European Journal of Radiology 191 (2025) 112358

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Auto-METRICS: LLM-assisted scientific quality control for radiomics research

José Guilherme de Almeida [a,*], Nickolas Papanikolaou [a,b]

[a] Champalimaud Foundation, Lisbon, Portugal
[b] Royal Marsden Hospital, Sutton, UK

ARTICLE INFO

Keywords:
Large language models
Artificial intelligence
Radiomics
METRICS

ABSTRACT

Purpose: The quality of radiomics research is critical for reliable clinical translation, yet methodological flaws remain prevalent. This study evaluates whether large language models (LLMs) can reliably assess radiomics methodological quality using the METhodological RadiomICs Score (METRICS).

Methods: We compared a commercial cloud-based LLM (Gemini Flash 2.0) METRICS assessments for 46 articles with those of radiologists using two reproducibility studies (ADA2025 and K2025, with 6 radiologist groups and 3 radiologists, respectively, with varying degrees of experience). Cohen's kappa (κ) and METRICS Pearson's correlation (PC), and error rates between LLMs and human raters were evaluated. Prompt clarifications to METRICS were suggested to improve human-LLM agreement. Twenty four privacy-preserving open LLMs were compared with Gemini Flash 2.0.

Results: In ADA2025, the commercial LLM achieved inter-rater agreements with human raters comparable to those between human raters (average $\kappa = 0.48$ vs. average $\kappa = 0.48$, respectively, Wilcoxon rank-sum test $p = 0.41$), leading to similar correlation values in METRICS scoring (average PC = 0.62 vs. average PC = 0.56, Wilcoxon rank-sum test $p = 0.11$). This was confirmed with K2025 (mean human-LLM $\kappa = 0.58$ vs. human–human $\kappa = 0.57$, Wilcoxon rank-sum test $p = 0.28$), with no evidence for correlation differences (PC = 0.68 vs. PC = 0.51, respectively, Wilcoxon rank-sum test $p = 0.55$). Phi4-Reasoning, an open model which can be run locally, performed comparably to Gemini Flash 2.0 (median ranking = 1 vs. median ranking = 3, respectively, across all raters).

Conclusion: LLMs can assist in standardized radiomics quality assessment. Open privacy-preserving models can offer comparable performance to commercial cloud-based LLMs, suggesting their utility in supporting human raters for evaluating radiomics research integrity.

1. Introduction

Machine-learning (ML) studies can inflate their conclusions due to methodological biases or errors [1]. Coupled with fast-paced innovation in the world of artificial intelligence (AI), this can lead to excessive confidence in the capacities of AI and ML [2]. This is the case in clinical AI: Maleki *et al.* showed how three major methodological pitfalls (violation of the independence assumption, model evaluation with inappropriate metrics or baselines, and batch effect) led to unrealistic performance overestimates in head and neck computed tomography (CT), lung CT, chest radiography and histopathology data [3].

Radiomics — the subfield of clinical AI focusing on extracting patterns from radiological images — is affected by similar biases. Importantly, radiomics literature is biased towards positive results, leading to inflated perceptions of the efficacy of these methods [4]. Lu *et al.* showed that such conclusions can be attributable to uncontrolled confounders such as the inclusion of features associated with slice thickness or tumour size [5]. Cannella *et al.* showed how different data splits could lead to serendipitous performance estimates in microvascular invasion prediction in hepatocellular carcinoma [6]. The application of different feature selection strategies (typically recommended in radiomics research) influences performance [7], while CT acquisition parameters can negatively impact radiomic feature reproducibility [8]. Gillies *et al.* claimed that, for radiomics, "images are more than pictures, they are data" [9]. However, multiple poorly designed publications have raised questions about the reliability of the field [10].

* Corresponding author.
E-mail address: jose.almeida@research.fchampalimaud.org (J.G. de Almeida).

https://doi.org/10.1016/j.ejrad.2025.112358
Received 9 May 2025; Received in revised form 16 July 2025; Accepted 5 August 2025
Available online 7 August 2025

As noted in the literature, radiomics cannot be applied in the clinic without proper method assessment [11,12]. To make radiomics research more robust and interpretable, two radiomic quality metrics were developed in recent years: the Radiomics Quality Score (RQS) [13] and the METhodological RadiomICs Score (METRICS) [14]. While other guidelines exist for radiomics (such as the CheckList for EvaluAtion of Radiomics research [15]), only RQS and METRICS are used to evaluate methodological quality. While both are relatively straightforward, recent studies showed that the latter is more reproducible and accurate [16,17]. Akinci D'Antonoli *et al.* (2025) [18] and Kocak *et al.* (2025) [17] released reproducibility estimates for METRICS. Their excellent reporting provides individual answers for METRICS across 6 rater groups and 3 raters, respectively.

Large language models (LLMs) are recent technological advances with remarkable human-like text generation capacities in scientific topics such as molecular property prediction [19] or experimental result prediction [20]. One such topic is that of assessing scientific articles, where LLMs have been successfully used to analyse scientific summaries [21], provide feedback [22], screen systematic reviews [23], and potentially increase review quality through feedback [24].

Here, we make use of recent literature on METRICS reproducibility to determine whether inter-rater agreements between LLMs are similar to those observed between expert raters. To achieve this, we use a protocol combining detailed prompting with structured output generation. We analyse inter-rater agreements between raters/rater groups and a commercially available LLM, showing how differences between human raters are similar to those observed between human raters and the commercially available LLM. We conduct similar assessments for 24 open LLMs, determining that some perform comparably to Gemini Flash 2.0.

2. Methods

2.1. Data collection

We collect the answers for all 30 items and 5 conditions used in METRICS and described in [14] from two different sources: Akinci D'Antonoli *et al.* (2025) [18] (ADA2025) and Kocak *et al.* (2025) [17] (K2025). Both articles contain assessments performed by radiologists, guaranteeing that all METRICS evaluations are provided by a cohort of experts.

We collect the data from ADA2025 available as Supplement 1 in the original publication [18]. These data have 6 distinct conditions: 2 treatment conditions (with and without training on how to use METRICS as a quality assessment tool) and 3 levels of expertise (1, 2 and 3, considering both years practicing and academic experience in producing/evaluating radiomics studies). These conditions were populated by 12 different radiologists, equally divided into one of two treatment conditions (received/did not receive training) and into one of three levels of expertise.

For K2025, which analyses articles on radiomics analyses of glioma, we retrieved the individual METRICS scores from three supplementary figures (Figs. S1, S2 and S3 in the original publication), one provided for each radiologist [17].

For ADA2025 and K2025, we retrieved 27 and 19 full-text articles by copy-pasting from each article webpage out a total of 34 and 27 articles, respectively. Articles were excluded due to lack of institutional access or restrictive copyright licenses. The complete list of articles, annotated for whether they were retrieved is presented in Supplementary Table 1. The ratings in ADA2025 and K2025 were used to recalculate human–human inter-rater agreements using exclusively the retrieved full-text articles and calculate human-LLM inter-rater agreements.

2.2. Large language model experiment

For our approach, which we call "Auto-METRICS" (Automated METRICS), we use Gemini Flash 2.0 accessed through its API on June 24th, 2025. We made use of the Gemini paid plan to ensure that no data is used to train or improve Google models or products. The used generation parameters were temperature = 0.0, top-P = 0.001, top-K = 5, and presence/frequency penalties of 0. While there is the possibility of stochasticity, this maximises reproducibility (with low temperatures and top-P, and top-K) while also reducing biasing the generation through presence/frequency penalties. The choice of this particular model was purely heuristic — while different models can yield different results, we focused on a single model to showcase whether LLMs have the potential to assist in automatic standardised assessment of scientific production. We make the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM) checklist available as Supplementary Table 2.

2.2.1. Prompt construction and clarifications

To construct the prompt we use the prompt detailed in the Supplementary Methods. Summarily, we provide:

1. Instructions to the prompt, explaining how it should analyse the text of each manuscript;
2. Details on what METRICS is, what it tries to analyse and a brief description (i.e. 30 questions and 5 conditions defining whether some questions should be answered)
3. The evaluation rubric for each question/condition: choose between yes, no and n/a (if a condition allows a question to be skipped) and provide a reason for each evaluation
4. The input format: an entire scientific article
5. The output format: a structured output format containing a summary for the scientific publication and the answers to each of the 30 questions and 5 conditions, as well as a reason for each rating.
6. The complete guidelines available in the METRICS online tool (https://metricsscore.github.io/metrics/METRICS.html, accessed on March 30th, 2025) [14]. To do so, we copied all items and conditions and their respective explanations available in each tooltip to characterise each item/condition in free-text format.

We used constrained generation (i.e. structured output) generation as made available in the Google Gemini API [25], ensuring that the output format specified above was consistent.

To potentially improve the prompt, we tested Auto-METRICS on the scientific articles analysed in ADA2025 and checked whether Gemini 2.0 Flash output was systematically in disagreement with human raters in specific items. The protocol to do this consisted in analysing manuscripts where the average agreement between the LLM and the raters was below 1 (the agreement was 1 when the LLM and the human rater group agreed, 0 if the item was not considered by either, −1 when they disagreed). The prompt improvement strategy consisted in analysing the "reason" provided by the LLM for each rating. If the reason for recurring disagreements was systematic (the LLM tended to disagree with human raters due to misinterpretations of the prompt, poor contextual understanding, or other discernible causes), we expanded the prompt with clarifications.

We term the initial LLM prompt outputs as "LLM" and the prompt with clarifications as "LLM + changes". We then applied both prompts to K2025, using this to validate our improved prompt.

2.2.2. Comparison with open models

We compare Auto-METRICS with variants using open LLMs and open large reasoning models (LRMs) through Ollama version 0.6.6,[1] a lightweight framework to run LLMs and LRMs locally. We test 18 open LLMs and 6 LRMs, adapting prompts and output formats in LLMs and LRMs to improve text generation stability and better support for reasoning

[1] https://ollama.com/, accessed on April 24th, 2025.

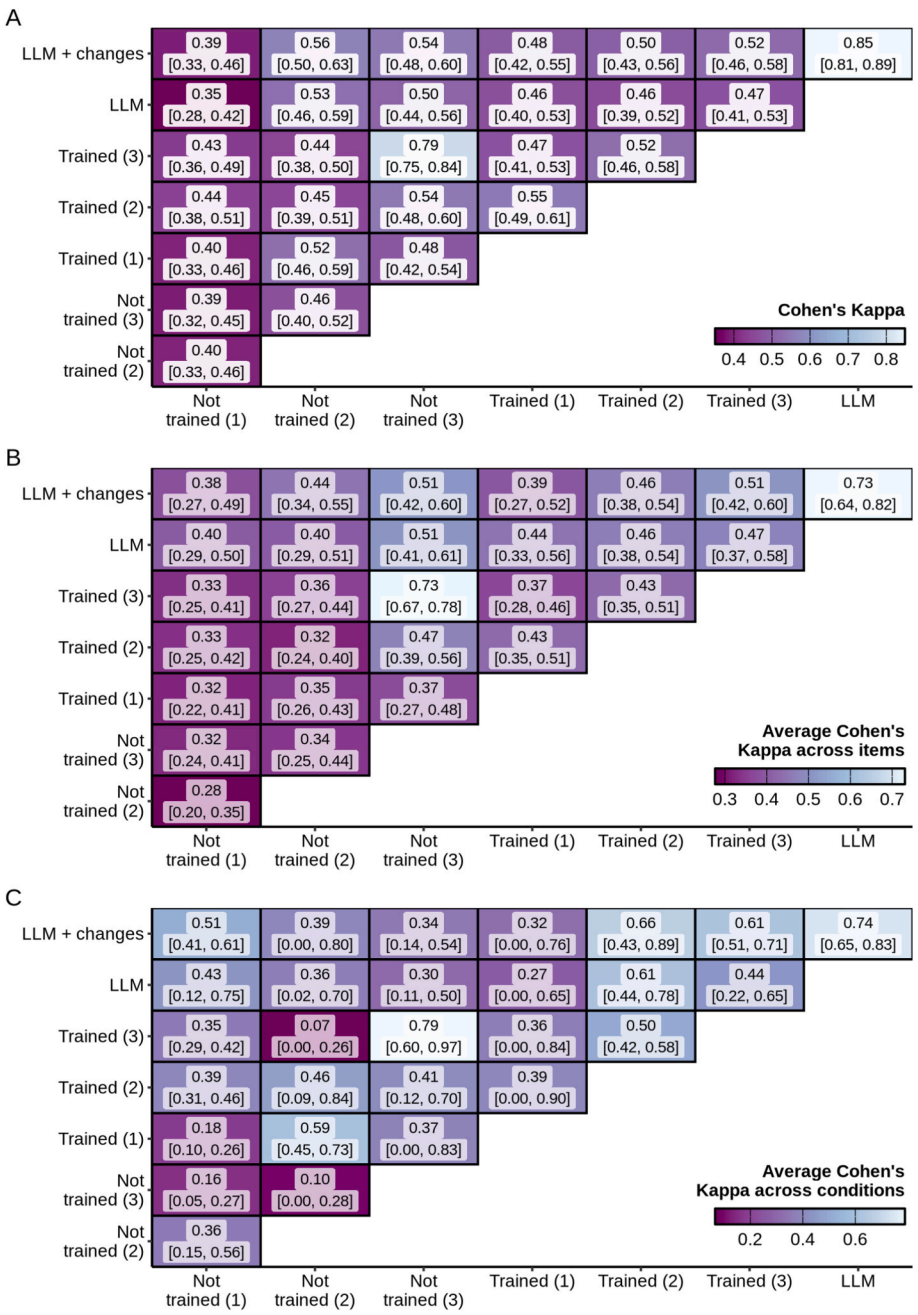

Fig. 1. Inter-rater agreement as measured by Cohen's Kappa between all human rater groups, the large language model (LLM) and the large language model upon iterative alterations seeking to improve the LLM performance. Measured for Akinci D'Antonoli et al. [18]. A. Cohen's Kappa calculated using all items and conditions. B. Average item Cohen's Kappa. C. Average condition Cohen's Kappa. For A-C, colours represent the Cohen's Kappa/average Cohen's Kappa. The text inside each cell represents Cohen's Kappa/average Cohen's Kappa and its 95% confidence interval in brackets.

(Supplementary Methods). The reason behind this assessment lies in the fact that open models tend to underperform commercial models [26].

2.3. Statistical analysis

To compare different raters (human groups, human or LLM-based) we make use of Cohen's Kappa (κ) as a measure of inter-rater agreement. We make comparisons considering three strata: i) considering all items and conditions, ii) considering the average Cohen's Kappa across items and iii) considering the average Cohen's Kappa across conditions. These estimates are compared with one another using their confidence intervals, considering that statistically significant differences exist whenever two 95 % confidence intervals are non-overlapping. To calculate 95 % confidence intervals for the average Cohen's Kappa calculated across items/conditions, we consider its standard error and use it to calculate the upper and lower bounds of the estimate (approximately 1.96 * SE, where SE is the standard error). Human-LLM and human–human Kappa estimates are compared using Wilcoxon rank sum tests. We construct a pseudo-consensus for both datasets which is given as the majority rating for each item across all raters, and use this to calculate LLM-pseudo-consensus Cohen's Kappa and METRICS score Pearson's correlation. Whenever an item had no majority (i.e. the same number of positive and negative answers), the negative answer was considered.

We fit a mixed effects linear model where the METRICS score is the dependent variable, the rater/rater groups are the fixed effects, and the

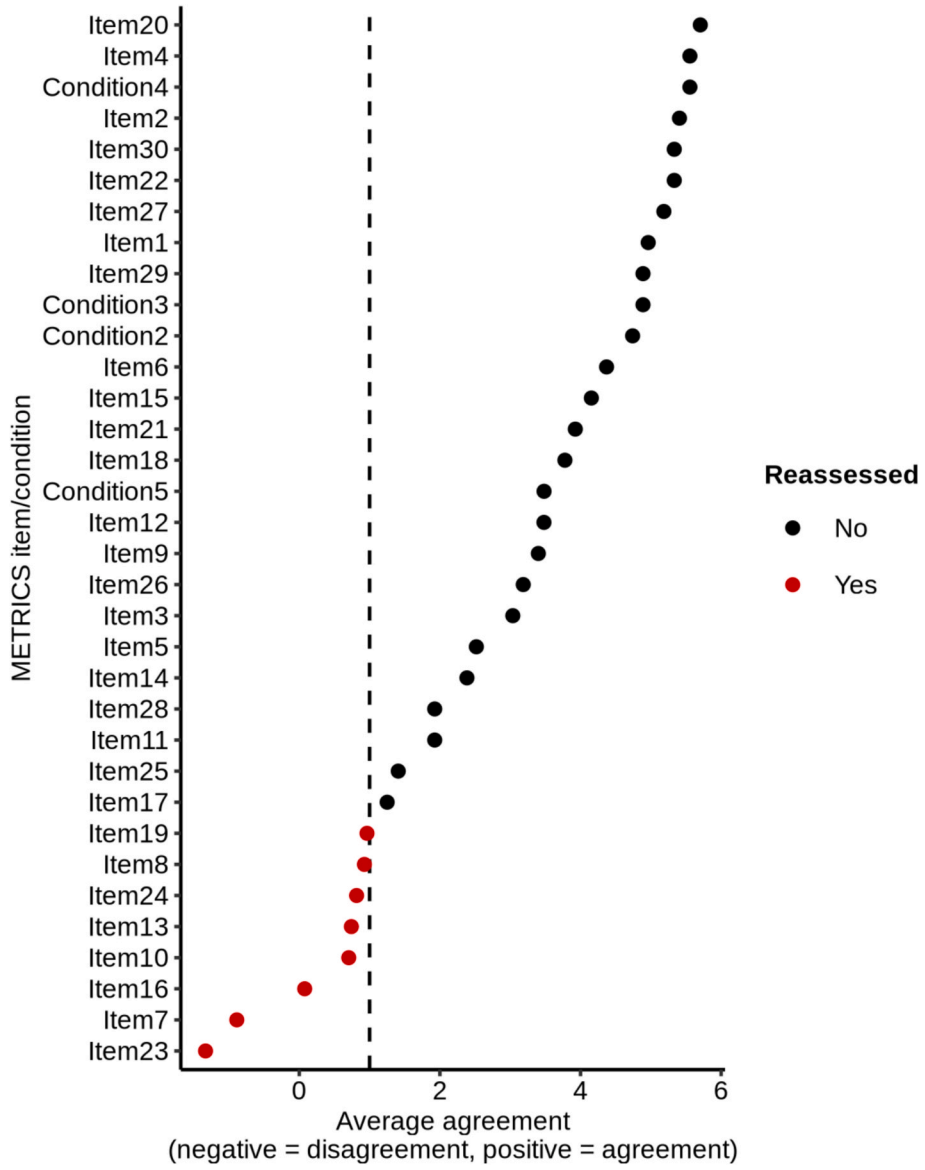

Fig. 2. Average agreement between the LLM and all rater groups. Negative values imply a higher proportion of disagreements, whereas higher values imply a higher proportion of agreements.

scientific article is the random effect. We then perform pairwise comparisons between the expected marginal means of each rater/rater groups and adjust for multiple comparisons using Tukey's method (analogous to Tukey's Honestly Significant Differences in post-hoc analyses of variance). The analysis for both rater sets (ADA2025 and K2025) are performed independently.

To compare open with commercially available models, we calculate the overall Cohen's kappa between all open models and each rater/rater group. We then calculate an average ranking (first for most similar to human rater, last for least similar) for each model across all raters/rater groups. All open model experiments were executed in a single computer with two NVIDIA RTX A6000 GPU cards.

2.4. Code availability

We make all of the code and prompts used for this manuscript in github.com/josegcpa/auto-metrics. In summary, we used Python version 3.13 and Google Gen AI Python SDK (google-genai) version 1.8.0 to perform LLM queries to Gemini Flash 2.0 and Pydantic version 2.11.1 to specify the structured output format. To assist researchers in using Auto-METRICS, we provide an easily accessible website at https://auto-metrics.netlify.app/ where users can use Auto-METRICS with their Google AI Studio API keys.

3. Results

3.1. Similarities between large language models and human raters

Using the inter-rater agreements collected from ADA2025 [18], we calculated human–human, human-LLM and LLM-LLM inter-rater agreements. As visible in Fig. 1 and similarly to the original publication [18], there is limited inter-rater agreement between human rater groups. There is a single exception — the value for Cohen's Kappa measured between expert raters with and without training is considerably higher than what is observed for other groups ($\kappa = 0.79$ for the entire METRICS questionnaire, 95 % CI = [0.75, 0.83]; Fig. 1A). When analysing the average inter-rater agreement for items and conditions, this is also the case: the experienced rater groups rated scientific publications similarly (Fig. 1B,C). In general and excluding comparisons between experienced rater groups, the observed scores align with those

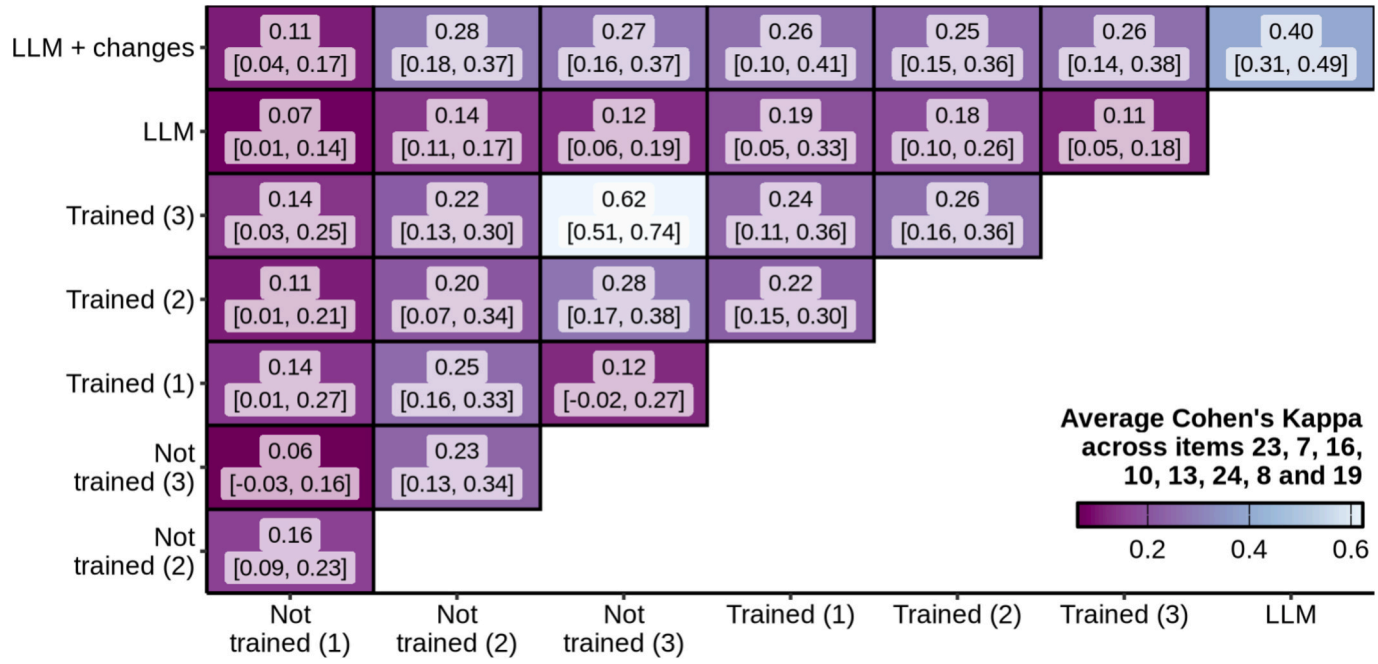

Fig. 3. Inter-rater agreement as measured by the average Cohen's Kappa for items 23, 7, 16, 10 and 19 between all human rater groups, the large language model (LLM) and the large language model upon iterative alterations seeking to improve the LLM performance. Measured for Akinci D'Antonoli *et al.* (Akinci D'Antonoli *et al.* 2025).

reported in the original publication, i.e. moderate agreement. Agreement is higher within individuals with training, highlighting how training improves the reproducibility of METRICS.

When considering the scores between raters and the LLM, this picture is similar — routinely, the LLM-rater agreement is almost exclusively moderate (0.4–0.6), with the exception of the comparison between the LLM and the rater group with the lowest experience and no training ($\kappa = 0.35$, CI 95 %=[0.28, 0.42]). While lower in general, average Cohen's Kappa calculated across all items between the LLM and rater groups are similar to those between other rater groups. Finally, LLM-rater group comparisons for conditions tend to show lower inter-rater agreements ($\kappa = 0.27$ between the LLM and low-expertise raters with training, CI 95 %=[0.00, 0.65]). These are, however, not the lowest observed inter-rater agreements for METRICS conditions ($\kappa = 0.08$ between mid-expertise raters without training and high expertise raters with training, CI 95 %=[0.00, 0.27], $\kappa = 0.10$ between mid-expertise raters without training and high expertise raters without training, CI 95 %=[0.00, 0.28]).

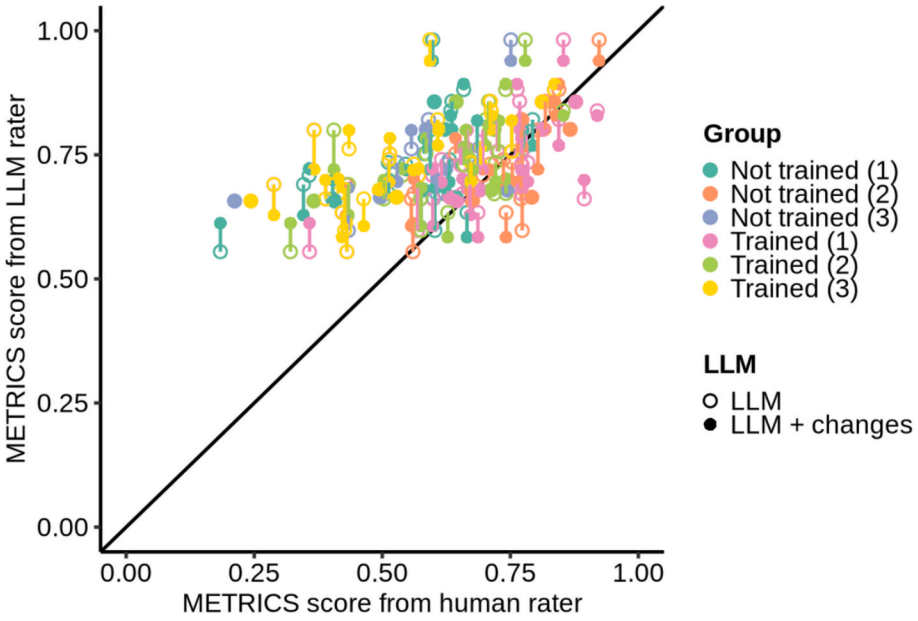

Fig. 4. Comparison of METRICS score between all human rater groups, the large language model (LLM) and LLM outputs for Akinci D'Antonoli et al. [18]. Colours represent different human rater groups, and shapes represent whether the used LLM was with or without changes (empty circle and full circle, respectively). The black diagonal line represents the identity.

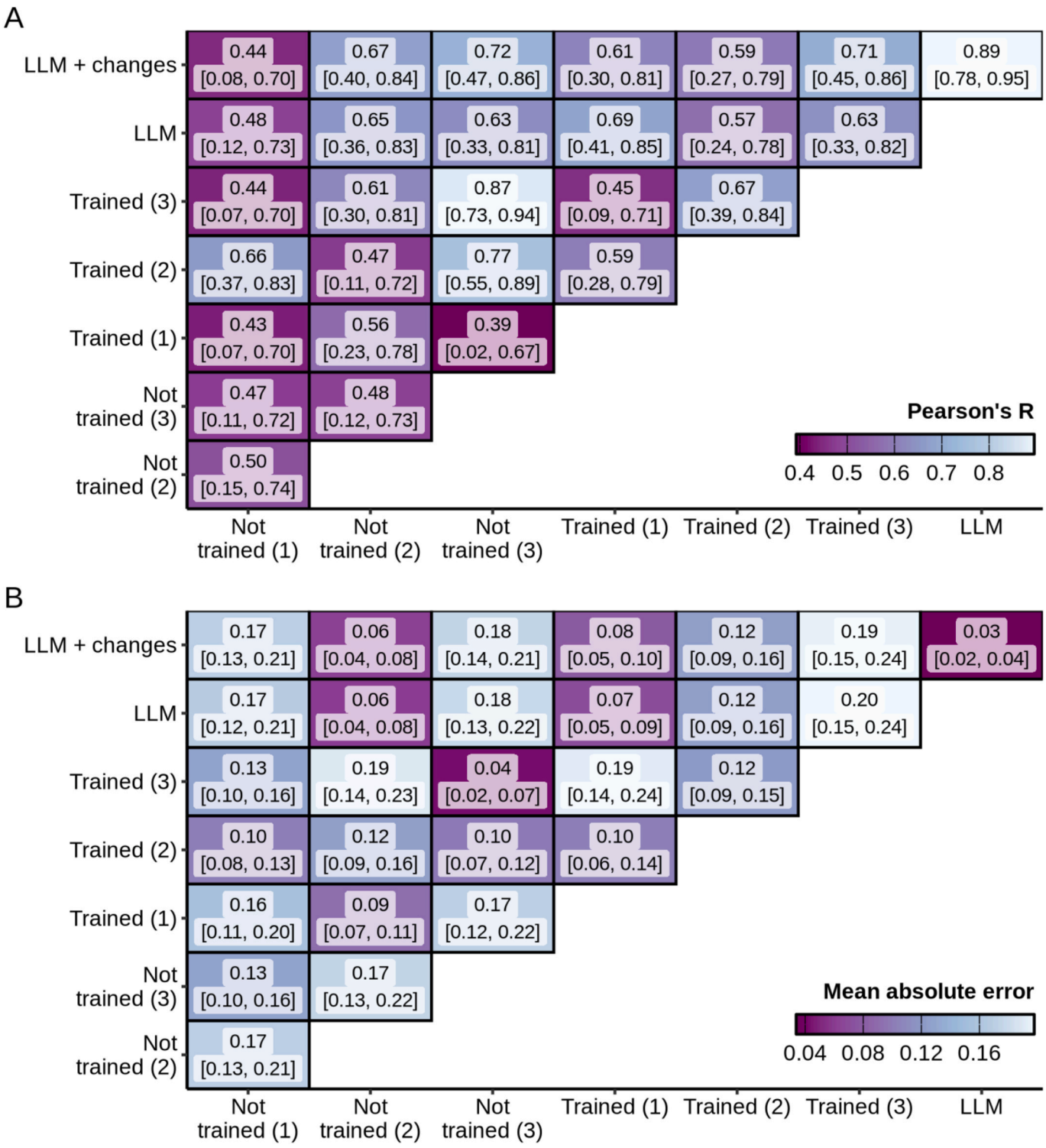

Fig. 5. METRICS score Pearons's correlation (R) (A) and mean absolute error (B) between all human rater groups, the large language model (LLM) and the large language model upon iterative alterations seeking to improve the LLM performance. Measured for Akinci D'Antonoli et al. [18]. The text inside each cell represents each metric and its 95% confidence interval in brackets.

3.2. Analysing recurring failures

To understand and improve cases where human-LLM inter-rater agreement was low, we conducted a systematic assessment of items where the human-LLM agreement was below 1 (Items 23, 7, 16, 10, 13, 24, 8 and 19; Fig. 2) and inspected, at least, the five scientific articles where the average disagreement for each item was lowest. The Auto-METRICS output includes, for each item/condition, a "reason" which highlights the logic behind the output of an LLM. In Supplementary Table 3 we present the analysed papers for these items, and in the Supplementary Results we present how disagreements between LLMs and human raters were analysed and provide the item-specific clarifications applied to our initial prompt.

With these prompt clarifications, we achieve consistent

Table 1
Pairwise differences between estimated marginal means for rater groups in a mixed effects linear model where the METRICS score was the dependent variable, the rater group/LLM was the fixed effect and the article was the random effect. This table refers to ADA2025. Stars (*) indicate statistically significant comparisons.

Comparison group	Comparison	Difference (standard error)	Adjusted p-value
Human-Human	Not trained (1) vs. Not trained (2)	−0.156 (0.023)	<0.0001*
	Not trained (1) vs. Not trained (3)	0.009 (0.023)	0.9999
	Not trained (1) vs. Trained (1)	−0.143 (0.023)	<0.0001*
	Not trained (1) vs. Trained (2)	−0.054 (0.023)	0.2486
	Not trained (1) vs. Trained (3)	0.033 (0.023)	0.8312
	Not trained (2) vs. Not trained (3)	0.165 (0.023)	<0.0001*
	Not trained (2) vs. Trained (1)	0.013 (0.023)	0.9989
	Not trained (2) vs. Trained (2)	0.102 (0.023)	0.0001*
	Not trained (2) vs. Trained (3)	0.189 (0.023)	<0.0001*
	Not trained (3) vs. Trained (1)	−0.151 (0.023)	<0.0001*
	Not trained (3) vs. Trained (2)	−0.062 (0.023)	0.1030
	Not trained (3) vs. Trained (3)	0.024 (0.023)	0.9627
	Trained (1) vs. Trained (2)	0.089 (0.023)	0.0020*
	Trained (1) vs. Trained (3)	0.175 (0.023)	<0.0001*
	Trained (2) vs. Trained (3)	0.086 (0.023)	0.0031*
Human-LLM	Not trained (1) vs. (LLM + changes)	−0.162 (0.023)	<0.0001*
	Not trained (1) vs. LLM	−0.165 (0.023)	<0.0001*
	Not trained (2) vs. (LLM + changes)	−0.006 (0.023)	1.0000
	Not trained (2) vs. LLM	−0.009 (0.023)	0.9999
	Not trained (3) vs. (LLM + changes)	−0.170 (0.023)	<0.0001*
	Not trained (3) vs. LLM	−0.173 (0.023)	<0.0001*
	Trained (1) vs. (LLM + changes)	−0.019 (0.023)	0.9903
	Trained (1) vs. LLM	−0.022 (0.023)	0.9765
	Trained (2) vs. (LLM + changes)	−0.108 (0.023)	<0.0001*
	Trained (2) vs. LLM	−0.111 (0.023)	<0.0001*
	Trained (3) vs. (LLM + changes)	−0.194 (0.023)	<0.0001*
	Trained (3) vs. LLM	−0.198 (0.023)	<0.0001*
LLM-LLM	LLM vs. (LLM + changes)	0.003 (0.023)	1.0000

improvements across the overall, item average and condition average Cohen's Kappa (Fig. 1). However, these improvements, albeit recurrent, are not statistically significant (i.e. there is significant confidence interval overlap). Calculating the average Cohen's Kappa for the eight items targeted by our clarifications further confirms this (Fig. 3).

3.3. Human-LLM rater METRICS correlations

We calculated for all human and LLM raters the METRICS score, as well as the relative and absolute error between raters. In Fig. 4 and Fig. 5, we show that METRICS scores are comparable between one another and highlight greater similarities between high-expertise raters with and without training. We perform pairwise comparisons between the METRICS score of each rater group while controlling for the scientific article (Table 1). To understand whether human raters disagree between themselves more frequently than with an LLM, we compare the proportion of statistically significant human–human and human-LLM comparisons. We show that 66 % (8/12) of Human-LLM and 60 % (9/15) of human–human comparisons are statistically significant, indicating no evidence for differences between proportions ($p = 1.0$ for a two-sample two-way test for equality of proportions). Of note, LLM ratings tend to be higher than human ratings.

3.4. Generalisation of conclusions to other surveys

To further validate our results, we reproduced our analysis on K2025 [17] (Fig. 6). We observe good Human-LLM inter-rater agreement. However — while there is a positive trend — no statistically significant changes were found between the LLM and the LLM + changes. Similarly to what was performed for ADA2025, we calculated METRICS scores for all human and LLM raters. Pairwise comparisons between rater groups for the expected METRICS scores while controlling for scientific article (Table 2) showed that 66 % (2/3) of human–human comparisons and 100 % of human-LLM comparisons (6/6) were statistically significant, once again indicating no evidence for differences ($p = 0.71$ for a two-sample two-way test for equality of proportions).

As a final confirmation of our results, we compare average kappa and METRICS correlation values, observing no evidence for differences between human-LLM and human–human comparisons (Table 3). Additionally, we calculate the Cohen's Kappa between the LLM and the LLM + changes and a pseudo-consensus (the majority rating across all items in each publication), showing moderate-to-substantial inter-rater agreement and relatively high Pearson correlations (Table 4).

3.5. Gap between open models and commercial model

We test 18 local LLMs and 6 local LRMs, introducing relevant changes to prompting and output format as relevant (Methods, Supplementary Methods). Mistral-7b did not produce JSON-compliant output and its results were discarded, while Gemma3-4b produced 11 outputs which were not JSON-compliant but which could be coherced to JSON.

Most open models underperform when compared with Gemini Flash 2.0 (Fig. 7A,B). However, Phi4-Reasoning ranks as having the highest Cohen's kappa across multiple rater/rater groups (median rank = 1, compared with median ranking = 3 for Gemini Flash 2). Small model size appears to be associated with worse performance within model providers. As expected, Gemini Flash 2.0 is the fastest model as computations are performed in the Google Cloud Platform (average time = 14.6 s, range = [11.5, 22.3]; Fig. 7C). With local models time increases with model size and with model reasoning. Phi4-Reasoning, for example, is the 5th slowest of 25 models with an average execution time of 132.5 s (range = [98.4, 163.8]). Finally, prompt improvements did not affect local models ($p = 0.29$ for a paired Wilcoxon test comparing LLM with LLM + changes prompts using local models Wilcoxon, **Supplementary** Fig. 1).

4. Discussion

Here, we show that Auto-METRICS, an automatic assessment of METRICS using LLMs, is a feasible complement to those provided by radiologists when assessing the scientific quality of radiomics publications. We show how this assessment shows levels of inter-rater agreement which are similar to those observed between radiologists, with open models demonstrating comparable performance.

Measuring the quality of scientific production is not trivial. While impact factors or h-indices [27] are straightforward metrics, they can only be applied retrospectively and do not measure methodological quality. METRICS focuses on evaluating the methodological quality of individual publications, aligning its analysis with those from other fields such as epidemiology and clinical trials, where concerns such as sample size, confounders, methodological transparency, and reproducibility are apparent [28,29]. Alongside RQS [13] and other frameworks emphasizing radiomic feature repeatability [30], it serves as a valuable tool for

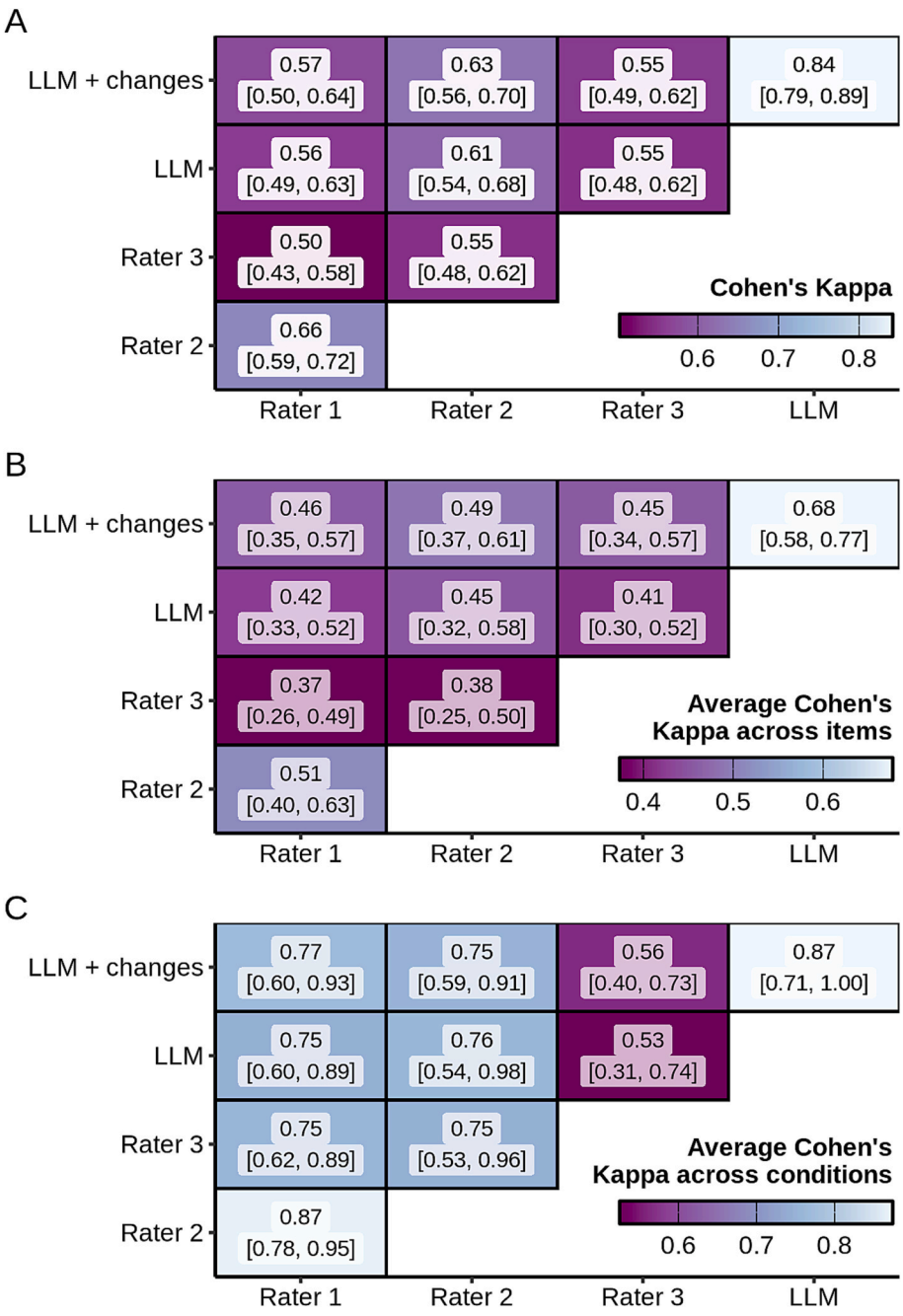

Fig. 6. Inter-rater agreement as measured by Cohen's Kappa between all human rater groups, the large language model (LLM) and the large language model upon iterative alterations seeking to improve the LLM performance. Measured for Kocak et al. [17]. A. Cohen's Kappa calculated using all items and conditions. B. Average item Cohen's Kappa. C. Average condition Cohen's Kappa. For A-C, colours represent the Cohen's Kappa/average Cohen's Kappa. The text inside each cell represents Cohen's Kappa/average Cohen's Kappa and its 95% confidence interval in brackets.

radiomics research quality assessment. Such instruments may aid in identifying robust findings suitable for clinical translation, with automation potentially enabling more systematic and efficient discovery of clinically actionable research. Indeed, Auto-METRICS can reasonably allow for such scaling: the complete set of queries performed for this study (46 scientific articles with two prompts) amounted to no more than 0.20$, with each query taking only a few seconds.

Indeed, the field of automatic scientific article reviews using LLMs has become increasingly popular. Some studies demonstrate promising results from LLM-based peer review [22,31]. Others employ tool-based methods to retrieve and analyze related literature, assessing consistency with prior work [32–35]. Others use LLM-LLM or human-LLM interaction to improve review quality [24,36]. Mese and Kocak showed that commercially available chatbots can reliably estimate RQS and METRICS [37,38]. Auto-METRICS adds to this body of research, showing that standardized automated assessments of methodological quality are feasible in radiomics research with good inter-rater agreement with experts.

Mese and Kocak [38] in particular evaluate the inter-rater agreement for METRICS between a rater group (through consensus) and two commercially available LLMs. Our work performs a similar analysis in terms of inter-rater reliability, expanding it to inter-rater reliability estimates between different expert raters/rater groups and LLM raters, while also using structured outputs, guaranteeing that LLM outputs can be easily parsed and displayed. They show slight-to-moderate human-LLM agreement (ChatGPT vs. human Fleiss's kappa = 0.407;

Table 2
Pairwise differences between estimated marginal means for rater groups in a mixed effects linear model where the METRICS score was the dependent variable, the rater group/LLM was the fixed effect and the article was the random effect. This table refers to K2025. Stars (*) indicate statistically significant comparisons.

Comparison group	Comparison	Difference (standard error)	Adjusted p-value
Human-Human	Rater 1 vs. Rater 2	−0.026 (0.023)	0.8020
Human-Human	Rater 1 vs. Rater 3	0.069 (0.023)	0.0229
Human-Human	Rater 2 vs. Rater 3	0.095 (0.023)	0.0004*
Human-LLM	Rater 1 vs. (LLM + changes)	−0.089 (0.023)	0.0010*
Human-LLM	Rater 1 vs. LLM	−0.098 (0.023)	0.0002*
Human-LLM	Rater 2 vs. (LLM + changes)	−0.064 (0.023)	0.0444*
Human-LLM	Rater 2 vs. LLM	−0.072 (0.023)	0.0148*
Human-LLM	Rater 3 vs. (LLM + changes)	−0.158 (0.023)	<0.0001*
Human-LLM	Rater 3 vs. LLM	−0.167 (0.023)	<0.0001*
LLM-LLM	LLM vs. (LLM + changes)	0.008 (0.023)	0.9962

NotebookLM vs. human Fleiss's kappa = 0.173), similarly to our results showing fair-to-moderate agreement between humans and LLM raters. They also show that ChatGPT-4o takes between 2.9 and 3.5 min to evaluate an article using RQS [37], whereas ChatGPT-4o takes a median time of 3 min and NotebookLM takes a median time of 4 min for METRICS [38]. We show that Gemini Flash 2.0 and the best open model (Phi4-Reasoning) running in an NVIDIA Quadro RTX A6000 (similar to a consumer-grade NVIDIA RTX 4090) both offer faster performance (average time = 14.6 s, range = [11.5, 22.3] and average time = 132.5 s, range = [98.4, 163.8], respectively). While impossible to fully explain without additional details, our faster inference speeds likely stem from the fact that the number of output tokens is relatively small due to our adherence to structured outputs and reduction of necessary reasoning (for Phi4-Reasoning). This is in opposition to ChatGPT-4o, which is trained to be relatively verbose and comprehensive in its explanations, and NotebookLM, which was designed to be relatively comprehensive in its ellaborations and research/reasoning capabilities.

With Auto-METRICS, human-LLM and human–human METRICS ratings are comparable. However, there are some systematic disagreements. Two such items – 16 and 19 – were considered hard to interpret and understand by radiologists [18], showcasing how human and LLM-based problems with interpretation can be aligned. This can potentially be used to programatically finetune similar assessments. Furthermore, while Auto-METRICS tends to overestimate METRICS scores when compared with human raters, these are still closely aligned.

Open LLMs enable local deployment without externalizing data, facilitate version control, preserve sensitive information, and ensure consistent outputs. These considerations can be critical when considering high-risk or privacy-preserving applications. Our comparative analysis of open models shows that there are viable alternatives to commercial LLMs: Phi4-Reasoning offered comparable performance to Gemini Flash 2.0, a cloud-based commercial solution. Interestingly, while Phi4-Reasoning vastly outperforms Phi4, improvements stemming from reasoning capabilities are inconsistent as Qwen-32B and QwQ perform similarly.

4.1. Limitations

While the potential of Auto-METRICS is demonstrated, we note some limitations. Selection biases are likely to affect this analysis as the data stems from two publications with overlapping authors. Addressing these issues would involve expanding the number of analysed publications and selecting publications across different impact factors as perfomed by Mese and Kocak [38]. Our analysis can be further improved by testing additional prompting strategies [39] or output formats. Furthermore, the attribution of reasons to each METRICS evaluation by LLMs is useful but does not correspond to a direct attribution in the article – future work should focus on extracting specific parts of the article to justify ratings. The "black box" nature of LLMs (particularly proprietary models) makes the inspection of their outputs considerably more complicated.

While Auto-METRICS achieves inter-rater agreements similar to those observed between human raters, it focuses exclusively on text; with multimodal LLMs, other data types such as figures and tables could more easily be included. In addition, direct parsing of web pages or PDF files (as opposed to Auto-METRICS which uses only text) could increase the usability of this tool and make it more appropriate for academic use. While METRICS has been widely adopted by the radiology community, it represents a quantitative assessment, obfuscating finer methodological aspects. Auto-METRICS allowed us to inspect why each rating was produced by appending a reason to each individual item or condition rating. This was helpful in our failure analysis of Auto-METRICS, but the lack of an analog for human ratings reduced our ability to better understand human-LLM disagreements. A future version of METRICS could include an auxilliary field encompassing the reason for different ratings.

As noted earlier, Auto-METRICS can be deployed at scale during review processes provided author consent, or to determine potentially interesting research directions for translational research. However, further performance assessments using larger sample sizes are warranted. Automated assessments can empower researchers which would otherwise not have the expertise to adequately assess their research, but due to potential errors they can also provide excessive and unwarranted confidence in poor methodological practices. Reviewers can be similarly affected: while Auto-METRICS may enable reviewers to accelerate the review process, those who are not confident in their ability to judge a publication or are under time contraints could overly rely on such systems [40]. Finally, the utility of Auto-METRICS must be confirmed with studies comparing how users rate radiomics publications with and without access to Auto-METRICS. Further usability tests are required to attest to the utility of Auto-METRICS (and other similar approaches) in

Table 4
Comparison of pseudo-consensus (calculated as the majority vote for each item in each paper) with LLM and LLM + changes for both datasets.

	Akinci D'Antonoli 2025		Kocak 2025	
	LLM	LLM + changes	LLM	LLM + changes
Cohen's Kappa (95% CI)	0.57 ([0.51, 0.62])	0.62 ([0.57, 0.67])	0.66 ([0.60, 0.72])	0.68 ([0.63, 0.74])
Pearson's correlation (95% CI)	0.65 ([0.36, 0.83])	0.68 ([0.40, 0.84])	0.83 ([0.61, 0.93])	0.70 ([0.36, 0.88])

Table 3
Comparison of average estimates for human-LLM and human-human comparisons of Cohen's Kappa and Pearson's correlation across both studies. p-values were calculated for a two-way Wilcoxon rank sum test.

	Akinci D'Antonoli 2025			Kocak 2025		
	Human-LLM	Human-Human	p-value	Human-LLM	Human-Human	p-value
Cohen's Kappa	0.48	0.49	0.41	0.58	0.57	0.28
Pearson's correlation	0.62	0.56	0.11	0.68	0.51	0.55

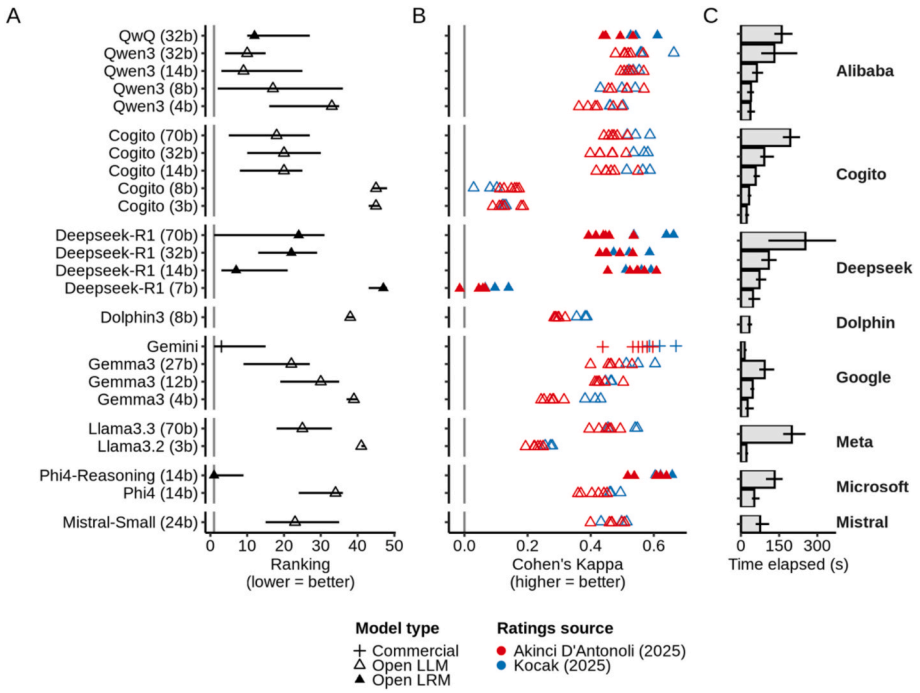

Fig. 7. Comparison of Cohen's Kappa with different raters/rater groups and elapsed time for open large language models (LLMs), large reasoning models (LRMs) and Gemini (commercial alternative). **A –** Median ranking (triangles or crosses) and minimum/maximum rankings (horizontal line) across all rater/rater groups. **B** – Human-LLM Cohen's kappa for each rater/rater group. **C** – Average (bars) and minimum/maximum (horizontal lines) elapsed time in seconds for all models. For **A** and **B**, shapes relate to whether the model was commercial, an open LLM or an open LRM, while colours (red and blue) refer to the paper group. The size of all open models is specified between parentheses after the name of each model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

peer-review or editorial processes, particularly to understand what sort of biases these can introduce. AI assistance can easily lead to "automation bias", a phenomenon where AI users tend to rely excessively on AI systems. The opposite of automation bias – "algorithm aversion bias", where users tend to uncritically reject AI outputs – can also happen across multiple contexts [41]. Automation bias can be mitigated through training and emphasising user accountability [42], and algorithm aversion through technological readiness [43]. However, how this can impact the editorial and peer-review process should be further studied.

5. Conclusion

By achieving inter-rater agreement comparable to human raters and competitive performance across both commercial and open LLMs, Auto-METRICS represents a promising step toward scalable, structured, and transparent assessments of methodological quality in radiomics research. Limitations remain, particularly regarding interpretability, potential biases, and reliance on textual input. However, future work such as continued validation, careful implementation, and attention to usability and human-AI dynamics can significantly migitate these issues.

CRediT authorship contribution statement

José Guilherme de Almeida: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nickolas Papanikolaou:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2025.112358.

Data availability

Data will be made available on request.

References

[1] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns (n y). 4 (2023) 100804.
[2] AAAI 2025 Presidential Panel on the Future of AI Research. In: AAAI [Internet]. 28 Feb 2025 [cited 7 Apr 2025]. Available: https://aaai.org/about-aaai/presidential-panel-on-the-future-of-ai-research/.
[3] F. Maleki, K. Ovens, R. Gupta, C. Reinhold, A. Spatz, R. Forghani, Generalizability of Machine Learning models: Quantitative evaluation of three methodological pitfalls, Radiol. Artif. Intell. 5 (2023) e220028.
[4] B. Kocak, E. Bulut, O.N. Bayrak, A.A. Okumus, O. Altun, Z. Borekci Arvas, et al., NEgatiVE results in Radiomics research (NEVER): a meta-research study of publication bias in leading radiology journals, Eur. J. Radiol. 163 (2023) 110830.
[5] L. Lu, F.S. Ahmed, O. Akin, L. Luk, X. Guo, H. Yang, et al., Uncontrolled confounders may lead to false or overvalued radiomics signature: a proof of concept using survival analysis in a multicenter cohort of kidney cancer, Front. Oncol. 11 (2021) 638185.
[6] R. Cannella, J. Santinha, A. Bèaufrere, M. Ronot, R. Sartoris, F. Cauchy, et al., Performances and variability of CT radiomics for the prediction of microvascular invasion and survival in patients with HCC: a matter of chance or standardisation? Eur. Radiol. 33 (2023) 7618–7628.
[7] A. Demircioğlu, Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics, Insights Imaging. 12 (2021) 172.
[8] R. Berenguer, M. del R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M. V. Villas, F. Mansilla Legorburo, et al., Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters, Radiology 288 (2018) 407–415.

[9] R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: Images are more than pictures, they are data, Radiology 278 (2016) 563–577.
[10] D. Pinto Dos Santos, M. Dietzel, B. Baessler, A decade of radiomics research: are images really data or just patterns in the noise? Eur. Radiol. 31 (2021) 1–4.
[11] N. Horvat, N. Papanikolaou, D.-M. Koh, Radiomics beyond the hype: a critical evaluation toward oncologic clinical use, Radiol. Artif. Intell. 6 (2024) e230437.
[12] C.S. Moskowitz, M.L. Welch, M.A. Jacobs, B.F. Kurland, A.L. Simpson, Radiomic analysis: Study design, statistical analysis, and other bias mitigation strategies, Radiology 304 (2022) 265–273.
[13] J.E. Park, H.S. Kim, D. Kim, S.Y. Park, J.Y. Kim, S.J. Cho, et al., A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features, BMC Cancer 20 (2020) 29.
[14] B. Kocak, T. Akinci D'Antonoli, N. Mercaldo, A. Alberich-Bayarri, B. Baessler, I. Ambrosini, et al., METhodological RadiomICs score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII, Insights Imaging. 15 (2024) 8.
[15] B. Kocak, B. Baessler, S. Bakas, R. Cuocolo, A. Fedorov, L. Maier-Hein, et al., CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII, Insights Imaging. 14 (2023) 75.
[16] B. Koçak, T. Akinci D'Antonoli, R. Cuocolo, Exploring radiomics research quality scoring tools: a comparative analysis of METRICS and RQS, Diagn. Interv. Radiol. 30 (2024) 366–369.
[17] B. Kocak, I. Mese, K.E. Ates, Radiomics for differentiating radiation-induced brain injury from recurrence in gliomas: systematic review, meta-analysis, and methodological quality evaluation using METRICS and RQS, Eur. Radiol. 1–16 (2025).
[18] T. Akinci D'Antonoli, A.U. Cavallo, B. Kocak, A. Borgheresi, A. Ponsiglione, A. Stanzione, et al., Reproducibility of methodological radiomics score (METRICS): an intra- and inter-rater reliability study endorsed by EuSoMII, Eur. Radiol. 1–13 (2025).
[19] Y. Zheng, H.Y. Koh, J. Ju, A.T.N. Nguyen, L.T. May, G.I. Webb, et al., Large language models for scientific discovery in molecular property prediction, Nat. Mach. Intell. 7 (2025) 437–447.
[20] X. Luo, A. Rechardt, G. Sun, K.K. Nejad, F. Yáñez, B. Yilmaz, et al., Large language models surpass human experts in predicting neuroscience results, Nat. Hum. Behav. 9 (2025) 305–315.
[21] J. Evans, J. D'Souza, S. Auer, Large Language Models as evaluators for scientific synthesis. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2407.02977.
[22] W. Liang, Y. Zhang, H. Cao, B. Wang, D.Y. Ding, X. Yang, et al., Can large language models provide useful feedback on research papers? A large-scale empirical analysis, NEJM AI. 1 (2024), https://doi.org/10.1056/aioa2400196.
[23] C. Cao, J. Sang, R. Arora, R. Kloosterman, M. Cecere, J. Gorla, et al., Prompting is all you need: LLMs for systematic review screening, Available: Health Informatics. Medrxiv (2024) https://www.medrxiv.org/content/10.1101/2024.06.01.24308323v1.
[24] N. Thakkar, M. Yuksekgonul, J. Silberg, A. Garg, N. Peng, F. Sha, et al., Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025. arXiv [cs.AI]. 2025. Available: http://arxiv.org/abs/2504.09737.
[25] Generate structured output with the Gemini API. In: Google AI for Developers [Internet]. [cited 7 Apr 2025]. Available: https://ai.google.dev/gemini-api/docs/structured-output?lang=python.
[26] Cottier B. How Far Behind Are Open Models? In: Epoch AI [Internet]. 4 Nov 2024 [cited 24 Jun 2025]. Available: https://epoch.ai/blog/open-models-report.
[27] M. Kumar, Evaluating scientists: Citations, impact factor, h-index, online page hits and what else? IETE Tech. Rev. 26 (2009) 165.
[28] J.P. Vandenbroucke, E. von Elm, D.G. Altman, P.C. Gøtzsche, C.D. Mulrow, S. J. Pocock, et al., Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration, PLoS Med. 4 (2007) e297.
[29] X. Liu, S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension, Nat. Med. 26 (2020) 1364–1374.
[30] E. Pfaehler, I. Zhovannik, L. Wei, R. Boellaard, A. Dekker, R. Monshouwer, et al., A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features, Phys. Imaging Radiat. Oncol. 20 (2021) 69–75.
[31] Z. Robertson, GPT4 is slightly helpful for peer-review assistance: A pilot study. arXiv [cs.HC]. 2023. Available: http://arxiv.org/abs/2307.05492.
[32] Q. Wang, Q. Zeng, L. Huang, K. Knight, H. Ji, R.NF. ReviewRobot, Explainable paper review generation based on knowledge synthesis, arXiv [cs.CL]. (2020). http://arxiv.org/abs/2010.06119.
[33] E. Chamoun, M. Schlichktrull, A. Vlachos, Automated focused feedback generation for scientific writing assistance, arXiv [cs.CL]. (2024). http://arxiv.org/abs/2405.20477.
[34] M. D'Arcy, T. Hope, L. Birnbaum, D. Downey, MARG: Multi-Agent Review Generation for Scientific Papers. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2401.04259.
[35] J. Yu, Z. Ding, J. Tan, K. Luo, Z. Weng, C. Gong, et al., Automated peer reviewing in paper SEA: Standardization, evaluation, and Analysis. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2407.12857.
[36] Z. Gao, K. Brantley, T. Joachims, Reviewer2: Optimizing review generation through prompt generation. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2402.10886.
[37] I. Mese, B. Kocak, ChatGPT as an effective tool for quality evaluation of radiomics research, Eur. Radiol. 35 (2025) 2030–2042.
[38] I. Mese, B. Kocak, Large language models in methodological quality evaluation of radiomics research based on METRICS: ChatGPT vs NotebookLM vs radiologist, Eur. J. Radiol. 184 (2025) 111960.
[39] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv [cs.CL]. 2022. Available: http://arxiv.org/abs/2201.11903.
[40] W. Liang, Z. Izzo, Y. Zhang, H. Lepp, H. Cao, X. Zhao, et al., Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2403.07183.
[41] I. Filiz, J.R. Judek, M. Lorenz, M. Spiwoks, The extent of algorithm aversion in decision-making situations with varying gravity, PLoS One 18 (2023) e0278751.
[42] K. Goddard, A. Roudsari, J.C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, J. Am. Med. Inform. Assoc. 19 (2012) 121–127.
[43] H. Mahmud, A.K.M.N. Islam, R.K. Mitra, What drives managers towards algorithm aversion and how to overcome it? Mitigating the impact of innovation resistance through technology readiness, Technol Forecast Soc Change. 193 (2023) 122641.