Just Ask the Model: One-Shot LLM Research Evaluation and Structured Expert Review

Valentin Klotzbücher David Reinstein Lorenzo Pacchiardi
Tianmai Michael Zhang

2026-11-10

Abstract

Peer review is strained, and AI tools generating referee-like feedback are already adopted by researchers and commercial services—yet field evidence on how reliably frontier LLMs can evaluate research remains scarce. We compare structured one-shot evaluations by GPT-5 Pro against paid expert review packages from The Unjournal, an open evaluation platform covering economics and social-science working papers, where both humans and the model rate papers on seven percentile criteria with uncertainty intervals and provide narrative critiques. Treating human evaluations as a high-quality but noisy reference signal, we find that GPT-5 Pro approaches the agreement levels observed among human evaluators themselves on several criteria, while exhibiting consistent failure modes: compressed rating scales, uneven criterion coverage, and variable identification of expert-flagged concerns. A minimal one-shot setup—a single prompt with a fixed rubric, no iteration or retrieval augmentation—yields ratings associated with the human panel, supporting further tests of supplementary use. Uncertainty in the human benchmark and assumptions underlying the reliability adjustment prevent an equivalence claim. Appendix results describe five additional models with differing sample coverage.

Introduction

i What the date above means

The date is the last time any page of this paper changed. It is not the date of the main analysis.

- **July 2025 to March 2026:** main analysis and writing by the authors. The GPT-5 Pro and GPT-5.2 Pro runs used in the main results were done in January 2026.
- **July 2026:** exploratory headless Codex pilot; signposting between the project sites.
- **September 2026:** corrections to numerical checks and comparisons.
- **October 2026:** AI-generated related-work note (OpenAI Codex), labelled where it appears.
- **November 2026:** provisional, AI-scored persistent-miss diagnostic (Claude Opus 5.5), added as an appendix and not yet reviewed by the authors.

Since September 2026, most changes have been made by AI agents at David Reinstein's request. AI-written material is labelled where it appears.

Choose the right project view

- **Current working paper:** this is the authoritative version of the statistical analysis and academic presentation. It is available at GitHub Pages and the primary Netlify URL.
- **Full project workspace and grant proposal:** see the broader project site for project goals, team and funding context, earlier analysis, and appendices. Grant reviewers can go directly to the proposal. This material is retained for context and is not the authoritative source for the latest results.
- **LLM–human critique concordance:** see the public concordance evidence site for the exploratory issue-mapping judgments and methods notes.

We are also maintaining a small **headless Codex evaluation pilot** as an exploratory extension. It is deliberately separated from the main API-based evidence because subscription-backed headless runs are useful for low-cost exploration but are less standardized and less directly replicable than the native API workflow.

A collaboration with The Unjournal (55+ open evaluation packages for global-priorities research). Funding: Survival and Flourishing Fund, Long Term Future Fund, EA Funds.

Can the model find the issues experts find?

On 30 papers, a one-shot Claude Opus 5.5 report raised 61% of an individual evaluator's issues by a GPT judge and 35% by a stricter Claude judge; a second evaluator raised 36% and 22% of the same issues. Both are far above a placebo that pairs one paper's evaluations with the model's report on another paper (2% and 1%). So the model overlaps with an individual evaluator more than a second evaluator does, under both judges, though the two judges disagree on the level and no human has checked the matches yet. An earlier analysis of 12 papers against a short curated consensus list reported 80% coverage, but that judge was GPT-5.2 Pro grading its own critiques; a blinded re-judging gives 44%–60% at the main level, and about 80% only when any criticism of the same result counts.

Headline figures and caveats · Per-paper coverage

Related work: King and Worthington (2026) · AI-generated note, October 23, 2026

King and Worthington (2026) report strong AI review performance against paid rapidPeer experts. This substantially overlaps with our high-quality human comparison; our clearest distinctions concern structured research ratings, uncertainty, and proposed prospective and human–AI studies. See the AI-generated comparison and research priorities and discussion.

Peer review is under strain. Reviewers are hard to find, turnaround times are lengthening, and the system costs an estimated $1.5 billion per year in the United States alone (Aczel, Szaszi, and Holcombe 2021). At the same time, generative AI lowers the cost of producing polished manuscripts; in at least some fields, editors report submission growth that exceeds reviewer capacity, and explicitly link this trend to LLM-assisted writing (Spitzer 2026). This combination creates demand for automated support in editorial and pre-submission workflows.

Commercial AI reviewer products such as Refine (Refine n.d.) and QED Science (QED Science 2026) already market automated referee-like feedback directly to authors, while disclaiming that

they substitute for human peer review. Their internal architectures are not publicly documented. Other tools are open. IsItCredible, run by The Catalogue of Errors Ltd, sold automated "Reviewer 2" reports before retiring its hosted service, and has released that pipeline and its prompts as open-source code (The Catalogue of Errors Ltd 2026; IsItCredible.com n.d.). OpenAIReview (Hsu and Tan 2026) is an open-source pipeline whose "progressive" method reads a paper in sequence while keeping a running summary of its definitions, equations, and key claims, at roughly $4 per paper. Meanwhile, publishers are formalizing policies that restrict reviewer use of general-purpose AI tools[1] while permitting controlled in-house applications for screening tasks (Leung 2026; Elsevier 2026). Our study complements these efforts with a baseline: how closely the simplest possible setup—a single prompt to a frontier model, with no bespoke pipeline—matches structured expert evaluations, both in its ratings and in the issues it raises.

These developments make the evidentiary gap salient: funders, editors, and policymakers need to know when AI evaluation outputs are trustworthy enough to use, and when they are unstable, biased, or manipulable. Recent work documents three interlocking concerns.[2] Reproducibility can be "jagged" across models and time (Thomas, Romasanta, and Pujol Priego 2026), and subtle task reframings can induce systematic output shifts reminiscent of specification search (Asher et al. 2026). Adversarial manipulation is not hypothetical: invisible prompt-injection text can inflate LLM review scores in simulated peer review (Choi et al. 2026). And even without manipulation, AI reviews tend to be less thematically diverse and less focused on interpretation and originality than human reviews (Rajakumar et al. 2026), while LLM scoring exhibits range restriction and halo effects that distort agreement metrics (Wang et al. 2025). The most direct recent comparison is more favorable to AI: on 124 rapidPeer submissions, King and Worthington (2026) find that AI reviews win every decided overall comparison against paid, carefully selected human reviews when an AI model judges, and 26 of 30 judgments by three former journal editors (see our comparison with that study).

The central question we address is therefore: how reliably can frontier LLMs evaluate research, relative to expert peer review and under realistic levels of rater disagreement?[3] We study this question in a setting designed to make "expert judgment" observable and multi-dimensional rather than implicit.

We use The Unjournal's structured human evaluations as a reference signal. We prompt GPT-5 Pro—a frontier reasoning model—with the same rubric and guidelines used by human evaluators, then compare its ratings with available expert evaluations. The main run covers 60 economics and social-science working papers, with 47 human-matched papers; critique overlap is a separate GPT-5.2 Pro analysis of 12 eligible pairings. We ask whether a frontier LLM evaluation can approximate

[1] For example, Elsevier's policy states that "reviewers must not upload submitted manuscripts, or any information from them, into AI Tools" (Elsevier 2026, Reviewers FAQ).

[2] King and Worthington (2026) summarize a wider set: hallucinated claims and fabricated sources (Ye et al. 2024), failure to recognize novelty (Shin et al. 2025), bias toward prestigious authors and institutions (Hosseini and Horbach 2023; Macharla Vasu et al. 2026), excessive leniency (Russo Latona et al. 2024), hidden-prompt manipulation (Lin 2025), and confidentiality breaches (Naddaf 2025). Field evidence is also growing: the AAAI-26 conference added one labeled AI review to each of 22,977 papers sent for full review (Biswas et al. 2026); randomized studies examine LLM feedback to reviewers (Thakkar et al. 2026) and reviewer AI-use policies at ICML 2026 (Kim et al. 2026); and blinded editors at a cardiology journal rated AI reviews above human reviews on five of seven criteria (Zancanaro et al. 2026). We cite these as King and Worthington describe them and have not yet assessed each study ourselves.

[3] A second, more exploratory question is how LLMs' research reasoning and priorities compare with those of human experts: which criteria and features of a paper they weight, and whether they would choose the same research to evaluate. The grant proposal describes planned work on this, including comparing AI prioritization with The Unjournal's internal prioritization decisions.

expert judgment, where systematic differences arise, and whether the model reveals characteristic AI preferences over research. GPT-5 Pro shows positive rank agreement with the human panel, alongside compressed ratings and uneven coverage of expert concerns. A paired bootstrap places uncertainty around an approximate comparison with human inter-rater agreement[4]; it does not establish equivalence to an additional expert. Appendix A reports results from five additional LLMs spanning different capability and cost tiers, with sample availability varying by model. Because the model ranks papers broadly like the human panel but compresses its ratings and misses some expert concerns, it may be more useful alongside human evaluators than in place of them. Testing this directly would mean, for example, checking whether adding a model's ratings to a human panel improves agreement with held-out evaluations or with later publication outcomes.

Our approach is minimal by design: each model receives the same PDF and a fixed rubric in a single prompt, with no iteration, retrieval augmentation, chain-of-thought scaffolding, or multi-step agentic loop. This provides a baseline for testing more elaborate evaluation procedures. Structured measurement tools,[5] steps that verify each flagged problem,[6] and prompt-sensitivity controls[7] could improve performance, but added complexity could also amplify shared errors or introduce new biases. Their incremental value requires comparison on held-out papers.

The Unjournal setting is particularly well suited for this comparison. It commissions paid expert evaluations using a structured rubric covering seven percentile criteria with 90% credible intervals plus journal-tier predictions, and publishes the resulting packages openly rather than making binary accept/reject decisions—which may increase reviewer effort through accountability and transparency. The resulting ratings and critiques still exhibit substantial inter-rater variation;[8] accordingly, we treat human evaluations as a high-quality but noisy reference signal, not ground truth. The rich, multi-dimensional data allow us to compare the priorities and calibration of humans and AI models across criteria and domains, while eventual publication outcomes for journal-tier predictions provide an external validation opportunity[9] enabling a human-vs-LLM horse race. Finally, The Unjournal's pipeline of future evaluations allows for clean out-of-training-data predictions, serving as a live testing lab for prospective validation.

[4]We resample papers with replacement 2,000 times. In each draw we recompute both the model–human and the human–human correlation on the same resampled papers, and take their difference after a reliability adjustment. The spread of that difference across draws gives the interval. See Results and Methods.

[5]Asirvatham, Mokski, and Shleifer (2026) present GABRIEL, a package that uses GPT to rate attributes of text, and report performance generally indistinguishable from human annotators across more than 1,000 tasks. Their applications are not research evaluation; we cite the paper as evidence that carefully structured LLM rating can match human raters, not as a tested way to improve research evaluation.

[6]Zhang and Abernethy (2025) treat LLMs as manuscript quality checkers that list a paper's most critical problems, scored against the stated reasons for withdrawn arXiv papers. They test single-pass checking and suggest a propose-then-verify workflow, in which flagged problems are checked again, for future work.

[7]Asher et al. (2026) run AI coding agents on the same empirical analyses with neutral prompts and with prompts that nudge toward statistically significant results, to see whether the reported estimates shift.

[8]On the overall rating, two human evaluators of the same paper have a rank correlation of 0.33 (95% CI 0.02 to 0.59; 39 papers) and a mean absolute difference of 11.6 percentile points; see Results. This is in line with journal peer review generally: a meta-analysis of 48 studies found a mean inter-rater reliability of about 0.34 (intraclass correlation or r^2) and a mean Cohen's kappa of 0.17 (Bornmann, Mutz, and Daniel 2010), though these measures are not identical to ours.

[9]These represent verifiable publication outcomes, not statements about the "true quality" of the paper.

Unjournal impact: author engagement evidence

Do authors engage with and respond to Unjournal evaluations? An April 2026 tracking file covers 57 evaluations, but adjustment status had been assessed for 22 rather than the full set. Within that partly assessed snapshot, 16 papers have a recorded formal written response, at least 5 are coded as showing clear evidence of substantive revision, and 7 authors stated an intention to update. A full staff reassessment remains an ongoing goal.
See the **full tabulation: did authors adjust their papers?** for the interactive table and public author statements. For broader context, see **Evidence: do authors engage with evaluations?** in The Unjournal's knowledge base.

Results

We evaluate GPT-5 Pro against human expert reviews from The Unjournal on 47 matched papers (papers with both GPT-5 Pro and human evaluations). The model receives the same PDF, system prompt mirroring The Unjournal rubric, and JSON schema requiring a diagnostic summary plus numeric midpoints and 90% credible intervals for every metric. Results for five additional models are reported in Appendix A. Full methodological details appear in Methods.

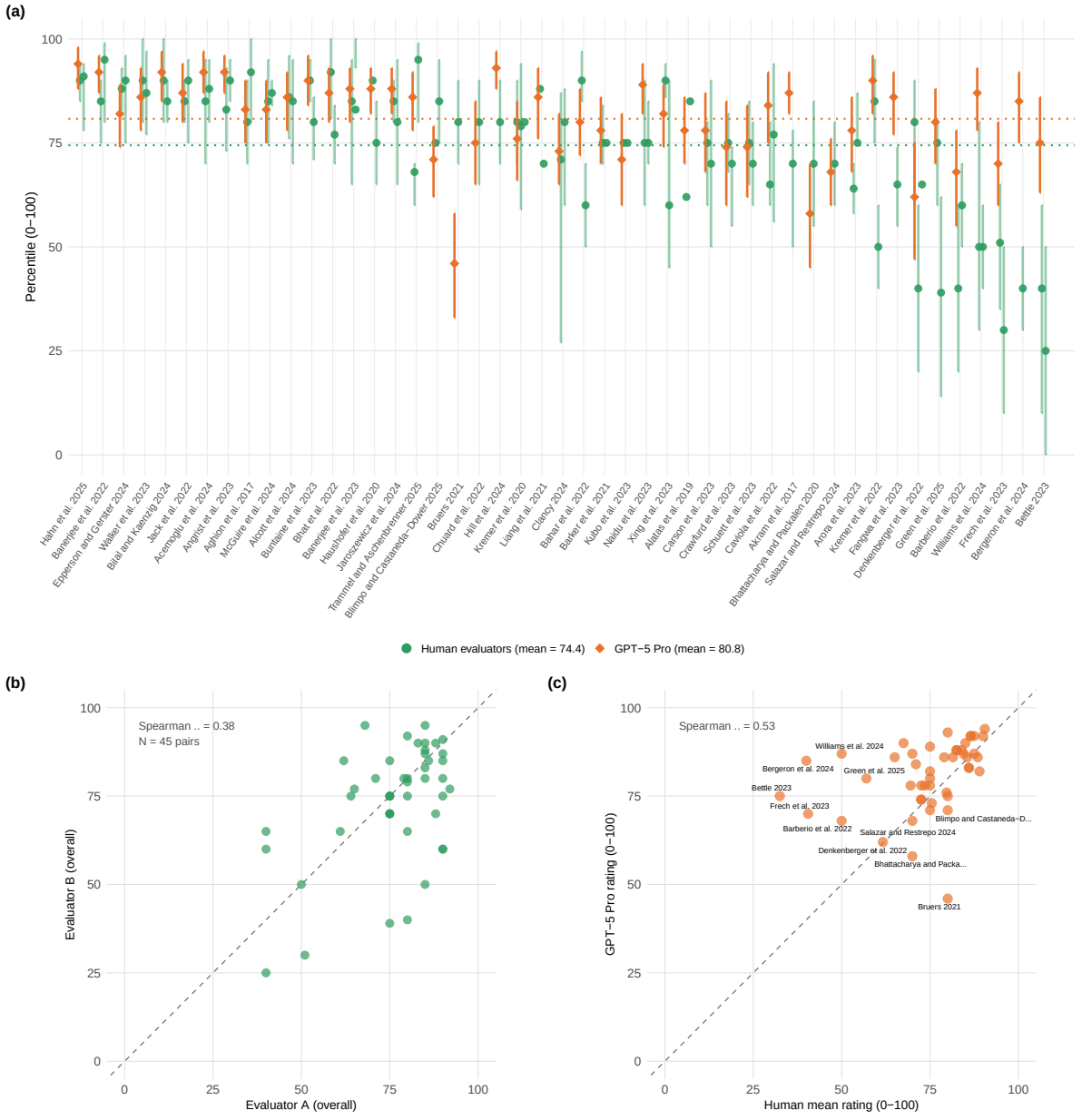
We do not treat human ratings as ground truth. Quantitative percentile scoring is genuinely difficult: even domain experts disagree, and individual scores reflect both signal about paper quality and idiosyncratic tendencies (severity, topic familiarity, interpretation of the scale). Our question is whether an LLM provides signal *comparable to an additional expert rater*. The Human–Human baseline row in Table 1 provides this reference. **Caution**: the primary comparison computes the LLM's ρ against the *mean* of 2.1 human raters, which reduces noise and inflates apparent agreement relative to the individual-vs-individual ρ_HH. The Spearman-Brown column provides an approximate sensitivity adjustment. Its classical reliability assumptions (exchangeable raters and independent errors around a common target) need not hold here, and applying it to rank correlations with unequal panel sizes is not an exact correction. Krippendorff's α_{HH} in the human-baseline table in Appendix A provides the criterion-level reference.

Per-paper overview. Figure 1 presents three complementary views of overall (0–100 percentile) ratings. Panel (a) displays individual human evaluator ratings alongside GPT-5 Pro (orange diamonds) for each paper, revealing inter-rater variability—the self-reported 90% credible intervals from individual evaluators often span 20–40 percentile points. In most cases the LLM falls within the range of human opinions, though several papers show substantial divergence. Panel (b) plots all pairwise human evaluator combinations, showing the observed human-human agreement baseline. Panel (c) compares GPT-5 Pro ratings against human mean ratings with per-paper labels.

Panel (b) displays human-human disagreement, with differences of as much as 30–40 percentile points on individual papers. Panel (c) uses a human mean, so the two scatterplots compare different targets and do not by themselves establish equivalent rater performance. GPT-5 Pro clusters around the identity line in the 40–80 range but diverges more at the extremes, compressing ratings toward the centre of the scale relative to humans—a descriptive pattern whose mechanism is not identified by this comparison. Where humans rate a paper very highly or very harshly, the LLM typically pulls toward the middle. Full agreement metrics for all six models appear in Appendix A.

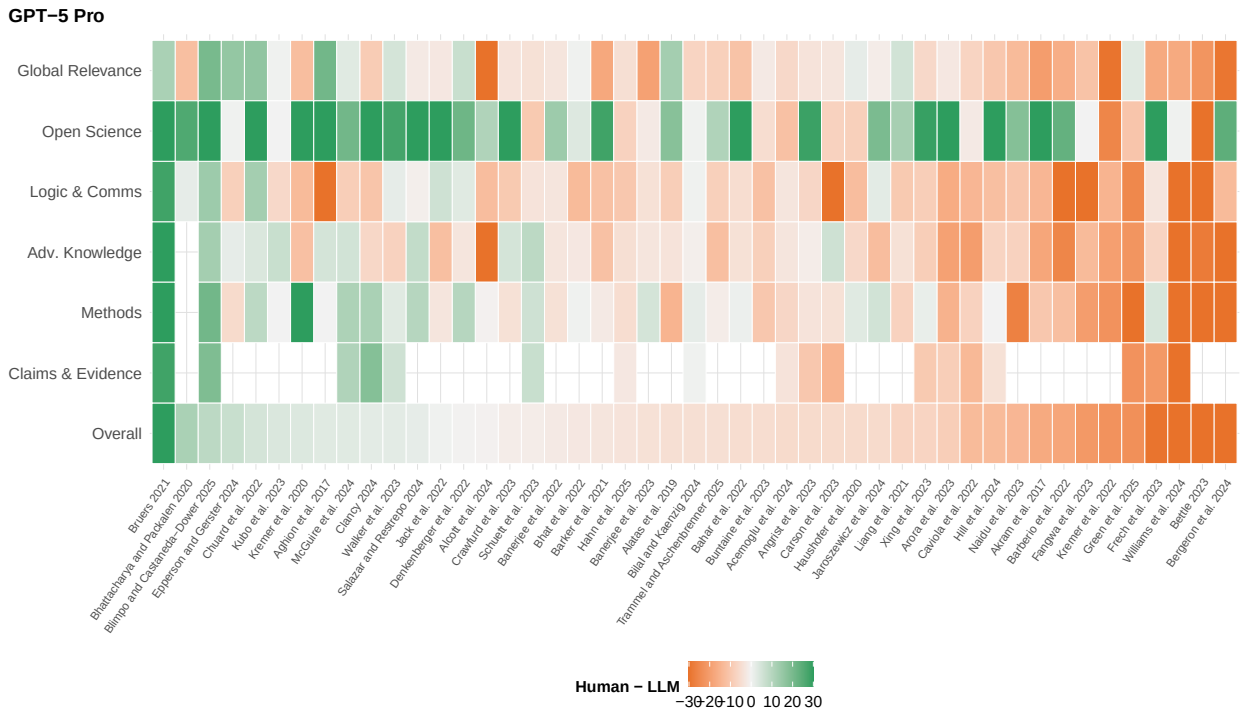
Rating differences by paper and criterion. Figure 2 unpacks Human − GPT-5 Pro differences across all seven criteria simultaneously. Each column is a paper, each row a criterion, and tile colour encodes the signed difference (human mean minus GPT-5 Pro midpoint). Green tiles indicate the

Figure 1: Per-paper overall ratings (0–100 percentile). **(a)** Individual human evaluator midpoints (green circles) with each evaluator's self-reported 90% credible interval (reflecting their own uncertainty about the true score; the vertical separation between green dots per paper reflects inter-rater disagreement) and GPT-5 Pro (orange diamonds with CI), sorted by descending human mean. Dotted horizontal lines show grand means. **(b)** Pairwise human evaluator agreement: each point is one evaluator pair (papers with 3 raters contribute 3 points). **(c)** Human mean vs GPT-5 Pro overall rating; dashed diagonal is the identity line. Panel (c) uses an averaged human target, so tighter scatter does not by itself establish equivalence to an individual evaluator.



human panel rated the paper higher; orange tiles indicate GPT-5 Pro was more generous. Papers are sorted by overall difference.

Figure 2: Human minus GPT-5 Pro rating difference for every paper (columns) and criterion (rows). Green tiles indicate the human mean was higher; orange tiles indicate GPT-5 Pro rated the paper higher. Colour is clamped to ±30 percentile points. Papers are sorted by overall difference.

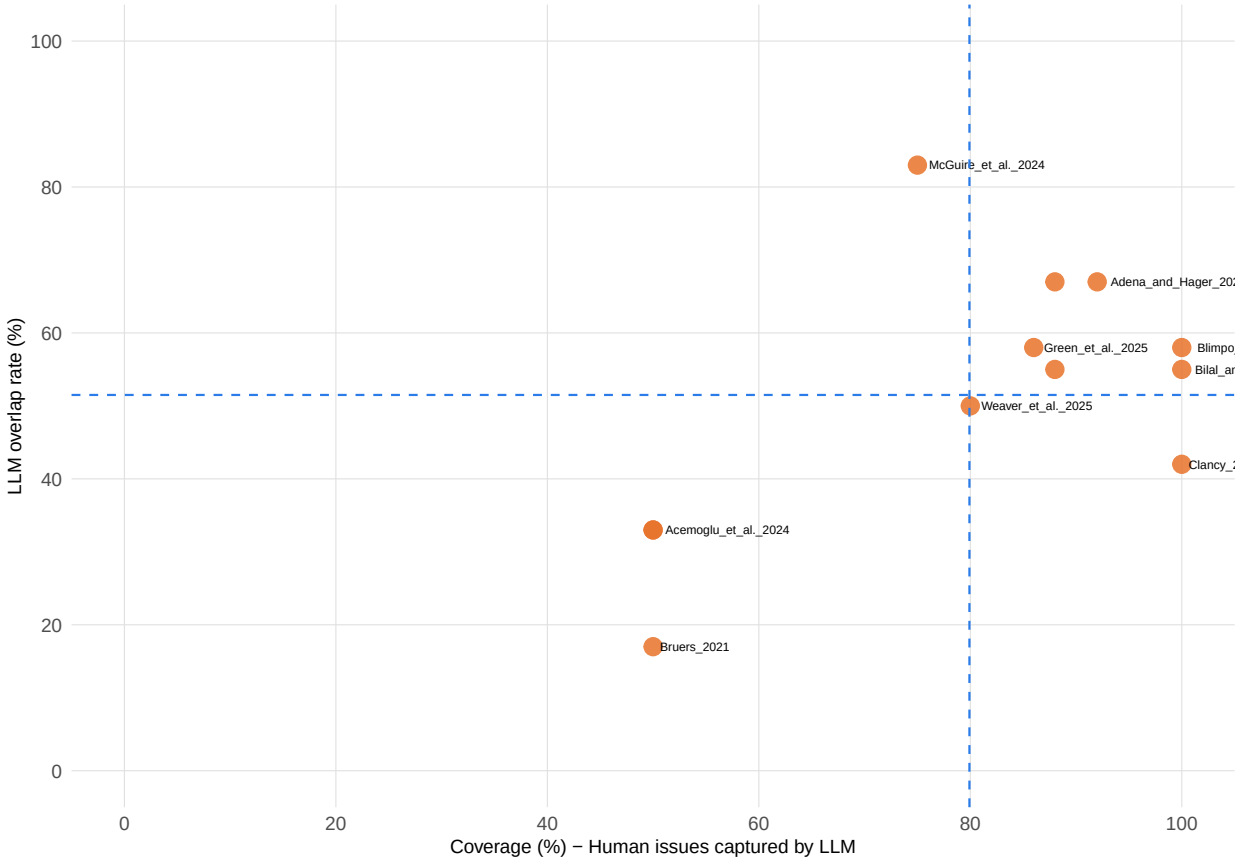


Rows with uniformly green or orange tiles indicate papers where humans and GPT-5 Pro disagree systematically across all criteria, not just on one dimension. Columns with consistent colour suggest criteria-level biases—for instance, if "Open Science" is green across nearly all papers, humans may systematically reward data-sharing practices more than the model does. Multi-model comparisons in Appendix A reveal whether these disagreement patterns are GPT-5 Pro–specific or shared across frontier LLMs.

Does the model find the issues experts find?

Beyond numeric ratings, we compare the substantive critiques each model raises against the consensus issues identified by human experts. The figure below plots coverage—the fraction of human-identified concerns that the LLM also raised in some form—against the LLM overlap rate—the fraction of LLM-raised issues that correspond to a substantive human concern. These metrics are assessed by GPT-5.2 Pro in a separate judging call (the same model that generated the focal critiques) (see Methods for the judging protocol). Two of the 14 originally curated pairs are excluded for manuscript-identity mismatches: Peterman et al. 2025 and Bénabou et al. 2023. The latter was identified in the July 2026 concordance audit: its human critiques concern *Willful Ignorance and Moral Behavior*, while the PDF and model critiques concern *Ends versus Means: Kantians, Utilitarians, and Moral Decisions*. The remaining 12 pairs contribute to the analysis. These exclusions concern critique pairings and do not by themselves establish an error in the separate quantitative ratings joins.

Figure 3: Coverage (% of human-identified issues that the LLM also raised) vs LLM overlap rate (% of LLM-raised issues that match a substantive human concern) for each paper, as assessed by GPT-5.2 Pro acting as judge. Dashed lines mark the cross-paper mean for each axis. Papers in the upper-right quadrant show strong alignment between human and LLM critiques.



Main result: per-evaluator issues on 30 papers. On 30 papers, a one-shot Claude Opus 5.5 report raised 61% of an individual evaluator's issues by a GPT judge and 35% by a stricter Claude judge; a second evaluator raised 36% and 22% of the same issues. The difference is 26 [17, 35] points under the GPT judge and 13 [7, 20] under the Claude judge (95% bootstrap intervals over papers), and a placebo pairing with another paper's model report matches 2% and 1%. All matches are AI-judged and await human checks (details). The figures below are from the earlier 12-paper analysis against curated consensus lists.

Coverage and the LLM overlap rate vary substantially across papers. Averaged over the 12 focal papers, the model captures 80% of human consensus issues (range 50–100%), while 52% of model-raised issues match a substantive human concern. Pooled over these papers, the curated human lists contain 63 issues, of which the model missed 10; the model raised 142 issues, of which 69 matched no curated human concern and have not yet been checked (headline figures and caveats). A low LLM overlap rate partly reflects the reference set: the human consensus list contains only the issues evaluation managers prioritised, so model issues flagged by individual evaluators but not curated into the consensus reduce the overlap rate. An unmatched model concern is not necessarily incorrect. Because these metrics are themselves LLM-assessed, they should be interpreted with the caveat that an LLM judge may systematically over- or under-credit matches relative to a human annotator. A revised, blinded judge (a different model, both lists rewritten into neutral statements, a rubric based on the result criticised and the reason) gives lower figures: mean coverage of 44% when only the same concern counts, and 64% when criticism of the same result for a different reason also counts. In 8 control cases pairing critiques of different papers it found no matches. A GPT-5.6 judge on the same blinded inputs is more lenient, and its loosest level reproduces the original estimate, so we read that estimate as an upper bound rather than the headline (details).

> Examine what was found and missed
>
> The per-paper coverage figure shows how coverage and overlap vary across papers. Every matched human–model pair, with the judge's explanation, is in the repository (`results/key_issues_comparison_results.json` for the original judge, `results/issue_match_judge_v2_opus55_2026-11-10/` for the revised one).

Detailed paper-by-paper comparisons—including matched issue pairs with severity labels, structural difference tables, and per-evaluator breakdowns—appear in Appendix B: Critiques & Key Issues. Extended quantitative analysis including all six models, per-criterion correlations, bootstrap confidence intervals, tier prediction accuracy, and cost-quality trade-offs is reported in Appendix A: Results Ratings. The full LLM reasoning traces and assessment summaries are available in Appendix C: LLM Traces.

Agreement summary. Table 1 shows GPT-5 Pro's agreement split into a **primary matched group** (same 39 papers as the H–H baseline, enabling like-for-like comparison) and a **secondary full-sample group** (all 47 matched papers). In the Spearman ρ column, the H–H row shows individual-vs-individual ρ (the reference, already on the natural scale); GPT-5 Pro rows show ρ vs. the human mean, which is upward-biased. The ρ **adj.** column applies the Spearman-Brown correction to GPT-5 Pro rows only, as an assumption-dependent sensitivity analysis. Six-model comparison in Appendix A, Table A2.

The Fisher-z intervals in Table 1 quantify uncertainty in each correlation separately; to make an uncertainty statement about the *comparison itself* we add a paired bootstrap over papers (2,000

Table 1: GPT-5 Pro overall-rating agreement. **Primary (matched) group**: same 39 papers as the H-H baseline (papers with at least 2 human raters). **Secondary group**: all 47 matched papers. **Spearman rho column**: the H-H row shows individual-vs-individual rank correlation (the reference, already on the natural scale); GPT-5 Pro rows show correlation vs. the human mean, which is upward-biased relative to H-H. **rho adj.**: approximate Spearman-Brown sensitivity adjustment using each row's mean panel size (matched-group factor 0.81). **95% CI**: approximate Fisher-z interval for the raw correlation, not the adjusted value. Bias = LLM minus Human. Full six-model comparison in Appendix A.

Model	N	Spearman ρ	ρ adj.	95% CI	Pearson r	Mean bias	MAE
Matched sample: same 39 papers as H-H baseline (primary)							
Human-Human	**39**	**0.331**	**-**	**[0.02, 0.59]**	**0.520**	**+3.7**	**11.6**
GPT-5 Pro	39	0.627	0.508	[0.39, 0.79]	0.577	+6.6	8.4
Full sample: all 47 GPT-5 Pro matched papers (secondary)							
GPT-5 Pro	47	0.526	0.436	[0.28, 0.71]	0.348	+6.4	10.1

Table 2: GPT-5 Pro agreement with human mean by criterion (N = 47 papers; N = 39 for H-H). **H-H rho**: pairwise human evaluator rank correlation (reference; computed on 39 papers with at least 2 raters; raw LLM rho is upward-biased vs. H-H rho, see Table 1). Positive bias = GPT-5 Pro rates higher on average. Note the variation across criteria: human evaluators agree strongly on some dimensions and barely at all on others (Open Science), setting expectations for LLM performance.

Criterion	H-H ρ	Spearman ρ	Pearson r	Mean bias	RMSE
Overall	**0.331**	0.526	0.348	+6.4	15.1
Claims & Evidence	**0.400**	0.388	0.213	+4.0	18.3
Methods	**0.375**	0.562	0.356	+4.5	18.1
Adv. Knowledge	**0.148**	0.454	0.304	+7.9	17.3
Logic & Comms	**0.179**	0.287	0.193	+9.5	16.2
Open Science	**-0.065**	0.150	0.177	-16.8	27.3
Global Relevance	**0.114**	0.424	0.341	+5.8	15.7

resamples of the 39 matched-group papers). Each draw recomputes the GPT-5 Pro–human-mean rank correlation, applies that draw's own Spearman-Brown correction, and takes the difference against the human–human pairwise correlation on the same draw. The mean panel size is recomputed within each draw. The estimated 95% interval for the adjusted difference (ρ adj. $- \rho_{\text{HH}}$) is [-0.12, 0.43], based on 1965 of 2000 draws. The remaining 35 draws have an undefined correlation or non-positive human reliability and cannot support this adjustment; the interval is conditional on valid draws. This analysis estimates uncertainty in an approximate comparison. An interval containing zero does not establish equivalence or non-inferiority to an additional expert. Such a claim requires a justified margin specified before testing and an interval narrow enough to rule out a practically important deficit. The present bootstrap also conditions on the observed evaluators and first-two-rater ordering; it does not measure uncertainty from recruiting a new evaluator panel.

Table 2 breaks the same agreement metrics down by evaluation criterion for GPT-5 Pro. The **H-H** ρ column shows pairwise human-human Spearman ρ for each criterion as a reference — note that raw LLM ρ is upward-biased relative to H-H ρ by the mean-vs-individual asymmetry (see Table 1). Criteria where H-H ρ is itself low indicate genuine expert disagreement; these rows alone cannot distinguish reference disagreement from model failure.

Journal-tier predictions and verifiable outcomes. Beyond percentile ratings, both evaluator

Table 3: Papers with resolved publication outcomes: predicted vs realized journal tier (0–5 scale). Realized tier maps The Unjournal's recorded publication status: '~top journal' = 4.5, 'decent journal' = 3.0, field journal = 2.5. Ambiguous statuses, R&R/forthcoming, and book chapters are treated as unresolved and excluded. Human prediction is the evaluator mean.

Paper	Recorded outcome	Realized tier	Human prediction	GPT-5 Pro predicti
Alatas et al. 2019	Published, ~top journal	4.5	3.50	3
Barker et al. 2021	Published, ~top journal	4.5	4.00	3
Buntaine et al. 2023	Published, ~top journal	4.5	4.70	4
Hill et al. 2024	Published, ~top journal	4.5	4.50	4
Jack et al. 2022	Published, ~top journal	4.5	4.00	4
Fangwa et al. 2023	published, decent journal	3.0	3.60	4
Kremer et al. 2020	published, decent journal	3.0	4.00	3
Caviola et al. 2022	Published: Psychology journal	2.5	4.35	4

types predict where each paper *will* publish on The Unjournal's 0–5 journal-tier scale. Unlike percentile scores, these predictions eventually resolve against an observable outcome: the venue where the paper actually publishes.

Human and GPT-5 Pro tier predictions agree at Spearman $\rho = 0.75$ (N = 42 papers with both predictions)—stronger agreement than on any percentile criterion, suggesting that "where will this publish" is a better-anchored question than "what percentile is this paper." For 8 of these papers, The Unjournal's records already show a resolved publication outcome that we can map to the tier scale (Table 3). On this small resolved set, mean absolute error against the realized tier is 0.62 tiers for GPT-5 Pro and 0.71 for the human panel mean—both track outcomes to within roughly two-thirds of a tier, and the sample is far too small to rank the two.

Two caveats apply. First, most of these papers were published before the model's training cutoff, so GPT-5 Pro may partly *recall* venues rather than predict them; human predictions were made before the publication outcome was known, which makes this comparison structurally generous to the model. Second, the remaining 34 papers with registered predictions from both evaluator types are still unresolved (working papers, R&R, or forthcoming). These predictions are locked in this repository and can be scored as outcomes accrue—a prospective validation design that sidesteps the contamination concern entirely, since future publication decisions cannot appear in any current model's training data.

All appendices, interactive figures, and full code are available at the companion website: https://llm-uj-research-eval.netlify.app. Source code and data are hosted at https://github.com/valentinklotzbuecher/llm-uj-research-eval.

Discussion

GPT-5 Pro shows positive rank agreement with human expert evaluations. The comparison with human inter-rater agreement remains uncertain: the paired bootstrap uses an approximate reliability adjustment, and failure to detect a difference does not establish equivalence to an additional expert. Central compression—the tendency to pull extreme ratings toward the middle of the scale—is the most consistent pattern, whose mechanism this observational comparison cannot identify. Journal-tier predictions are the dimension where human and model judgments align most closely, and on the small set of papers whose publication outcomes have already resolved, both predict realized venues with broadly similar accuracy. Qualitative coverage varies widely across papers: on some,

the model captures nearly all consensus human concerns; on others, it misses key critiques or raises issues absent from the expert consensus. Appendix results describe five additional models, but differences in paper availability and run configurations limit comparisons across models.

Limitations. Several caveats temper these conclusions. Our sample comprises 60 social-science papers evaluated by the model, of which 47 have published human evaluation packages for direct comparison; all were specifically selected by The Unjournal for evaluation, not a random draw from the research literature; performance may differ in other fields or on less polished manuscripts. Human evaluations are themselves a noisy reference signal rather than ground truth, with substantial inter-rater variation; agreement with a panel mean need not be bounded by pairwise human agreement. We cannot fully rule out knowledge contamination: while we instruct the model to ignore prior knowledge about authors, institutions, or publication history in the system prompt, the models' training data may include fragments of these papers or related discussions. Robustness checks with models whose training cutoffs predate the papers, or out-of-time validation on papers entering The Unjournal's pipeline after model training, would help address this concern. Prompt wording, reference-group interpretation, training, and shrinkage toward typical scores are possible explanations for rating patterns; this design does not separate them. Self-reported interval width alone does not establish calibration. All LLM evaluations are single-run; aggregating across multiple runs or temperature settings could change the picture. Finally, the qualitative human-issue coverage and LLM overlap-rate estimates are themselves LLM-assessed (GPT-5.2 Pro as judge), introducing a further layer of model dependence.

Implications. Even a reasoning-capable model costing several dollars per paper is orders of magnitude cheaper than human expert review. The qualitative gaps we observe—missed critiques, generic issues, and central compression of ratings—argue against full automation of peer review. AI evaluation appears most promising as a supplement: providing fast structured feedback, flagging potential concerns for human reviewers, and enabling systematic comparison across large paper sets that would be infeasible with human effort alone.

Governance and attack surface. As AI review tools move from research prototypes to deployed products, the attack surface expands. Prompt-injection techniques—embedding hidden instructions in a manuscript's metadata, footnotes, or even white-on-white text—could steer model outputs toward inflated ratings or suppressed critiques. Because our pipeline (and similar commercial services) routes unpublished manuscripts through third-party APIs, confidentiality depends on provider access, retention and training policies, contractual protections, and deployment configuration; encryption in transit alone does not prevent access during inference. Over-reliance on AI scores introduces a further governance risk: if editorial decisions weight model ratings, authors may optimise papers for the model rather than for scientific rigour, creating a Goodhart dynamic. Finally, current evaluations reflect a single model checkpoint; model updates, alignment changes, or fine-tuning can shift ratings in ways that are invisible to users. We recommend that any operational deployment include adversarial red-teaming of prompts, formal confidentiality agreements with API providers, transparent disclosure of AI involvement in review, and periodic re-calibration against fresh human evaluations.

Future directions. All LLM evaluations reported here are single-run; multi-run robustness and prompt-sensitivity analysis would directly address the key open methodological question. Prospective validation should archive the input PDF, model identifier, prompt, and prediction timestamp before human evaluations and publication outcomes become available. Entry into The Unjournal's pipeline after a training cutoff is insufficient if an earlier manuscript or related review was already public. The qualitative human-issue coverage and LLM overlap-rate estimates are them-

selves LLM-assessed (GPT-5.2 Pro as judge); human validation of these scores is needed to close the model-dependence loop. Finally, the tier-prediction comparison in Results is a first step toward externally verifiable accuracy measurement: roughly thirty papers with registered predictions from both evaluator types have not yet resolved, and scoring those predictions as publication outcomes accrue can strengthen external validation. Eligibility requires checking that each prediction predates any publicly known outcome or venue cue. A fixed follow-up window and a rule for unresolved papers are needed to avoid selectively scoring early publications.

Related evidence: rapidPeer

AI-generated note by OpenAI Codex, October 23, 2026, added at David Reinstein's request. Interpretation and proposed directions; not new findings or an agreed change to the research plan.

King and Worthington (2026) report that AI-generated reviews outperform paid rapidPeer experts in overall AI-judged comparisons and in 26 of 30 judgments by three former journal editors assessing ten review pairs. Their checks include manuscript-based claim audits, a human audit sample, and controls for length, style, model choice, and submission year. This is substantial evidence for the quality of constructive AI feedback in their setting, and overlaps with our use of incentivized expert evaluations. Our older one-shot results should not be read as a limit on current AI review capability.

The outcomes differ: their review-quality comparisons do not establish calibrated ratings of research, sound research prioritization, or improvements in subsequent manuscripts. Our structured ratings and uncertainty estimates address part of that gap, while our small, LLM-judged critique comparison still needs independent validation. Human agreement is a reference signal, not a correctness criterion; an AI concern absent from the human reports may be valuable. The strongest proposed extensions are prospective scoring, blinded adjudication of critique validity and importance, and randomized tests of the additional value and cost of human evaluation after AI assistance. The detailed AI-generated comparison separates current evidence from these proposals and discusses the study's strengths and limits.

Methods

This chapter describes the data sources, evaluation pipeline, prompt design, and statistical methods used throughout. All Python code underlying the pipeline is preserved in the collapsible blocks below and can be executed to reproduce the evaluation runs.

Sample and human reference data. We draw on published evaluations from The Unjournal, an open-access platform that commissions expert reviews of policy-relevant research without requiring journal submission. Each paper is typically assessed by two independent evaluators (occasionally one or three), who provide a written critique resembling a referee report together with quantitative ratings on seven criteria scored as percentiles (0–100) relative to "all serious research in the same area encountered in the last three years."[10] Evaluators additionally predict the journal tier in which the work "should" and "will" publish, using a 0–5 continuous scale anchored to familiar venue categories (0 = unpublishable, 5 = top-5 journal). For each metric, evaluators report a midpoint (median of their belief distribution) and a 90% credible interval that expresses their epistemic uncertainty.

The sample comprises working papers spanning 2017–2025 in development economics, health policy, environmental economics, and related fields that The Unjournal identified as high-impact. The quantitative human-comparison sample is restricted to papers with published evaluation packages and available matched ratings; the broader model-evaluated set also contains papers without a usable human match. For our anal-

[10] The seven criteria are: *overall assessment*, *claims and evidence*, *methods*, *advancing knowledge and practice*, *logic and communication*, *open and collaborative science*, and *relevance to global priorities*. Full definitions are given in The Unjournal's guidelines for evaluators.

ysis we extracted the individual evaluator scores, aggregated them by taking the arithmetic mean per paper per criterion, and retained the range of individual scores to characterise inter-rater spread.

Comparison with an individual evaluator. The main overall-rating table compares the model with the human mean and reports human-human correlation using the first two available evaluator rows on the matched papers. Its Spearman-Brown adjustment is a sensitivity analysis based on classical reliability assumptions, extended approximately to rank correlations and unequal panel sizes. The mean panel size is calculated for the relevant model row and recomputed within each paired bootstrap draw. Draws with non-positive human reliability or undefined correlations are reported as invalid. An interval including zero for the adjusted difference is not an equivalence test. A stronger design would compare the model and a held-out human against the same remaining panel, with a prespecified non-inferiority margin and uncertainty clustered by paper.

LLM evaluation pipeline. We evaluate six frontier models spanning three providers: GPT-5 Pro and GPT-5.2 Pro (OpenAI, reasoning-capable), GPT-4o-mini (OpenAI, lightweight), Claude Sonnet 4 and Claude Opus 4.6 (Anthropic), and Gemini 2.0 Flash (Google). Each model receives the identical PDF file, system prompt, and output schema. We pass the PDF directly to the model's native multimodal input rather than extracting text, preserving tables, figures, equations, and layout cues that ad-hoc scraping could mangle. A single API call per paper avoids hand-offs and summary loss from multi-stage pipelines.

The output is constrained by a strict JSON Schema enforcing the same nine fields that human evaluators complete: seven percentile metrics (each with `midpoint`, `lower_bound`, `upper_bound`) and two journal-tier predictions (each with `score`, `ci_lower`, `ci_upper`). Additionally, the model produces an `assessment_summary` of approximately 1,000 words that must precede scoring—a "think first, score second" protocol designed to ground numeric ratings in specific textual evidence. The extended schema used for our GPT-5.2 Pro focal run adds a `key_issues` array of concise, ranked issue statements for downstream critique comparison.

Agreement and reliability statistics. We quantify human–LLM agreement using several complementary measures. *Pearson's r* captures linear association between paired ratings; it is invariant to positive linear rescaling and constant offsets, so it does not detect all forms of rating bias. *Spearman's* ρ measures rank-order agreement and is robust to outliers and non-linear monotone relationships. *Mean bias* (LLM minus human) indicates the direction and magnitude of any systematic offset. *Root mean squared error* (RMSE) and *mean absolute error* (MAE) measure typical prediction error in the original scale units; RMSE penalises large deviations more heavily. To contextualise human–LLM agreement we report *Krippendorff's alpha* (α), a chance-corrected reliability coefficient that generalises across varying numbers of raters, accommodates missing data, and applies to any measurement level (nominal, ordinal, interval, or ratio). An α of 1 indicates perfect agreement, 0 indicates agreement no better than chance, and values below 0 indicate systematic disagreement. We compute α_{HH} (among human evaluators only) as a descriptive reference. The appendix also reports α_{HL} for the panel formed by adding each LLM to the individual human evaluators. This describes the reliability of the augmented panel, not a stand-alone model-human coefficient. The reference is not a mathematical ceiling, and a similar augmented-panel value does not establish that a model is interchangeable with a human evaluator. Model availability varies within the appendix paper set. For the qualitative key-issue comparison, *coverage* denotes the fraction of human-identified issues that received a match score $\geq 30\%$ from the LLM judge. The *LLM overlap rate* denotes the fraction of LLM-generated issues that matched at least one human issue; unmatched LLM concerns are not necessarily incorrect.

JSON Schema. The output schema enforces that every paper is scored on identical fields with identical types and bounds. Credible intervals are required (paralleling the human protocol) so that the model can express genuine uncertainty rather than suggest false precision.

System prompt design. The system prompt is assembled from modular components and concatenated before each API call. It opens with a role definition instructing the model to act as an expert research evaluator, followed by a debiasing block that explicitly prohibits use of author identity, institutional prestige, publication venue, or any extrinsic information—the model must base all judgments on the PDF content alone. A diagnostic-summary instruction requires the model to produce a roughly 1,000-word assessment identifying methodological, evidential, and interpretive issues *before* any scoring, implementing a "think first,

score second" protocol intended to anchor numeric ratings in specific textual evidence.

Percentile ratings are anchored to the reference group "all serious research in the same area encountered in the last three years," following The Unjournal's guidelines for evaluators. The prompt defines each of the seven criteria with emphasis on global priorities and practical relevance over pure academic novelty, mirroring the weight structure that human Unjournal evaluators are asked to apply.

Subsequent prompt components instruct the model on constructing 90% credible intervals—the smallest interval the evaluator believes is 90% likely to contain the true value—encouraging calibrated uncertainty rather than artificially narrow bounds. The prompt requests journal-tier predictions on a 0–5 continuous scale anchored to familiar venue categories (0 = unpublishable through 5 = top-5 journal), providing an externally verifiable reference point for papers that eventually publish. A validation block then requires the model to verify internal consistency: numeric scores must align with the written assessment, credible intervals must be non-degenerate, and high or low ratings must be explicitly justified in the assessment summary.

The components above are concatenated into a single system prompt string before each API call:

Submission and collection. The `evaluate_paper` function uploads a PDF to the API and submits a background job for evaluation. File IDs are cached by path, size, and modification time so that re-running on the same PDF reuses the previously uploaded file.

Each model receives the full PDF in a single reasoning call, avoiding hand-offs and summary loss from multi-stage ingestion. We submit one background job per paper to the OpenAI Responses API with "high" reasoning effort and server-side JSON-Schema enforcement, recording the response ID, model, file ID, status, and timestamps in a jobs index. No external sources or cross-paper material are retrieved; the evaluation is anchored entirely in the manuscript itself.

A polling loop then checks each job's status and, for completed jobs, retrieves the raw JSON response object and writes it to disk alongside reasoning-trace metadata (token counts, reasoning summary).

Multi-model evaluation. To assess whether systematic biases or calibration differences emerge across architectures, we collect evaluations from six models spanning three providers. For OpenAI models that lack background-job or extended-reasoning support (e.g., GPT-4o-mini), we use a synchronous call variant:

Anthropic. Anthropic's API accepts PDFs as base64-encoded document content rather than uploaded file IDs, and all calls are synchronous. We use Claude's native PDF support to preserve the same multimodal evaluation approach:

Google. Google's Gemini API accepts PDFs via a file-upload endpoint with MIME-type tagging:

A unified runner dispatches evaluations across all configured providers and writes per-model output directories:

Focal run with key-issue extraction. For a subset of 14 papers with rich human critiques, we ran an extended evaluation using GPT-5.2 Pro. The schema for this focal run adds a `key_issues` array—a ranked list of concise issue statements identifying the most important methodological, interpretive, or evidential concerns—alongside the standard metrics. This structured output enables direct comparison between machine-generated and human-identified issues. Two curated pairings (Peterman et al. 2025 and Bénabou et al. 2023) were subsequently excluded because their human and model critiques concerned different manuscripts; the coverage/overlap analysis therefore covers 12 papers. The Bénabou mismatch was identified in the July 2026 concordance audit. Raw responses are retained, with exclusions applied consistently in both results chapters.

Key-issue comparison with human critiques. To validate how well the LLM identifies substantive concerns, we compare its `key_issues` output against human expert critiques drawn from The Unjournal's Coda database. These human critiques—produced by paid domain experts and synthesized by evaluation managers—provide a high-quality but noisy reference standard for issue identification; accordingly, we treat them as a comparative signal rather than ground truth.

The comparison proceeds in two stages. First, we parse and align the data sources: LLM key issues are extracted from the focal-run JSON responses, while human critiques are drawn from a manually curated markdown document pairing each paper's machine and expert assessments. Second, we use an LLM judge (GPT-5.2 Pro with schema-enforced structured output) to systematically assess the degree of alignment between each human-identified issue and the set of machine-generated issues, producing issue-by-issue match scores, coverage estimates (fraction of human issues captured), and LLM overlap-rate estimates (fraction of LLM issues that match the curated human concern set).

The comparison pipeline takes a manually curated markdown file pairing each paper's LLM issues with human critiques and produces two structured outputs: a parsed JSON dataset (`key_issues_comparison.json`) and an LLM-assessed alignment report (`key_issues_comparison_results.json`) containing per-paper coverage, LLM overlap rate, matched-pair explanations, and an overall quality rating. Together, these outputs feed the qualitative analysis presented in Appendix B: Critiques & Key Issues.

Headless Codex Pilot

This page records a small exploratory extension to the main evaluation pipeline: running the same broad research-evaluation scaffold through `codex exec` using ChatGPT/Codex subscription access rather than metered OpenAI API calls.

This is not part of the main evidence base. The primary results in this working paper use API-based model calls with controlled inputs, recorded model identifiers, and native PDF ingestion where available. The headless Codex pilot is useful for low-cost exploration, prompt development, and deciding whether further evaluations are worth running, but it should be treated as a separate input modality and execution surface.

What we tried

We added a small runner:

- Script: `scripts/run_headless_codex_eval.py`
- Pilot run: `results/codex_gpt55_headless_pilot_jul2026/`
- Model surface: `gpt-5.5` via `codex exec`, authenticated through ChatGPT/Codex rather than an API key
- Prompt version: `v4_assessment_current_text_extraction`
- Input modality: `pdftotext`-extracted text, truncated at 60,000 characters for the first pilot
- First paper: `Bruers_2021.pdf`

The runner deliberately marks the result as `pdftotext_extracted_text`, stores the source PDF hash, writes a `run_config.json`, and emits JSON in a compatibility shape that the existing analysis code can read. The aggregation helper now reads both older OpenAI response payloads and the newer `parsed` / `output_text` shape used by this pilot.

The one-paper pilot produced the expected artifacts:

- `json/Bruers_2021.response.json`
- `metrics_long.csv`
- `tiers_long.csv`
- `combined_long.csv`
- `assessment_summaries.csv`

For this first run, the combined metrics were:

Metric	Midpoint	Lower	Upper
Overall	46	32	61
Claims and evidence	42	27	58
Methods	34	20	50
Advancing knowledge	57	40	72
Logic and communication	62	48	75
Open science	36	20	55
Global relevance	70	55	82
Tier should	2.3	1.5	3.2
Tier will	2.4	1.4	3.4

The model also correctly noted in its assessment that the text had been truncated and widened uncertainty accordingly. This is the behavior we want from an exploratory runner: it should surface input limitations rather than presenting the output as equivalent to a full native-PDF evaluation.

Terra → Sol concordance follow-up

We also tested a two-stage concordance workflow for comparing separately enumerated human issues with LLM-generated issues. A high-effort Terra pass checked input alignment and proposed mappings; only three ambiguous pair judgments were escalated to a high-effort Sol pass. The workflow used stable issue IDs, retained a historical overlap threshold of 30, recorded exact Codex telemetry, and kept all outputs separate from the confirmatory API results.

We fixed the candidate set as the first three records in source order before inspecting concordance. A manual title/PDF/issue screen excluded Bénabou et al. before judgment: the recorded title and human issues concerned *Willful Ignorance and Moral Behavior*, while the local PDF and LLM issues concerned *Ends versus Means: Kantians, Utilitarians, and Moral Decisions*. Acemoglu et al. and Adena–Hager passed the screen and were processed together in one Terra batch.

Result	Acemoglu et al.	Adena–Hager	Pooled eligible issues
Human-issue coverage	2/3 (66.7%)	6/7 (85.7%)	8/10 (80.0%)
LLM overlap rate	3/12 (25.0%)	6/12 (50.0%)	9/24 (37.5%)

The LLM overlap rate is the fraction of unique LLM issue IDs with at least one non-distinct mapping scoring 30 or higher. It is not an estimate of whether unmatched LLM concerns are incorrect. Sol changed one threshold-sensitive mapping out of the three escalated pairs: an Adena–Hager mapping moved from 31 to 27, lowering the pooled LLM overlap rate from 41.7% to 37.5%. Human-issue coverage was unchanged because that human issue retained another qualifying match.

The original Terra and Sol calls used 41,993 input tokens and 4,001 output tokens, including 1,229 reasoning-output tokens. An identical verification run reused both completed artifacts and made no new model calls. The compact methods note, eligibility screen, and structured final judgments are retained in the repository.

This increment supports preparing a larger preregistered-style study, but not running the full corpus yet. The remaining inputs should first receive the same alignment audit, and the protocol should freeze eligibility, compound-issue handling, aggregation, ambiguity ranking under an escalation cap, and treatment of ambiguous pairs that are not escalated.

Replicability and consistency

The main concern is real: this approach is less standardized than the API-based approach.

The API workflow is better for strict replication because it gives us a cleaner contract: explicit model identifiers, structured-output enforcement, direct request and response payloads, file IDs or native PDF inputs, usage metadata, and a provider surface designed for programmatic runs. It is also easier to re-run from a clean environment, pin defaults, and audit failures.

The headless Codex workflow is attractive for exploration because it can use subscription access and avoid metered API costs. But several details make it less directly replicable:

- The execution surface is the Codex CLI, not the model API.
- Subscription entitlements and rate limits can vary by account and over time.
- Default behavior may change with Codex releases unless we pin command flags and record the CLI version.
- The pilot currently uses extracted text, not native PDF input, so figures, tables, equations, and layout can be degraded.
- The first run intentionally truncated the extracted text; this must be treated as a pilot constraint, not a production setting.
- `codex exec` is an agentic wrapper around a model call. Even when prompted not to use tools, it is not as minimal as a direct API request.

This does not make the approach unusable. It means we should use it with a different evidential status. It is well suited for:

- cheap preliminary screening;
- prompt iteration before API spending;
- exploratory robustness checks;
- identifying candidate papers for deeper API-based evaluation;
- running larger informal batches where exact comparability is not the goal.

It is less suited for:

- headline quantitative comparisons with human evaluators;
- claims about model-vs-model performance;
- strict reproduction of the native-PDF GPT-5 Pro runs;
- archival baselines where future re-runs must be as close as possible to the original call.

Practical guardrails

For any larger headless batch, we should keep the following guardrails:

- Record the Codex CLI version, model flag, auth mode, run timestamp, prompt version, PDF hash, text-extraction command, input character count, and truncation status.
- Keep the output in a separate `results/<run_id>/` directory with `input_modality` clearly marked.
- Do not merge headless/text-extraction results into the main model-comparison plots without a visible modality label.
- Use API-based runs for confirmatory comparisons or for any result we expect others to reproduce exactly.
- Avoid browser automation or ChatGPT web scraping; use documented Codex non-interactive execution instead.

The appropriate interpretation is therefore: headless Codex is a useful exploration and cost-control tool, not a drop-in replacement for the API-based evidence pipeline.

References

Aczel, Balazs, Barnabas Szaszi, and Alex O Holcombe, "A billion-dollar donation: Estimating the cost of researchers' time spent on peer review," *Research integrity and peer review*, 6 (2021), 1–8 (Springer).

Asher, Samuel G. Z., Janet Malzahn, Jessica M. Persano, Elliot J. Paschal, Andrew C. W. Myers, and Andrew B. Hall, "Do claude code and codex p-hack? Sycophancy and statistical analysis in large language models," 2026.

Asirvatham, Hemanth, Elliott Mokski, and Andrei Shleifer, "GPT as a measurement tool," {NBER} Working Paper, 2026 (National Bureau of Economic Research).

Biswas, Joydeep, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins, "AI-assisted peer review at scale: The AAAI-26 AI review pilot," 2026.

Bornmann, Lutz, Rüdiger Mutz, and Hans-Dieter Daniel, "A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants," *PLoS ONE*, 5 (2010), e14331.

Choi, Byungjin, Tae Joon Jun, Joung Won Sung, Il Woo Park, Jeong-Moo Lee, Soo Ick Cho, Hyung Jun Park, Ro Woon Lee, and Jungyo Suh, "Invisible text injection and peer review by AI models," *JAMA Network Open*, 9 (2026), e2552099.

Elsevier, "Generative AI policies for journals" (Nov. 5, 2026).

Hosseini, Mohammad, and Serge P. J. M. Horbach, "Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review," *Research Integrity and Peer Review*, 8 (2023), 4.

Hsu, Chao-Chun, and Chenhao Tan, "OpenAIReview: Open-source AI-assisted academic paper reviewing," 2026.

IsItCredible.com, "Is it credible?" (Feb. 19, 2026).

Kim, Sunnie S. Y., Wesley Hanwen Deng, Jennifer Wortman Vaughan, Buxin Su, Weijie Su, Alekh Agarwal, Sharon Li, Martin Jaggi, Daniel G. Goldstein, Nihar B. Shah, and Miroslav Dudík, "Use and effects of LLMs in peer review: A randomized experiment and survey at ICML 2026," 2026.

King, Gary, and Steven Worthington, "AI Article Reviews Dominate Human Peer Reviews at Their Best: Evidence and Policy Choices," 2026.

Leung, Tiffany I., "LLMs in peer review—how publishing policies must advance," *JAMA Network Open*, 9 (2026), e2552042.

Lin, Zhicheng, "Hidden prompts in manuscripts exploit AI-assisted peer review," 2025.

Macharla Vasu, Sai Suresh, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte, and Mario Fritz, "Justice in judgment: Unveiling (hidden) bias in LLM-assisted peer reviews," in *Findings of the association for computational linguistics: ACL 2026*, 2026.

Naddaf, Miryam, "AI is transforming peer review — and many scientists are worried," *Nature*, 639 (2025), 852–854.

QED Science, "QED science: Critical thinking AI for research," 2026.

Rajakumar, Hamrish Kumar, Kailash Abhishek Sankaran, Manasi Pillai Ashok, and Srinivas Rachoori, "Peer review in the age of artificial intelligence: A comparative study of human and AI-generated review reports," *Postgraduate Medical Journal*, (2026), qgag005.

Refine, "FAQ - refine" (Feb. 19, 2026).

Russo Latona, Giuseppe, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West, "The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates," 2024.

Shin, Hyungyu, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim, "Mind the blind spots: A focus-level evaluation framework for LLM reviews," in *Proceedings of the 2025 conference on empirical methods in natural language processing*, 2025.

Spitzer, Markus Wolfgang Hermann, "The emerging submission crisis in behavioral science," *Trends in Neuroscience and Education*, 42 (2026), 100276.

Thakkar, Nitya, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou, "A large-scale randomized study of large language model feedback in peer review," *Nature Machine Intelligence*, 8 (2026), 326–336.

The Catalogue of Errors Ltd, "Reviewer 2: Automated peer review pipeline" (Nov. 5, 2026).

Thomas, Llewellyn D. W., Angelo Kenneth G. Romasanta, and Laia Pujol Priego, "Jagged competencies: Measuring the reliability of generative AI in academic research," *Journal of Business Research*, 203 (2026), 115804.

Wang, Yuehan, Jinyan Huang, Lun Du, Yuxin Guo, Ying Liu, and Rong Wang, "Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity," *Computers and Education: Artificial Intelligence*, 9 (2025), 100481.

Ye, Rui, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen, "Are we there yet? Revealing the risks of utilizing large language models in scholarly peer review," 2024.

Zancanaro, Edoardo, Andreas Giannopoulos, Alessia Gimelli, and Karl-Patrik Kresoja, "AI as a peer reviewer: A blinded comparative study of LLM-generated and human reviews in a cardiology journal," *European Heart Journal – Imaging Methods and Practice*, 4 (2026), qyag097.

Zhang, Tianmai M, and Neil F Abernethy, "Reviewing scientific papers for critical problems with reasoning LLMs: Baseline approaches and automatic evaluation," *arXiv preprint arXiv:2505.23824*, (2025).

Results: Ratings

Extended ratings analysis (per-criterion correlations, agreement metrics, human baseline, cost–quality trade-offs) is available in the online supplement at https://llm-uj-research-eval.netlify.app/results_ratings.html.

Results: Critiques & Key Issues

Detailed paper-by-paper critique comparisons (matched issue pairs, severity labels, structural differences) are available in the online supplement at https://llm-uj-research-eval.netlify.app/results_critiques.html.

Persistent misses: a provisional Claude Opus 5.5 diagnostic

⚠ Provisional, AI-scored diagnostic · November 4, 2026 (Williams leakage item corrected November 6)

This page reports a small follow-up diagnostic, run and scored by Claude Opus 5.5 at David Reinstein's request. *All scores are a provisional first pass. David Reinstein has not yet reviewed them, or the two critiques the model disputes (below).* The diagnostic is separate from the one-shot benchmark: its results are not pooled with, and do not change, the ratings or issue-coverage numbers reported in the main chapters and in Results: Critiques & Key Issues.

Question. The earlier one-shot runs (GPT-5.2 Pro, Claude Opus 4.6, Claude Sonnet 4) missed some critiques that human evaluators raised (see Results: Critiques & Key Issues). Do these misses persist with a current model? Where they do, is the cause:

- failing to *generate* the point;
- failing to *prioritise* it;
- failing to *recognise* it when told?

The aim is a short ledger of cases where an expert caught something and the model did not. Such cases can seed harder evaluation items with known answers.

Design

Papers. We used six papers: Acemoglu et al. (2024), Williams et al. (2024), Dullaghan and Zhang (2022), Frech et al. (2023), Clancy (2024) and Green et al. (2025). Each was given as the same PDF used in the historical runs. Bénabou et al. (2023) and Peterman et al. (2025) remain excluded, as elsewhere in this paper.

Model. `claude-opus-5-5` was called through the Anthropic Messages API at effort "high", with adaptive thinking, the PDF as a native document, no tools and no refusal fallback. Each condition was run once, and no call was refused. The total cost was about US$14, including one condition-C rerun on November 6 (see the Williams leakage note below).

Prompt conditions.

- **A, standard:** the combined evaluation prompt with the key-issues list from Methods.
- **B, prioritisation:** "What is the single most important problem with this paper's main claims? Then give your top five, ranked."

- **C, recognition:** the evaluator's critique stated as a claim: "An evaluator argues X. Is this right, and how much does it matter?" C was run only for critiques missed or only partly caught in both A and B.

Reference set. The target critiques were re-checked against the *published* Unjournal evaluations, not only the internal curation used in Results: Critiques & Key Issues. Two corrections followed:

- The Acemoglu point about the prior-election control came from the evaluation manager, who called it "not a major limitation".
- The Frech point about the $40,000-per-QALY estimate is only partly supported by the published evaluations.

Results (provisional)

No longer missed. The two clearest earlier misses did not persist:

- **Acemoglu et al.** Both A and B note that the field experiment's statistical significance depends on controlling for the 2018 vote share. A adds that the control is "pre-specified and sensible", which is the balanced answer.
- **Williams et al.** Both conditions catch the gap between the headline 215 Mha and the 31.6 Mha of regrowth observed in the paper's training data. B ranks it first. A calls it a "~7-fold multiple of the observed 16-year regrowth" that is "not reconciled".

Also caught.

- Acemoglu: spillovers and the partisan framing of the field treatment.
- Williams: socioeconomic confounding.
- Frech: the mechanical GDP regression.
- Clancy: the omission of AI risk (rank 4 in B).
- Green: the imputation of 0.01 for unreported effects, and the non-systematic search.

Persistent or disputed.

Table 6: Persistent or disputed misses. "A / B" gives the provisional score in each condition; C is the model's own verdict when shown the critique.

Paper	Critique	A / B	C: model's verdict	Failure type	Status
Clancy (2024)	The ~2% discount rate is open to the same objection the paper uses to reject forecasters' lower rates	partly / partly	right; important	internal consistency	persists

Paper	Critique	A / B	C: model's verdict	Failure type	Status
Dullaghan & Zhang (2022)	Consumer-acceptance answers conditional on >10 kt produced are read as unconditional	missed / missed	partly right; minor	conditional-probability logic, then prioritisation	persists
Dullaghan & Zhang (2022)	1000× units error (FGF2, per kg vs per gram) in the table given to forecasters	missed / missed	cannot assess from the paper	needs external source	persists; not a fair one-shot test
Dullaghan & Zhang (2022)	Cost studies (TEAs) presented as predictions rather than conditional estimates	partly / partly	partly right; important	needs external source; prioritisation	partly persists
Williams (2024)	"Validation" compares 2030 predictions with 2000–2016 outcomes	missed / missed	partly right; important	needs external source (the table is in the supplement, not in the PDF used)	persists
Williams (2024)	Predictors measured over the outcome period (leakage), including soil data that is in the final model	partly / missed; soil leg not raised in either	partly right; minor (rerun with the full published wording)	generation (soil leg), then prioritisation	persists
Acemoglu (2024)	Use post-minus-pre change scores in the online experiment	missed / missed	**disputes it**: the headline specification already controls for the baseline outcome	none established	to be adjudicated

Paper	Critique	A / B	C: model's verdict	Failure type	Status
Green (2025)	Excluding studies with n<25 limits publication-bias detection	missed / missed	partly right; minor: the paper's bias methods do not rely on small-study asymmetry	specification judgement	to be adjudicated
Frech (2023)	$40,000/QALY estimate stated without sensitivity analysis	partly / partly	partly right; minor	prioritisation	weak item

Correction to the Williams leakage item (November 6). The first condition-C prompt paraphrased the critique from our internal curation and named only net primary production, burned area and road density. The model judged that version wrong: none of the three is in the final ten-variable model (Extended Data Fig. 2). But the published critique also names soil characteristics. The evaluators' joint synthesis lists them under Primary Concern 1, and evaluator 2's leakage section leads with the SoilGrids data. Soil organic carbon *is* in the final model; it ranks third by Gini importance, and the paper itself attributes its effect to carbon being "higher within and near forests". Rerun with the full published wording, the model rates the critique **partly right, minor**. It accepts a "legitimate, partial leakage risk via SOC that the paper does not test", rejects the other legs for the final model, and asks for a refit without the soil variables. It judges the point minor because forest density and distance to forest dominate. That last step is its own untested claim. The original output is kept, with the ledger marking it as superseded.

Failure types

- **Internal consistency (Clancy).** The model calls the regime-exit parameter "judgmental" but does not notice that the paper's own reason for rejecting the forecasters' implied rates also undercuts its ~2% rate. When told, it agrees and would rank the point mid-to-high. This is the cleanest case of a point the model fails to generate but recognises.
- **Conditional-probability logic (Dullaghan and Zhang).** Neither A nor B mentions the conditioning on >10 kt. When told, the model accepts a narrower version: the conditioning makes the null correlation non-diagnostic, and the low bar selects worlds where demand already existed. It then rates the point minor.
- **Evidence outside the PDF (Dullaghan and Zhang; Williams).** The units error, and the validation table, sit in reference material or a supplement that the one-shot input did not include. The earlier models received the same PDFs, so these were never fair one-shot tests.
- **Prioritisation.** For several items the model engages once prompted but ranks the point below other concerns. For Frech, the published evaluations make the same call. For the Williams soil-data leakage, which the evaluators list as a primary concern, the model does not.
- **Wording of the critique (a design lesson).** In the recognition condition, the verdict depends on how the critique is phrased. A paraphrase that dropped the strongest leg of the

Williams leakage critique drew a confident "wrong"; the full wording drew "partly right".
- **Magnitude sanity checks** now appear to be handled. For Williams, the model also proposes a mechanism: uncalibrated, class-balanced random-forest scores are summed into an "expected area".

Possible test items

Each persistent miss suggests an item with a known answer:

- **Clancy.** Ask whether the reason for rejecting the forecasters' implied discount rates also applies to the paper's own rate. Known answer: yes. About 1.6 points of the ~2% come from a short-horizon AGI forecast applied as a permanent hazard.
- **Dullaghan and Zhang.** Ask what the conditional acceptance questions can and cannot show. Grade on whether the answer spots the selection created by the conditioning event.
- **Williams.** With the supplement attached, ask what period the validation table tests against. Known answer: 2030 predictions against 2000–2016 regrowth, so it does not test forward skill.
- **Williams, leakage.** Ask whether any predictor in the final model could encode the outcome, and how to test it. Grade on naming soil organic carbon, noting that the other contemporaneous predictors are not in the final model, and proposing a refit without the soil variables.

The two disputed critiques (Acemoglu change scores, Green small-study exclusion) become items only after they are adjudicated. If the model is right, they become "reject a mistaken critique" items.

Contamination

These papers and their published evaluations were public before Claude Opus 5.5's training cutoff (about mid-2026). The model may therefore have seen them, so a catch is weaker evidence than a miss.

Two points argue against simple recall:

- the model still missed points that appear in the published evaluations, such as Clancy's discount rate and the Dullaghan conditional forecasts. This is the stronger point: recall should have produced them;
- what it did catch can be found from the paper alone, such as the roughly 7× gap from the paper's own figures, and the prior-election control in the main results table.

A weaker point: in condition C the model was *shown* published critiques, so it plainly saw them, and it disputed some rather than deferring. That only weakly suggests it was not recalling the evaluations, since a model can disagree after re-reading the paper either way.

Some influence on the catches cannot be ruled out. Cleaner tests would use papers evaluated after the cutoff, or disguised variants that change names, setting and numbers while keeping the flaw. Disguised variants are also how these cases could become test items the model cannot have memorised.

Other caveats

- **Run-to-run variation.** Each condition was run once, so catches and misses ranked low in a list could flip on a rerun.
- **Scorer.** The provisional scorer is the same model family as the model tested.
- **Comparability.** Condition A matches the prompt of the GPT-5.2 Pro key-issues run, but the model and interface differ, so this is not a like-for-like comparison with the earlier runs.

- **Clancy coverage.** One of the Clancy evaluation reports was not in our exports, so that paper's critique set is incomplete.

Materials. The runner is `scripts/run_headless_claude_misses.py`. Model outputs, raw API responses, the claims used in condition C and the provisional ledger with verbatim quotes are in `results/claude_opus_5_5_high_misses_2026-11-04/`. A longer note is in `docs/persistent_misses_opus55_2026-11-04.md`.

Supplementary Material

Full LLM reasoning traces, assessment summaries, related work, and extended analyses are available in the online supplement at https://llm-uj-research-eval.netlify.app/appendix_llm_traces.html.

Did Authors Respond to Unjournal Evaluations?

The interactive tables in this appendix are available in the online version at https://valentinklotzbuecher.github.io/llm-uj-research-eval/paper_response_analysis.html.

> **Note**
>
> This page reports an April 2026 tracking snapshot covering 57 Unjournal evaluations. Adjustment status had been assessed for 22 of them, not the full set; completing and re-validating the remaining staff assessment is an ongoing goal. The page also includes an exploratory LLM comparison of eight paper pairs and the more careful July 2026 checks described below. See Combined Evidence.

> **Comment or flag a correction**
>
> This page has Hypothes.is enabled. Select text and open the annotation sidebar from the upper-right edge of the page to leave an inline comment. We especially welcome corrections to document/version matching, missing public revisions, or our interpretation of whether a change reflects a suggestion. A free Hypothes.is account is required to post.

Overview

The Unjournal commissions open, structured evaluations of working papers in economics and social science. A natural question is whether this process changes research: do authors engage with feedback and revise their work?

This page draws on a tracking file covering 57 evaluations. In the April snapshot, 22 had an assessed adjustment status and 35 remained unassessed. For a subset, we also use automated PDF comparison and exploratory LLM screening. Background: The Unjournal's knowledge base.

```
library(dplyr)
library(ggplot2)
library(readr)
library(DT)
library(tidyr)
library(stringr)
library(glue)
library(jsonlite)
```

```r
df <- read_csv("data/author_adjustment_manual.csv", show_col_types = FALSE)
reviewed_outcomes <- read_csv(
  "data/paper_response_reviewed_outcomes.csv",
  show_col_types = FALSE
)

stopifnot(
  nrow(reviewed_outcomes) == 22,
  sum(reviewed_outcomes$public_version_status == "no_observed_public_change") == 13,
  sum(reviewed_outcomes$public_version_status == "later_public_version_reviewed") == 9
)

df <- df |>
  rename(
    paper_title    = label_paper_title,
    pubpub_link    = dup_pubpub_final_links,
    adj_status     = `Adjusted_paper?`,
    updated_manual = `Updated since UJ report -- manual confirmation`,
    deposit_after  = `deposit date > unjournal pub date`,
    research_area  = main_cause_cat_abbrev,
    nb_link        = notebookLM_link,
    pub_status     = publication_status,
    wp_date        = working_paper_release_date,
    uj_pub_date    = publication_date_unjournal
  )

adj_labels <- c(
  "Evidence of updating", "Stated intention to update",
  "Mixed / minor updating", "Unlikely to update", "Not yet assessed"
)

df <- df |>
  mutate(
    adj_status_clean = case_when(
      adj_status == "evidence of updating"          ~ "Evidence of updating",
      adj_status == "Stated intention to update"    ~ "Stated intention to update",
      adj_status == "mixed evidence/minor updating" ~ "Mixed / minor updating",
      adj_status == "Unlikely to update"            ~ "Unlikely to update",
      TRUE                                          ~ "Not yet assessed"
    ),
    adj_status_clean = factor(adj_status_clean, levels = adj_labels),
    author_response_clean = case_when(
      author_response == "Formal response" ~ "Formal response",
      author_response == "Informal"        ~ "Informal response",
      author_response == "None?"           ~ "No response",
      TRUE                                 ~ "Not yet coded"
    )
```

```
  )

theme_uj <- function(base_size = 12) {
  theme_minimal(base_size = base_size) +
    theme(panel.grid.minor = element_blank(),
          axis.text = element_text(size = base_size * 0.85),
          plot.caption = element_text(size = base_size * 0.75, color = "grey50"),
          legend.position = "right")
}
```

Combined Evidence of Paper Updating

Two evidence streams are shown together: April tracking categories for 57 papers, of which 22 had an assessed adjustment status, and an exploratory Claude Opus 4.6 screen for eight papers where before/after PDFs were available and meaningful text changes were detected. The latter is a way to prioritize inspection, not a causal estimate.

> **!** Current validation work (July 2026)
>
> The April figures below remain the historical, partly assessed snapshot. A fresh July 18 source refresh captured 32 public current-paper versions and created 13 distinct before/after comparison records. Nine records passed document, identity, evaluation-mapping, and timing checks. Four superficially usable comparisons were excluded because the supposed later document predates the evaluation. We also audited all 17 cases where the initially selected input and current public PDF were byte-identical. Thirteen support a limited "no observed public document change" finding. Two apparent no-change cases were repaired by locating the correct later endpoint, one revised endpoint remains unavailable, and one record had no completed evaluation. These are coverage figures for a workflow, not findings that an evaluation caused a change.
> For each candidate, the revised workflow preserves PDF snapshots and hashes, checks that the documents are the same paper and in the right order, records page- and section-level changes, and requires human review before any public claim of evaluation-linked updating. A newer PDF, a textual difference, or a model suggestion is not enough on its own.

Quick updated takeaways from the stricter screen

- **The substantive rates do not change.** The refresh found two additional different PDFs, but both predate their Unjournal evaluations. The reviewed evidence therefore remains 13 cases with no observed public document change and nine genuine later versions.
- **Whole-document controls matter.** A first card-level pass produced three medium-priority lexical matches. Checking the entire earlier paper downgraded all of them: the final deterministic pass found 17 low-priority review leads and no medium- or high-priority link. This does not overturn the manually reviewed correspondences below; it shows that simple lexical matching should not be used to create them.
- **The most informative low-priority leads illustrate different outcomes.** The macroclimate revision adds an appendix explicitly titled "Additional Robustness Checks," consistent with a broad request for more checks but without identifying the evaluation as its source. The resilient-education revision adds questionnaire and protocol material about speakerphone use

and other children, which documents the possible spillover issue but does not establish that only one student benefited. An apparent Urban Forests match disappears because alternative upwind-cone checks were already in the earlier appendix.

- **Coverage remains the binding constraint.** Among the 57 tracked records, 21 still need timeline validation and 15 current endpoints failed even after headless-browser fallback. Those cases cannot be treated as either updated or unchanged.

The automated screen is deliberately conservative. It requires an eligible version timeline, exact evaluation-document mapping, an evaluator quote with a valid line anchor, language that is new relative to the aligned change card, a whole-earlier-document preexistence check, and a cross-paper placebo benchmark. Passing produces a review lead only; it never assigns influence or enters the response-rate numerator automatically.

Expanded case review (updated July 18, 2026)

For each of the initial seven cases, separate passes extracted exact suggestions, classified paper changes without seeing the evaluations, tried to disprove suggested links, and adjudicated the surviving candidates. We then manually checked the strongest apparent matches. Across these cases, the process examined 90 mechanically valid suggestion items and 99 changes classified as substantive. For *Urban Forests*, the bounded classifier reviewed 31 of 68 deterministic change cards; 16 of those 31 were substantive. Two additional cases were reviewed after correcting stale endpoint selections.

Paper	Evidence items	Change cards	Current assessment
The Environmental Effects of Economic Production	4	28	**One plausible aligned revision, attribution unconfirmed.** A new taxa-heterogeneity figure includes birds and partly matches a suggestion to assess productivity effects for bird taxa. The new analysis concerns agricultural income and eight taxa, and there is no author statement connecting it to the evaluation.

Paper	Evidence items	Change cards	Current assessment
Building Resilient Education Systems	10	10	**Possible partial correspondence, attribution unconfirmed.** A request for stronger robustness checks overlaps with newly reported multiple-testing-adjusted results. Newly appended questionnaire and protocol material also documents speakerphone use and educational activity with other children, but it does not show that only one student benefited per call or rule out spillovers. These are standard design/reporting steps, and no response letter links them to the evaluation.

Paper	Evidence items	Change cards	Current assessment
The Macroeconomic Impact of Climate Change	6	7	**Circumstantial correspondence only.** The revision adds an "Additional Robustness Checks" appendix and a new cross-sample robustness statement, both topically consistent with the evaluator's small-sample concern. The request was broad, related robustness material already existed, and no revision note identifies the source.
A Welfare Analysis of Policies Impacting Climate Change	25	6	**No supported link at present.** One automated candidate appeared to connect a request to clarify "international policies" with international examples. Manual checking found that those examples were already present in the earlier version, so we rejected the match.

Paper	Evidence items	Change cards	Current assessment
Zero-Sum Thinking, the Evolution of Effort-Suppressing Beliefs, and Economic Development	8	21	**No positive link found.** Three possible pairings reflected timing only; five suggestions remained unmatched.
Water Treatment and Child Mortality	14	11	**One plausible aligned revision, attribution unconfirmed.** A new formal cost-effectiveness model, including cost per DALY and net DALYs averted, closely matches requests to deepen the cost analysis. No author response or revision memo links the addition to the evaluation.

Paper	Evidence items	Change cards	Current assessment
Urban Forests: Environmental Health Value and Disparities	23	16	**Possible topical correspondences, attribution unconfirmed.** New wind-direction checks overlap with a confounding concern, and an NDVI footnote partly overlaps with a request about alternative vegetation measures. The strongest automated site-selection match was excluded because its source quote failed validation. A newer lexical match about alternative upwind measures was also downgraded after the whole-document check found alternative upwind-cone tests in the earlier appendix.

Paper	Evidence items	Change cards	Current assessment
Artificial Intelligence and Economic Growth	8	1	**Direct documented evaluation-linked correction.** The public author response says the evaluation caught a significant proof error. The maintained document marks the exact invalid derivation and points to the corrected proof. The whole-document format change produced only one coarse change card, so this conclusion rests on the explicit response and document annotation rather than model matching.

Paper	Evidence items	Change cards	Current assessment
Adaptability and the Pivot Penalty in Science	8	1	**No supported substantive implementation; one weak framing overlap.** A manual, section-aligned review found real revisions, but seven of eight suggestions were not implemented. A new caveat that the measured outcomes do not exhaust research quality weakly acknowledges the concern about omitted benefits without adding a benefit outcome. Several apparently responsive items were already in the evaluated version.

The nine reviewed comparisons now include one direct documented correction, two reasonably specific content alignments with attribution unconfirmed, three weaker correspondences, and three cases with no supported substantive match. Only the proof correction has explicit public evidence connecting the evaluation to the change.

Overall and conditional response rates

The 22 reviewed endpoints support a more useful distinction than "changed" versus "unchanged." The five outcome categories below are mutually exclusive.

Reviewed endpoint outcome	Cases	Share of 22
No observed public document change	13	59%
Later version, but no supported substantive implementation	3	14%

Reviewed endpoint outcome	Cases	Share of 22
Later version with weaker correspondence	3	14%
Later version with a reasonably specific alignment; attribution unconfirmed	2	9%
Explicit public author link between evaluation and correction	1	5%

Percentages are rounded; the counts are primary.

Overall, six of 22 reviewed endpoints (27%) show at least some correspondence between an evaluation point and a later change. Three of 22 (14%) show either a specific alignment or an explicit link, and one of 22 (5%) has an explicit public author statement connecting the evaluation to the correction.

Conditional on the nine papers with a genuine later public version, six (67%) show at least some correspondence, three (33%) show a specific or explicit alignment, and one (11%) has an explicit author link. Three of nine (33%) have no supported substantive implementation in the versions we reviewed.

These are descriptive rates for a small, selected, source-available set, not estimates for all 57 tracked evaluations. "Later version, but no supported implementation" is also not equivalent to "the authors ignored the suggestions." Authors may have considered and rejected a suggestion, responded privately, revised a document we have not found, or changed the paper for unrelated reasons.

The identity checks now follow normalized DOIs where available and permit explicit, manually reviewed title aliases. They do not depend on exact title equality or infer aliases from a similar title. The evidence record stores the canonical title, any reviewed aliases, and which title matched the PDF; this matters for small changes such as "vs." to "versus" and for future substantive retitling.

Exploratory differences across research areas

The table counts only the 22 reviewed endpoints. "Any correspondence" includes weak correspondences and therefore should not be read as demonstrated influence.

Tracked research area	Reviewed endpoints	Later public version	Any correspondence overall	Any correspondence, conditional on update
Environment	3	3	2/3	2/3

Tracked research area	Reviewed endpoints	Later public version	Any correspondence overall	Any correspondence, conditional on update
Dev. economics / governance (LMICs)	6	3	2/6	2/3
Social impact of technology / AI	2	1	1/2	1/1
Innovation and meta-science	4	1	0/4	0/1
Global health (LMICs)	6	1	1/6	1/1
Economics, welfare, miscellaneous	1	0	0/1	—

Environment papers were more likely to have a later public version in this reviewed set, while global-health papers were less likely. Conditional response patterns cannot be ranked credibly: most area-specific denominators are one or three updated papers, and case selection partly reflects whether usable public versions could be recovered. The table is best treated as a prompt for stratified sampling as coverage expands.

What kinds of evaluation points aligned with changes?

For an initial qualitative comparison, we assigned one primary evaluation aspect to each of the nine updated cases. This simplifies evaluations that often raised several distinct issues.

Primary evaluation aspect	Cases	Strongest observed result	Pattern in this review
Correctable proof error	1	Explicit documented link	The clearest result was precise, checkable, and acknowledged publicly.
Specified additional analysis	2	Specific alignment, attribution unconfirmed	Both later papers added a closely related analysis, but neither linked it publicly to the evaluation.

Primary evaluation aspect	Cases	Strongest observed result	Pattern in this review
Robustness, confounding, or measurement check	3	Weaker correspondence	Later checks overlapped with the concern, but standard-practice explanations remained plausible.
Broad theory, mechanisms, welfare, framing, or interpretation	3	No supported substantive implementation	The later versions changed, but the broad requested extensions were not clearly carried out.

The tentative pattern is that concrete, falsifiable requests are easier both to act on and to verify. This does not show that they cause more responsiveness: specificity is confounded with feasibility, observability, and the cost of the requested work. A credible next analysis should code all validated suggestions into a common taxonomy, double-code a sample, and compare response rates within updated papers.

Pilot check for later related work

We also ran a targeted search for later work extending three updated papers where the reviewed version showed no supported substantive implementation. This is a separate outcome from revision of the evaluated paper.

Evaluated case	Public continuation checked	Pilot result
A Welfare Analysis of Policies Impacting Climate Change	The July 2026 AER version of record and the October 2025 NBER full text	The AER item is the same paper, not a separate extension. Its full text was not publicly accessible in our check, so the content assessment still rests on the public NBER revision. We did not identify a separate later extension by the author team.
The Pivot Penalty in Research	The May 2025 Nature paper, supplement, and related records	This is the final version of the evaluated research line, not later separate work. It was accepted before the public Unjournal evaluation, and we did not identify a subsequent author-team extension addressing the unmatched suggestions.

Evaluated case	Public continuation checked	Pilot result
Zero-Sum Environments, the Evolution of Effort-Suppressing Beliefs, and Economic Development	The August 2025 NBER revision and targeted title/author searches	The NBER revision is the same paper and is already included in the comparison above. We did not identify clearly attributable later work by the overlapping author team that implements the unmatched suggestions.

This pilot is deliberately phrased as "not identified," not "does not exist." Related-work tracking is harder than version tracking: titles, author teams, and empirical settings can all change. A scalable extension should record search dates and queries, require meaningful author overlap and a documented intellectual link, and keep "same paper," "later extension," and "unrelated related literature" as separate categories.

Detailed aligned review: *The Pivot Penalty in Research*

The initial comparison was too coarse because the later paper reorganized the manuscript and supplement. We therefore split the evaluated August 2024 bundle at the actual main-text/supplement boundary, aligned later sections by content rather than section number, and checked each validated suggestion against the February 2025 manuscript and the Nature main text and supplement. We also checked the Nature reporting summary, peer-review file, and source-data package for provenance and context; these are not treated as an Unjournal author response.

Evaluation suggestion	What the later documents show	Verdict
Examine promotion or funding outcomes	The evaluated version already checked whether papers acknowledged funding. Later versions do not study promotion or funding received as an outcome.	Not addressed
Account for potential benefits omitted by the outcome measures	Later discussion newly notes that citation, publication, novelty, conventionality, and applied-value measures do not exhaust a multidimensional concept of research quality. It adds no benefit outcome.	Weak framing acknowledgment only
Examine decision-making behind pivoting	The evaluated version already discussed researchers weighing pivots and making sequential choices. Later versions add no choice model or new evidence on the decision process.	Not newly addressed

Evaluation suggestion	What the later documents show	Verdict
Analyze welfare or possible misallocation	Portfolio and pre-positioning policy discussion was already present. No welfare or misallocation analysis was added.	Not addressed
Test longer horizons, including the COVID/mRNA setting	Alternative citation windows and time-trend figures were already present. No new COVID/mRNA long-horizon test was added.	Not addressed as requested
Test when in a career or sequence researchers should pivot	Existing career-stage and sequential-dynamics material does not provide the requested timing-heterogeneity test.	Not addressed
Adjust for evolving journal preferences	We found no aggregate journal-preference adjustment in the later documents.	Not addressed
Define paper pivots by fields rather than journals	The evaluated supplement already reconstructed the paper-pivot measure using L1 fields and still found a penalty. Later versions retain that robustness check and keep journals as the main definition.	Already present; main choice unchanged

The later documents do contain substantive revisions: an updated retraction sample, explicit COVID sample sizes and effect changes, reorganized methods and supplementary material, improved replication links, and the new multidimensional-quality caveat. The principal estimates and conclusions remain stable. The February manuscript predates both Nature acceptance and the public Unjournal evaluation; the Nature paper was accepted on April 18, 2025, while the public evaluation is dated May 21. There is no posted Unjournal author response. Without private transmission dates, none of these revisions should be attributed to the evaluation. We also lack the separate February supplementary files, so supplementary-material changes can be established only from August 2024 to the Nature version.

Audit of byte-identical cases (updated July 18, 2026)

An identical current PDF is potentially informative, but only if the local input is credibly the document that was evaluated. Repository history was not enough: all 17 local inputs entered this repository after their evaluations. We therefore checked official version histories, archived pre-evaluation PDFs, distinctive page-level references in the evaluations, and fresh publisher downloads.

Thirteen cases now support the limited statement **"no observed public document change as of July 18, 2026."** This does not imply that authors ignored the evaluation. They may have revised another manuscript, made changes that are not publicly available, or decided that no revision was

warranted.

Paper	Evidence binding the evaluated and current documents
Pharmaceutical Pricing and R&D as a Global Public Good	The official NBER record retains its May 2023 issue date with no revision date; a fresh publisher download has the same hash.
Towards Best Practices in AGI Safety and Governance	arXiv lists only version 1 (May 11, 2023); the current PDF is byte-identical.
Choose Your Moments: NIH Peer Review and Scientific Risk Taking	The official NBER record retains its June 2023 issue date with no revision date; a fresh publisher download has the same hash.
The Returns to Science in the Presence of Technological Risks	The evaluation identifies arXiv version 2 (May 9, 2024); arXiv lists no later version.
Effects of Emigration on Rural Labor Markets	The evaluation quotes distinctive page-level text in the October 2017 NBER PDF; the current file has the same hash.
Legalizing Entrepreneurship	The evaluation cites distinctive page-specific text, including an unusual typo, in the March 2023 revision; the current file has the same hash.
The Benefits and Costs of Guest Worker Programs	A pre-evaluation Wayback capture and the current NBER PDF are byte-identical.
The Comparative Impact of Cash Transfers and a Psychotherapy Program	A pre-evaluation Wayback capture and the current NBER PDF are byte-identical.
Maternal Cash Transfers for Child Development	A distinctive page-level quotation binds the evaluation to the August 2025 NBER revision; the current publisher PDF has the same hash.
Stagnation and Scientific Incentives	A pre-evaluation Wayback capture and the current NBER PDF are byte-identical.
Intergenerational Child Mortality Impacts of Deworming	Distinctive page and table references bind the evaluation to the original April 2023 NBER PDF; NBER lists no later revision and the current file has the same hash.
Economic versus Epidemiological Approaches to Policy Evaluation	A distinctive figure-level comparison binds the evaluation to the July 2022 NBER PDF; NBER lists no later revision and the current file has the same hash.
The Long-Run Effects of Psychotherapy on Depression, Beliefs, and Economic Outcomes	A pre-evaluation Wayback capture and a version-specific appendix reference bind the evaluation to the May 2022 NBER PDF; the current NBER file is byte-identical. An author-hosted January 2023 draft was ruled out because its corresponding appendix table contains a different analysis.

The other four initially identical pairs should not be counted as no-change findings. Two have now

been repaired by selecting the author-shared or corrected endpoint and appear in the comparison table above. One paper is known to have been revised, but the revised manuscript is not publicly available; one record had no completed evaluation. These failures show why a hash comparison without source-history checks can give the wrong answer.

Pipeline and caveats

Current scripts: `scripts/paper_response_evidence.py` for source snapshots, validation, and comparison cards; `scripts/screen_paper_response_matches.py` for deterministic, placebo-benchmarked suggestion-to-change leads; `scripts/run_paper_response_agents.py` for optional bounded, review-only model triage.

- **Historical screen shown below**: before version = PDF sent to evaluators; after version = then-current NBER/arXiv PDF, fetched April 2026. Eight pairs had meaningful text changes and a matched evaluation file.
- **Current validation**: the refresh keeps immutable PDF snapshots; verifies paper identity and chronology; produces page- and section-anchored change cards; and puts ambiguous cases into human review. A transient endpoint failure now preserves the last captured comparison with an explicit cached-snapshot warning instead of erasing it from the queue.
- **Automated correspondence check**: suggestion cues must have valid source anchors. Matches are discounted when the same concepts occur in a local window anywhere in the earlier paper and are benchmarked against matches to other papers' changes. The output is a priority queue, not a response label.
- **Models**: the historical Opus screen used bounded extracted text, so it could not reliably assess figures, tables, or all document context. The current agent triage is also deliberately non-decisive: it can identify evidence to inspect, but cannot label causal influence.
- **Interpretation**: a newer version does not imply that changes were evaluation-driven. Some papers are already being revised, and a substantial revision can have unrelated causes.

```
nn <- function(x, default = NA) if (!is.null(x)) x else default

llm_results_path <- "data/paper_change_llm_results.json"

# Signal tier labels used throughout - define once
tier_levels <- c(
  "Confirmed, LLM >=40%",
  "Confirmed, LLM <40%",
  "LLM >=30% (not confirmed)",
  "LLM 10-29%",
  "LLM <10%"
)

tier_colors_llm <- c(
  "Confirmed, LLM >=40%"      = "#1a7a47",
  "Confirmed, LLM <40%"       = "#2D9D5E",
  "LLM >=30% (not confirmed)" = "#5AE08A",
```

```
  "LLM 10-29%"                = "#a8d5b5",
  "LLM <10%"                  = "#d0e8d8"
)

if (file.exists(llm_results_path)) {
  llm_raw <- fromJSON(llm_results_path, simplifyVector = FALSE)

  llm_df <- lapply(llm_raw, function(r) {
    oa <- r$overall_assessment
    data.frame(
      paper_title      = nn(r$paper_title, ""),
      adj_status_csv   = nn(r$adj_status, ""),
      deposit_after_uj = isTRUE(r$deposit_after_uj),
      text_chg_pct     = nn(r$text_length_change_pct, NA_real_),
      additions        = nn(r$additions_count, NA_integer_),
      deletions        = nn(r$deletions_count, NA_integer_),
      n_major_changes  = length(nn(r$major_changes, list())),
      n_suggestions    = length(nn(r$evaluator_suggestions, list())),
      pct_influenced   = if (!is.null(oa)) nn(oa$pct_likely_influenced, NA_real_) else NA_real
      attr_confidence  = if (!is.null(oa)) nn(oa$confidence, NA_real_) else NA_real_,
      narrative        = if (!is.null(oa)) nn(oa$narrative, "") else "",
      eval_found       = !is.null(r$eval_file) && nchar(nn(r$eval_file, "")) > 0,
      skipped          = !is.null(r$skipped_reason),
      stringsAsFactors = FALSE
    )
  }) |> bind_rows()

  llm_analyzed <- llm_df |>
    filter(!skipped & eval_found & n_major_changes > 0 & !is.na(pct_influenced))

  llm_enriched <- llm_analyzed |>
    left_join(df |> select(paper_title, research_area, pubpub_link), by = "paper_title") |>
    mutate(
      manual_tier = case_when(
        adj_status_csv == "evidence of updating"          ~ "Confirmed update",
        adj_status_csv == "Stated intention to update"    ~ "Stated intention",
        adj_status_csv == "mixed evidence/minor updating" ~ "Mixed / minor",
        adj_status_csv == "Unlikely to update"            ~ "Unlikely",
        TRUE                                              ~ "Unclassified"
      ),
      combined_tier = case_when(
        manual_tier == "Confirmed update" & pct_influenced >= 40 ~ tier_levels[1],
        manual_tier == "Confirmed update"                        ~ tier_levels[2],
        pct_influenced >= 30                                     ~ tier_levels[3],
        pct_influenced >= 10                                     ~ tier_levels[4],
        TRUE                                                     ~ tier_levels[5]
      ),
```

```
      combined_tier = factor(combined_tier, levels = tier_levels)
    )

  n_analyzed      <- nrow(llm_enriched)
  n_total_fetched <- nrow(llm_df)
  med_pct         <- median(llm_enriched$pct_influenced, na.rm = TRUE)
  n_high          <- sum(llm_enriched$pct_influenced >= 30, na.rm = TRUE)
} else {
  llm_df <- llm_analyzed <- llm_enriched <- data.frame()
  n_analyzed <- n_total_fetched <- n_high <- 0
  med_pct <- NA_real_
}
```

```
if (nrow(llm_enriched) > 0) {
  n_confirmed <- sum(llm_enriched$manual_tier == "Confirmed update")
  n_llm_high  <- sum(llm_enriched$manual_tier == "Unclassified" &
                       llm_enriched$pct_influenced >= 30)
  cat(glue::glue(
    "Of {n_total_fetched} papers with fetchable DOIs, {n_analyzed} had meaningful text ",
    "changes and a matched evaluation. {n_confirmed} are manually confirmed updates; ",
    "of the remaining {n_analyzed - n_confirmed}, the LLM finds {n_llm_high} with at least 30% ",
    "of changes likely driven by evaluator feedback (median across all: {round(med_pct)}%)."
  ))
}
```

Of 22 papers with fetchable DOIs, 8 had meaningful text changes and a matched evaluation. 3 are manually confirmed updates; of the remaining 5, the LLM finds 2 with at least 30% of changes likely driven by evaluator feedback (median across all: 22%).

```
if (nrow(llm_enriched) > 0) {
  # Build one row per paper with tier assignment
  waffle_llm <- llm_enriched |>
    select(paper_title, combined_tier) |>
    mutate(group = as.character(combined_tier))

  waffle_manual <- df |>
    filter(!paper_title %in% llm_enriched$paper_title) |>
    mutate(group = case_when(
      adj_status_clean == "Evidence of updating"       ~ "Confirmed (manual only)",
      adj_status_clean == "Stated intention to update" ~ "Stated intention",
      adj_status_clean == "Mixed / minor updating"     ~ "Mixed / minor",
      adj_status_clean == "Unlikely to update"         ~ "Unlikely",
      TRUE                                             ~ "Not yet assessed"
    )) |>
    select(paper_title, group)

  waffle_all_groups <- c(tier_levels,
    "Confirmed (manual only)", "Stated intention",
```

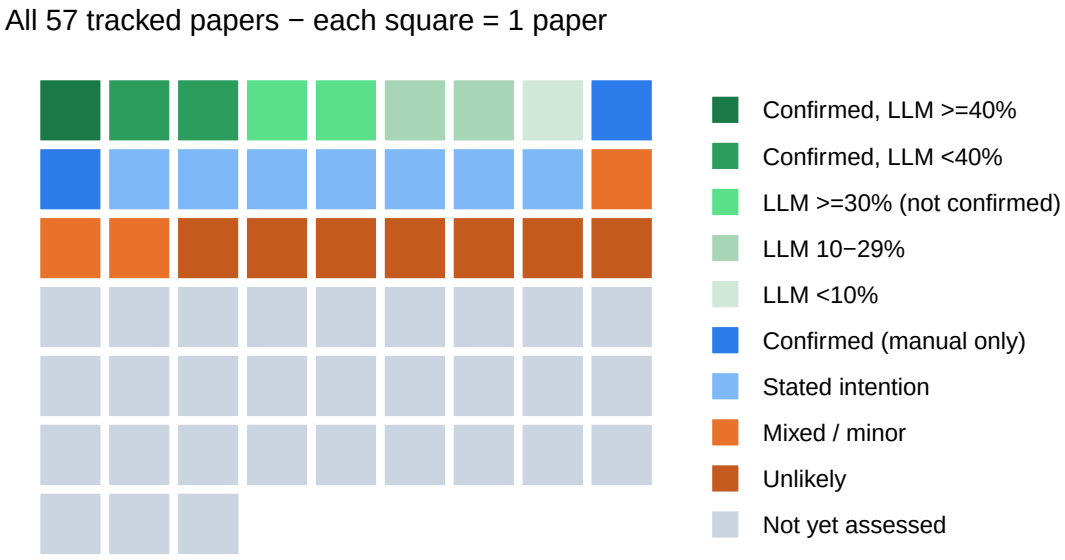
```
    "Mixed / minor", "Unlikely", "Not yet assessed")

  waffle_colors <- c(
    tier_colors_llm,
    "Confirmed (manual only)" = "#2B7CE9",
    "Stated intention"        = "#7EB8F7",
    "Mixed / minor"           = "#E8722A",
    "Unlikely"                = "#C45A1E",
    "Not yet assessed"        = "#CBD5E1"
  )

  waffle_data <- bind_rows(waffle_llm, waffle_manual) |>
    mutate(group = factor(group, levels = waffle_all_groups)) |>
    arrange(group) |>
    mutate(rank = row_number(),
           x = (rank - 1) %% 9,
           y = -((rank - 1) %/% 9))

  ggplot(waffle_data, aes(x = x, y = y, fill = group)) +
    geom_tile(color = "white", linewidth = 1.5) +
    coord_equal() +
    scale_fill_manual(values = waffle_colors, name = NULL,
                      drop = FALSE) +
    theme_void() +
    theme(legend.position = "right",
          legend.text = element_text(size = 9),
          plot.title = element_text(size = 11, face = "plain",
                                    margin = margin(b = 6)),
          plot.caption = element_text(size = 8, color = "grey40",
                                      hjust = 0, margin = margin(t = 8))) +
    labs(title = "All 57 tracked papers - each square = 1 paper",
         caption = "LLM % = share of major post-evaluation changes attributed to evaluator feed
}
```

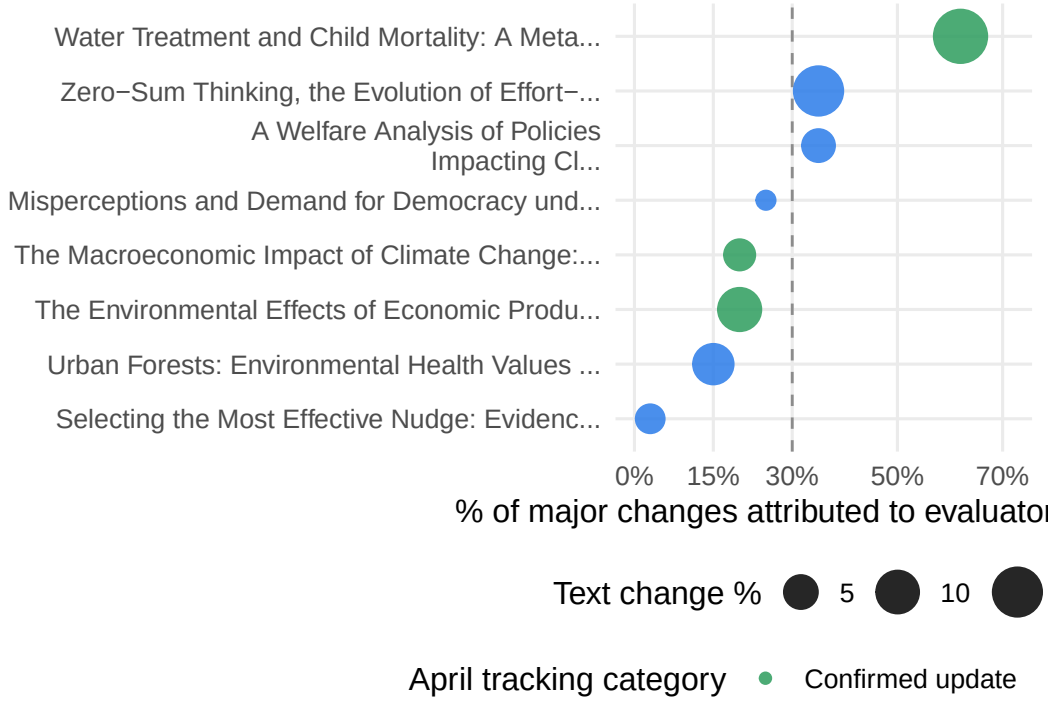
Figure 4: April tracking snapshot for all 57 papers (each square = 1 paper); adjustment status had been assessed for 22. Green tones: papers where LLM analysis was run, shaded by attribution score and recorded confirmation status. Blue: recorded updates without LLM analysis. Orange/grey: stated intention, weak signal, or not yet assessed.




```
if (nrow(llm_enriched) > 0) {
  dot_colors <- c("Confirmed update" = "#2D9D5E", "Unclassified" = "#2B7CE9")

  llm_enriched |>
    mutate(paper_short = str_trunc(paper_title, 46)) |>
    ggplot(aes(x = pct_influenced, y = reorder(paper_short, pct_influenced),
               color = manual_tier, size = text_chg_pct)) +
    geom_vline(xintercept = 30, linetype = "dashed", color = "grey55", linewidth = 0.5) +
    geom_point(alpha = 0.85) +
    scale_color_manual(values = dot_colors, name = "April tracking category") +
    scale_size_continuous(range = c(3, 9), name = "Text change %") +
    scale_x_continuous(limits = c(0, 72), labels = \(x) paste0(x, "%"),
                       breaks = c(0, 15, 30, 50, 70)) +
    labs(x = "% of major changes attributed to evaluator feedback", y = NULL) +
    theme_uj() +
    theme(legend.position = "bottom", legend.box = "vertical",
          legend.margin = margin(0, 0, 0, 0))
}
```

Figure 5: LLM attribution for the 8 papers with pre/post versions compared. x-axis: estimated % of major changes attributed to evaluator feedback (Claude Opus 4.6, conservative). Point size scales with text change magnitude. Green = manually confirmed update; blue = not yet manually classified. Dashed line at 30%.



Summary Statistics

```
n_total      <- nrow(df)
n_assessed   <- df |> filter(adj_status_clean != "Not yet assessed") |> nrow()
n_evidence   <- df |> filter(adj_status_clean == "Evidence of updating") |> nrow()
n_intention  <- df |> filter(adj_status_clean == "Stated intention to update") |> nrow()
n_any_positive <- df |>
  filter(adj_status_clean %in% c("Evidence of updating",
                                 "Stated intention to update",
                                 "Mixed / minor updating")) |> nrow()
n_formal  <- df |> filter(author_response_clean == "Formal response") |> nrow()
n_any_resp <- df |>
  filter(author_response_clean %in% c("Formal response", "Informal response")) |> nrow()
```

Of 57 tracked evaluations, 22 have an assessed adjustment status. Among these: 5 are coded as showing clear evidence of substantive updating, 7 as stating an intention to update, and 15 as showing at least some positive signal. Response status is recorded for only part of the set: 19 papers have a recorded formal or informal response, of which 16 are formal written responses.

```
resp_colors <- c(
  "Formal response"   = "#2B7CE9",
  "Informal response" = "#5DA3F5",
  "No response"       = "#CBD5E1",
  "Not yet coded"     = "#E2E8F0"
```

Table 14: Papers with LLM attribution results, sorted by % of changes attributed to evaluator feedback. 'Confirmed' = manually verified by Unjournal staff. Signal tier: 'Confirmed, LLM >=40%' = manually confirmed AND LLM finds strong corroboration; 'Confirmed, LLM <40%' = confirmed but weaker LLM signal; 'LLM >=30% (not confirmed)' = unconfirmed but LLM finds likely influence.

```
if (nrow(llm_enriched) > 0) {
  tier_bg <- c(
    "Confirmed, LLM \u226540%"      = "#c8e6c9",
    "Confirmed, LLM <40%"           = "#d4edda",
    "LLM \u226530% (not confirmed)" = "#e8f5e9",
    "LLM 10\u201329%"               = "#fff3e0",
    "LLM <10%"                      = "#f8fafc"
  )

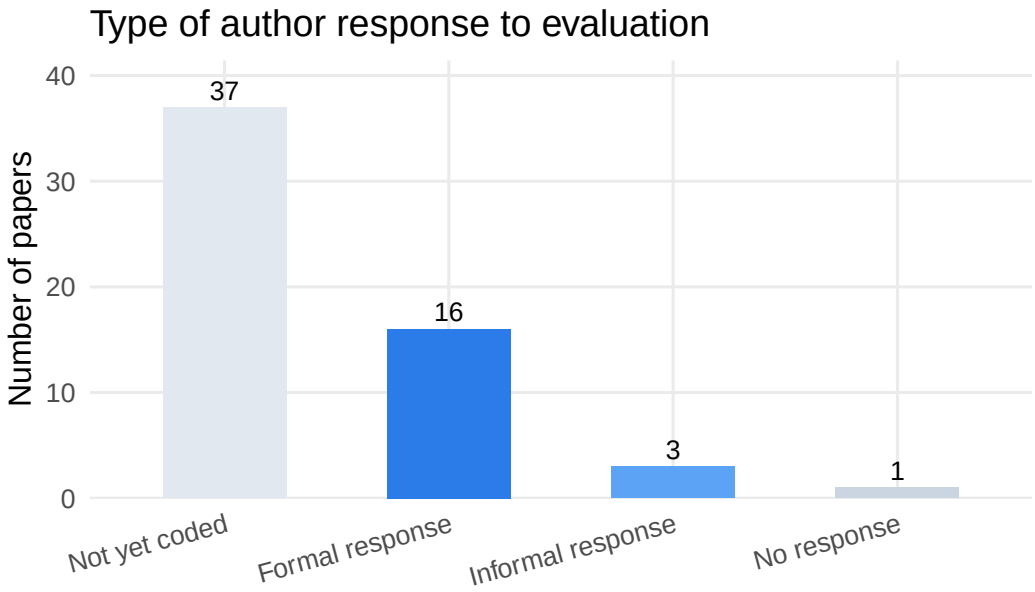
  tbl_llm <- llm_enriched |>
    arrange(desc(pct_influenced)) |>
    mutate(
      Title = ifelse(
        !is.na(pubpub_link) & nchar(pubpub_link) > 0,
        paste0('<a href="', pubpub_link, '" target="_blank">',
               str_trunc(paper_title, 60), '</a>'),
        str_trunc(paper_title, 60)
      ),
      `Manual`      = manual_tier,
      `LLM %`       = paste0(round(pct_influenced), "%"),
      `Conf.`       = paste0(attr_confidence, "/5"),
      `Signal tier` = combined_tier,
      `LLM narrative` = str_trunc(narrative, 350)
    ) |>
    select(Title, Manual, `LLM %`, Conf., `Signal tier`, `LLM narrative`)

  datatable(
    tbl_llm,
    escape   = FALSE,
    rownames = FALSE,
    options  = list(
      pageLength = 5,
      dom        = "tip",
      scrollX    = TRUE,
      columnDefs = list(
        list(width = "18%", targets = 0),
        list(width = "11%", targets = 1),
        list(width = "6%",  targets = 2),
        list(width = "5%",  targets = 3),
        list(width = "16%", targets = 4),
        list(width = "44%", targets = 5)
      )
    )
  ) |>
  formatStyle("Signal tier",
    backgroundColor = styleEqual(names(tier_bg), unname(tier_bg)))
} else {
  cat("*No LLM attribution results available yet.*")
}
```

```
)

df |>
  count(author_response_clean) |>
  ggplot(aes(x = reorder(author_response_clean, -n), y = n,
             fill = author_response_clean)) +
  geom_col(width = 0.55, show.legend = FALSE) +
  geom_text(aes(label = n), vjust = -0.4, size = 3.5) +
  scale_fill_manual(values = resp_colors) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.12))) +
  labs(x = NULL, y = "Number of papers",
       title = "Type of author response to evaluation") +
  theme_uj() +
  theme(axis.text.x = element_text(angle = 15, hjust = 1))
```

Figure 6: Type of author response to Unjournal evaluation. 'Not yet coded' indicates the response status has not been assessed for those papers (37 blanks in data); this is distinct from confirmed non-response (only 1 paper is explicitly coded as no response).



Full Tabulation

All 57 tracked evaluations

Click column headers to sort; use the search box to filter. Adjustment status is color-coded: green = positive signal, orange = minor/uncertain, grey = unlikely or not yet assessed.

Table 15

```
tbl <- df |>
  mutate(
    title_linked = if_else(
      !is.na(pubpub_link) & pubpub_link != "",
      paste0('<a href="', pubpub_link, '" target="_blank">',
             str_trunc(paper_title, 72), '</a>'),
      str_trunc(paper_title, 72)
    ),
    authors_short = str_trunc(authors, 40),
    nb_icon = if_else(
      !is.na(nb_link) & nb_link != "",
      paste0('<a href="', nb_link, '" target="_blank" title="NotebookLM">&#128214;</a>'),
      ""
    )
  ) |>
  select(Paper = title_linked, Authors = authors_short,
         Area = research_area, Response = author_response_clean,
         Adjustment = adj_status_clean, `Pub.` = pub_status, NB = nb_icon)

adj_bg <- c(
  "Evidence of updating"        = "#d4edda",
  "Stated intention to update"  = "#e8f5e9",
  "Mixed / minor updating"      = "#fff3e0",
  "Unlikely to update"          = "#fce4d4",
  "Not yet assessed"            = "#f8fafc"
)

datatable(tbl, escape = FALSE, rownames = FALSE, filter = "top",
  options = list(
    pageLength = 8, autoWidth = FALSE, scrollX = TRUE, dom = "lfrtip",
    columnDefs = list(
      list(width = "36%", targets = 0),
      list(width = "18%", targets = 1),
      list(width = "8%",  targets = 2),
      list(width = "11%", targets = 3),
      list(width = "16%", targets = 4),
      list(width = "8%",  targets = 5),
      list(width = "3%",  targets = 6)
    )
  ),
  caption = "All tracked Unjournal evaluations. NB = NotebookLM analysis link."
) |>
  formatStyle("Adjustment",
    backgroundColor = styleEqual(names(adj_bg), unname(adj_bg)))
```

King and Worthington (2026): implications for this project

AI-generated research note · October 23, 2026

Written by OpenAI Codex at David Reinstein's request, based on the October 22, 2026 paper and supplementary appendix. This is an AI-generated interpretation and proposed research agenda, not a new empirical result or an agreed revision of the team's plans.
Updated November 5, 2026 by Claude (Opus 5.5), following David Reinstein's annotations on the paper: sample sizes, the editors' expertise, the model generation, task differences, and a comparison with our own agreement data.

Assessment. The study by King and Worthington (2026) substantially overlaps with this project's motivation and weakens novelty claims based on comparing AI with unusually good, paid human reviewers. The existing Unjournal study still contributes a transparent baseline for structured research ratings and critique agreement. The strongest case for further work is to establish when AI judgments are correct, calibrated, and useful for consequential research decisions, and what additional human effort buys.

What the study establishes

King and Worthington analyze 124 rapidPeer submissions from 2019–2026, with 134 commissioned human reviews. They report generating reviews with Claude Opus 4.8 at maximum reasoning effort, using either the original reviewer instructions without tools or a tailored prompt with web retrieval. GPT-5.5 judges the comparisons; these model names and settings are as reported in the paper. Three former journal editors also judge ten randomly selected review pairs. The human reviewers were paid, carefully selected, and asked for comprehensive, constructive feedback (main text, sections 2–3).

Only 10 of the 124 submissions received two human reviews, so most comparisons set one AI review against a single human review. The scale is similar to ours: 47 human-matched papers, but 39 of them have two or more evaluators (2.1 on average), which is what allows the human–human comparisons below. Claude Opus 4.8 has since been superseded. Their results can be read as a lower bound on what current models achieve, though that is an interpretation rather than something they test. The same point applies more strongly to our January 2026 GPT-5 Pro runs.

The AI reviews win every decided overall comparison in the headline AI-judged analysis and 26 of 30 editor judgments (87%). Those 30 judgments concern ten pairs assessed by three editors, not 30 independent manuscripts. The editors see the reviews without the manuscripts (supplement, Appendix A). Their choices therefore support the perceived quality of the reviews, but do not independently verify their technical claims against the underlying research. Former journal editors may also lack the field knowledge to judge which technical critiques in a specialized paper are correct or important. That concern applies less to the authors, 88% of whom (of 74 responding) rated their human review as high or very high quality, but the authors rated only the human review, not the comparison with AI.

King and Worthington also report that the editors agreed with one another on 80% of judgments, and with the AI judge on 87% on average. Our data allow a related check on ratings of the papers themselves rather than on which review is better. On the 39 papers with two or more evaluators, GPT-5 Pro's overall rating has a rank correlation of 0.63 with the human mean, against 0.33 between individual human evaluators; an approximate adjustment for comparing against a mean rather than an individual gives 0.51 (Results). The adjusted difference has a wide interval that

includes zero, so this does not show the model agrees with humans more than humans agree with each other.

The evidence goes beyond preferences over fluent prose. Checks address review length, rewriting human reviews in AI style, judging with manuscript access, submission year, stochastic variation, and different generating and judging model families. A blinded audit flags 5.5% of AI claim assessments and 15.9% of human claim assessments as contradicted by the manuscript; a human check agrees on 87 of 90 sampled claims. Appendix F.7 and Table S6 define this error measure narrowly: opinions, suggestions, and claims not verifiable from the manuscript remain in the denominator without counting as errors. It is not a measure of the validity or usefulness of every critique. Moreover, 46 of 124 AI reviews contain at least one flagged error. The controlled-error experiment shows detection of explicit inserted contradictions; it does not establish equivalent sensitivity to subtle failures of identification or external validity.

Version detail. The main text reports 132/132 decided overall AI comparisons, whereas Figure 1 labels this 134/134. This note preserves the common qualitative finding and flags the denominator discrepancy rather than silently resolving it. Sources: paper, sections 2–4 and Figure 1; supplement, Appendices A and F.4–F.10.

Where the Unjournal project differs

Ratings of research and their uncertainty. Both services ask for constructive, whole-paper feedback. rapidPeer asks reviewers mainly for comprehensive suggestions to help authors revise. The Unjournal also asks evaluators for percentile ratings with uncertainty intervals, journal-tier predictions, and an assessment of the paper's main claims; the claim assessments may overlap with ordinary improvement suggestions, but the ratings have no rapidPeer counterpart. The working paper compares seven percentile ratings, uncertainty intervals, and journal-tier predictions under the same rubric. King and Worthington's seven headline areas assess the quality of reviews; they are not seven ratings of the underlying papers. A convincing report can coexist with compressed scores, unstable rankings, or poorly calibrated forecasts. Our current agreement estimates describe these patterns; they do not establish true research quality or calibrated uncertainty.

Correct, important critiques and complementary information. Our critique analysis asks which human concerns the model recovers and which model concerns have no human match. This is a different outcome from overall review preference, but it is not yet a decisive advantage: the current analysis has only 12 eligible pairings and uses an LLM judge. An unmatched model critique can be valuable, and a shared critique can be wrong. The useful extension is blinded, independent adjudication of correctness, severity, actionability, and genuinely new information for both human and AI concerns, including checks of data or code where relevant. Their claim audit makes a stronger validation standard necessary for ours.

Prospective predictions and decisions. The proposed pipeline can register predictions before human evaluations or publication outcomes are available. It should also check earlier public versions and reviews; a new intake date alone does not prevent contamination. Publication venues are verifiable outcomes, with their own selection mechanisms and biases, rather than ground truth about scientific merit. Research prioritization adds a separate question: which work deserves scarce evaluation effort because it could inform important decisions? Agreement with internal prioritization is descriptive evidence, not proof of social value.

Human–AI collaboration and downstream effects. King and Worthington explicitly leave publication decisions, improvements in subsequent revisions, and collective effects unmeasured. Pro-

posed Unjournal trials could compare unaided human evaluation, AI-only evaluation, and human evaluation with AI assistance. Relevant outcomes include independently validated issue detection, harmful false positives, evaluator time, author revisions, and usefulness to research users. Author satisfaction and textual revision alone would not establish scientific improvement. These trials and broader impact measures remain proposed work.

The combination of these questions is a credible contribution. Paid expert reports, broad social-science relevance, or a simple no-tools baseline alone are weak differentiators: rapidPeer already provides substantial overlap on those features. The current study's smaller sample and older model runs also limit claims about today's best achievable review quality. Its findings and rapidPeer's can both hold because the tasks, models, and outcome measures differ.

Proposed priorities for further research

1. **Validate a focused set of consequential critiques.** Use independent domain experts, blinded to review source, to adjudicate AI-only, human-only, and shared concerns. Preserve disagreements and distinguish factual contradictions from contested methodological judgments. Prioritize serious errors and useful additions over another overall preference score.
2. **Register a prospective comparison.** Freeze manuscripts, prompts, model identifiers, access to tools and supplementary material, predictions, and scoring rules. Compare a current fixed-rubric baseline with a stronger retrieval-assisted approach on held-out papers. Cluster uncertainty by paper when multiple reviews or judges share a manuscript; report ties, missing judgments, and sensitivity to judge choice.
3. **Estimate the value of human effort after AI assistance.** In a randomized pilot, measure additional validated issues and useful decisions per unit of evaluator time and total cost, including adjudication and correction costs. Specify contamination controls, assignment units, and primary outcomes before collecting results.
4. **Test preferences and resistance to manipulation.** Vary prestige cues, prompt framing, and adversarial instructions while holding research content fixed. Measure whether assistance homogenizes judgments or suppresses useful minority objections. These address concerns King and Worthington also raise; they are motivated extensions, not untouched topics.

Funding implication (AI assessment). A smaller, well-validated prospective study would now be easier to justify than a large expansion devoted mainly to repeating human–LLM agreement comparisons. The existing proposal's structured measurement, adjudication, prospective validation, prioritization, and collaboration components remain useful, but their value depends on executing them. This note does not change the proposal's budget or claim those studies have been completed.

See the current working paper, discussion, and broader project and grant proposal. The working paper remains the source for current results; the broader site preserves earlier project context.