived identifiers are included in the reproducibility package. These subset sizes were pragmatic stress tests intended to detect substantial instability rather than powered equivalence studies.

For each model, exact agreement, overgrading, undergrading, and absolute ordinal error were recorded, as well as the proportions of disagreements one, two, or three levels from the reference. Per-category precision, recall, and F1 were calculated descriptively. Model–reference discrepancies of at least two ordinal levels were identified for case-level semantic review, including consideration of contextual misweighting and potentially contested rubric boundaries.

All three model–reference discrepancies of at least two ordinal levels underwent post-hoc review by the attending radiologist after the primary reference standard had been locked. This review was used solely to characterize the error mechanism and did not alter the original consensus labels or any quantitative analysis. Because the attending radiologist contributed to the original reference standard, this review constitutes confirmation of the locked consensus rather than independent adjudication.

Available billing records were not sufficiently comparable to support a normalized cross-model inference-cost comparison. The July Gemini batch workflow additionally recorded prompt, output, thinking, and total token counts. Detailed billing provenance and expenditure records are retained in the supplementary cost record.

2.6 Statistical analysis

The primary endpoint was Cohen's quadratic weighted kappa (QWK) between each model and the three-reader consensus. Ninety-five percent percentile bootstrap CIs were calculated from 1,000 report-level resamples with seed 42. Because model and lexicon predictions were made on the same reports, paired bootstrap distributions of the QWK difference were used for comparative analyses. The paired Claude–Gemini difference was also estimated; aggregate equivalence was not inferred from overlapping or non-overlapping marginal CIs.

Human pairwise reliability was summarized with QWK. Multi-reader nominal agreement was summarized using unweighted Fleiss kappa and unweighted Gwet AC1. Because the rubric is ordered, Gwet AC2 was additionally computed using linear and quadratic weights. These coefficients differ in weighting and chance-agreement assumptions and were reported as complementary statistics rather than as a formal diagnosis of a prevalence paradox.

Sensitivity analyses used (i) an attending-reference standard generated by the senior-attending rule with attending resolution of disagreements, which is attending-equivalent by construction; (ii) exclusion of the seven complete-disagreement reports while retaining the primary consensus; (iii) the combination of the attending reference and exclusion of those seven reports; and (iv) exclusion of the single validation report (ID 1000) that overlapped with early prompt development, with model QWK and paired Gemini-minus-Claude QWK recalculated. The overlap-exclusion analysis was secondary; the prespecified primary cohort remained n = 100. Wilson intervals were used for discovery score proportions. Discovery-cohort analyses were descriptive only because no independent human reference was available. Analyses were performed in Python with a fixed seed of 42.

3. Results

3.1 Cohort and reference-standard reliability

The final validation reference distribution was Score 0 in 40 reports, Score 1 in 31, Score 2 in 21, and Score 3 in 8 (Table 1). Complete three-way disagreement occurred in 7/100 reports. Human pairwise QWK ranged from 0.421 to 0.659 (Table 2). The senior resident's QWK against the consensus was 0.576, indicating that the rubric was not uniformly intuitive across training levels. The consensus coincided with the attending score in 89/100 reports. The attending-consensus QWK of 0.902 is therefore partly structural because the attending contributed to the majority and supplied the adjudicated score in all complete disagreements.

Table 1. Study cohorts, reference distribution, and model snapshots

Characteristic	Value
Source corpus	Open-I Indiana University Chest X-ray Collection
Source reports	3,955
Excluded	28 reports with neither Findings nor Impression text
Eligible	3,927
Validation cohort	100; 25 reports from each original lexicon stratum (0, 1, 2, ≥ 3 terms)
Discovery cohort	3,827; no human reference standard
Validation consensus distribution	Score 0: 40; Score 1: 31; Score 2: 21; Score 3: 8
Readers	Postgraduate Year 2 resident; senior resident (approx. 4 years); attending radiologist (23 years)
Claude validation	claude-opus-4-5-20251101; temperature 0; locked April 2026 validation run
Claude discovery	Same Application Programming Interface (API) identifier; later full-corpus source used only for discovery IDs. Six validation predictions differed between sources; locked validation scores were not overwritten (Supplementary Material 4).
Gemini validation	gemini-2.5-pro-preview-05-06; temperature 0; April 2026
Gemini robustness/discovery	gemini-2.5-pro; temperature 0; July 2026; thinking budget 1,024 documented for discovery Batch run only

Validation and later Gemini robustness/discovery analyses used different dated Application Programming Interface snapshots. The six Claude validation prediction differences are described in Methods 2.2, Results 3.4, and Supplementary Material 4. Discovery has no human reference standard.

Table 2. Human reference-standard reliability

Agreement comparison/coefficient	n	Estimate	95% bootstrap confidence interval
PGY-2 vs. senior resident	100	0.421	0.246–0.572
PGY-2 vs. attending	100	0.659	0.506–0.782
Senior resident vs. attending	100	0.497	0.322–0.639
Fleiss kappa (unweighted)	100	0.273	0.181–0.357
Gwet AC1 (identity/unweighted)	100	0.334	0.256–0.422
Gwet AC2 (linear weights)	100	0.541	0.460–0.625
Gwet AC2 (quadratic weights)	100	0.694	0.602–0.775

Pairwise comparisons use Cohen's quadratic weighted kappa. Fleiss kappa and Gwet AC1 are unweighted; AC2 uses the indicated ordinal weights. All confidence intervals are report-level bootstrap percentile intervals (1,000 resamples; seed 42). PGY-2, Postgraduate Year 2.

Unweighted Fleiss kappa was 0.273 (95% CI, 0.181–0.357) and unweighted Gwet AC1 was 0.334 (95% CI, 0.256–0.422). Ordered-category estimates were higher: linear-weighted AC2 was 0.541 (95% CI, 0.460–0.625) and and quadratic-weighted AC2 was 0.694 (95% CI, 0.602–0.775). The differences reflect both the ordinal weighting and the chance model; they do not by themselves establish a prevalence-paradox explanation. Pairwise reader and reader-to-consensus estimates, including the structural-dependence warning for reader-to-consensus comparisons, are provided in Supplementary Table 7.

3.2 Validation performance and category-specific behavior

Claude yielded QWK: 0.807 (95% CI, 0.722–0.876) and Gemini QWK: 0.814 (95% CI, 0.735–0.877) against the consensus (Table 3). Exact agreement was 68/100 for Claude and 71/100 for Gemini. The paired Gemini-minus-Claude QWK difference was 0.007 (95% CI, −0.062 to 0.079), providing no evidence of a meaningful aggregate difference in this sample.

Aggregate QWK concealed a difference in exact Score-3 classification (Figure 2). Claude correctly identified 5/8 Score-3 reports (recall, 62.5%) and assigned Score 3 to 11 reports in total, giving Score-3 precision of 5/11 (45.5%). Gemini identified 1/8 Score-3 reports (recall, 12.5%) and assigned Score 3 to 3 reports, giving Score-3 precision of 1/3 (33.3%). For Scores 0, 1, and 2, Gemini recall was 87.5%, 64.5%, and 71.4%, respectively; Claude recall was 77.5%, 58.1%, and 66.7%. Of the 8 reference Score-3 reports, Claude assigned all 8 to Score 2 or 3, whereas Gemini assigned 7 to Score 2 or 3. Among the 7 reports not classified exactly as Score 3 by Gemini, 6 were assigned Score 2 and only 1 was assigned Score 1. Gemini's lower exact Score-3 recall therefore largely reflected adjacent-category disagreement rather than a complete failure to recognize linguistic uncertainty. Near-equivalent aggregate QWK should not be interpreted as equivalent exact-category behavior, and recall alone should not be treated as a complete measure of high-category utility.

Because only 8 reference-standard reports were assigned Score 3, reclassification of a single case would change recall by 12.5 percentage points; no formal inferential comparison of Score-3 recall or precision was performed. Claude disagreed with the consensus in 32 reports: 25 overgrades and 7 undergrades. Thirty of 32 errors were one level, and two were two levels. Gemini disagreed in 29 reports: 17 overgrades and 12 undergrades; 28 were one level, and one was two levels. No three-level error occurred. The proportion of all reports within one category of consensus was 98% for Claude and 99% for Gemini. Detailed overall and category-specific metrics and the two model confusion matrices are provided in Supplementary Tables 8–11 and Supplementary Figures 2, 3.

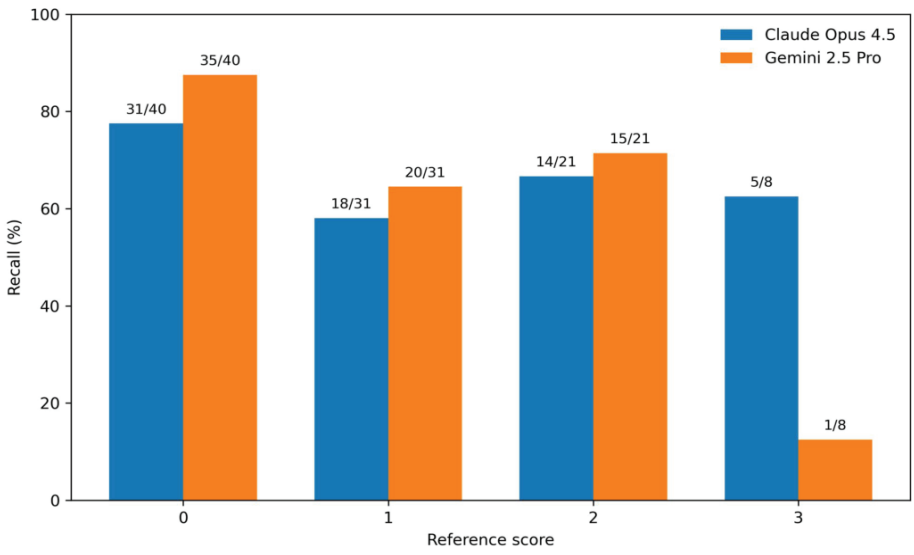


Figure 2. Per-category recall against the multi-reader consensus. Bars show exact-category recall and correct/reference counts for Claude Opus 4.5 and Gemini 2.5 Pro. For Score 3 (5/8 vs. 1/8), six of Gemini's seven exact misses were classified as adjacent Score 2.

3.3 Lexicon audit and paired comparisons

The legacy 43-term lexicon reproduced the archived validation scores in 100/100 reports and yielded QWK: 0.622 (95% CI, 0.518–0.712). Because the baseline score and sampling stratum were the same variable, its validation predictions were fixed at 25 reports per category and should be interpreted as a design-linked floor.

This linkage also explains the floor's apparently favorable Score-3 recall of 6/8 in Table 3: Twenty-five validation reports were assigned to the highest category by construction against a reference support of only 8, at the cost of a Score-1 recall of 8/31. Per-category values in the lexicon rows therefore describe the sampling frame and are not comparable with model recall.

The corrected longest-match occurrence implementation changed 32/100 validation scores and yielded QWK: 0.572 (95% CI, 0.440–0.677). In the discovery cohort, occurrence counts changed in 231/3,827 reports and ordinal scores changed in 213/3,827 (5.6%); across all 3,927 eligible reports, 245 scores changed (6.2%) (Supplementary Tables 4 and 5; Supplementary Figure 1). The audit demonstrates that overlap and boundary handling materially affect a subset of reports but does not retrospectively change the sampled validation cohort.

In paired bootstrap analysis, Claude exceeded the legacy lexicon by 0.185 QWK (95% CI, 0.103–0.287) and Gemini by 0.192 (95% CI, 0.097–0.297). Compared with the corrected occurrence implementation, the corresponding differences were 0.235 (95% CI, 0.141–0.353) and 0.242 (95% CI, 0.146–0.353) (Table 4). These estimates quantify within-report differences, but the design-linked baseline remains a floor rather than a fully independent benchmark. Detailed lexicon confusion matrices are provided in Supplementary Tables 12 and 13.

3.4 Sensitivity, prompt robustness, and stochasticity

The attending-reference sensitivity analysis yielded QWK: 0.831 for Claude and 0.832 for Gemini. Excluding the seven complete-disagreement reports yielded QWK: 0.816 and 0.810, respectively. Combining the attending reference with exclusion of those seven reports yielded QWK: 0.841 and 0.829 (Table 4). These analyses did not materially change the aggregate result, although the attending-reference analyses are not independent of the attending reader.

Exclusion of the single validation report that overlapped with early prompt development (report ID 1000; n = 99) did not materially alter performance: Claude QWK was 0.799 (95% CI, 0.714–0.866), Gemini QWK was 0.814 (95% CI, 0.731–0.873), and the paired Gemini-minus-Claude difference was 0.015 (95% CI, −0.061 to 0.086). Counts of model–reference discrepancies of at least two levels remained two for Claude and one for Gemini. Because ID 1000 was a consensus Score-3 report classified as Score 3 by Claude and Score 2 by Gemini, Score-3 recall in this secondary analysis was 4/7 for Claude and 1/7 for Gemini (Supplementary Tables 14 and 15).

Across the 50-report robustness subset, pairwise QWK ranged from 0.818 to 0.899 for Claude and from 0.861 to 0.913 for Gemini (Table 5; Supplementary Table 16; Supplementary Figure 4). Across three temperature-0 runs, all three scores were identical for 29/30 Claude reports (96.7%) and 28/30 Gemini reports (93.3%). Temperature 0 therefore reduced but did not eliminate output variability.

In a Claude provenance sensitivity analysis, substituting the six differing validation predictions from the later full-corpus source changed QWK from 0.807 to 0.803. Exact agreement remained 68/100, over-/undergrading remained 25/7, and Score-3

Table 3. Validation performance against the multi-reader consensus

Method	QWK (95% CI)	Exact	Within 1	MAE	Recall 0	Recall 1	Recall 2	Recall 3
Claude Opus 4.5	0.807 (0.722–0.876)	68/100	98/100	0.34	31/40	18/31	14/21	5/8
Gemini 2.5 Pro	0.814 (0.735–0.877)	71/100	99/100	0.30	35/40	20/31	15/21	1/8
Legacy 43-term lexicon	0.622 (0.518–0.712)	46/100	87/100	0.67	22/40	8/31	10/21	6/8
Corrected lexicon: occurrence	0.572 (0.440–0.677)	45/100	89/100	0.66	22/40	12/31	7/21	4/8

Recall is shown as correct/reference support. The legacy lexicon distribution was fixed at 25 reports per score by design; its recall values are not an independent performance comparison. QWK, quadratic weighted kappa; MAE, mean absolute ordinal error; CI, confidence interval.

Table 4. Sensitivity analyses and paired quadratic weighted kappa (QWK) comparisons

Panel A. Reference-standard 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)

Panel A note: The attending-reference analyses are attending-equivalent by construction because the attending resolved senior-attending disagreements. QWK, quadratic weighted kappa; CI, confidence interval.

Panel B. Paired bootstrap difference in QWK

Comparison (A vs. B)	Delta QWK (A–B)	95% paired bootstrap CI
Claude vs. legacy lexicon	+0.185	+0.103 to +0.287
Gemini vs. legacy lexicon	+0.192	+0.097 to +0.297
Claude vs. corrected occurrence lexicon	+0.235	+0.141 to +0.353
Gemini vs. corrected occurrence lexicon	+0.242	+0.146 to +0.353
Claude vs. corrected distinct lexicon	+0.222	+0.122 to +0.350
Gemini vs. corrected distinct lexicon	+0.229	+0.129 to +0.349
Gemini vs. Claude	+0.007	−0.062 to +0.079

Panel B note: Predictions were paired at report level (1,000 bootstrap resamples; seed 42). QWK, quadratic weighted kappa; CI, confidence interval.

recall remained 5/8. The number of ≥2-level discrepancies remained two, although one such report changed from ID 2002 to ID 2188 (Supplementary Material 4).

3.5 Semantic error assessment

No fully fabricated key phrase was identified in the source-text audit. Three model–reference discrepancies of at least two ordinal levels were identified (Figure 3). Claude assigned Score 3 to two reports with consensus Score 1. Report 1557 is presented as a contested rubric-boundary case: Multiple qualifiers and competing differential formulations 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. In the locked validation source, Claude's other ≥2-level discrepancy was report 2002, which combined a possible rib fracture with a conditional recommendation for dedicated rib views. Gemini's sole ≥2-level discrepancy was report 397, assigned Score 1 despite a consensus Score 3 and an explicit computed tomography (CT) recommendation. This was its only reference Score-3 report assigned below Score 2.

Six of Gemini's seven exact Score-3 misses were assigned Score 2; thus, its lower exact Score-3 recall largely reflected the Score-2/3 boundary rather than a distinct generalized failure to recognize uncertainty. Claude made 25 assignments above and 7 below the locked reference score, but the interpretation of report 1557 as a model error is contested. Figure 3 uses the locked Claude validation source; in the later full-corpus Claude source, six validation predictions differed by one category, and the identity of one ≥2-level discrepancy changed from report 2002 to report 2188, whereas aggregate results changed little. Post-hoc attending re-review maintained the locked scores but was not independent adjudication because the attending contributed to the original reference standard. Case details and provenance information are provided in Supplementary Table 17 and Supplementary Material 4.

3.6 Discovery-cohort deployment

Both model families produced outputs for all 3,827 discovery reports. Claude assigned Score 0, 1, 2, and 3 to 2,649 (69.2%), 802 (21.0%), 316 (8.3%), and 60 (1.6%) reports, respectively. Gemini assigned the same categories to 2,871 (75.0%), 635 (16.6%), 299 (7.8%), and 22 (0.6%) reports (Table 5; Figure 4). Claude classified 376/3,827 reports (9.8%) as Score 2 or 3, compared with 321/3,827 (8.4%) for Gemini.

Descriptive discovery inter-model QWK was 0.858, exact agreement was 87.6%, and 99.8% of reports were within one category. Nine reports differed by two categories and none by three; all nine two-category differences were in the direction of a higher Claude score. These values describe output convergence only; without human annotation, they do not estimate clinical prevalence or accuracy.

Gemini batch inference processed 3,827/3,827 requests without failure. Mean total token use was 967.4 per report (median, 903; interquartile range, 824.5–1,121.0), comprising a mean of 331.6 prompt, 72.7 output, and 563.1 thinking tokens. Total recorded token use was 3,702,372 (Supplementary Tables 18–20). The corresponding invoice-defined expenditure and limitations of cost comparison are documented in Supplementary Table 21.

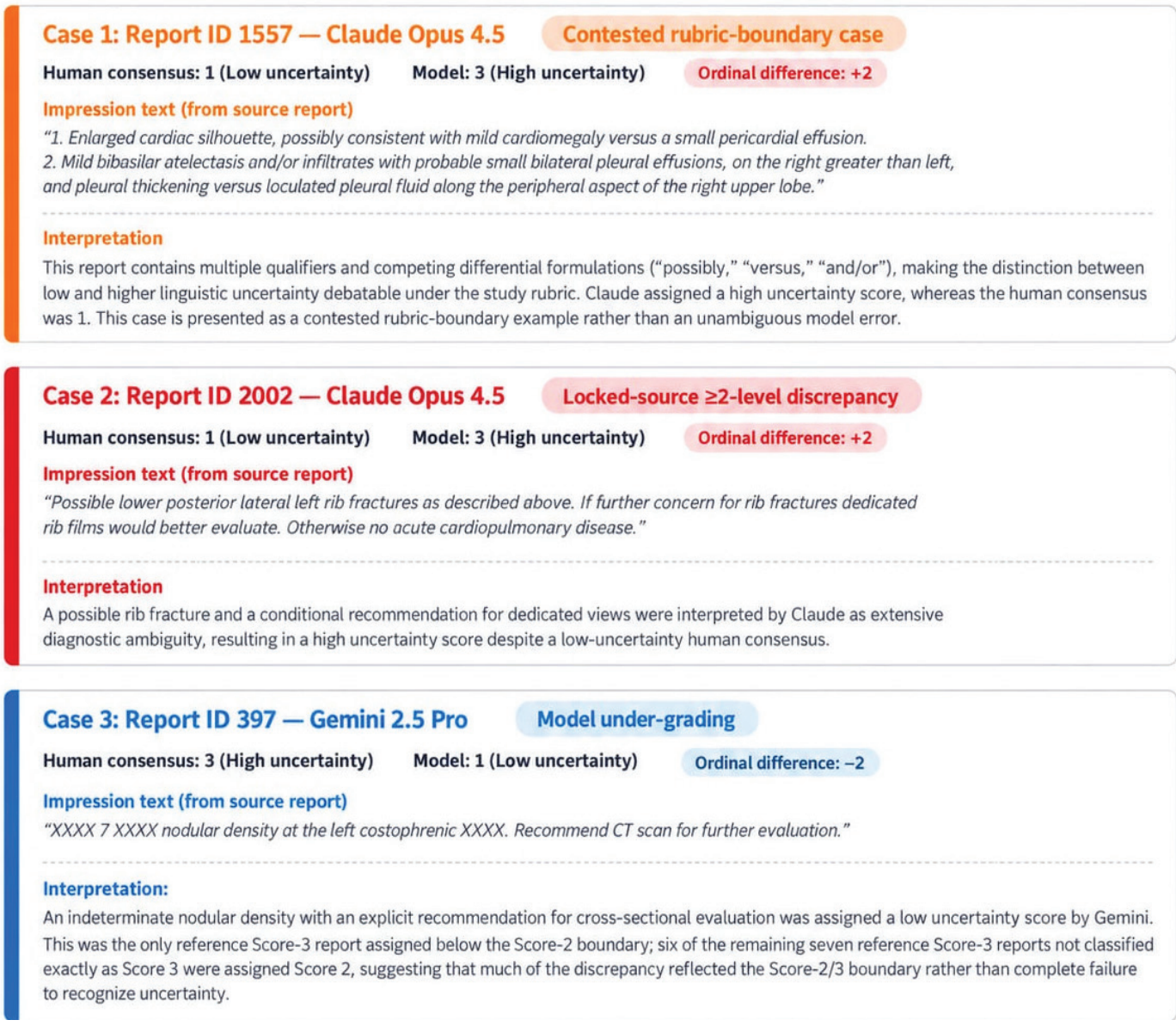


Figure 3. All three model–reference discrepancies of at least two ordinal levels, including one contested rubric-boundary case. Claude assigned Score 3 to two reports with consensus Score 1 (IDs 1557 and 2002); the interpretation of report 1557 is contested under the study rubric. Gemini assigned Score 1 to one consensus Score-3 report (ID 397). The figure uses the locked Claude validation source. In the later full-corpus Claude source, six validation predictions differed by one category, and the identity of one ≥2-level discrepancy changed from report 2002 to report 2188, whereas aggregate performance was essentially unchanged (Supplementary Material 4). Post-hoc attending re-review maintained the locked consensus scores but was not independent adjudication because the attending contributed to the original reference standard. The locked labels were retained, and no quantitative results were altered.

Table 5. Discovery deployment and robustness

Panel A. Model-derived discovery score distribution (n = 3,827)

Model	Score 0	Score 1	Score 2	Score 3
Claude Opus 4.5	2,649 (69.2%)	802 (21.0%)	316 (8.3%)	60 (1.6%)
Gemini 2.5 Pro	2,871 (75.0%)	635 (16.6%)	299 (7.8%)	22 (0.6%)

Panel B. Prompt sensitivity and three-run consistency

Model	Prompt-variant pairwise QWK range	All three runs identical
Claude Opus 4.5	0.818–0.899	29/30 (96.7%)
Gemini 2.5 Pro	0.861–0.913	28/30 (93.3%)

Discovery distributions are model-derived, with no independent human reference. Prompt sensitivity used 50/3,927 reports (49 outside validation); repeated-run analysis used 30 reports (no validation overlap). Gemini discovery thinking-budget metadata do not establish comparable settings for other runs. QWK, quadratic weighted kappa.

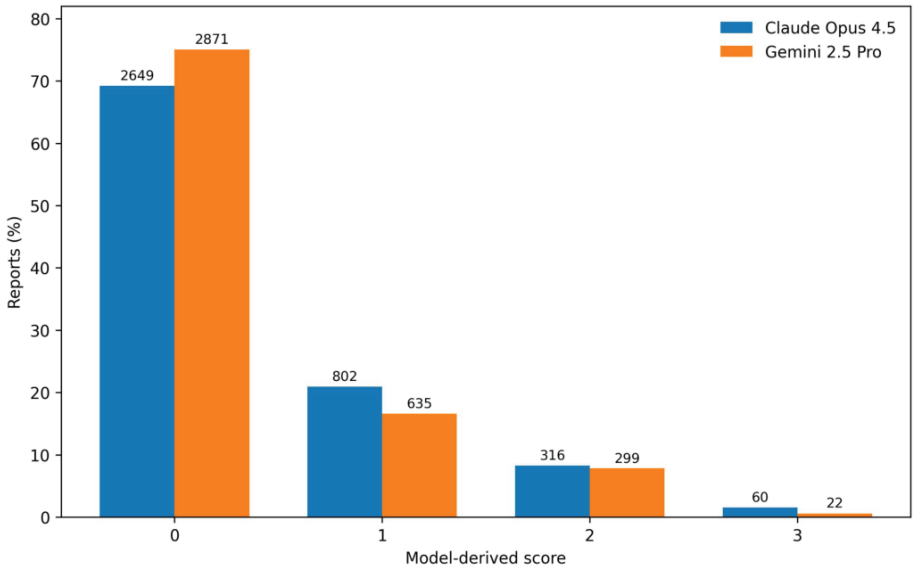


Figure 4. Model-derived uncertainty distribution in the 3,827-report discovery cohort. Claude and Gemini distributions are descriptive outputs without human adjudication and should not be interpreted as clinical uncertainty prevalence.

4. Discussion

This study demonstrates that a REFINE-aligned pipeline for zero-shot ordinal grading of linguistic uncertainty can be executed reproducibly across two LLM families, but it also shows why aggregate agreement alone is insufficient. Claude and Gemini had nearly identical QWK against the 100-report consensus, yet their exact Score-3 classification differed. In this validation sample, Gemini identified 1/8 reference Score-3 reports exactly, whereas Claude identified 5/8; Claude's Score-3 precision was 5/11, and Gemini's was 1/3. However, six of Gemini's seven exact Score-3 misses were adjacent Score-2 assignments, and Claude assigned all eight reference Score-3 reports to Score 2 or 3. The difference therefore mainly concerned the moderate/high category boundary. For corpus-curation workflows, missed high-category reports and false-positive high-category flags create different manual-review burdens. Model selection should therefore be based on use-specific error costs and category-level behavior, not a single QWK or a single category-level metric.

The reference standard itself is a central limitation rather than a neutral ground truth. Pairwise agreement was moderate to substantial by conventional benchmarks;[16] seven reports produced complete disagreement, and the senior resident's agreement with the consensus was only 0.576. Although three readers were used to reduce dependence on any single annotator, the final consensus matched the attending score in 89/100 reports, and the attending also supplied the adjudicated score in all complete disagreements. The resulting standard therefore remained strongly attending-influenced rather than fully independent of a single expert. These findings suggest that the rubric requires formal calibration using representative examples and explicit discussion of routine recommendations, multifocal findings, and defensive language. Future studies should measure pre- and post-calibration reliability and, for the highest category, use enriched complex reports with independent expert adjudication.

The reliability analysis also illustrates why coefficients must be named precisely. Unweighted Fleiss kappa and unweighted AC1 were relatively close, whereas weighted AC2 values were higher because near-miss ordinal disagreements receive partial credit and the chance model differs. Reporting the higher weighted coefficient as AC1 or treating the difference as proof of a prevalence paradox would be misleading. A transparent panel of unweighted and ordinally weighted metrics is more informative.

The lexicon analysis serves two purposes. First, it provides an interpretable context-blind floor. Second, its audit exposes how implementation choices propagate into both cohort construction and comparator scores. The original 43-term list, not the previously documented 59-term list, was the executable source. Longest-match boundary correction changed 5.6% of discovery ordinal scores. Nevertheless, because the validation sample was selected from the original strata and the score mapping was identical to those strata, neither legacy nor corrected lexicon results represent a calibrated independent benchmark. The paired bootstrap demonstrates within-report improvement of the LLMs over the lexicon floor without relying on invalid CI overlap claims.

The models' divergence from lexicon outputs supports context-sensitive processing, not independence from pretraining exposure. The Open-I text may have been encountered during pretraining, but the report is supplied directly in each prompt, and the 0–3 study labels did not exist before this annotation project. Corpus familiarity could still aid recognition of style, template structure, or common phrasing. Accordingly, the present estimates are feasibility bounds rather than model-unseen generalization estimates; external evaluation on later institutional reports remains essential.

The score should not be interpreted as a clinical risk measure. Linguistically similar qualifiers can accompany low-stakes atelectasis or possible malignancy, and a recommendation for CT can be routine, defensive, or driven by a consequential unresolved finding. A Score-3 report is not necessarily more dangerous than a Score-1 report, and the present rubric must not be used for autonomous risk triage, urgency assignment, or clinical decision support. The single Gemini Score-3-to-Score-1 discrepancy illustrates why an ordinal linguistic label cannot substitute for clinical assessment.

A concrete research-stage use case is corpus curation under radiologist supervision. An institution could use ordinal scores to construct evaluation sets that deliberately include definitive, mildly hedged, differential-rich, and highly ambiguous reports, then manually verify those strata. Utility would be demonstrated not by label generation alone but by a prespecified downstream experiment: for example, comparing extraction or imaging-model performance, calibration, and error rates when training/evaluation data are sampled with vs. without uncertainty stratification. Similar labels could support audits of reporting-style variation or targeted review of inconsistent templates. Table 6 sets out these candidate applications, the failure mode each inherits from the results

Table 6. Intended use, current evidence, and prerequisites for deployment

Candidate application	Who acts on the output	What the score would add	Failure mode inherited from this study	What must be demonstrated first
Building uncertainty-stratified evaluation or training corpora for report-based artificial intelligence	Researcher, under radiologist supervision	Replaces manual reading of thousands of reports to assemble sets that deliberately span definitive, mildly hedged, differential-rich, and highly ambiguous language	Exact Score-3 recall differed (Gemini 1/8, Claude 5/8), with six of Gemini's seven misses assigned Score 2; category-boundary assignments may affect stratum composition. Claude assigned scores above the locked reference in 25 reports, including a contested rubric-boundary case	A prespecified downstream comparison of extraction or imaging-model performance, calibration, and error rate for sets built with vs. without uncertainty stratification
Departmental audit of reporting language	Quality or education lead	A quantitative summary of hedging patterns across readers, subspecialties, or report templates	The score reflects wording, not diagnostic difficulty, so differences in case mix can be misread as differences in reporting style	Case-mix adjustment, agreement with a radiologist panel on a departmental sample, and acceptability testing before any reader-level feedback
Targeted review of inconsistent report templates	Reporting committee	Flags templates whose hedging is systematically higher or lower than comparable templates	Routine defensive language (e.g., a recommendation for dedicated views) can be overgraded, inflating the number of flagged templates	Precision at the chosen flagging threshold, measured against radiologist review of the flagged templates
Report triage, urgency assignment, or clinical decision support	Not recommended	Nothing; the score is linguistic	Linguistically similar qualifiers accompany low-stakes and high-stakes findings alike, so a Score-3 report is not necessarily more dangerous than a Score-1 report	Not attempted. This use would require an outcome-anchored reference standard rather than a linguistic one and is outside the scope of the present rubric

The present study does not validate any application. Each row identifies evidence needed before use; report triage, urgency assignment, and clinical decision support are not recommended under the present linguistic rubric.[12]

above, and what would need to be demonstrated before use. Referring-physician summaries and real-time report triage are more demanding applications and were not evaluated.

Practical feasibility includes computational and economic considerations, but the available billing records differed in workload definition, attribution, model snapshot, and request mode and therefore did not support a normalized cross-model inference-cost comparison. Future economic evaluations should record model identifiers, access dates, token counts, billing modes, retries, and failures for clearly defined workloads. Detailed billing provenance is provided in Supplementary Table 21.

Generalizability beyond chest radiography is uncertain. Chest radiograph reports are usually short and centered on a limited set of thoracic findings. Magnetic resonance imaging and CT reports may contain multiple organs, incidental findings, staged differentials, and several recommendations with different stakes. Their longer context and mixed-certainty statements may increase both reader disagreement and model aggregation errors. Multimodality and multilingual validation should therefore precede broader claims.

Limitations

• The validation cohort was small and included only eight Score-3 reports; category-specific estimates are therefore imprecise, and no clinically enriched edge-case cohort was available.

• Because validation was deliberately balanced across four lexicon-derived sampling strata, aggregate agreement estimates are conditional on this enriched case-mix and should not be interpreted as expected performance in an unselected clinical population with natural uncertainty prevalence.

QWK is sensitive to marginal score distributions, and no prevalence reweighting was attempted because the discovery cohort lacks an independent human reference standard.

• The consensus was derived from three readers with variable experience; complete disagreements were adjudicated by the attending, and one resident also operated the API workflow. Formal rubric calibration and independent prompt operation were not performed.

• One of five preliminary prompt-development examples (report ID 1000) overlapped with the validation cohort, so complete pre-annotation separation cannot be claimed for that report. Excluding it in a secondary (n = 99) analysis did not materially alter aggregate performance.

• Open-I may have appeared in pretraining corpora. Gemini validation and later discovery/robustness analyses used different dated model identifiers, and directly comparable reasoning-budget metadata were unavailable across runs. In addition, six Claude validation predictions differed between the locked validation file and a later full-corpus source despite the same recorded API identifier. The archived metadata could not explain the cause, leaving provenance variability unexplained; source and configuration differences limit longitudinal attribution.

• The discovery cohort had no human reference standard; distributions and inter-model agreement describe model behavior, not true uncertainty prevalence.

• The phrase audit tested source-text presence and selected semantic errors but did not exhaustively evaluate every rationale or contextual attribution.

• Costs were observed for specific APIs and dates and are not durable price estimates. No prospective workflow, patient outcome, referring-physician comprehension, or downstream artificial intelligence benefit was measured.

Conclusion

Zero-shot LLMs can be incorporated into a reproducible methodology for ordinal grading of linguistic uncertainty in chest radiograph reports, with aggregate QWK near 0.81 against a multi-reader consensus. However, similar aggregate agreement concealed a difference in exact Score-3 recall, largely involving the adjacent Score-2/3 boundary, and the human rubric itself showed meaningful variability. The primary contribution is therefore the audited workflow: explicit rubric limitations, reader-reliability reporting, paired comparisons, per-category error analysis, prompt and model provenance, and full-cohort computational accounting.

The output is a linguistic annotation, not a clinical-risk score. Before any deployment, the framework requires calibrated multi-reader external validation on model-unseen, modality-diverse data, enrichment of rare high-uncertainty cases, and demonstration of measurable downstream utility under radiologist supervision.

Acknowledgements

A large language model assisted with language editing and document structure. The authors verified all numerical results, citations, methodological statements, and clinical interpretations, and take full responsibility for the content of the manuscript.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

No external funding was received. API costs were covered by the corresponding author.

Data and code availability

Analysis scripts, prompts, audit outputs, and derived validation data are provided in the accompanying reproducibility package. Raw Open-I report text is not redistributed.

Supplementary Materials: https://d2v96fxpocvxx.cloudfront.net/cf9d60d6-523c-458a-a2e6-78728d3ffbb0/content-images/27f37e53-efbe-472b-8e68-213bcbe57bdf.pdf

Supplementary Tables: https://d2v96fxpocvxx.cloudfront.net/cf9d60d6-523c-458a-a2e6-78728d3ffbb0/content-images/a47c9f0b-60d9-4b50-aae7-368cb6a1c559.pdf

Supplementary Figures: https://d2v96fxpocvxx.cloudfront.net/87200e80-fa4b-4309-8816-eff4c4e638c8/content-images/30d280c4-dc09-4655-9c55-596b0e4eaed1.pdf

References

1. Callen AL, Dupont SM, Price A, et al. Between always and never: evaluating uncertainty in radiology reports using natural language processing. *J Digit Imaging.* 2020;33(5):1194-1201. [Crossref]
2. Hanauer DA, Liu Y, Mei Q, Manion FJ, Balis UJ, Zheng K. Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. *AMIA Annu Symp Proc.* 2012;2012:321-330. [Crossref]
3. Liu F, Zhou P, Baccei SJ, et al. Qualifying certainty in radiology reports through deep learning-based natural language processing. *AJNR Am J Neuroradiol.* 2021;42(10):1755-1761. [Crossref]
4. Reiner BI. Using analysis of speech and linguistics to characterize uncertainty in radiology reporting. *J Digit Imaging.* 2012;25(6):703-707. [Crossref]
5. Xu SQ, Yang HR, Roa T, Ahmad ZY, Zech JR, Wong TT. The impact of experience on radiology report uncertainty for early-career musculoskeletal radiologists. *Skeletal Radiol.* 2026;55(9):2363-2368. [Crossref]
6. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform.* 2001;34(5):301-310. [Crossref]
7. Peng Y, Wang X, Lu L, Bagheri M, Summers RM, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Jt Summits Transl Sci Proc.* 2018;2017:188-196. [Crossref]
8. Sugimoto K, Wada S, Konishi S, et al. Classification of diagnostic certainty in radiology reports with deep learning. *Stud Health Technol Inform.* 2024;310:569-573. [Crossref]
9. Banerjee I, Madhavan S, Goldman RE, Rubin DL. Intelligent word embeddings of free-text radiology reports. *AMIA Annu Symp Proc.* 2018;2017:411-420. [Crossref]
10. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med.* 2022;5(1):194. [Crossref]
11. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. *NPJ Digit Med.* 2023;6(1):210. [Crossref]
12. Bhayana R, Nanda B, Dehkharghanian T, et al. Large language models for automated synoptic reports and resectability categorization in pancreatic cancer. *Radiology.* 2024;311(3):e233117. [Crossref]
13. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol.* 2008;61(Pt 1):29-48. [Crossref]
14. 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]
15. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc.* 2016;23(2):304-310. [Crossref]
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174. [Crossref]