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

Valentin Klotzbücher David Reinstein Lorenzo Pacchiardi
Tianmai Michael Zhang

2026-04-27

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. Our results suggest that even a minimal one-shot setup—a single prompt with a fixed rubric, no iteration or retrieval augmentation—yields LLM ratings comparable to an additional expert rater, though central compression and uneven qualitative coverage indicate clear limitations. Appendix results for five additional models confirm the pattern across capability tiers.

Introduction

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

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.), IsItCredible (IsItCredible.com n.d.), and QED Science (QED Science 2026)—already market automated referee-like feedback directly to authors, though they explicitly disclaim substituting for human peer review and their internal architectures remain undocumented. OpenAIReview (Hsu and Tan 2026) takes a different approach: an open-source, transparent pipeline that uses progressive prompting to generate detailed critiques at roughly $4 per paper. Meanwhile, publishers are formalizing policies that restrict reviewer use of general-purpose AI tools while permitting controlled in-house applications for screening tasks (Elsevier 2025; Leung 2026). Our study complements these efforts by asking how

far the simplest possible setup—a single prompt to a frontier model, with no bespoke pipeline—can go.

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. 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 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? 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 the resulting quantitative ratings and qualitative critiques against expert evaluations for 60 economics and social-science working papers. We ask whether a frontier LLM evaluation can approximate expert judgment, where systematic differences arise, and whether the model reveals characteristic AI preferences over research. Our headline finding is that GPT-5 Pro matches or exceeds pairwise human inter-rater rank agreement on overall quality—that is, the LLM-human correlation is comparable to the human-human correlation, which is the appropriate benchmark given substantial disagreement among human evaluators themselves. Appendix results for five additional models spanning different capability and cost tiers confirm the pattern. This suggests that top reasoning models can currently serve as supplementary raters in structured evaluation pipelines, even under our minimal one-shot setup.

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 makes our results a conservative lower bound on what LLM-based evaluation can currently achieve. If frontier models already yield meaningful agreement with expert reviewers under the simplest possible setup, more sophisticated pipelines—structured measurement schemas (Asirvatham, Mokski, and Shleifer 2026), iterative quality-checking workflows (Zhang and Abernethy 2025), or the kind of prompt-robustness engineering motivated by specification-search concerns (Asher et al. 2026)—should improve further. Quantifying how much headroom remains above this one-shot baseline, and which pipeline elements unlock it, is a key direction for future work.

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; 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[1] 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.

> **i** Unjournal impact: author engagement evidence
>
> Do authors engage with and respond to Unjournal evaluations? Manual tracking of 57 evaluations finds that 16 papers received formal written responses and at least 5 show clear evidence of substantive revision in response to the feedback—including one paper with over 3,000 net line changes. A further 7 authors stated an intention to update their paper.
> 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 45 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.

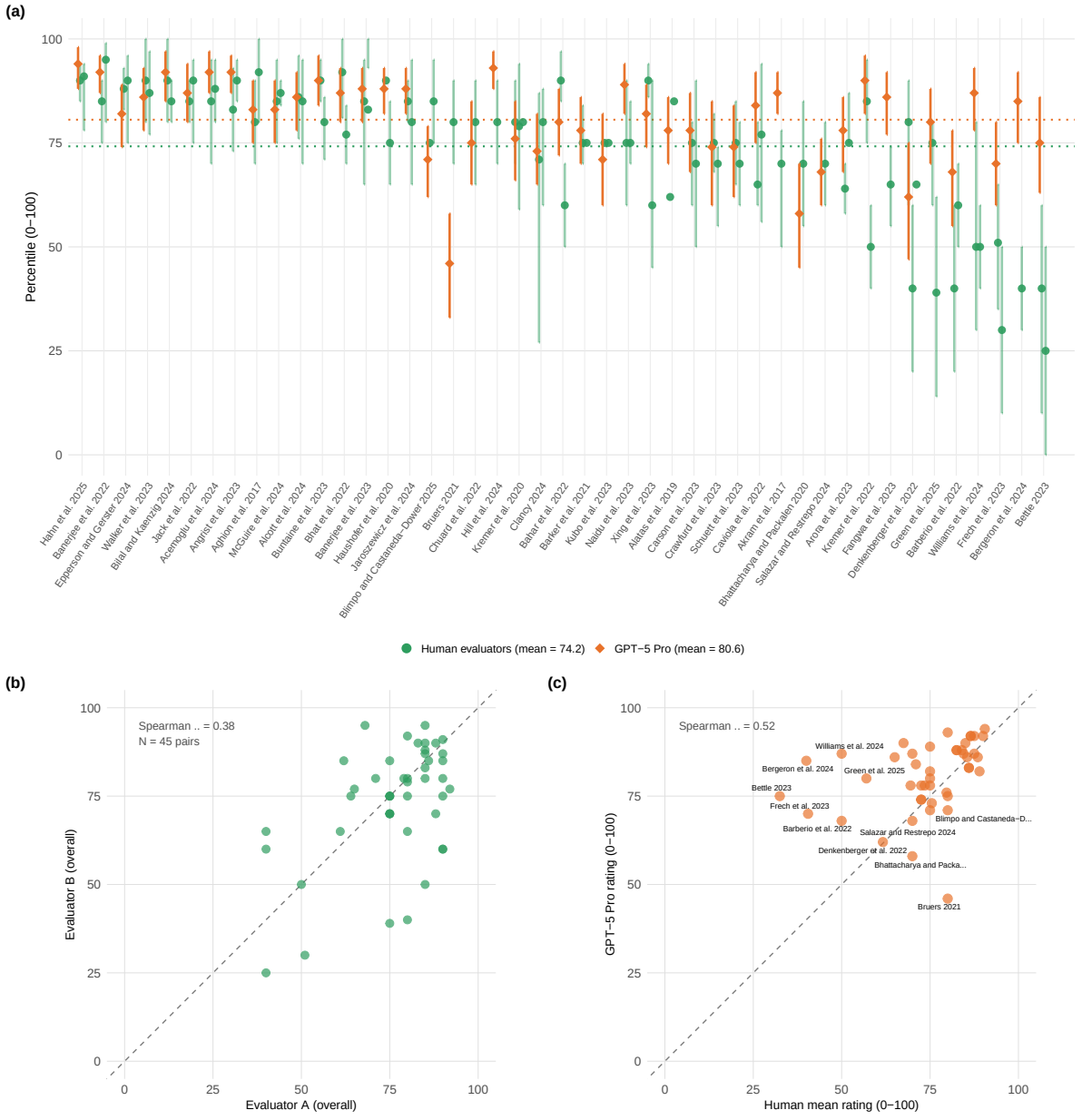
We do not treat human ratings as ground truth. Quantitative percentile scoring is genuinely difficult: even domain experts disagree, and individual scores reflect both signal about paper quality and idiosyncratic tendencies (severity, topic familiarity, interpretation of the scale). Our question is whether an LLM provides signal *comparable to an additional expert rater*. The Human–Human baseline row in Table 1 provides this reference. **Caution**: the LLM's is computed against the *mean* of 1.9 human raters, which reduces noise and inflates apparent agreement relative to the individual-vs-individual _HH. The Spearman-Brown adjusted column corrects for this; the fair comparison is adj. vs. _HH. Krippendorff's $_{\text{HH}}$ 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, making the human-human agreement ceiling directly visible. Panel (c) compares GPT-5 Pro ratings against human mean ratings with per-paper labels.

Panel (b) makes the human-human ceiling directly visible: the scatter of evaluator pairs is no tighter than the GPT-5 Pro–human scatter in panel (c), and individual pairs disagree by as much as 30–40 percentile points on individual papers. 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 pattern consistent with alignment training that discourages extreme outputs. 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.

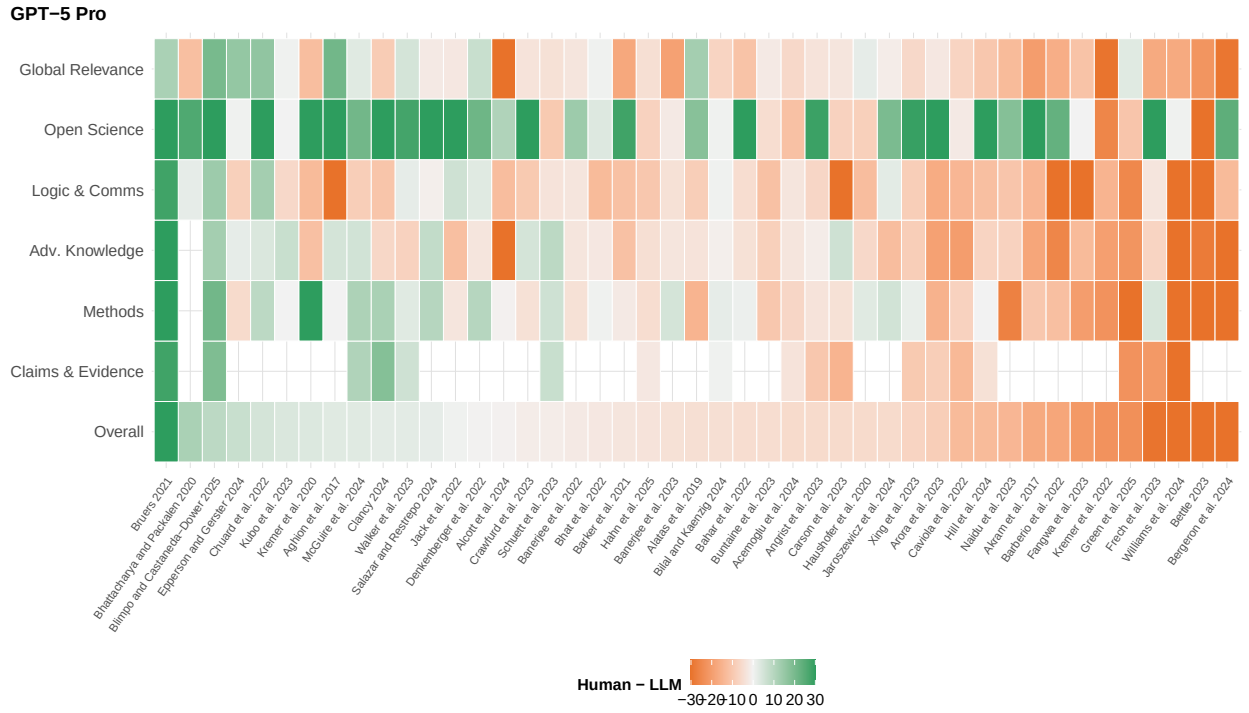
[1] These represent verifiable publication outcomes, not statements about the "true quality" of the paper.

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. Compare panels (b) and (c) directly to see whether LLM-human scatter is tighter than human-human scatter.



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 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.

Qualitative critique comparison. 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 precision—the fraction of LLM-raised issues that correspond to a substantive human concern. These metrics are assessed by GPT-5.2 Pro acting as an independent judge (see Methods for the judging protocol).

Coverage and precision vary substantially across papers. Some papers achieve high scores on both dimensions, indicating strong alignment between human and LLM critiques, while others reveal the LLM missing key human concerns or raising issues absent from the expert consensus. 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; manual validation

Table 1: GPT-5 Pro agreement with human mean overall rating (N = `r n_focal` matched papers). **Human–Human** row (bold) shows pairwise individual-vs-individual Spearman as a reference. **(vs. mean)**: raw Spearman between GPT-5 Pro and human *mean*—upward-biased because the mean suppresses noise. **adj. (SB)**: Spearman-Brown corrected to individual-rater-equivalent (×`r sprintf('%.2f', sb_factor)` for k `r round(k_raters, 1)` raters/paper); compare this to Human–Human = `r hh_spearman`. **95% CI**: Fisher-z. Bias = LLM − Human.

Model	N	(vs. mean)	adj. (SB)	95% CI	Pearson r	Mean bias	MAE
Human–Human	**37**	**0.432**	—	**[0.13, 0.66]**	**0.565**	**+4.1**	**11.0**
GPT-5 Pro	45	0.517	0.444	[0.26, 0.70]	0.342	+6.4	10.3

Table 2: GPT-5 Pro agreement with human mean by criterion (N = `r n_focal` matched papers). **H-H** shows pairwise human evaluator Spearman (the reference for each criterion; see Table 1 note about mean-vs-individual bias). Positive bias = GPT-5 Pro rates higher on average. Note the variation: human evaluators agree strongly on some dimensions and barely at all on others (Open Science).

Criterion	H-H	Spearman	Pearson r	Mean bias	RMSE
Overall	**0.432**	0.517	0.342	+6.4	15.4
Claims & Evidence	**0.400**	0.388	0.213	+4.0	18.3
Methods	**0.362**	0.536	0.339	+4.5	18.5
Adv. Knowledge	**0.179**	0.459	0.303	+7.9	17.6
Logic & Comms	**0.225**	0.276	0.187	+9.6	16.5
Open Science	**-0.033**	0.125	0.166	-17.0	27.7
Global Relevance	**0.130**	0.429	0.345	+6.1	16.0

through our annotation tool is underway and preliminary results are consistent with the automated scores.

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 with the human mean on overall ratings, alongside the Human–Human baseline as a reference. The Spearman-Brown adjusted column (adj.) is the appropriate comparator to _HH; see Appendix A for all six models. Full model comparison is in Appendix A.

Table 2 breaks the same agreement metrics down by evaluation criterion for GPT-5 Pro. The **H-H** column shows pairwise human-human Spearman for each criterion as a reference — note that raw LLM is upward-biased relative to H-H by the mean-vs-individual asymmetry (see Table 1). Criteria where H-H is itself low indicate genuine expert disagreement; low LLM agreement on those criteria is therefore expected rather than a model failure.

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 achieves moderate-to-strong agreement with human expert evaluators, approaching the agreement levels observed among human evaluators themselves. Central compression—the tendency to pull extreme ratings toward the middle of the scale—is the most consistent pattern, likely reflecting alignment training that discourages confident extreme outputs. 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 for five additional models confirm these patterns across capability tiers, with reasoning-capable models outperforming lightweight ones.

Limitations. Several caveats temper these conclusions. Our sample comprises roughly 50 social-science papers 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 that caps achievable 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. Alignment training likely contributes to score inflation and narrower credible intervals than humans provide. All LLM evaluations are single-run; aggregating across multiple runs or temperature settings could change the picture. Finally, the qualitative coverage and precision metrics 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 cannot be guaranteed without end-to-end encryption or on-premise deployment. 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. Out-of-time validation on papers entering The Unjournal's pipeline after model training cutoffs would

eliminate residual contamination concerns. The qualitative coverage and precision metrics are themselves LLM-assessed (GPT-5.2 Pro as judge); human validation of these scores is needed to close the model-dependence loop. Finally, comparing human and LLM tier predictions against verified publication venues would offer a rare opportunity for externally verifiable accuracy measurement.

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."[2] 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. All papers have completed The Unjournal's full evaluation process, meaning the authors received evaluations that have been publicly posted. For our analysis 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.

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 and is sensitive to proportional biases but not to constant offsets. *Spearman's* 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 ceiling: if human evaluators agree with each other at only $\alpha = 0.5$ on a given

[2]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.

criterion, expecting an LLM to exceed that level would be unrealistic. We then report α_{HL} (between the human mean and each LLM) to assess how close machine ratings come to this ceiling. 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, and *precision* denotes the fraction of LLM-generated issues that matched at least one human issue.

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.

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 precision estimates (fraction of LLM issues that are genuinely substantive).

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, precision, matched-pair explanations, and an overall quality rating. Together, these outputs feed the qualitative analysis presented in Appendix B: Critiques & Key Issues.

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).

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" (Feb. 19, 2026).

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

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

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

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).

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

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.

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.

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?

> **Note**
>
> 57 Unjournal evaluations tracked as of April 2026, drawing on two evidence streams: (a) manual staff assessment of author responses and paper revisions; (b) Claude Opus 4.6 change-attribution for 8 papers where pre/post PDFs were available. See Combined Evidence below.

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 manual classification of 57 tracked evaluations and, for a subset, automated PDF diffing and LLM attribution. Background: The Unjournal's knowledge base.

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

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

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 combined: manual classification for all 57 papers, and Claude Opus 4.6 LLM attribution for 8 papers where before/after PDFs were available and meaningful text changes were detected.

> **i** Pipeline and caveats
>
> **Scripts**: `scripts/fetch_latest_papers.py` → `scripts/run_paper_change_llm.py`
> - **Before version**: PDF sent to Unjournal evaluators (`papers/`); **After version**: latest

NBER/arxiv version, fetched April 2026

- **Model**: Claude Opus 4.6, temperature 0.2; conservative instructions — only flags "direct" or "indirect" attribution with clear conceptual alignment to a specific evaluator suggestion
- **Coverage**: 30 papers had fetchable DOIs; 8 had meaningful changes (15 line edits or 0.5% size change) and a matched evaluation file; the 3 manually-confirmed updates were also LLM-analyzed
- **Temporal caveat**: a newer version does not imply changes were evaluation-driven; some papers were already in revision
- **LLM limitations**: text truncated at 60,000 chars/version; figures and tables not visible to the model

```
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 \u226540%",
  "Confirmed, LLM <40%",
  "LLM \u226530% (not confirmed)",
  "LLM 10\u201329%",
  "LLM <10%"
)

tier_colors_llm <- c(
  "Confirmed, LLM \u226540%"      = "#1a7a47",
  "Confirmed, LLM <40%"           = "#2D9D5E",
  "LLM \u226530% (not confirmed)" = "#5AE08A",
  "LLM 10\u201329%"               = "#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(
```

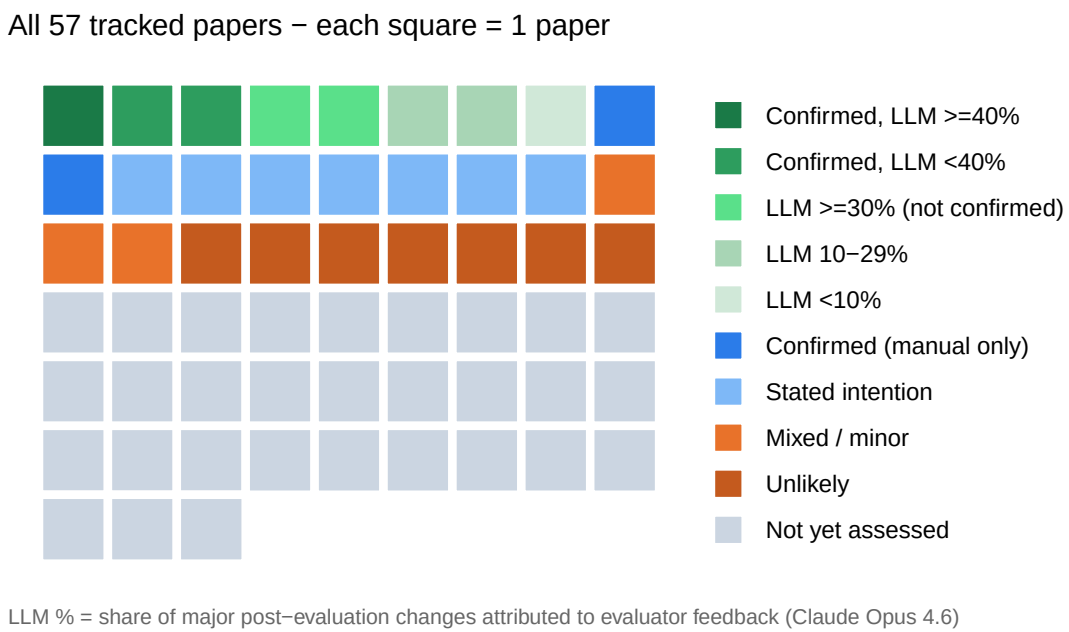
```
    "Of {n_total_fetched} papers with fetchable DOIs, {n_analyzed} had meaningful text ",
    "changes and a matched evaluation. {n_confirmed} are manually confirmed updates; ",
    "of the remaining {n_analyzed - n_confirmed}, the LLM finds {n_llm_high} with \u226530% ",
    "of changes likely driven 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 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
}
```

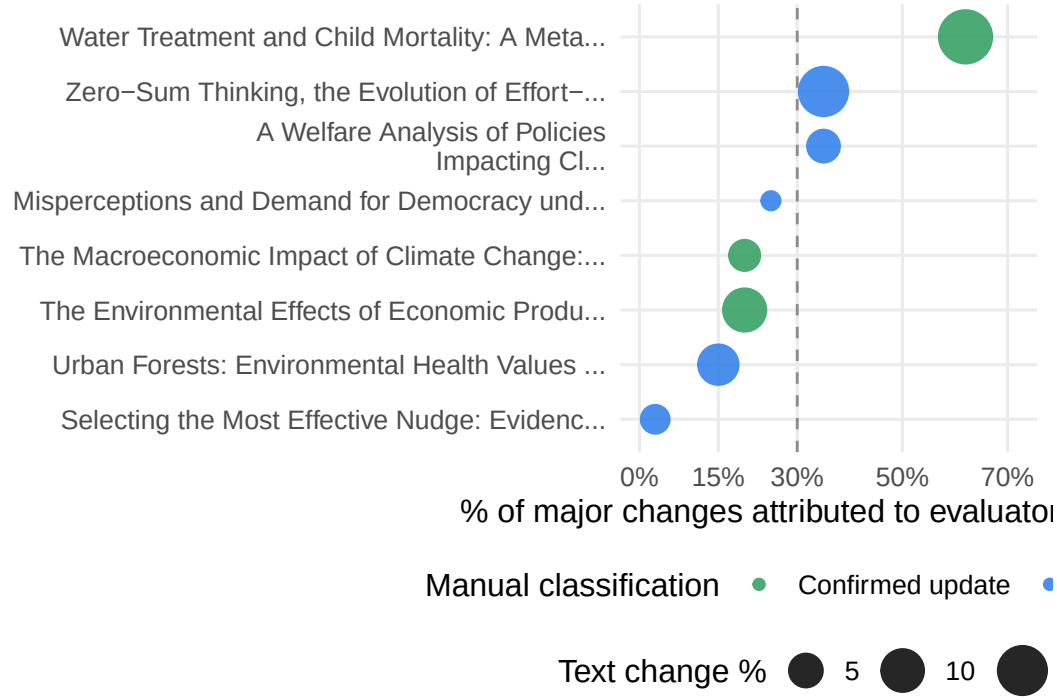
Figure 3: All 57 tracked papers by combined evidence tier (each square = 1 paper). Green tones: papers where LLM analysis was run, shaded by attribution score and manual confirmation status. Blue: manually confirmed updates without LLM analysis. Orange/grey: stated intention, weak signal, or not yet 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 = "Manual classification") +
    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 4: 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%.




```
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.*")
}
```

Summary Statistics

```
n_total      <- nrow(df)
n_assessed   <- df |> filter(adj_status_clean != "Not yet assessed") |> nrow()
```

Table 3: 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.

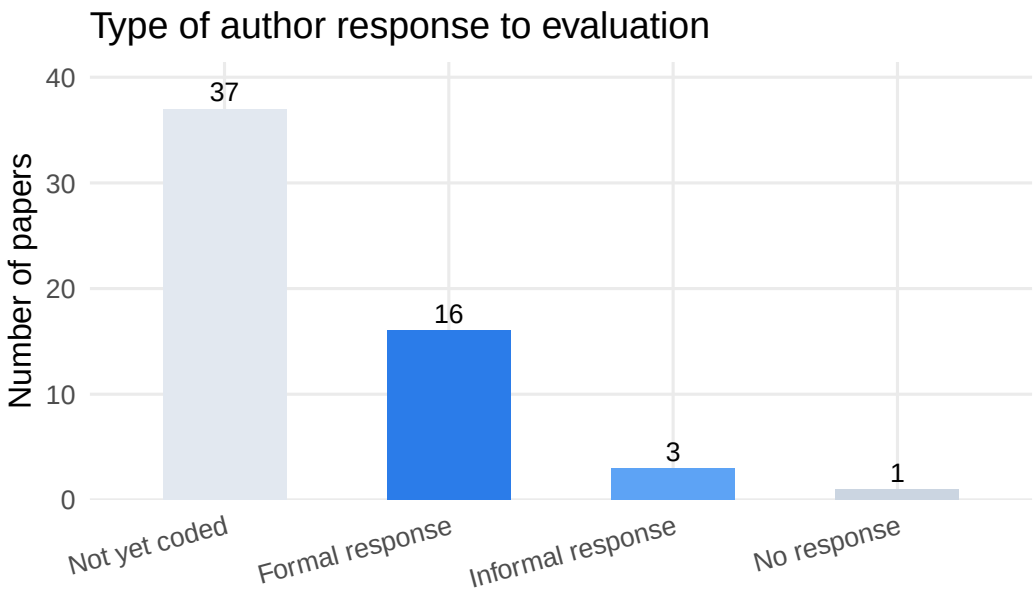
Title	Manual LLM %	Conf. Signal tier	LLM narrative

```
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 been manually assessed. Among these: 5 show clear evidence of substantive updating, 7 authors stated intention to update, and 15 show at least some positive signal. On engagement: 19 papers received a formal or informal author response, of which 16 were formal written responses.

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

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 5: 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.

```r
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)))
```

Table 4

All tracked Unjournal evaluations. NB = NotebookLM analysis link.

Paper	Authors	Area	Response	Adjustment	Pub.	NB
All	All	All	All	All	All	All