fortune Journals

Research Article

JOURNAL OF ORTHOPAEDICS AND SPORTS MEDICINE

ISSN: 2688-5115

Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro

Derick Rodríguez-Reyes, MD; Joseph Salem-Hernández, MD*; Norman Ramírez, MD, FAAOS-FAAP; Rafael Fernández-Soltero, MD, FAAOS; José Bossolo-Flores, MD, FAAOS

Abstract

Purpose: To compare the diagnostic accuracy, confidence calibration, and test-retest reliability of three general-purpose multimodal large language models on binary wrist radiograph classification.

Methods: All 237 wrist studies from the MURA v1.1 validation set (97 abnormal, 140 normal) were submitted twice to each model via REST API in May 2026 using an identical zero-shot prompt. Accuracy, sensitivity, specificity, predictive values, and F1 score were calculated with 95% confidence intervals and compared by McNemar's test with Holm correction. Calibration was assessed with expected calibration error (ECE), Brier score, and calibration intercept and slope; test-retest reliability with Cohen's kappa and the intraclass correlation coefficient (ICC).

Results: Claude Opus 4.6 classified 231 of 237 studies (97.5%) as abnormal, yielding 100% sensitivity (95% CI, 96.2-100) with 4.3% specificity (2.0-9.0) and the lowest accuracy (43.5%; 37.3-49.8). GPT-4o and Gemini 2.5 Pro achieved similar accuracy (59.7% and 59.9%) and did not differ significantly on any metric after correction. All three were severely miscalibrated (ECE 0.291-0.454; all intercepts excluded zero), and none returned a probability estimate between 0.30 and 0.70 for any study. Test-retest reliability ranged from almost perfect (Claude, kappa 0.853) to moderate (GPT-4o, 0.486), with confidence reproducibility following the same ordering (ICC 0.884 and 0.529).

Conclusions: Under zero-shot prompting on this benchmark, the three models showed divergent diagnostic biases, uniformly poor calibration, and variable output stability. Stated confidence did not correspond to observed accuracy and should not be used to weight model outputs. These findings apply to a single anatomic region, dataset, and prompting strategy.

Level of Evidence: Diagnostic Level III.

Keywords: Artificial intelligence; Large language models; Multimodal models; Wrist radiographs; Fracture detection; Diagnostic accuracy; Confidence calibration; Test-retest reliability

Abbreviations: AI: Artificial Intelligence; API: Application Programming Interface; CI: Confidence Interval; ECE: Expected Calibration Error; FN: False Negative; FP: False Positive; ICC: Intraclass Correlation Coefficient; LLM: Large Language Model; MURA: Musculoskeletal Radiographs; NPV:

Affiliation:
Department of Orthopaedic Surgery, Ponce Health Sciences University, 2 Rius Rivera Avenue, Ponce, Puerto Rico 00716, USA

***Corresponding Author:**
Joseph Salem-Hernández, MD, Orthopaedic Surgery Research Fellow, Department of Orthopaedic Surgery, Ponce Health Sciences University, Ponce, Puerto Rico, USA.

Citation: Derick Rodríguez-Reyes, Joseph Salem-Hernández, Norman Ramírez, Rafael Fernández-Soltero, José Bossolo-Flores. Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro. Journal of Orthopedics and Sports Medicine. 8 (2026): 264-272.

Received: August 10, 2026
Accepted: August 17, 2026
Published: August 21, 2026

Rodríguez-Reyes D, et al., J Ortho Sports Med 2026
DOI:10.26502/josm.511500281

Negative Predictive Value; OR: Odds Ratio; PPV: Positive Predictive Value; STARD: Standards for Reporting of Diagnostic Accuracy; TN: True Negative; TP: True Positive.

Introduction

Large language models (LLMs) have demonstrated broad capability in synthesizing clinical information and supporting medical decision-making. In orthopedic and hand surgery, these tools have been evaluated on board certification examinations [1-3] and, more recently, on image-based diagnostic problems [4-6]. Reported performance varies widely: a systematic review of artificial intelligence for hand and wrist fracture detection found sensitivity and specificity estimates spanning more than 40 percentage points across studies [7], while purpose-built imaging models perform substantially more consistently [8]. Recent work has extended multimodal LLM evaluation to skeletal maturity staging from hand-wrist radiographs [9], distal radius fracture detection [10], and fracture classification from radiology text reports [11], and parallel evaluations in dentistry and orthodontics report comparable model-dependent variability on craniofacial imaging and clinical decision tasks [12-14]. The sources of this variability are unresolved; dataset composition, fracture complexity, prompting strategy, and model version have all been proposed, and the present study was not designed to distinguish among them.

Purpose-built radiology platforms operate through narrow, regulatory-cleared algorithms prospectively validated within institutional infrastructure. Claude (Anthropic), ChatGPT (OpenAI), and Gemini (Google) are publicly accessible consumer products without disease-specific imaging training, regulatory clearance, or institutional gatekeeping. Because they are accessible without clinical oversight, the properties determining whether their outputs can be safely interpreted differ from those relevant to institutionally deployed systems. Two such properties were examined in this study.

The first is confidence calibration, the correspondence between a model's stated probability and its observed rate of being correct. Guo et al. [15] established expected calibration error (ECE) as a standardized metric for this discordance and showed that modern neural networks are systematically overconfident. A model expressing 90% confidence while correct 40% of the time is misleading rather than merely imprecise. Contemporary guidance recommends assessing calibration with multiple complementary measures rather than a single summary statistic [16]. The second is test-retest reliability. Guler et al. [17] showed that the same model can return different verdicts on identical radiographs across independent queries. Because these systems are accessed at non-deterministic default settings, some variability is expected, but its magnitude across current models has not been quantified, nor is it known whether stated confidence is more stable than the verdict it accompanies.

No published study has characterized all three properties simultaneously across multiple general-purpose multimodal LLMs on a single benchmark. We did so using the MURA wrist validation dataset [18]. Three models were selected to represent the highest-capability publicly accessible multimodal tier offered by each of the three providers whose consumer products hold the largest user bases at the time of querying; no claim of exhaustiveness is made. We tested three prespecified hypotheses: that diagnostic accuracy would differ significantly between models; that ECE would exceed the 0.05 threshold conventionally taken to indicate adequate calibration for all three models; and that test-retest agreement would fall below the threshold for almost perfect agreement (kappa 0.81) for at least one model.

Materials and Methods

Dataset

Institutional review board approval was not required for this secondary analysis of a publicly available, fully de-identified dataset. No human participants were enrolled by the investigators; the study population consists of the archived, previously collected imaging studies described below, and no new enrollment period applies. We used the MURA (Musculoskeletal Radiographs) v1.1 wrist validation set from the Stanford Artificial Intelligence in Medicine and Imaging group [18], a held-out partition of 659 images across 237 studies from 207 patients, with radiologist-assigned binary labels (97 abnormal, 140 normal) derived from study folder names per standard MURA convention.

MURA was selected as the largest publicly available musculoskeletal radiograph benchmark with an established held-out partition, because its binary labels match the triage-level question a non-specialist user would pose, and because public availability permits exact replication. Its limitations for this purpose, namely the absence of fracture subtype, anatomic localization, and demographics, and the use of study-level rather than image-level labels, are addressed in the Limitations. Images per study ranged from one to five (1 image, n = 24; 2, n = 44; 3, n = 133; 4, n = 32; 5, n = 4).

Model querying

Claude Opus 4.6 (Anthropic, San Francisco, CA, USA), GPT-4o (OpenAI, San Francisco, CA, USA), and Gemini 2.5 Pro (Google DeepMind, Mountain View, CA, USA) were queried via their REST application programming interfaces (APIs) in May 2026. Each study was submitted with up to three images; where more were available, the first three in alphabetical filename order were used. This cap bounded cost and latency across 1,422 total queries and affected 36 of 237 studies (15.2%); a subgroup analysis restricted to the 201 unaffected studies is reported in the Results.

Citation: Derick Rodríguez-Reyes, Joseph Salem-Hernández, Norman Ramírez, Rafael Fernández-Soltero, José Bossolo-Flores. Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro. Journal of Orthopedics and Sports Medicine. 8 (2026): 264-272.

Rodríguez-Reyes D, et al., J Ortho Sports Med 2026
DOI:10.26502/josm.511500281

All three models received an identical zero-shot prompt specifying a binary normal or abnormal classification task and requesting a structured JSON response with a judgment field, an integer confidence score from 0 to 100, and a one-sentence reasoning field; the verbatim prompt and query script are provided in Supplementary Material 1. No clinical context, patient history, or demographic information was provided, and queries were stateless with no conversation history carried between studies.

Models were accessed using the alias identifiers claude-opus-4-6, gpt-4o, and gemini-2.5-pro. Provider default sampling parameters were used throughout; temperature, top-p, and random seed were not specified and no deterministic decoding mode was enabled. This was deliberate, as it corresponds to the conditions under which these products are accessed by non-specialist users. Maximum output was capped at 256 tokens for Claude Opus 4.6 and GPT-4o but not for Gemini 2.5 Pro. No retry logic was implemented: each study was submitted once per model per run, and queries returning no parseable output were recorded as missing and excluded from that run. Responses were parsed as JSON after removal of markdown code fences; any response that failed parsing or returned an empty body was classified as non-parseable. Response latencies were not logged.

Each study was queried in two independent sessions, Run 1 and Run 2, executed as separate invocations of the script with no shared state. Each run required approximately 40 to 50 minutes; the interval between runs was not recorded. Reporting follows the Standards for Reporting of Diagnostic Accuracy (STARD) 2015 guidelines [19]. A completed checklist with section-level cross-references is provided as Supplementary Material 2.

Statistical analysis

Study-level accuracy, sensitivity, specificity, predictive values, and F1 score were calculated against MURA labels. Ninety-five percent confidence intervals for all proportions were calculated by the Wilson score method; intervals for F1 by bootstrap resampling with 5,000 iterations.

Because all models were evaluated on the same studies, between-model comparisons are paired. Accuracy, sensitivity, and specificity were compared pairwise using McNemar's test with continuity correction, restricted for each comparison to studies with valid responses from both models, with sensitivity comparisons restricted to abnormal studies and specificity comparisons to normal studies. The resulting family of nine tests was corrected using the Holm procedure [20], with a two-sided threshold of $P < .05$.

For calibration, stated confidence was transformed to P(abnormal): confidence / 100 for abnormal judgments and (100 − confidence) / 100 for normal judgments. Studies were binned into 10 equal-width intervals spanning 0.0 to 1.0, following Guo et al. [15], and ECE was calculated as the weighted absolute difference between mean predicted P(abnormal) and observed fraction abnormal within each populated bin, with observations per bin reported. Calibration was additionally quantified by the Brier score [21] and by the calibration intercept and slope from logistic regression of the observed outcome on the logit of predicted probability, for which ideal values are 0 and 1 respectively [16]. Intervals for ECE and Brier score, and pairwise between-model ECE differences, were obtained by paired bootstrap resampling stratified by ground-truth label with 3,000 iterations, Holm-corrected across the three pairwise tests.

Test-retest reliability of verdicts used Cohen's kappa between runs, interpreted per Landis and Koch [22], with bootstrap confidence intervals and a permutation test against kappa = 0. Reproducibility of the continuous confidence scores used a two-way random-effects, absolute-agreement, single-measures intraclass correlation coefficient, ICC(2,1) [23], applied to both raw confidence and transformed P(abnormal), with the mean absolute between-run difference in P(abnormal).

Missingness was evaluated by Fisher exact tests against ground-truth label and image count. Sensitivity analyses assigned all missing responses to best-case and worst-case values and restricted the sample to studies from which no images had been discarded; Run 2 performance is reported as an independent replication.

Because the study analyzed a complete, fixed public benchmark rather than a sample drawn to a target size, no prospective sample size calculation was performed; achieved precision is reported instead. Analyses used Python 3.12 (Python Software Foundation, Wilmington, DE, USA) with the scipy, scikit-learn, and statsmodels libraries. Analysis code is provided as Supplementary Material 3.

Results

Completion and missing data

Valid, parseable responses were obtained for all 237 studies in both runs for Claude Opus 4.6 and Gemini 2.5 Pro. GPT-4o returned valid responses for 231 of 237 studies in Run 1 (97.5%) and 234 of 237 in Run 2 (98.7%); all nine failures were empty API responses. The failing studies differed between runs, so 228 studies had valid GPT-4o responses in both runs and formed the paired sample for its test-retest analysis.

Missingness was not associated with ground-truth label (odds ratio [OR] 1.39; Fisher exact P = .74) or with submission of three or more images (OR 0.80; P = .72); two of the six Run 1 failures involved single-image studies. These patterns are consistent with data missing completely at random. In sensitivity analysis, assigning all six Run 1

Citation: Derick Rodríguez-Reyes, Joseph Salem-Hernández, Norman Ramírez, Rafael Fernández-Soltero, José Bossolo-Flores. Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro. Journal of Orthopedics and Sports Medicine. 8 (2026): 264-272.

Rodríguez-Reyes D, et al., J Ortho Sports Med 2026
DOI:10.26502/josm.511500281

failures to correct classification yielded GPT-4o accuracy of 60.8% and assigning all to incorrect classification yielded 58.2%, against 59.7% observed. No conclusion changed under either extreme.

Diagnostic accuracy

The three models demonstrated markedly divergent diagnostic profiles (Table 1, Figure 1). Claude Opus 4.6 classified 231 of 237 studies (97.5%) as abnormal, yielding perfect sensitivity (100%; 95% CI, 96.2-100) and near-zero specificity (4.3%; 2.0-9.0). It missed no abnormal study but misclassified 134 of 140 normal studies, whereas GPT-4o and Gemini 2.5 Pro classified 60.6% and 63.3% of studies as abnormal.

Claude Opus 4.6 was significantly less accurate than both GPT-4o (43.7% vs 59.7% on the 231 shared studies; Holm-corrected $P = 5.4 \times 10^{-4}$) and Gemini 2.5 Pro (43.5% vs 59.9%; corrected $P = 2.3 \times 10^{-4}$), and its sensitivity advantage and specificity deficit were significant against each comparator after correction (all corrected $P < 10^{-5}$). GPT-4o and Gemini 2.5 Pro did not differ significantly on any metric (all corrected $P = 1.00$).

Claude's negative predictive value was 100%, but its 95% confidence interval extended from 61.0% to 100%, reflecting that only six studies received a normal verdict. All six were correctly classified

Confidence calibration

All three models were severely miscalibrated (Table 2). ECE ranged from 0.291 for GPT-4o to 0.454 for Claude Opus 4.6, all far exceeding the 0.05 threshold conventionally considered adequate [15], and Brier scores followed the same ordering.

Calibration intercepts excluded zero for all three models, indicating systematic overestimation of the probability of abnormality. Slopes were 0.32 for GPT-4o and 0.11 for Gemini 2.5 Pro, far below the ideal of 1, so variation in stated confidence corresponded to little variation in observed outcome. Claude's slope was 1.49, but derives from a distribution in which 94.1% of predictions fell in the two highest bins.

Pairwise ECE differences were significant after Holm correction: Claude Opus 4.6 was less well calibrated than GPT-4o (ΔECE +0.160; 95% CI, 0.100-0.219; corrected P = .001) and than Gemini 2.5 Pro (+0.078; 0.015-0.139; corrected P = .036), and GPT-4o was marginally better calibrated than Gemini 2.5 Pro (−0.073; −0.146 to −0.001; corrected P = .049).

No model returned a P(abnormal) between 0.30 and 0.70 for any study in the dataset. Gemini 2.5 Pro populated only two of ten bins, placing 150 studies (63.3%) in the 0.9-1.0 bin

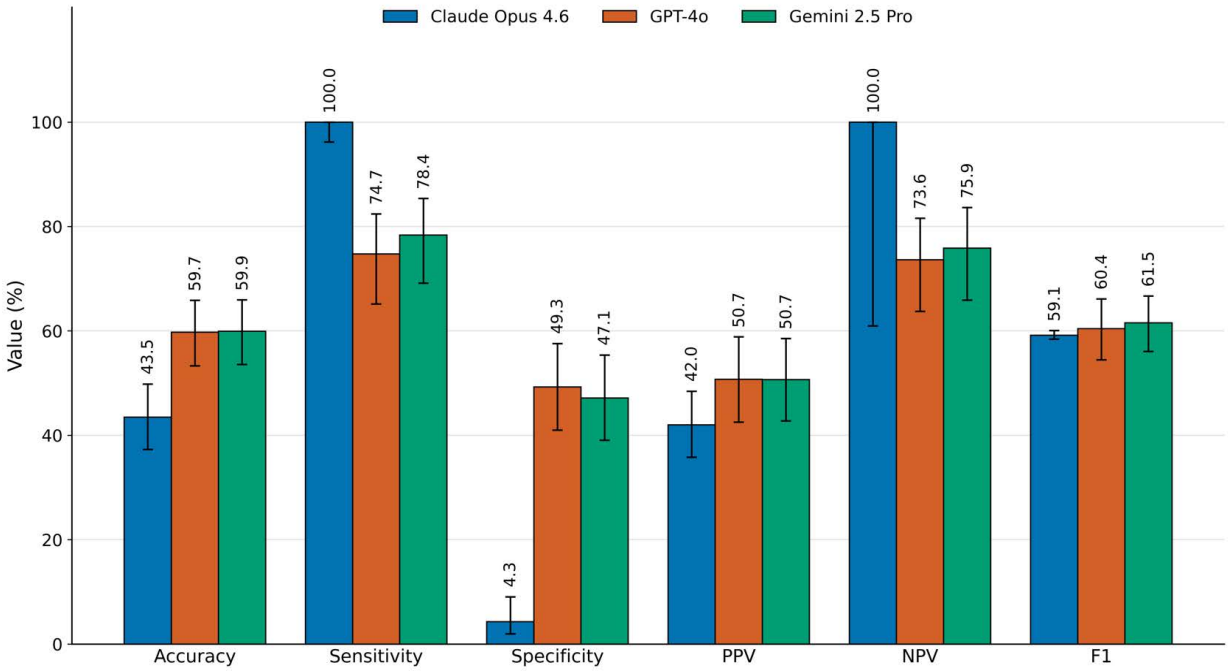

Figure 1: Diagnostic performance by model. Accuracy, sensitivity, specificity, positive and negative predictive value, and F1 score across all three models on the MURA wrist validation set (n = 237 for Claude Opus 4.6 and Gemini 2.5 Pro; n = 231 for GPT-4o), with 95% confidence intervals. Claude Opus 4.6 achieved perfect sensitivity with 4.3% specificity, reflecting near-universal abnormal classification. GPT-4o and Gemini 2.5 Pro demonstrated more balanced profiles with statistically indistinguishable accuracy.

Citation: Derick Rodríguez-Reyes, Joseph Salem-Hernández, Norman Ramírez, Rafael Fernández-Soltero, José Bossolo-Flores. Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro. Journal of Orthopedics and Sports Medicine. 8 (2026): 264-272.

Rodríguez-Reyes D, et al., J Ortho Sports Med 2026
DOI:10.26502/josm.511500281

Table 1: Diagnostic performance metrics by model (Run 1).

Metric	Claude Opus 4.6 (n = 237)	GPT-4o (n = 231)	Gemini 2.5 Pro (n = 237)
Accuracy (95% CI)	43.5% (37.3 to 49.8)	59.7% (53.3 to 65.9)	59.9% (53.6 to 65.9)
Sensitivity (95% CI)	100.0% (96.2 to 100)	74.7% (65.2 to 82.4)	78.4% (69.2 to 85.4)
Specificity (95% CI)	4.3% (2.0 to 9.0)	49.3% (41.0 to 57.6)	47.1% (39.1 to 55.4)
PPV (95% CI)	42.0% (35.8 to 48.4)	50.7% (42.5 to 58.9)	50.7% (42.7 to 58.6)
NPV (95% CI)	100.0% (61.0 to 100)	73.6% (63.7 to 81.6)	75.9% (65.9 to 83.6)
F1 score (95% CI)	0.591 (0.584 to 0.601)	0.604 (0.545 to 0.661)	0.615 (0.561 to 0.667)
TP / TN / FP / FN	97 / 6 / 134 / 0	71 / 67 / 69 / 24	76 / 66 / 74 / 21

CI: Confidence Interval; FN: False Negative; FP: False Positive; NPV: Negative Predictive Value; PPV: Positive Predictive Value; TN: True Negative; TP: True Positive. Intervals for proportions were calculated by the Wilson score method; intervals for F1 by bootstrap resampling with 5,000 iterations.

Table 2: Confidence calibration by model (Run 1).

Measure	Claude Opus 4.6	GPT-4o	Gemini 2.5 Pro
ECE (95% CI)	0.454 (0.439 to 0.466)	0.291 (0.232 to 0.352)	0.375 (0.316 to 0.433)
Brier score (95% CI)	0.426 (0.410 to 0.441)	0.314 (0.264 to 0.359)	0.381 (0.325 to 0.439)
Calibration intercept (95% CI)	−3.53 (−4.57 to −2.48)	−0.58 (−0.88 to −0.28)	−1.08 (−1.47 to −0.68)
Calibration slope (95% CI)	1.49 (1.01 to 1.96)	0.32 (0.19 to 0.46)	0.11 (0.07 to 0.15)

CI: Confidence Interval; ECE: Expected Calibration Error. Values of ECE below 0.05 are considered indicative of adequate calibration. Ideal values are 0 for the calibration intercept and 1 for the calibration slope. All three models showed severe miscalibration, and all three intercepts excluded zero. Intervals were obtained by paired bootstrap resampling stratified by ground-truth label with 3,000 iterations.

Table 3: Test-retest reliability between Run 1 and Run 2.

Measure	Claude Opus 4.6	GPT-4o	Gemini 2.5 Pro
Paired studies, n	237	228	237
Cohen's kappa (95% CI)	0.853 (0.537 to 1.000)	0.486 (0.369 to 0.598)	0.673 (0.576 to 0.767)
Raw agreement (95% CI)	99.2% (97.0 to 99.8)	75.4% (69.5 to 80.6)	84.8% (79.7 to 88.8)
ICC(2,1), P(abnormal) (95% CI)	0.884 (0.712 to 0.991)	0.529 (0.419 to 0.632)	0.686 (0.587 to 0.783)
ICC(2,1), raw confidence	0.952	0.505	0.721
Mean absolute change in P(abnormal)	0.012	0.202	0.144

CI: Confidence Interval; ICC: Intraclass Correlation Coefficient. Kappa is interpreted per Landis and Koch. Bootstrap confidence intervals used 3,000 iterations; all kappa values significantly exceeded zero (permutation P < .001). ICC(2,1) is a two-way random-effects, absolute-agreement, single-measures coefficient.

and 87 (36.7%) in the 0.0-0.1 bin. Claude Opus 4.6 placed 223 of 237 studies (94.1%) in the two highest bins. Within Gemini's 0.9-1.0 bin, mean predicted probability was 0.988 while the observed fraction abnormal was 0.507.

Test-retest reliability

Test-retest reliability varied substantially across models (Table 3, Figure 2). Claude Opus 4.6 demonstrated almost perfect verdict agreement (kappa 0.853; 99.2% agreement across 237 paired studies), though its interval is wide because only two studies were discordant. Gemini 2.5 Pro showed substantial agreement (kappa 0.673; 84.8%) and GPT-4o only moderate agreement (kappa 0.486; 75.4% across 228 paired studies). All kappa values significantly exceeded zero (permutation P < .001).

Confidence-score reproducibility followed the same ordering: ICC(2,1) for P(abnormal) was 0.884 for Claude Opus 4.6, 0.686 for Gemini 2.5 Pro, and 0.529 for GPT-4o, with mean absolute between-run change of 0.012, 0.144, and 0.202 respectively. For GPT-4o, therefore, neither the verdict nor the accompanying confidence value was reproducible on repeat submission of identical images.

Replication and subgroup analyses

Run 2 closely replicated Run 1 for Claude Opus 4.6 (accuracy 44.3%, sensitivity 100%, specificity 5.7%) and GPT-4o (59.4%, 74.0%, 49.3%). Gemini 2.5 Pro shifted more, to 64.1%, 83.5%, and 50.7%, a change of 4.2 percentage points in accuracy and 5.1 in sensitivity.

Citation: Derick Rodríguez-Reyes, Joseph Salem-Hernández, Norman Ramírez, Rafael Fernández-Soltero, José Bossolo-Flores. Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro. Journal of Orthopedics and Sports Medicine. 8 (2026): 264-272.

Rodríguez-Reyes D, et al., J Ortho Sports Med 2026
DOI:10.26502/josm.511500281

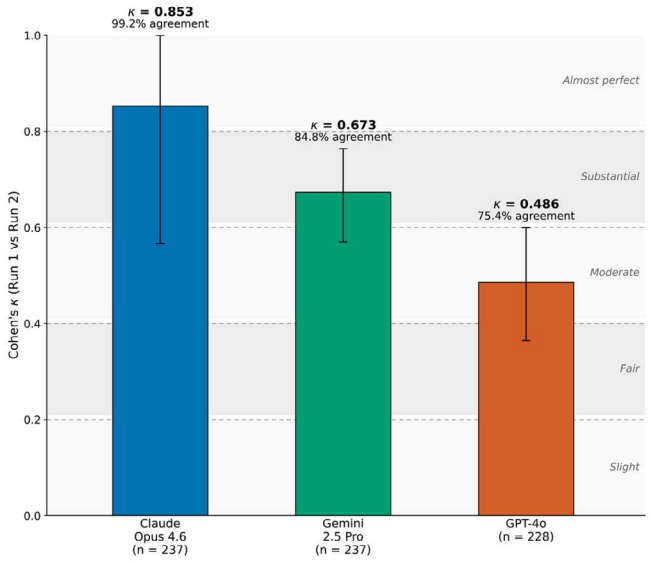

Figure 2: Test-retest reliability. Cohen's kappa for agreement between Run 1 and Run 2 with 95% bootstrap confidence intervals (n = 237 paired studies; 228 for GPT-4o). Horizontal dashed lines indicate standard kappa interpretation thresholds per Landis and Koch [22]. Claude Opus 4.6 achieved almost perfect agreement, Gemini 2.5 Pro substantial agreement, and GPT-4o only moderate agreement, indicating that approximately one study in four received a discordant verdict on resubmission of identical images.

In the 201 studies for which no images were discarded by the three-image cap, accuracy was 40.8% for Claude Opus 4.6, 59.0% for GPT-4o, and 59.7% for Gemini 2.5 Pro, compared with 43.5%, 59.7%, and 59.9% in the full sample. Accuracy in the 36 truncated studies did not differ significantly for GPT-4o (P = .71) or Gemini 2.5 Pro (P = 1.00). The difference for Claude Opus 4.6 approached significance (58.3% vs 40.8%; P = .067) but is attributable to case mix rather than truncation: abnormal prevalence was 55.6% among studies with more than three images versus 38.3% among the remainder, and a model classifying almost all studies as abnormal necessarily achieves higher accuracy where prevalence is higher.

Achieved precision

With 237 studies (97 abnormal, 140 normal), the half-width of the 95% confidence interval was approximately 6.2 percentage points for accuracy near 60%, 8.6 points for sensitivity near 75%, and 8.3 points for specificity near 50%. The study had 95% power to detect the observed 16-point accuracy difference between Claude Opus 4.6 and the other models, and roughly 80% power for differences of 10 to 12 points; smaller differences could not be reliably excluded.

Discussion

Wrist radiographs are among the most commonly obtained studies in hand surgery, and the question we posed to these three models, whether a film is normal or abnormal, is the question a patient or a non-specialist clinician is most likely to ask of them. None of the three answered it in a way that would be usable in practice, and the manner of their failure differed enough between models to matter clinically.

Claude Opus 4.6 called 231 of 237 studies abnormal. It therefore missed no fracture, which is superficially the error a screening tool ought to make, and its negative predictive value of 100% might be read as reassuring. Neither figure survives inspection. Perfect sensitivity obtained by calling nearly everything abnormal reflects a decision threshold rather than an ability to recognize a fracture, and the negative predictive value rests on six films, with a confidence interval extending down to 61%. With a specificity of 4.3%, an abnormal verdict from this model tells the clinician almost nothing about the film in front of them. The behavior was consistent between runs (kappa 0.853), so this is a stable property of the model rather than erratic output. Why its threshold sits where it does cannot be determined from these data, and would require access to model internals that commercial interfaces do not provide.

GPT-4o and Gemini 2.5 Pro produced the more familiar pattern of a non-experienced reader. Each identified roughly three-quarters of the fractures present while calling abnormal about half of the normal films, and their accuracy was indistinguishable from one another. In clinical terms, a reader who overlooks one fracture in four and refers half of uninjured wrists for further evaluation is of little use at either end of the pathway: not reliable enough to exclude injury, and not specific enough to make triage worthwhile. That the two models did not differ statistically should not be read as equivalence, since our sample could only have detected differences of roughly 10 percentage points.

The confidence values accompanying these readings deserve separate attention, because they are the part of the output a clinician or patient is most likely to act upon. A verdict qualified as 95% confident invites more trust than one offered tentatively. In our data these numbers did not behave like probabilities. Not one of the 237 films received a probability between 0.30 and 0.70 from any model; each system was either almost certain a fracture was present or almost certain it was absent. Gemini 2.5 Pro assigned a probability of 0.9 or higher to 63% of the films it read, with a mean of 0.99, and was correct for half of them. For GPT-4o and Gemini the relationship between stated confidence and being correct was nearly flat, which is the practically important point: a flat relationship cannot be corrected by rescaling, because the number contains no information to recover. Guo et al. [15] described systematic overconfidence as a general property of neural networks; what we observed is a more complete failure, in which confidence and correctness are close to unrelated.

Nor did repetition help. Resubmitting identical films to GPT-4o produced a different verdict for approximately one

Citation: Derick Rodríguez-Reyes, Joseph Salem-Hernández, Norman Ramírez, Rafael Fernández-Soltero, José Bossolo-Flores. Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro. Journal of Orthopedics and Sports Medicine. 8 (2026): 264-272.

Rodríguez-Reyes D, et al., J Ortho Sports Med 2026
DOI:10.26502/josm.511500281

study in four, and the confidence value accompanying it was equally unstable. A clinician who repeated a query hoping to confirm an uncertain reading would frequently receive the opposite answer, with nothing to indicate which to believe. This extends the observation of Guler et al. [17] that identical radiographs can elicit different verdicts on repeated submission. The instability reflects the default sampling settings under which these products are accessed rather than any deliberate configuration, and it is not unavoidable: Claude Opus 4.6 agreed with itself for 99.2% of studies under identical conditions.

These findings should be read against a hand surgery literature in which model performance on imaging has been notably inconsistent. Bulut [6] found 57% sensitivity for scaphoid fracture detection with one model and 9% with another applied to the same cases, and Mert et al. [10] reported that ChatGPT-4 performed well on distal radius radiographs yet remained below both a hand surgery resident and a dedicated fracture-detection algorithm. Purpose-built musculoskeletal imaging models continue to outperform general-purpose systems by a wide margin [7,8]. The apparent tension with reports that these models can pass the American Society for Surgery of the Hand self-assessment examination [1-3] is resolved by recognizing that those are different tasks: reasoning about a written vignette draws on knowledge well represented in text, whereas deciding whether a cortical step-off is a fracture line is a perceptual judgment. Nguyen et al. [24] observed the same divergence between text and imaging performance, the range across upper extremity imaging studies is correspondingly wide [4,5], and comparable model dependence has been reported in craniofacial and dental applications [9,12-14].

For the hand surgeon the practical implications are narrow but worth stating. These products are already available to patients and trainees without clinical gatekeeping, and our data provide no support for treating either a verdict or its accompanying confidence as informative about an individual radiograph. Because the three models failed in different directions, performance published for one product cannot be assumed to describe another, and a report that a model called a film normal carries no information that should modify clinical assessment. The more constructive reading is that these tools require evaluation beyond aggregate accuracy: a model's operating point, the calibration of its confidence, and the reproducibility of its output each determine whether its answers can be acted upon, and none of the three is visible in an accuracy figure.

Implications for future research

Three directions follow directly from these results. First, evaluation of these systems should report operating point, calibration, and test-retest reliability alongside accuracy, since our data show that the three vary independently and that a single accuracy figure conceals all of them. Second, whether contextualized prompting, including mechanism of injury, age, and acuity, shifts the classification threshold or improves calibration is testable and untested here. Third, because calibration failure was near-total rather than merely imperfect, post hoc recalibration on locally representative films should be evaluated before any deployment, together with studies of how patients and non-specialist clinicians actually interpret a model verdict and its stated confidence.

Limitations

The evaluation was restricted to a single anatomic region, a single publicly available dataset, and a single binary classification task. Performance on other anatomic regions, on datasets with different case mix or image quality, or on tasks requiring fracture localization or characterization cannot be inferred from these results.

Only a zero-shot prompting strategy was evaluated. Clinical queries frequently include mechanism of injury, patient age, or acuity of presentation, and prompt formulation is known to influence LLM output. Whether contextualized prompting improves accuracy, reduces miscalibration, or alters the classification threshold is untested here, and these findings should not be taken to characterize the models' capabilities under other prompting conditions.

Several aspects of the query configuration limit exact reproducibility. Models were accessed through alias rather than pinned snapshot identifiers, and the builds to which these resolved were not captured at query time and cannot be verified retrospectively. Provider default sampling parameters were retained rather than fixed deterministic settings, and the interval between runs was not logged. Response latencies were not recorded and no retry logic was implemented, so transient API failures were treated as missing rather than repeated. Model behavior may change with version updates, and these findings apply to the models as accessed in May 2026.

The three-image cap discarded images in 36 of 237 studies. Subgroup analysis found no evidence this affected performance, but the possibility that a discarded view contained the only visible abnormality cannot be excluded. An output token cap of 256 was applied to Claude Opus 4.6 and GPT-4o but not to Gemini 2.5 Pro, an asymmetry that may have contributed to GPT-4o's empty responses.

MURA provides study-level binary labels without fracture subtype, anatomic location, or demographics, precluding subgroup analysis by injury pattern, sex, or age; the sex and gender dimensions of performance could not be assessed. Labels derive from folder names assigned during original dataset construction, so any source labeling errors would propagate here. No prospective registration was performed.

Citation: Derick Rodríguez-Reyes, Joseph Salem-Hernández, Norman Ramírez, Rafael Fernández-Soltero, José Bossolo-Flores. Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro. Journal of Orthopedics and Sports Medicine. 8 (2026): 264-272.

Rodríguez-Reyes D, et al., J Ortho Sports Med 2026
DOI:10.26502/josm.511500281

Finally, this study measured model outputs, not clinical outcomes. It did not evaluate how patients or clinicians interpret these outputs, whether they influence care-seeking behavior, or what follows from acting on them. Any inference to clinical impact is therefore indirect.

Conclusions

On a single public benchmark of wrist radiographs, read with a simple zero-shot prompt at default settings, three widely used multimodal language models proved unreliable in ways that differed by model. Claude Opus 4.6 called almost every film abnormal, missing no fracture but flagging nearly every normal wrist. GPT-4o and Gemini 2.5 Pro erred in both directions, overlooking roughly one fracture in four while calling half of normal films abnormal. None produced confidence values that tracked whether it was right, and none expressed intermediate uncertainty about any film, so the numbers attached to their readings should not be used to weight them.

The clinical message is not that these tools have no future in hand surgery, but that accuracy alone is an inadequate basis on which to judge them. Operating point, calibration, and reproducibility varied independently between models here, and any one of the three can render an otherwise accurate reader unusable. We did not test contextualized prompting, deterministic settings, other anatomic regions, or workflows in which a clinician reviews the output, and these results should not be extended to those situations. Assessment of calibration on locally representative films would be a reasonable prerequisite before any of these systems is allowed to influence a clinical decision.

Acknowledgement

The authors thank the Stanford Machine Learning Group for making the MURA dataset publicly available for research use.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. All costs associated with the study, including application programming interface usage fees, were borne by the authors.

Data availability: The MURA v1.1 dataset is publicly available from the Stanford Machine Learning Group. The query script, model outputs, and analysis code are provided as Supplementary Material 1 and 3.

Conflicts of Interest: The authors declare that they have no conflicts of interest. No benefits in any form have been received or will be received related directly or indirectly to the subject of this article. The authors have no financial, personal, or professional relationship with Anthropic, OpenAI, or Google DeepMind.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

Three commercially available large language models (Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro) were used as the experimental systems under investigation. These models generated the diagnostic classifications, confidence scores, and explanatory outputs that constituted the study dataset through standardized API queries. The authors independently designed the study protocol, selected the dataset, developed the prompting strategy, performed the statistical analyses, interpreted the results, and wrote the manuscript. No generative AI tools were used to draft, edit, or revise the manuscript text beyond serving as the objects of investigation.

Authorship Contribution Statement:

Derick Rodríguez-Reyes, MD: Data curation, formal analysis, investigation, methodology, writing – original draft, writing – review & editing.

Joseph Salem-Hernández, MD: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, writing – original draft, writing – review & editing.

Norman Ramírez, MD, FAAOS-FAAP: Conceptualization, investigation, methodology, supervision, writing – review & editing.

Rafael Fernández-Soltero, MD: Data curation, investigation, writing – review & editing.

José Bossolo-Flores, MD: Data curation, investigation, writing – review & editing.

IRB / Ethics Statement: Institutional review board approval was not required for this secondary analysis of a publicly available, fully de-identified dataset (MURA v1.1). All procedures were conducted in accordance with the World Medical Association Declaration of Helsinki.

Reporting Guideline: This diagnostic accuracy study was reported in accordance with the Standards for Reporting of Diagnostic Accuracy (STARD) guidelines.

Ethics declaration: The author(s) declare(s) that the study does not involve humans nor animal subjects.

References

1. Rakauskas TR, Da Costa A, Moriconi C, et al. Evaluation of Chat Generative Pre-Trained Transformer and Microsoft Copilot Performance on the American Society of Surgery of the Hand Self-Assessment Examinations. Journal of Hand Surgery Global Online 7 (2025): 23-28.

2. Arango SD, Flynn JC, Zeitlin J, et al. The Performance

of ChatGPT on the American Society for Surgery of the Hand Self-Assessment Examination. Cureus 16 (2024): e58950.

3. Ghanem D, Nassar JE, El Bachour J, et al. ChatGPT Earns American Board Certification in Hand Surgery. Hand Surgery and Rehabilitation 43 (2024): 101688.
4. Hiredesai AN, Martinez CJ, Anderson ML, et al. Is Artificial Intelligence the Future of Radiology? Accuracy of ChatGPT in Radiologic Diagnosis of Upper Extremity Bony Pathology. Hand 21 (2026): 73-80.
5. Erginoglu SE, Ulgen NK, Yigit N, et al. Multimodal Large Language Model for Fracture Detection in Emergency Orthopedic Trauma: A Diagnostic Accuracy Study. Diagnostics 16 (2026): 476.
6. Bulut B. Diagnostic Capabilities of Large Language Models in the Detection of Scaphoid Fractures in the Emergency Department. Ulusal Travma ve Acil Cerrahi Dergisi 31 (2025): 987-994.
7. Alosaimi H, Al-Mutairi A, Alshehri F, et al. Evaluating Artificial Intelligence for Accurate Detection of Hand and Wrist Fractures: A Systematic Review and Meta-Analysis. F1000Research 14 (2025): 1062.
8. Wong CR, Zhu A, Baltzer HL. The Accuracy of Artificial Intelligence Models in Hand/Wrist Fracture and Dislocation Diagnosis. JBJS Reviews 12 (2024).
9. Yıldırım A, Cicek O, Genç YS. Can AI-Based ChatGPT Models Accurately Analyze Hand-Wrist Radiographs? A Comparative Study. Diagnostics 15 (2025): 1513.
10. Mert S, Stoerzer P, Brauer J, et al. Diagnostic Power of ChatGPT 4 in Distal Radius Fracture Detection Through Wrist Radiographs. Archives of Orthopaedic and Trauma Surgery 144 (2024): 2461-2467.
11. Mergen M, Ryan MK, Sonnow L, et al. Leveraging Large Language Models for Accurate AO Fracture Classification From CT Text Reports. Journal of Imaging Informatics in Medicine (2025).
12. Yıldırım A, Cicek O. Assessment of AI-Driven Large Language Models for Orthodontic Aesthetic Scoring Using the IOTN-AC. Diagnostics 15 (2025): 3048.
13. Erdem R, Yıldırım A, Genç YS, et al. Comparative Performance Analysis of AI-Based Large Language Models in Assessing Cervical Vertebral Maturation Stages on Lateral Cephalometric Radiographs. BMC Oral Health 26 (2026).
14. Cicek O, Yıldırım A. Evaluating the Reliability of Artificial Intelligence in Clinical Decisions on Early Maxillary Expansion: A Comparative Study of Eight Large Language Models. BMC Oral Health 26 (2026): 715.
15. Guo C, Pleiss G, Sun Y, et al. On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning 70 (2017): 1321-1330.
16. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: The Achilles Heel of Predictive Analytics. BMC Medicine 17 (2019): 230.
17. Guler I, Grieb G, Kraus A, et al. Diagnostic Accuracy and Stability of Multimodal Large Language Models for Hand Fracture Detection: A Multi-Run Evaluation on Plain Radiographs. Diagnostics 16 (2026): 424.
18. Rajpurkar P, Irvin J, Bagul A, et al. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. arXiv (2017): 1712.06957.
19. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. BMJ 351 (2015): h5527.
20. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 6 (1979): 65-70.
21. Brier GW. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review 78 (1950): 1-3.
22. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 33 (1977): 159-174.
23. Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin 86 (1979): 420-428.
24. Nguyen C, Carrion D, Badawy MK. Comparative Performance of Anthropic Claude and OpenAI GPT Models in Basic Radiological Imaging Tasks. Journal of Medical Imaging and Radiation Oncology (2025).

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license 4.0

Citation: Derick Rodríguez-Reyes, Joseph Salem-Hernández, Norman Ramírez, Rafael Fernández-Soltero, José Bossolo-Flores. Diagnostic Bias, Confidence Miscalibration, and Test-Retest Reliability of Three General-Purpose Multimodal Large Language Models for Wrist Radiograph Interpretation: A Comparison of Claude Opus 4.6, GPT-4o, and Gemini 2.5 Pro. Journal of Orthopedics and Sports Medicine. 8 (2026): 264-272.