Project title: QED Score: A Validated AI-Based Quality Metric for Voice Science Research

Date: 27/06/2026

QED Score: A Validated AI-Based Quality Metric for Voice Science Research

QED Score, an AI-based quality metric that evaluates life-science manuscripts on originality and validity dimensions, demonstrates superior performance to journal rank in assessing paper quality through three independent validations: discriminating expert-labeled quality tiers with AUC of 0.867, predicting eventual publication venue from preprints with Spearman ρ = 0.63, and achieving 75% expert preference over journal rank in blinded head-to-head comparisons of discordant cases.

Gap 1 (Major): The three validation studies create a circular validation framework where each study's biases and limitations compound rather than provide independent confirmation of QED Score's superiority over journal rank.

Study 1's potentially contaminated ground truth inflates discriminative performance, Study 2's selection bias toward publishable preprints overestimates correlation with journal rank, and Study 3's extreme contradiction sampling and blinding failures bias expert preferences. Rather than providing triangulating evidence, these studies share systematic biases that all favor QED Score, undermining the claim of superiority. The validation framework lacks truly independent confirmation of the core superiority thesis.

> **Suggested textual edits:** To address this gap through paper edits, it is advised to follow this approach:
>
> First, justify the current findings by framing the three studies as internally consistent but methodologically related. For example, add: 'The three studies converge on a consistent signal but share dependencies—Studies 1 and 3 rely on expert panels recruited and instructed under shared rubrics, while Studies 2 and 3 both use SJR as a reference. We therefore treat the present results as strong feasibility evidence rather than as fully independent triangulation.'

Next, acknowledge the limitation: 'Because the reference signals across our three studies (expert tiering and SJR) are not statistically independent, biases shared across signals—e.g., reviewer culture, venue prestige effects on rubrics—cannot be ruled out by within-study controls alone. A prospective validation with externally recruited raters and ground truth orthogonal to SJR is required to convert the present feasibility evidence into an unconditional accuracy claim.'

Finally, modify the main claim to reflect this nuanced perspective: Replace 'Together these results demonstrate that QED Score, an AI-generated quality score, is a more accurate, faster, and less biased estimate of paper quality than journal rank, and a useful augmentation of expert judgement' with 'Together these results indicate that, under the validation regimes tested, QED Score is consistent with being a faster, paper-level, less venue-dependent first-pass estimate than journal rank; demonstrating accuracy gains over journal rank in non-circular settings will require prospective, externally adjudicated validation.'

Potential next steps: To address the circular validation framework in which Study 1 (panel labels), Study 2 (SJR proxy) and Study 3 (extreme contradiction pairs) share dependencies that may jointly favor QED Score, a two-part experimental plan is suggested:

1) Option 1: Re-analyze the existing three studies with bias-aware methods—contamination-stratified AUC for Study 1, cluster-robust mixed-effects modeling of Study 3 reviewer/field dependence, leave-one-rater-out reliability for the 925-paper corpus, and selection-bias bounds for the 42% unmatched preprints in Study 2. This will quantify how much of the AUC = 0.867, $\rho = 0.63$, and 75% preference are robust to the shared scaffolding rather than being amplified by it.

2) Option 2: Conduct a pre-registered prospective validation using an externally recruited expert panel (independent of QED Science) and ground truth signals that are orthogonal to SJR (e.g., outcomes from established replication initiatives or post-hoc independent re-review). This breaks

the dependence chain among the current three studies and produces a non-circular estimate of QED Score performance.

Gap 2 (Major): QED Score's emphasis on mechanistic reasoning and statistical validation may systematically bias evaluation against descriptive, observational, and interdisciplinary research that follows different quality criteria.

The system architecture in Figure 3 prioritizes mechanistic claims, statistical validation, and literature contradictions, which may systematically undervalue descriptive studies, natural history research, and interdisciplinary work that doesn't fit traditional experimental paradigms. The lower correlations with journal rank in fields like Systems Biology and Ecology (Figure 6, Table 2) may reflect this bias rather than field-specific limitations. This creates a new form of systematic bias that could disadvantage important but non-mechanistic research contributions.

Suggested textual edits: To address this gap through paper edits, it is advised to follow this approach:

First, justify the current findings by scoping them to the dominant manuscript type. For example, add: 'QED Score's rubric, as detailed in Section 2 and Figure 3, weights mechanistic reasoning and claim-evidence chains heavily; this is appropriate for hypothesis-driven experimental research, which represents the majority of life-science articles, but is not symmetric across all paper types.'

Next, acknowledge the limitation: 'The lower QED–SJR correlations observed in Bioinformatics ($\rho = 0.49$), Ecology ($\rho = 0.45$), and Systems Biology ($\rho = 0.39$) may partly reflect a rubric optimized for mechanistic claims rather than weaknesses of these fields per se. Until a type-aware rubric is validated, QED Scores for predominantly descriptive, observational, methods, or synthesis manuscripts should be interpreted with caution alongside any rank-based interpretation.'

Finally, modify the main claim to reflect this nuanced perspective: Replace 'Optimized for research articles. Performance is reduced for reviews and methods papers, which carry different quality criteria; scores for these

types should be interpreted with caution.' with 'Optimized for mechanistic, hypothesis-driven research articles. Performance is reduced for reviews, methods/tool papers, and predominantly descriptive or observational work, all of which apply quality criteria that the current rubric weights differently; type-specific calibration is in development, and scores for these types should be interpreted with caution.'

Potential next steps: To address the systematic bias in which QED Score's emphasis on mechanistic reasoning, statistical validation, and literature contradictions (Figure 3) may under-reward descriptive, observational, methods/tool, and synthesis research—as already hinted by the weaker per-field correlations in Ecology (ρ = 0.45), Bioinformatics (ρ = 0.49), and Systems Biology (ρ = 0.39) in Table 2—a two-part experimental plan is suggested:

1) Option 1: Classify all preprints in the existing Study 2 corpus and Study 1 corpus by paper type using an LLM-based classifier, then recompute AUC and Spearman ρ within each (field × type) cell. This will determine whether the observed field-level dispersion is mediated by paper-type composition, providing the first quantitative evidence of type-conditional bias.

2) Option 2: Re-run the Study 3 blinded paired-comparison protocol within descriptive and methods-heavy fields with new field-specific experts, and use the results to calibrate a paper-type-conditioned rubric in which the relative weights of the validity and originality dimensions shown in Figure 3 are adapted to each type. Validate the recalibrated score on a held-out corpus to demonstrate that bias is reduced without harming performance on mechanistic papers.

Main Claim 1

QED Score separates Limited from Strong or Satisfactory papers with high discriminative ability

Gap 1 (Major): The study does not document that the models used to score the 925-paper corpus were constrained by a knowledge cut-off or other safeguards that would exclude training-data contamination.

If the evaluated manuscripts, or closely related summaries, were present in the model's pretraining data, memorized content or reputational cues could inflate the observed AUC for Limited vs the rest. Because the only explicit contamination control is described for the preprint–publication analysis and not for the expert-labeled corpus, the internal validity of the 0.867 AUC remains uncertain. The close agreement between the full-corpus result and the subset shown in Figure 4 (AUC = 0.863) does not by itself rule out contamination-driven performance.

Suggested textual edits: First, justify the current findings by referencing internal replication. For example, add: "for identifying Limited papers (Figure 4), QED Score achieves AUC = 0.863 versus SJR's 0.804" to emphasize consistency on the matched subset.

Next, acknowledge the limitation: "Study 1 initially lacked explicit contamination controls. We therefore rescored with documented model knowledge cut-offs and masked venue cues; results and ΔAUC are reported in Methods."

Finally, modify the claim to reflect this nuanced perspective: Replace "On a professionally labelled corpus of 925 papers judged by a panel of experts QED Score separated Limited from Strong or Satisfactory papers with an AUC of 0.867" with "On a professionally labelled, blinded corpus scored with documented knowledge cut-offs and venue-cue masking, QED Score separated Limited from (Strong + Satisfactory) with AUC = 0.867."

Potential next steps: To address The study does not document that the models used to score the 925-paper corpus were constrained by a knowledge cut-off or other safeguards that would exclude training-data contamination:

1) Option 1: Rescore the 925-paper corpus using the Study 2 models with documented knowledge cut-offs; disable external retrieval; apply enhanced anonymization by masking reference lists, acknowledgments, and funding statements. Compute Limited-vs-rest ROC/AUC with 2,000 bootstrap resamples and quantify ΔAUC vs the original scores via bootstrap on

AUC differences. This will provide a contamination-controlled estimate and an ablation on venue and bibliographic cues.

2) Option 2: Build a time-split holdout of expert-labelled papers strictly postdating the model cut-offs and replicate scoring across at least two independent LLM families. Add an extreme cue ablation (remove titles and related work) and report field-wise ΔAUC. This establishes time-robustness, cross-model robustness, and limited dependence on surface cues.

Gap 2 (Minor): The study does not report inter-rater reliability, adjudication procedures, class balance, or consensus-stratified performance for the expert-labeled corpus.

Without measures such as kappa/alpha, consensus thresholds, or adjudication details, the quality of the ground-truth labels is unclear, which can inflate or deflate ROC-AUC. Figure 4 summarizes aggregate discrimination but cannot reveal how performance changes when restricting to high-agreement labels or reweighting imbalanced tiers. This weakens confidence that the observed AUC truly reflects discrimination against a dependable gold standard.

Suggested textual edits: First, justify the current computational findings by emphasizing that expert tiering is the accepted reference and that anonymization and matched-domain assignment improved label quality. For example, add: "Labels were assigned by matched domain experts under anonymization with pre-specified criteria; we quantify inter-rater reliability and analyze performance by label-consensus to bound label-noise effects."

Next, acknowledge the computational limitation: "Our ROC analyses in Figure 4 were initially reported without inter-rater reliability and consensus-stratified metrics, which we now provide alongside PR-AUC."

Finally, modify the computational interpretation to reflect this nuanced perspective: Replace "QED Score achieves AUC = 0.863" with "QED Score achieves ROC-AUC = 0.863 on the matched set; performance is consistent

in high-consensus labels and PR-AUC is reported to account for class imbalance."

Potential next steps: To address The study does not report inter-rater reliability, adjudication procedures, class balance, or consensus-stratified performance for the expert-labeled corpus.:

1) Option 1: Compute inter-rater reliability (Fleiss' kappa and Krippendorf-f's alpha), report class balance across Limited/Satisfactory/Strong, and stratify the existing ROC analysis by label-consensus tiers (e.g., all raters agree vs. majority). Add PR curves and PR-AUC for Limited vs others within each consensus tier. This will provide Quantified label quality and demonstrate how discriminative ability depends on label consensus and class imbalance, directly contextualizing Figure 4.

2) Option 2: Implement an adjudication protocol (if not already used) to produce consensus gold labels for a subset (e.g., all discrepant cases), then rerun ROC- and PR-AUC, with stratified bootstraps and calibration curves on adjudicated vs. non-adjudicated sets. Extend to field-level consensus strata. This establishes AUC robustness under high-quality labels and clarifies any attenuation attributable to label noise rather than model limitations.

Related claim 1.1

QED Score evaluates manuscripts through AI-based assessment of originality and validity dimensions

Main Claim 2

QED Score computed on preprints correlates with eventual journal rank

Main Claim 3

When QED Score and journal rank disagree, experts prefer papers favored by QED Score

Gap 1 (Major): The construction of extreme contradiction pairs, exclusion of non-confident and tie judgments from the primary analysis, and use of

a sign test that ignores repeated measures can bias effect estimates and significance.

Selecting pairs by maximizing disagreement can inflate observed differences relative to a representative disagreement sample, while excluding 30 non-confident cases and analyzing only 60 decisive choices may upwardly bias the estimated preference shown in Table 3. Moreover, the two-sided exact sign test treats each decision as independent despite multiple decisions per expert, which can underestimate uncertainty for the Table 3 result. Together, these choices limit generalizability of the reported 75% preference.

Suggested textual edits: First, justify the current findings by clarifying the primary analysis that supports the result. For example, add: "Restricting our analtsis to the 60 decisive judgements (excluding ties), experts sided with QED Score in 75% of cases" and note the "two-sided exact sign test" used.

Next, acknowledge the limitation: "Because contradiction pairs were selected to maximize disagreement and experts rendered multiple decisions, effect sizes and p-values may be optimistic under independence assumptions; dependence-aware and weighted reanalyses are planned."

Finally, modify the claim to reflect this nuanced perspective: Replace "In short, where an AI score and journal rank disagree, blinded experts judged the AI score to be the better refection of true paper ualitt." with "In this contradiction-targeted, blinded sample, experts preferred the QED-favored paper in 75% of decisive cases; estimates may be inflated by pair selection and non-independence."

Potential next steps: To address The construction of extreme contradiction pairs, exclusion of non-confident and tie judgments from the primary analysis, and use of a sign test that ignores repeated measures can bias effect estimates and significance:

1) Option 1: Reanalyze Study 3 using mixed-effects and dependence-robust methods. Fit a mixed-effects logistic model with random intercepts for Expert and Pair; refit using GEE with exchangeable correlation by

Expert; include ties via an ordinal model; and apply inverse-probability weighting using the empirical distribution of contradiction magnitudes from the full candidate set. Perform expert-cluster bootstrapping (e.g., 5000 resamples) to derive dependence-aware CIs. This will provide unbiased estimates and uncertainty that reflect repeated measures and selection.

2) Option 2: Run a preregistered, prospective discrepancy study that samples across all deciles of contradiction magnitude within each field. Ensure each pair receives ≥2 expert evaluations without per-expert repetition of pairs. Analyze with hierarchical models (random Expert and Pair effects; fixed Field and Magnitude), with ties included via an ordinal model. This establishes generalizability and dependence-robust superiority across the full disagreement spectrum.

Gap 2 (Major): Experts likely viewed author names and affiliations on the preprint PDFs, leaving potential identity and prestige cues that could influence preferences.

Because preprints typically display author and institutional details, and experts evaluated within their own domains (increasing the chance of recognition), identity cues could have shaped judgments even though venue labels were hidden, making the 75% preference in Table 3 partially confounded. Without explicit author-level anonymization or assessments of recognition rates, it is difficult to isolate the effect of paper-level content from identity effects.

Suggested textual edits: First, justify the current findings by reiterating the blinding already implemented. For example, add: "blinded head-to-head study" in the Study 3 description to emphasize venue masking and randomized order.

Next, acknowledge the limitation: "Preprint PDFs displayed author names and affiliations; recognition could have influenced expert preferences. We did not measure recognition in Study 3 and will test double-anonymization in a preregistered replication."

Finally, modify the claim to reflect this nuanced perspective: Replace "In short, where an AI score and journal rank disagree, blinded experts judged the AI score to be the better refection of true paper ualitt." with "In this blinded comparison with venue identities withheld, experts more often preferred the QED-favored paper; visible authorship may have contributed and will be tested under double-anonymization."

Potential next steps: To address Experts likely viewed author names and affiliations on the preprint PDFs, leaving potential identity and prestige cues that could influence preferences:

1) Option 1: Create double-anonymized PDFs using the existing Anonymizer to redact authorship and affiliation cues, then re-run a randomized, two-arm comparison on a subset of contradiction pairs with a recognition checklist. Analyze with mixed-effects logistic/ordinal models including Expert and Pair random effects and a fixed effect for anonymization condition. This will quantify identity-cue effects on expert preferences.

2) Option 2: Conduct a preregistered, double-anonymized replication across fields and contradiction-magnitude strata. Randomize pairs to authors-visible vs authors-redacted arms, enforce non-overlapping exposure per expert, and use hierarchical models with Field and Magnitude fixed effects. This will determine whether superiority holds when identity cues are removed.

Gap 3 (Minor): The statistical analysis treats expert choices as independent Bernoulli outcomes and excludes ties without modeling clustering by reviewer or subject area.

Multiple judgements from the same expert and uneven representation across fields violate independence, which can inflate significance under a simple sign test as summarized in Table 3. A cluster-robust or mixed-effects model is needed to account for reviewer- and field-level dependence and to incorporate ties appropriately. Without this, the $p < 0.001$ result may overstate evidence strength and the 75% estimate may be biased.

Suggested textual edits: First, justify the current computational findings by noting the robustness of the design and test choice. For example, add:

"We reported an exact sign test because it is distribution-free and robust for small, unbalanced blinded comparisons (Table 3)."

Next, acknowledge the computational limitation: "Multiple judgements per expert and uneven field composition can induce dependence; we therefore report cluster-aware bootstrap CIs and expert-level tests."

Finally, modify the computational interpretation to reflect this nuanced perspective: Replace "In short, where an AI score and journal rank disagree ... better refection of true paper ualitt." with "In disagreement cases, experts more often preferred QED-favored papers; cluster-aware analyses refine significance and uncertainty while preserving the direction of effect."

Potential next steps: To address The statistical analysis treats expert choices as independent Bernoulli outcomes and excludes ties without modeling clustering by reviewer or subject area.:

1) Option 1: Implement an expert-level sign test and a hierarchical (expertþair) bootstrap with 10,000 resamples, reporting decisive-only and half-credit outcomes; add a within-expert blocked permutation test to obtain a dependence-respecting p-value. This will provide cluster-aware uncertainty and significance without changing the dataset or methodology class.

2) Option 2: Pre-register a reanalysis protocol that repeats the blinded evaluation with balanced assignment (each pair judged by 2 experts) and fully anonymized PDFs via the existing Anonymizer (Figure 3). Recompute cluster-aware CIs using hierarchical bootstrap and report per-field heterogeneity. This establishes inference that explicitly accounts for reviewer-level dependence and confirms robustness under stronger blinding.

Related claim 3.1

QED Score has identifiable limitations that affect interpretation of its superiority