Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Performance of the StrataScore AI-Powered Research Appraisal Engine in Multidisciplinary Peer Review: a cross-sectional agreement study

Ben D. Macfarlane

2619211m@student.gla.ac.uk

University of Glasgow https://orcid.org/0009-0008-4546-1691

Denver Correia

University of Glasgow

Fedora Lau

University of Glasgow

Sam J. Neilson

University of Glasgow

Daniel Taylor-Sweet

University of Glasgow

Gillian Horne

University of Glasgow

Tural Alekberli

Dalhousie University

Keith Anderson

Foothills Medical Centre

Craig.Bowhay@gjnh.scot.nhs.uk

Golden Jubilee National Hospital

Christopher Carlin

University of Glasgow

Kirsty Crowe

University of Glasgow

Cara A. Flynn

University of Glasgow

Richard I.S. Good

University of Glasgow

Daniel Gordon

Queen Elizabeth Hospital

Christopher Grant

University of Glasgow

Suzanne Harrogate
University Hospitals Bristol NHS Foundation Trust

David W. Hewson
University of Nottingham

Akhila Kalidindi
University of Glasgow

Angelina Kancheva
University of Geneva

Rachel Kearns
University of Glasgow

Katie Lewis
NHS Tayside

Alan J.R. Macfarlane
University of Glasgow

Peter W. Macfarlane
University of Glasgow

Kalliopi Mavromati
Trinity College Dublin

Alistair McLaren
University of Glasgow

Mirko Piazzini
University of Florence

Matthew Rutherford
University of Glasgow

Kyrillus S. Shohdy
University of Glasgow

Pavlina Spiliopoulou
University of Glasgow

Harrison Stubbs
University of Glasgow

Robert Sykes
University of Glasgow

Dylan Tan
University of Glasgow

Philippa K. Traill
University of Dundee

Lucie Tvrdá
University of Glasgow

Heather Walker
University of Dundee
Mark White
University of Glasgow
Matthew R. Wilson
Beatson West of Scotland Cancer Centre
Danny J.N. Wong
Singapore General Hospital
Jesse Dawson
University of Glasgow

Research Article

Keywords: artificial intelligence, peer review, large language models, critical appraisal

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11816153/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: B.D.M. is the founder and a director of Strata Collective Ltd, which owns the StrataScore engine evaluated in this study. No other author has a financial relationship with Strata Collective Ltd or StrataScore. M.W. declares honoraria from Servier and Merck; and travel, accommodation and/or expenses from Servier, Takeda and Merck unrelated to this research.

Abstract

Objective

Peer review is slow and reviewer assessments vary. We evaluated the performance of StrataScore, an artificial intelligence (AI)-powered appraisal engine that scores manuscripts on six weighted domains using design-specific pathways, against expert human peer reviewers.

Materials and Methods

In this agreement study, 68 published manuscripts across six medical specialties and study designs were appraised by StrataScore and by 33 international expert reviewers (198 reviews, balanced incomplete block design). The primary outcome was agreement between the StrataScore overall score and the mean human score, measured with a two-way mixed-effects, absolute-agreement intraclass correlation coefficient (ICC). Secondary analyses examined specialty, study design, domain, human inter-rater reliability and three general-purpose large language models.

Results

StrataScore showed good to excellent agreement with mean human scores (ICC 0.77, 95% CI 0.66–0.85; pre-specified threshold ≥ 0.60), ranging from 0.71 to 0.86 across specialties and from 0.50 to 0.99 across study designs. General-purpose large language models agreed less closely (ICC 0.50–0.67). Human inter-rater reliability varied widely, from 0.17 in cardiology to 0.78 in respiratory medicine.

Discussion

Agreement met the pre-specified threshold and was numerically higher than for general-purpose models in an exploratory comparison, which suggests that a pre-specified appraisal framework may narrow the gap between automated and expert judgement. Subgroup estimates were imprecise.

Conclusion

StrataScore showed good to excellent agreement with mean expert reviewer scores in this selected corpus, while subgroup estimates were variable. Independent replication and prospective evaluation in editorial workflows are needed.

Registration

Open Science Framework, osf.io; https://doi.org/10.17605/OSF.IO/H9W3R

Background and Significance

Critical appraisal underpins the use of biomedical research in clinical practice [1]. The volume of published literature now exceeds what individuals or institutions can evaluate comprehensively [2]. Methods have also become more complex, with adaptive trial designs, multivariable modelling and heterogeneous meta-analyses that need specialist expertise to appraise [3, 4].

Peer review is the main way in which research quality is assessed before publication, yet its limitations are well documented. Submissions to journals have been rising by about 6% a year, contributing to a recognised reviewer recruitment crisis [5–7]. Agreement between reviewers in structured appraisal tasks is moderate, formal appraisal criteria are applied inconsistently, and methodological deficiencies in published work raise questions about the reliability of peer review [8–12].

Large language models (LLMs) and artificial intelligence (AI)-enabled tools have emerged as potential supports for peer review. Interest has grown after preliminary evaluations of research quality suggested partial agreement between LLMs and human judgement [13, 14]. However, general-purpose LLMs were not designed for structured appraisal tasks and their outputs are sensitive to prompt formulation [15]. LLMs are further limited by non-deterministic sampling resulting in variance across repeated queries, and are not grounded in a calibrated, domain-specific appraisal framework [15, 16]. To our knowledge, no published study has prospectively evaluated agreement between a structured AI-assisted research appraisal system and expert human scoring using standardised psychometric methods and a multi-specialty manuscript corpus.

StrataScore was designed to address these limitations. It applies a pre-specified framework of six weighted domains and returns scores on a 0–100 scale, rather than the free-text output typical of general-purpose LLM queries. It generates output at a temperature of 0 (a model setting that makes output as deterministic as possible) and is not updated by user feedback or user data. The tool is intended as a structured quality signal supporting peer review.

Objective

We conducted a prospective, pre-registered, multi-specialty agreement study with three primary aims: (1) to quantify agreement between StrataScore outputs and structured expert human appraisal at overall and domain level; (2) to determine whether agreement varies by clinical specialty or study design; and (3) to contextualise StrataScore performance against unstructured general-purpose LLM outputs as an exploratory benchmark.

Materials and Methods

Study Design

This is a pre-registered, prospective, multi-specialty cross-sectional agreement study comparing structured automated research appraisal outputs from StrataScore against blinded independent expert human reviewer scores. The study was registered prospectively prior to data collection (10.17605/OSF.IO/H9W3R).

Settings and Data Source

Manuscripts appraised in this study were drawn from the peer-reviewed biomedical literature across six clinical specialties: stroke medicine, cardiology, renal medicine, oncology, anaesthetics, and respiratory medicine. For each specialty, a specialty lead identified candidate manuscripts that met the inclusion criteria, aiming for a spread of study designs and quality. The manuscript corpus was therefore a purposive, non-random sample and the reviewer panel a convenience sample. Each manuscript was also rated for quality on a four-point scale ranging from very low (1) to high (4) (Supplementary Table 1).

Manuscript Corpus

Study designs included randomised controlled trials, cohort studies, cross-sectional studies, case-control studies, and systematic reviews or meta-analyses. Manuscripts were eligible if they were full-text publications in a peer-reviewed biomedical journal, reported primary empirical research or a systematic review or meta-analysis, and were relevant to the assigned clinical specialty. Editorials, commentaries, letters and opinion pieces without primary data, and non-English language publications, were excluded.

Reviewer Recruitment and Eligibility

Expert human reviewers were recruited from within each specialty through specialty leads and professional networks. Eligibility required: (a) a substantive clinical or academic post in the relevant specialty; and (b) documented experience in research appraisal, defined as peer review or an editorial position for a biomedical journal.

Reviewer background information was collected (Table 1) on year of primary medical qualification, highest research degree; publications, peer-reviews completed, frequency of engagement with primary research and self-rated familiarity with appraisal frameworks on a five-point Likert scale.

Reviewer Assignment

A balanced incomplete block design (BIBD) was used to ensure an even distribution of manuscripts between reviewers [17]. Within each specialty, manuscripts were allocated so that each reviewer was assigned six manuscripts and each manuscript was assigned to three reviewers (planned: 12 manuscripts and 6 reviewers per specialty). Respiratory medicine was added after registration (May 2026), with eight manuscripts and four reviewers. Conflicts of interest were checked prior to review. Where a reviewer withdrew, their outstanding manuscripts were reallocated to other reviewers in the same specialty.

Review Procedure and Blinding

All human reviewer templates were completed independently, blinded to other reviewer scores and to all StrataScore output. Automated StrataScore outputs were generated after all human reviewer templates had been received and timestamped. Reviewers received the full text of each assigned manuscript and a structured appraisal template. The template listed the six domains used by StrataScore (Statistical Rigour, Bias and Limitations, Study Design Quality, Generalisability, Impact and Significance, and Writing and Clarity) with their definitions (Supplementary Table 2). No formal calibration training was provided. Reviewers were asked to return templates within four weeks.

Each reviewer completed one structured appraisal template per assigned manuscript, providing a domain score (0–100) for each of the six domains, and a confidence rating for each domain score (Low / Moderate / High).

The StrataScore Engine

StrataScore (version 1.0) is a structured appraisal pipeline built around a general-purpose large language model. No model weights were trained or fine-tuned at any stage. The engine's behaviour is specified by the instructions sent to the model with each manuscript, which were written and iteratively revised between September 2025 and March 2026 and did not use any of the 68 manuscripts in this study.

Each appraisal has two steps. First, study design is classified by a separate model (Claude Haiku 4.5, Anthropic PBC, San Francisco, CA, USA) from the first 2,000 characters of the manuscript text, which typically contain the title and abstract. The design selects one of four framework-specific instruction sets: RoB 2 with a nine-item CONSORT check for randomised trials; ROBINS-I with the Newcastle-Ottawa Scale for cohort and case-control studies; JBI checklists for cross-sectional studies, case series and case reports; QUADAS-2 with a five-item STARD check for diagnostic accuracy studies; and AMSTAR 2, ROBIS, a 20-item PRISMA 2020 check and GRADE for meta-analyses. Second, the selected instruction set and the full manuscript text are sent to Sonnet 4.6 (Anthropic PBC) in a single call, and the model returns a structured JavaScript Object Notation (JSON) object. All calls were made at a sampling temperature of 0. Temperature controls the randomness of a model's word selection; setting it to 0 reduces variability between runs, although small differences can still occur.

The instruction set contains 2,400 to 4,600 words, depending on the pathway. It specifies the formal domain-level questions of the framework, the six domains and their three to six weighted sub-domains, 0–100 anchor criteria, and design-specific mappings from framework judgements to sub-domain scores (Supplementary Table 2). Only the verbatim instruction text is held as proprietary. It also states explicit rules that convert each framework judgement into a numeric score. For example, an overall RoB 2 judgement of low risk, some concerns or high risk sets the risk-of-bias sub-domain to 90, 62 or 35. Checklist results set the reporting sub-domain, GRADE certainty sets the novelty sub-domain, and I^2 thresholds set the robustness sub-domain. This ties each score to a named framework judgement, and it

is the main difference from the general-purpose comparators described below, which received domain names and definitions only.

Manuscript text is extracted from the PDF as plain text. Figure captions and text-based tables are included, but figures and image-embedded tables are not seen by the engine. The engine returns the six domain scores; sub-domain scores are not returned. It retrieves no external information, appraises only the text supplied to it, and is not updated by user feedback or user data. Alongside the scores it returns text outputs, which were not analysed in this study and are examined in a companion qualitative analysis.

Manuscripts for this study were run through StrataScore by A.Kal. via the online appraisal workflow.

Large Language Model Appraisals

Three general-purpose LLMs were evaluated. Comparators were the leading frontier models from three developers at the time of analysis: GPT-5.6 Sol (OpenAI), Claude Opus 5 (Anthropic), and Gemini 3.1 Pro (Google DeepMind). The models were used through each developer's standard chat interface, and all manuscripts were processed during August 2026. Use of large language models is reported using the GUIDE-LLM checklist (Supplementary Appendix 3) [18].

Each LLM received an identical pre-specified prompt (Supplementary Appendix 1), providing domain names and definitions. The full manuscript PDF was attached to the user input. Each model was required to return an integer score from 0 to 100 for each domain as a JSON object. The chat interfaces do not allow temperature to be set, so default settings were used for all three models. Each model was run once per paper.

Overall Score Derivation

Paper scores for both reviewers and StrataScore were derived post hoc by applying the pre-specified StrataScore domain weighting formula to each of the six domain scores:

Overall score = 0.30S + 0.20B + 0.18D + 0.12G + 0.10I + 0.10W

where S = Statistical Rigour, B = Bias and Limitations, D = Study Design Quality, G = Generalisability, I = Impact and Significance, and W = Writing and Clarity. The weights were set by the developers before data collection, based on their judgement of the relative importance of each domain. They were not derived from the study data. The human overall score for each paper was then calculated as the mean of the independent reviewers' weighted overall scores.

Outcome Definitions

Primary outcome

The primary outcome is the intraclass correlation coefficient (ICC) between the automated StrataScore and the mean human reviewer overall score, calculated across all manuscripts using a two-way mixed-effects model with absolute agreement for a single rating, denoted ICC(A,1). An a priori threshold of ICC ≥ 0.60 was the minimum criterion for acceptable agreement.

Secondary outcomes

Secondary outcomes were the domain-level ICC between automated and mean human subscores for each of the six domains; the ICC between automated and mean human overall scores, stratified by clinical specialty and by study design category; the mean difference (bias) between automated and mean human overall scores; and human inter-rater reliability within each specialty.

Exploratory outcomes

The exploratory outcome was the ICC between the overall score of each general-purpose large language model and the mean human overall score.

Sample Size Calculation

Sample size was determined a priori for the primary outcome of ICC-based agreement between the automated StrataScore overall score and the mean human overall score. For the power calculation, an ICC of 0.75 was used as the alternative hypothesis (H_1) and an ICC of 0.50 as the null hypothesis (H_0). The criterion for acceptable agreement in the primary analysis was ICC ≥ 0.60. Using the method of Donner and Eliasziw with a two-sided alpha of 0.05 and target power of 85%, a minimum of 55 manuscripts was required [19].

Statistical Analyses

The primary and secondary analyses were pre-specified in the Statistical Analysis Plan (SAP) pre-registered on OSF. Deviations from the registered plan were: respiratory medicine was added after registration (May 2026); and the test-retest analysis was added post hoc. No analyses were conducted until all human reviewer templates were received and locked, and automated StrataScore outputs had subsequently been generated. Statistical analyses were conducted in R (version 4.6.0). The human reference score for each paper was the mean of all completed reviews for that paper. The study is reported in accordance with the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (Supplementary Appendix 2) [20].

Primary analysis

Agreement between the automated overall StrataScore and the mean human overall score was assessed using the intraclass correlation coefficient under a two-way mixed-effects model with absolute agreement (ICC(A,1), single measures) across all manuscripts. ICC(A,1) follows the notation of McGraw and Wong [21] and quantifies absolute agreement, so systematic offsets between StrataScore and the

human mean lower the estimate. ICCs were calculated in R using analysis of variance (stats::aov) within custom functions. For agreement between StrataScore and the mean human score, 95% confidence intervals were approximated using Fisher's z transformation with a standard error of 1/√(n−3). For test-retest and human inter-rater reliability, 95% confidence intervals were derived from the F distribution. ICC point estimates were interpreted descriptively as poor (<0.40), fair (0.40–0.59), good (0.60–0.74) or excellent (≥0.75) [22].

StrataScore Test-Retest Reliability

To assess output stability, a subsample of 24 manuscripts (four per specialty, selected at random) was re-submitted to the StrataScore engine a second time, with study design classification held fixed to isolate scoring variability from classification variability. This post hoc analysis quantified test-retest reliability using ICC(A,1) between the two runs. The second run was performed one month after the first, using the same model version and settings.

Secondary analyses

Domain level agreement was assessed using ICC(A,1) between automated domain subscores and mean human domain subscores for each of the six domains, calculated across all manuscripts, and agreement across domains was compared descriptively. The pre-specified secondary hypothesis required ICC ≥ 0.50 for at least four of six domains.Agreement was calculated separately for each specialty, using ICC(A,1) between automated and mean human overall scores, and was also stratified by study design category.Mean difference and 95% limits of agreement between automated and mean human overall scores were calculated using the method of Bland and Altman [23]. Human inter-rater reliabilitywithin each specialty was calculated using a one-way random-effects model with absolute agreement reported as ICC(1,1).

Sensitivity Analysis

Specialties in which human inter-rater reliability fell below a pre-specified threshold of ICC(1,1) < 0.40 were identified, and the primary analysis was repeated excluding all manuscripts from those specialties.

Exploratory LLM Comparison

For each of the three general-purpose LLM comparators, ICC(A,1) (two-way mixed-effects, absolute agreement, single measures) was calculated between each LLM’s overall score and the mean human overall score across all manuscripts.

Ethical Approval

The College of Medical, Veterinary and Life Sciences (MVLS) Ethics Committee, University of Glasgow, reviewed the project application (StrataScore Validation Study) and confirmed that formal ethical review was not required.

Data Availability

Deidentified data supporting the study findings and the analysis code are available on reasonable request.

Results

Paper and Reviewer Flow

A total of 68 manuscripts were assessed across six clinical specialties (stroke medicine, cardiology, renal medicine, oncology, anaesthetics, and respiratory medicine), by 33 reviewers across six countries. Reviewer demographics are summarised in Table 1. In total, 198 reviews were completed: 50 manuscripts received three independent reviews, six received four and 12 received two. Five of the six planned cardiology reviewers completed the study. The manuscripts comprised 24 randomised controlled trials, 5 prospective cohort studies, 11 retrospective cohort studies, 4 case-control studies, 12 meta-analyses and 12 cross-sectional studies; quality ratings were 1 (very low) for 14 manuscripts, 2 (low) for 19, 3 (moderate) for 17 and 4 (high) for 18 (Supplementary Table 1).

Table 1 – Reviewer Panel Characteristics

Reviewer Characteristic	n = 33
Specialty, n (%)	
Stroke medicine	6 (18.2)
Cardiology	5 (15.2)
Renal medicine	6 (18.2)
Oncology	6 (18.2)
Anaesthetics	6 (18.2)
Respiratory Medicine	4 (12.1)
Country of primary affiliation, n (%)	
United Kingdom	27 (81.8)
Italy	1 (3.0)
Ireland	1 (3.0)
Canada	2 (6.1)
Singapore	1 (3.0)
Switzerland	1 (3.0)
Years since graduation from primary medical qualification, median (IQR)	
	13 (10.5–21.5)
Highest research degree obtained, n (%)	
DSc	1 (3.0)
PhD	13 (39.4)
MD	8 (24.2)
Master's	5 (15.2)
Bachelor's	1 (3.0)
MBChB	4 (12.1)
MPharm	1 (3.0)
Authored publications, median (IQR)	14 (7.0–38.5)
Peer reviews completed, median (IQR)	10 (3.3–43.8)
Research engagement frequency, n (%)	
Monthly	3 (9.1)
Weekly	17 (51.5)

Daily	13 (39.4)
Appraisal framework familiarity (1–5 scale)[a], median (IQR)	
CONSORT	4 (3–5)
STROBE	4 (3–5)
PRISMA	4 (3–4)
ROB 2	3 (2–4)
CASP	2 (1–3)
GRADE	3 (2–4)
ROBINS 1	2 (1–3)
QUADAS 2	2 (1–2)
Newcastle-Ottawa Scale	2 (1–4)
AMSTAR 2	2 (1–3)
ROBIS	2 (1–3)
Cochrane Heterogeneity	3 (2–4)
Time to return reviews (days), median (IQR)	41 (18.5–51.5)

[a] 1 = not at all familiar, 5 = extremely familiar.

IQR, interquartile range; DSc, Doctor of Science; PhD, Doctor of Philosophy; MD, Doctor of Medicine; CONSORT, Consolidated Standards of Reporting Trials; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; ROB 2, Risk of Bias 2; CASP, Critical Appraisal Skills Programme; GRADE, Grading of Recommendations Assessment, Development and Evaluation; ROBINS-I, Risk Of Bias In Non-randomised Studies of Interventions; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies 2; AMSTAR 2, A MeaSurement Tool to Assess systematic Reviews 2; ROBIS, Risk Of Bias In Systematic reviews

Primary Outcome

Agreement between the automated StrataScore and mean human overall score was 0.77 (95% CI 0.66–0.85). StrataScore showed excellent repeatability: agreement between the first and second run was 0.97 (95% CI 0.95–0.99).

Secondary Outcomes

Domain-level agreement between automated and mean human subscores ranged from 0.54 (Writing and Clarity) to 0.74 (Study Design Quality) (Figure 1), and the point estimate for all six domains met the pre-

specified threshold of ICC ≥ 0.50. Agreement stratified by clinical specialty ranged from 0.71 (anaesthetics and respiratory medicine) to 0.86 (cardiology) (Table 2a), and by study design category from 0.50 (cross-sectional) to 0.99 (case-control) (Table 2b).

The mean difference between automated and mean human overall scores was −3.9 points (95% limits of agreement −19.6 to 11.9) (Figure 2), indicating that StrataScore scored slightly lower than human reviewers on average. Single-reviewer reliability, ICC(1,1), within specialty ranged from 0.17 (cardiology) to 0.78 (respiratory) (Supplementary Table 3).

Table 2 – Agreement between StrataScore and mean human overall score, by (a) clinical specialty and (b) study design

(a)

Specialty	N	ICC (95% CI)
Stroke medicine	12	0.82 (0.47–0.95)
Cardiology	12	0.86 (0.58–0.96)
Renal medicine	12	0.75 (0.32–0.93)
Oncology	12	0.84 (0.52–0.95)
Anaesthetics	12	0.71 (0.22–0.91)
Respiratory medicine	8	0.71 (0.01–0.94)

(b)

Study design	N	ICC (95% CI)
Randomised controlled trial	24	0.80 (0.59–0.91)
Prospective cohort	5	0.83 (−0.21–0.99)
Retrospective cohort	11	0.62 (0.03–0.89)
Case-control	4	0.99 (0.77–1.00)
Meta-analysis	12	0.66 (0.15–0.90)
Cross-sectional	12	0.50 (−0.10–0.84)

ICC, intraclass correlation coefficient; CI, confidence interval.

Sensitivity Analysis

Human inter-rater reliability was below the pre-specified threshold (ICC < 0.40) in two specialties, cardiology (0.17) and renal medicine (0.23), comprising 24 manuscripts. Excluding these, agreement was 0.77 (95% CI 0.61–0.87), consistent with the primary analysis.

Large language model comparisons

ICC (95% CI) between mean human overall score and each general-purpose LLM was 0.64 (0.48–0.76) (GPT-5.6 Sol), 0.67 (0.51–0.78) (Claude Opus 5), and 0.50 (0.30–0.66) (Gemini 3.1 Pro) (Figure 3).

Discussion

Principal findings

In this prospective agreement study, a purpose-built AI critical appraisal engine showed good to excellent agreement with expert human appraisal. The point estimate (ICC 0.77) met the pre-specified acceptability threshold of 0.60 and the lower confidence limit was above that threshold. The estimate was numerically higher than those of three general-purpose large language models, although the confidence intervals overlapped. Subgroup estimates were variable. The limits of agreement (−19.6 to 11.9 points; Figure 2) show that individual manuscripts could differ materially from the mean human score, and the Bland–Altman plot suggests that the offset was not constant: StrataScore tended to score higher than the human mean for lower-scoring manuscripts and lower for higher-scoring ones. Agreement was highest for the domain with the most explicit appraisal criteria (Study Design Quality) and lowest for the domain that relies most on reviewer judgement (Writing and Clarity).

Framework Alignment

Agreement appeared to be higher for domains governed by a more explicit, established appraisal framework. Study Design Quality, whose sub-domains are anchored to design features that the pathway-specific instructions make explicit (for example blinding, allocation concealment, eligibility criteria and follow-up), showed the strongest domain-level agreement (ICC 0.74). Writing and Clarity, which depends on a more diffuse, less codified judgement of prose and presentation, showed the weakest (ICC 0.54). Agreement by study design varied widely, with wide, overlapping confidence intervals (Table 2b), so differences between designs should not be over-interpreted. Structured, criterion-based appraisal instruments were developed because unstructured quality scores proved unreliable and difficult to reproduce between raters [24], and the major risk-of-bias and systematic-review appraisal tools in current use were built around explicit, itemised criteria for this reason [25,26]. Our results are compatible with extending that logic to AI-assisted appraisal using an engine constrained to a pre-specified, framework-anchored scoring structure.

Comparison with General-Purpose Language Models

All three general-purpose LLM comparators agreed less closely with human reviewers (ICC 0.50–0.67; fair to good). This is consistent with a small but growing body of parallel work showing that LLMs, when asked to appraise study quality or risk of bias without a structured scaffold, achieve at best fair agreement with expert human assessors and frequently worse [27-30]. A recent comparative validation study across four contemporary LLMs found risk-of-bias agreement with Cochrane assessors ranging from kappa 0.06 to 0.39, concluding that none of the models tested were reliable enough for unsupervised use [27], and related work assessing ChatGPT-class models against Cochrane authors' own risk-of-bias judgements has reported similarly mixed results [28-30]. Alongside our own exploratory comparison, this suggests that the gap between StrataScore and the general-purpose comparators may partly reflect the presence of a structured appraisal scaffold.

Comparison with Previous Literature

The limitations of human peer review are well recognised. A multilevel meta-analysis pooling estimates across the peer-review literature found mean inter-rater reliability for manuscript quality ratings of only 0.34, meaning reviewer disagreement accounted for roughly three-quarters of the variance in how a manuscript was rated [31]. In a comparable setting, Rothwell and Martyn found agreement between reviewers on both acceptance decisions and priority ratings for neuroscience manuscripts was little greater than chance, with kappa as low as 0.08 for acceptance decisions at one journal and below zero for priority ratings at another [8]. Human inter-rater reliability in our study ranged from 0.17 to 0.78 across specialties, consistent with this wider literature. StrataScore's agreement with the mean human score (ICC 0.77) was higher than the single-rater reliability reported by Rothwell and Martyn and within the range of human inter-rater reliability observed here [8]. However, this is not a like-for-like comparison, since our primary outcome compares StrataScore with an average across multiple human reviewers per manuscript, which is expected to be more reliable than any single reviewer.

Given rising submission volumes and the difficulty of recruiting reviewers, an appraisal engine with demonstrated agreement could, if shown to be useful, reduce the burden that peer review places on reviewers [5-7]. An engine with this level of agreement with the mean human reviewer score might be considered for triage or first-pass screening, for example to flag manuscripts for expedited handling or to inform reviewer allocation, alongside expert judgement. Scores should not decide the fate of an individual manuscript without expert review.

Strengths and Limitations

This study was prospectively registered, with the primary outcome and analysis plan specified before analysis [32]. Reviewers spanned six clinical specialties and six countries, and were blinded to one another's scores and all StrataScore outputs. A post hoc test-retest assessment demonstrated high score stability over approximately one month.

Human inter-rater reliability was poor in two specialties, and estimates for some study design subgroups were imprecise. Manuscripts were purposively selected rather than sampled randomly, limiting

generalisability. All had already passed peer review and appeared in identifiable journals, so journal reputation may have influenced both reviewers and StrataScore; the corpus included low-rated papers from high-impact journals (Supplementary Table 1), but this was not tested. Several widely cited trials may have featured in the underlying models' training data. Cardiology showed the highest StrataScore-to-human agreement despite poor human inter-rater reliability, and respiratory medicine showed the reverse. These statistics answer different questions: single-reviewer reliability describes agreement between individual reviewers, whereas StrataScore was compared with the panel mean, in which reviewer-specific noise is partly averaged out. Both also depend on the spread of manuscript quality within each specialty sample, and were estimated from 8 to 12 manuscripts with wide confidence intervals, so the contrast may reflect sampling variation. Interpretation for cardiology is further limited by reviewer attrition, which left seven manuscripts with only two reviews. Agreement with mean human scores does not establish correctness, freedom from bias, or clinical validity, and the human panel is itself an imperfect reference: reviewers disagreed substantially, so the mean of two to four reviewers is a noisy estimate of manuscript quality. Potential bias arising from training data, appraisal frameworks or the developer-defined domain weights was not directly evaluated.

The overall score may obscure differences in the strengths and weaknesses identified by StrataScore and human reviewers; these are examined in a companion qualitative analysis. General-purpose models used default settings, whereas StrataScore used a temperature of zero, limiting direct comparison. Test-retest reliability does not establish stability across model versions, and agreement under alternative domain weights was not directly evaluated.

Implications and Future Directions

Our findings suggest that structured, framework-anchored appraisal may improve agreement with expert human judgement compared with an unconstrained language-model prompt, although the exploratory comparison does not establish superiority and requires independent confirmation.

Future studies should use more representative manuscript samples and assess performance against explicit methodological criteria as well as human ratings. Evaluation in live editorial settings should establish whether the tool improves triage or feedback without adding workload or encouraging misplaced confidence. The effects of changes to model versions, appraisal frameworks, and domain weights should likewise be monitored. These steps would help establish whether structured AI appraisal can provide a reliable and useful adjunct to expert peer review.

Conclusion

In this international, prospective, multidisciplinary agreement study, an AI appraisal engine built around a pre-specified, framework-anchored scoring structure showed good to excellent agreement with mean expert reviewer scores in a selected corpus, although subgroup estimates were variable and imprecise. These findings are compatible with framework-anchored scoring as a promising design principle for AI-

assisted research appraisal. Independent replication in more representative manuscript samples, and prospective evaluation in editorial workflows, are needed before any use in triage.

Declarations

Acknowledgments

We would like to acknowledge the specialty leads who helped with manuscript selection. Generative AI (Sonnet 5.5, Anthropic PBC) was used to edit the language of this manuscript and to check analysis code, as well as checking the study design classification of the appraised manuscripts. The authors reviewed all AI-assisted text and take responsibility for the content of this paper. StrataScore, the system under evaluation, is described in the Methods.

Author Contributions

B.D.M. conceived the study, developed StrataScore, designed and carried out the analysis and wrote the first draft. D.C. and F.L. helped write the first draft and interpret the findings. S.J.N., D.T.-S. and G.A.H. contributed to the design of the study and the interpretation of the findings. A.Kal. ran the manuscripts through StrataScore. J.D. supervised the study. All other authors completed expert reviews of the manuscripts, which form the data analysed here. All authors revised the manuscript and approved the final version.

Sources of Funding

No funding was obtained for this work.

Disclosures

B.D.M. is the founder and a director of Strata Collective Ltd, which owns the StrataScore engine evaluated in this study. No other author has a financial relationship with Strata Collective Ltd or StrataScore. M.W. declares honoraria from Servier and Merck; and travel, accommodation and/or expenses from Servier, Takeda and Merck unrelated to this research.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB (1996) Richardson. Evidence based medicine: what it is and what it isn't. BMJ 312(7023):71–72. 10.1136/bmj.312.7023.71
2. Bastian H, Glasziou P, Chalmers I (2010) Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med 7(9):e1000326. 10.1371/journal.pmed.1000326

3. Ioannidis JP (2016) The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. Milbank Q 94(3):485–514. 10.1111/1468-0009.12210
4. Bhatt DL, Mehta C (2016) Adaptive Designs for Clinical Trials. N Engl J Med 375(1):65–74. 10.1056/NEJMra1510061
5. Aczel B, Szaszi B, Holcombe AO (2021) A billion-dollar donation: estimating the cost of researchers' time spent on peer review. Res Integr Peer Rev 6(1):14. 10.1186/s41073-021-00118-2 [published Online First: 20211114]
6. Ware MM (2015) M. The STM Report: An overview of scientific and scholarly journal publishing. http://digitalcommons.unl.edu/scholcom/9
7. Publons (2018) Global state of peer review. https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf
8. Rothwell PM, Martyn CN, Brain (2000) ;123(Pt 9):1964–196910.1093/brain/123.9.1964
9. Jefferson T, Rudin M, Brodney Folse S, Davidoff F (2007) Editorial peer review for improving the quality of reports of biomedical studies. Cochrane Database Syst Rev 2007(2):Mr000016. 10.1002/14651858.MR000016.pub3 [published Online First: 20070418]
10. Smith R (2006) Peer review: a flawed process at the heart of science and journals. J R Soc Med 99(4):178–182. 10.1177/014107680609900414
11. Glasziou P, Altman DG, Bossuyt P et al (2014) Reducing waste from incomplete or unusable reports of biomedical research. Lancet 383(9913):267–276. 10.1016/s0140-6736(13)62228-x [published Online First: 20140108]
12. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC (2004) Altman. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 291(20):2457–2465. 10.1001/jama.291.20.2457
13. Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G (2021) AI-assisted peer review. Humanit Social Sci Commun 8(1):25. 10.1057/s41599-020-00703-8
14. Liang W, Zhang Y, Cao H et al (2024) Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. NEJM AI 1(8):AIoa2400196. 10.1056/AIoa2400196
15. Quantifying language models (2024) ' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting. International Conference on Learning Representations (ICLR)
16. Van Dis BJ, Zuidema EAM, van Rooij W, Bockting R (2023) ChatGPT: five priorities for research. Nat 2023 doi 614(7947):224–226
17. Huang EP, Shih JH (2021) Assigning readers to cases in imaging studies using balanced incomplete block designs. Stat Methods Med Res 30(10):2288–2312 : 10.1177/09622802211037074 [published Online First: 20210901]

18. Feuerriegel S, Barrie C, Crockett MJ et al (2026) A reporting checklist for large language models in behavioural science. Nat Hum Behav 10(7):1182–1186. 10.1038/s41562-026-02492-7
19. Donner A, Eliasziw M (1987) Sample size requirements for reliability studies. Stat Med 6(4):441–448. https://doi.org/10.1002/sim.4780060404
20. Kottner J, Audigé L, Brorson S et al (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64(1):96–106 [published Online First: 20100617]
21. McGraw WS (1996) Forming inferences about some intraclass correlation coefficients. Psychol Methods 1:30–46. 10.1037/1082-989X.1.1.30
22. D. V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess (1994) ;6(4):284–29010.1037/1040-3590.6.4.284
23. Bland JM (1986) Altman. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476):307–310
24. Jüni P, Witschi A, Bloch R, Egger M (1999) The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282(11):1054–1060. 10.1001/jama.282.11.1054
25. Sterne JAC, Savović J, Page MJ et al (2019) RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 366:l4898. 10.1136/bmj.l4898
26. Shea BJ, Reeves BC, Wells G et al (2017) AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 358:j4008. 10.1136/bmj.j4008 [published Online First: 20170921]
27. Nyrhi L, Ponkilainen V, Laaksonen J et al (2026) Large language models for risk-of-bias assessment in randomised clinical trials—a comparative validation study. eBioMedicine 126:106238. https://doi.org/10.1016/j.ebiom.2026.106238
28. Rose CJ, Bidonde J, Ringsten M et al (2025) Using a Large Language Model (ChatGPT-4o) to Assess the Risk of Bias in Randomized Controlled Trials of Medical Interventions: Interrater Agreement With Human Reviewers. Cochrane Evid Synth Methods 3(5):e70048. 10.1002/cesm.70048 [published Online First: 20250910]
29. Taneri PE (2025) Human Versus Artificial Intelligence: Comparing Cochrane Authors' and ChatGPT's Risk of Bias Assessments. Cochrane Evid Synth Methods 3(5):e70044. 10.1002/cesm.70044 [published Online First: 20250831]
30. Kuitunen I, Ponkilainen VT, Liukkonen R et al (2024) Evaluating the Performance of ChatGPT-4o in Risk of Bias Assessments. J Evid Based Med 17(4):700–702. 10.1111/jebm.12662 [published Online First: 20241215]
31. Bornmann L, Mutz R (2010) Daniel. A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. PLoS ONE 5(12):e14331. 10.1371/journal.pone.0014331
32. Nosek BA, Ebersole CR, DeHaven AC (2018) Mellor. The preregistration revolution. Proc Natl Acad Sci U S A 115(11):2600–2606. 10.1073/pnas.1708274114

Figures

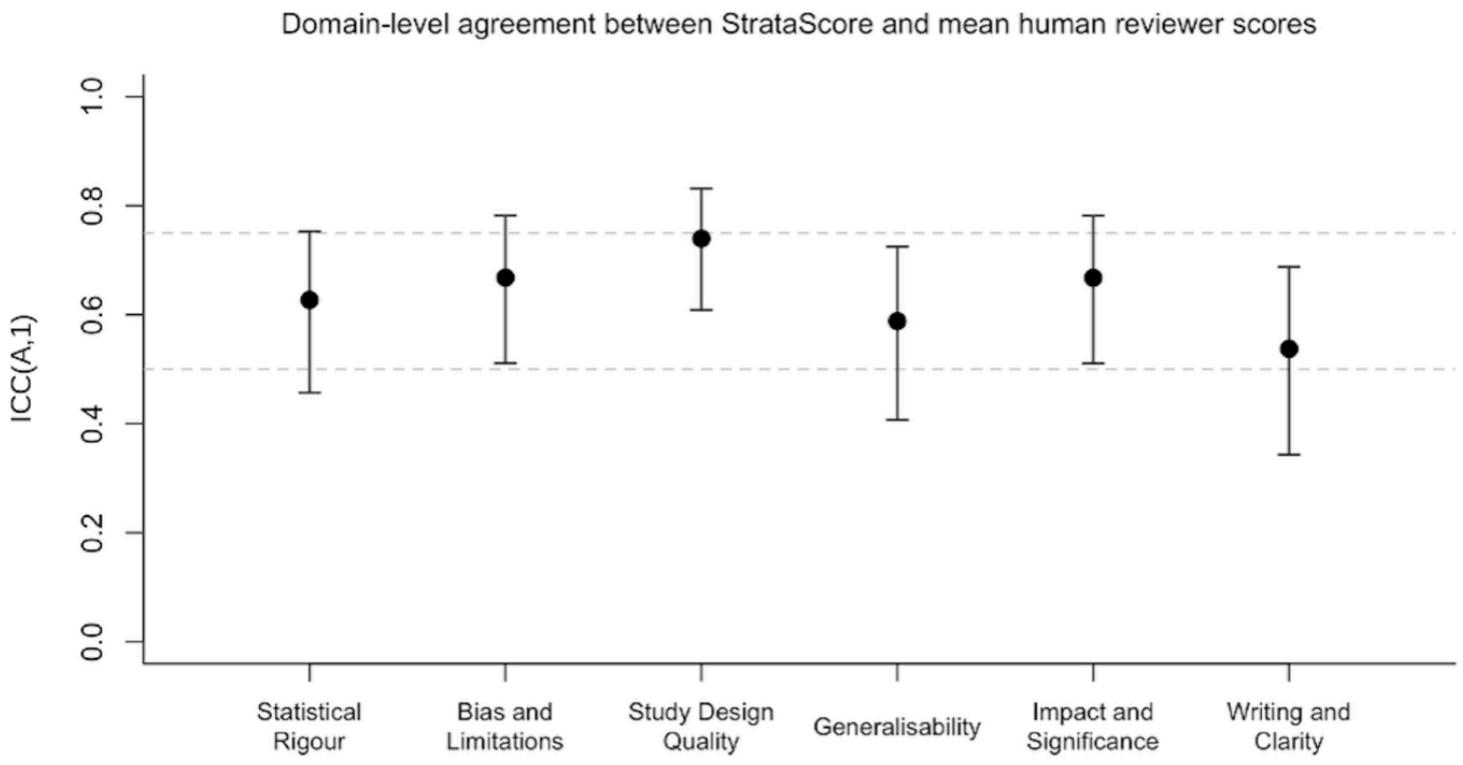

Figure 1

Domain-level agreement between StrataScore and mean human reviewer scores. Points show ICC(A,1) with 95% confidence intervals for each of the six domains; dashed lines mark 0.50 and 0.75.

Alt text: Point estimates and 95% confidence intervals for agreement between StrataScore and mean human reviewer scores across six appraisal domains. Intraclass correlation coefficients range from 0.54 for Writing and Clarity to 0.74 for Study Design Quality. The other four domains fall between these values. Dashed reference lines indicate ICC values of 0.50 and 0.75.

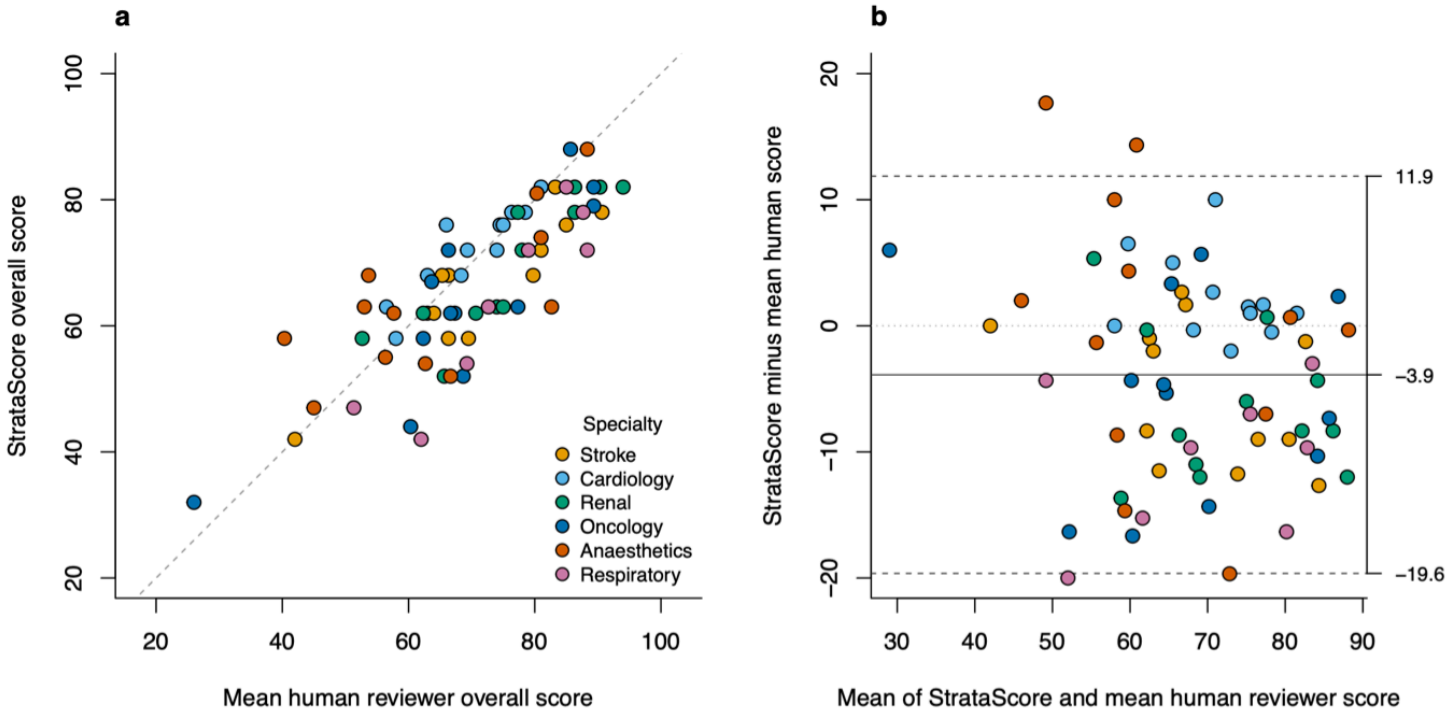

Figure 2

Individual manuscript scores. (a) StrataScore overall score against mean human reviewer overall score for each of the 68 manuscripts, with the line of identity; points are coloured by specialty. (b) Bland–Altman plot of the difference (StrataScore minus mean human score) against the mean of the two scores. In panel b, the solid line shows the mean difference and the dashed lines the 95% limits of agreement.

Alt text: Scatter plot and Bland–Altman plot for 68 manuscripts. StrataScore scores track mean human scores around the line of identity, with a mean difference of −3.9 points and limits of agreement from −19.6 to 11.9.

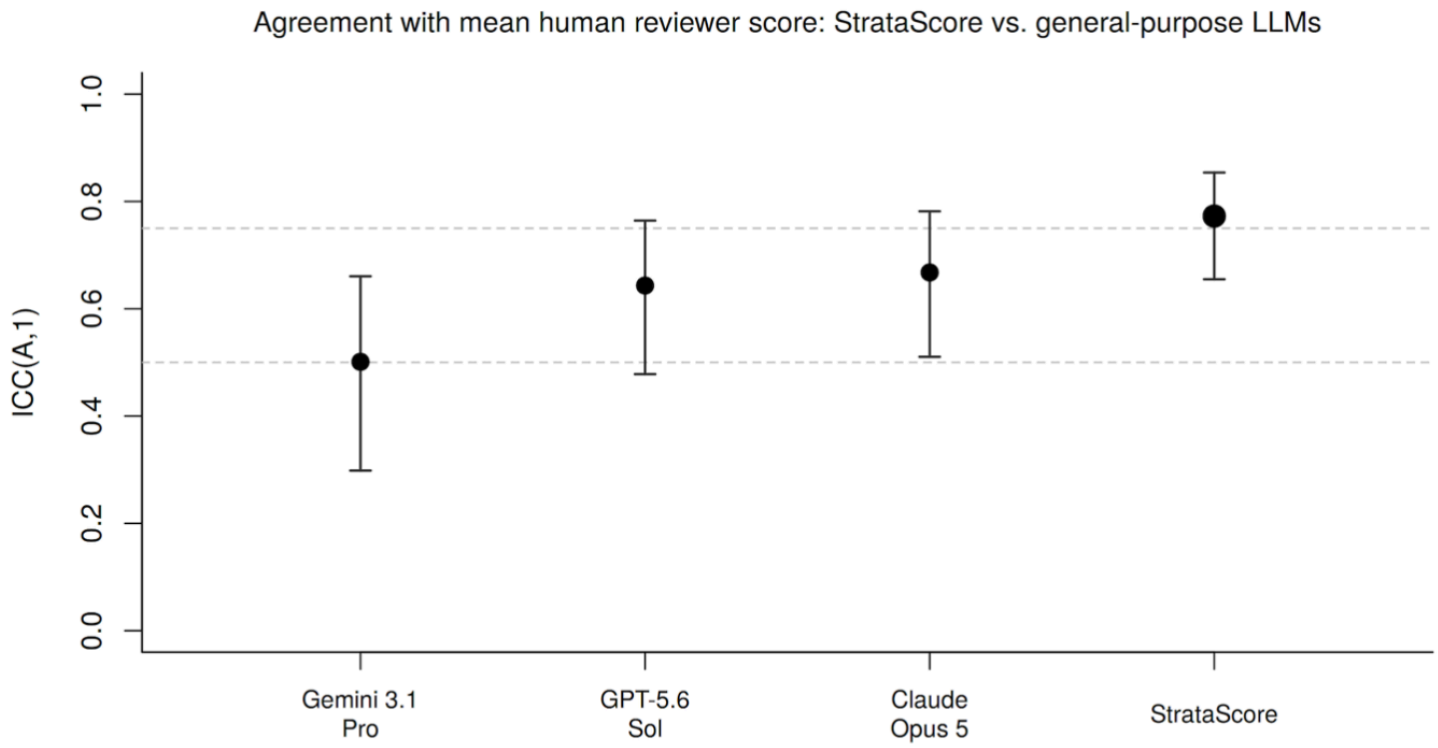

Figure 3

Agreement with mean human reviewer score: StrataScore vs. general-purpose LLMs. Points show ICC(A,1) with 95% confidence intervals; dashed lines mark 0.50 and 0.75.

Alt text: Point estimates and 95% confidence intervals for agreement with mean human reviewer scores, comparing StrataScore with three general-purpose language models. StrataScore has an ICC of 0.77 (95% CI 0.66–0.85), compared with 0.64 (0.48–0.76) for GPT-5.6 Sol, 0.67 (0.51–0.78) for Claude Opus 5, and 0.50 (0.30–0.66) for Gemini 3.1 Pro. Dashed reference lines indicate ICC values of 0.50 and 0.75.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx