# Comparing LLM and human reviews of social science research using data from Unjournal.org

David Reinstein     Valentin Klotzbücher     Tianmai Michael Zhang

2025-11-30

We will build and refine LLM tools to generate peer-reviews and ratings of impactful research, and compare these with human experts' work (esp. from Unjournal.org): to benchmark performance, understand AI's research taste, and develop tools to improve research evaluation and dissemination.

# Table of contents

# 1 Introduction

> ⚠ Work in progress
>
> Pages, metrics, and comparisons are under active development. Expect rough edges and frequent updates.

Is AI good at peer-reviewing? Does it offer useful and valid feedback? Can it predict how human experts will rate research across a range of categories? How can it help academics do this "thankless" task better? Is it particularly good at spotting errors? Are there specific categories, e.g. spotting math errors or judging real-world relevance, where it does surprisingly well or poorly? How does its "research taste" compare to humans?

If AI research-evaluation works it could free up a lot of scientific resources – perhaps $1.5 billion/year in the US alone Aczel, Szaszi, and Holcombe (2021)) – and offer more continual and detailed review, helping improve research. It could also help characterize methodological strengths/weaknesses across papers, aiding training and research direction-setting. Furthermore, a key promise of AI is to directly improve science and research. Understanding how AI engages with research evaluations may provide a window into its values, abilities, and limitations.

In this project, we are testing the capabilities of current large language models (LLMs), illustrating whether they can generate research paper evaluations comparable to expert human reviews. The Unjournal systematically prioritizes 'impactful' research and pays for high-quality human evaluations, structured quantified ratings, claim identification and assessment, and predictions. In this project, we use an AI (OpenAI's `GPT-5 Pro` model) to review social science research papers under the same criteria used by human reviewers for The Unjournal.

Each paper is assessed on specific dimensions – for example, the strength of its evidence, rigor of methods, clarity of communication, openness/reproducibility, relevance to global priorities, and overall quality. The LLM will provide quantitative scores (with uncertainty intervals) on these criteria and produce a written evaluation

Our initial dataset will include the 5 research papers that have existing Unjournal human evaluations. For each paper, the AI will generate: (1) numeric ratings on the defined criteria, (2) identification of the paper's key claims, and (3) a detailed review discussing the paper's contributions and weaknesses. We will then compare the AI-generated evaluations to the published human evaluations.

In the next phase, we will focus on papers currently under evaluation, i.e., where no human evaluation has been made public, to allow us to rule out any contamination.

### 1.0.1 Our work in context

Luo et al. (2025) survey LLM roles from idea generation to peer review, including experiment planning and automated scientific writing. They highlight opportunities (productivity, coverage of long documents) alongside governance needs (provenance, detection of LLM-generated content, standardizing tooling) and call for reliable evaluation frameworks.

Eger et al. (2025) provide a broad review of LLMs in science and a focused discussion of AI-assisted peer review. They argue: (i) peer-review data is scarce and concentrated in CS/OpenReview venues; (ii) targeted assistance that preserves human autonomy is preferable to end-to-end reviewing; and (iii) ethics and governance (bias, provenance, detection of AI-generated text) are first-class constraints.

Zhang and Abernethy (2025) propose deploying LLMs as quality checkers to surface critical problems instead of generating full narrative reviews. Using papers from WITHDRARXIV and an automatic evaluation framework that leverages "LLM-as-judge," they find the best performance from top reasoning models but still recommend human oversight.

Pataranutaporn et al. (2025) asked four nearly state-of-the-art LLM models (GPT-4o mini, Claude 3.5 Haiku, Gemma 3 27B, and LLaMA 3.3 70B) to consider 1220 unique papers "drawn from 110 economics journals excluded from the training data of current LLMs". They prompted the models to act "in your capacity as a reviewer for [a top-5 economics journal]" and make a publication recommendation using a 6-point scale ranging from "1 = Definite Reject..." to "6. Accept As Is...". They asked it to evaluate each paper on a 10-point scale for originality, rigor, scope, impact, and whether it was 'written by AI'. They also (separately) had LLMs rate 330 papers with the authors' identities removed, or replacing the names with fake male/female names and real elite or non-elite institutions (check this) or with prominent male or female economists attached.

They compare the LLMs' ratings with the RePEC rankings for the journals the papers were published in, finding general alignment. They find mixed results on detecting AI-generated papers. In the names/institutions comparisons, they also find the LLMs show biases towards named high-prestige male authors relative to high-prestige female authors, as well as biases towards elite institutions and US/UK universities. (Doublecheck the details here).

There have been several other empirical benchmarking projects, including work covered in LLM4SR: A Survey on Large Language Models for Scientific Research and Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation. (We will discuss these here.)

Zhang et al. (2025)

- AI conference paper data
- "employs LLM agents to perform pairwise comparisons among manuscripts"
- "significantly outperforms traditional rating-based methods in identifying high-impact papers" [by citation metrics]
- Some evidence of biases/~statistical discrimination based on characteristics like 'papers from established research institutions'

Our project distinguishes itself in its use of *actual* human evaluations of research in economics and adjacent fields, past and *prospective*, including both reports, ratings, and predictions.[1] The Unjournal's 50+ evaluation packages enable us to train and benchmark the models. Their pipeline of future evaluations allow for clean out-of-training-data predictions and evaluation. Their detailed written reports and multi-dimensional ratings also allows us to compare the 'taste', priorities, and comparative ratings of humans relative to AI models across the different criteria and domains. The 'journal tier prediction' outcomes also provides an external ground-truth[2] enabling a human-vs-LLM horse race. We are also planning multi-armed trials on these human evaluations (cf. Brodeur et al, 2025 and Qazi et al, 2025) to understand the potential for *hybrid* human-AI evaluation in this context.

Footnote, a fancier way to say this, from a grant application? Or from chatGPT?[3]

[1] Other work has relied on collections of research and grant reviews, including NLPEER, SubstanReview, and the Swiss National Science Foundation. That data has a heavy focus on computer-science adjacent fields, and iss less representative of mainstream research peer review practices in older, established academic fields. Note that The Unjournal commissions the evaluation of impactful research, often from high-prestige working paper archives like NBER, and makes all evaluations public, even if they are highly critical of the paper.

[2] About verifiable publication outcomes, not about the 'true quality' of the paper of course.

[3] Our approach differs from prior work by (i) focusing on structured, percentile-based quantitative ratings with credible intervals that map to decision-relevant dimensions used by The Unjournal; (ii) comparing those ratings to published human evaluations rather than using LLM-as-judge; and (iii) curating contamination-aware inputs (paper text extraction with reference-section removal and token caps), with a roadmap to add multi-modal checks when we score figure- or table-dependent criteria.

# 2 Data and methods

We draw on two main sources:

1) Human evaluations from The Unjournal's public evaluation data (PubPub reports and the Coda evaluation form export).

2) LLM-generated evaluations using a structured JSON-schema prompt with `gpt-5-pro-2025-10-06` (knowledge cut-off: 30 September 2024).

## 2.1 Unjournal.org evaluations

We use The Unjournal's public data for a baseline comparison. At The Unjournal each paper is typically evaluated (aka 'reviewed') by two expert evaluators[1] who provide quantitative ratings on a 0–100 percentile scale for each of seven criteria (with 90% credible intervals),[2] two "journal tier" ratings on a 0.0 - 5.0 scale,[3] a written evaluation (resembling a referee report for a journal), and identification and assessment of the paper's "main claim". For our initial analysis, we extracted these human ratings and aggregated them, taking the average score per criterion across evaluators (and noting the range of individual scores).

All papers have completed The Unjournal's evaluation process (meaning the authors received a full evaluation on the Unjournal platform, which has been publicly posted at unjournal.pubpub.org). The sample includes papers spanning 2017–2025 working papers in development economics, growth, health policy, environmental economics, and related fields that The Unjournal identified as high-impact. Each of these papers has quantitative scores from at least one human evaluator, and many have multiple (2-3) human ratings.

## 2.2 LLM-based evaluation

### 2.2.1 Quantitative ratings and journal-ranking tiers

Following The Unjournal's standard guidelines for evaluators and their academic evaluation form, evaluators are asked to consider each paper along the following dimensions: **claims & evidence**,

[1] Occasionally they use 1 or 3 evaluators.

[2] See their guidelines here; these criteria include "Overall assessment", "Claims, strength and characterization of evidence", "Methods: Justification, reasonableness, validity, robustness", "Advancing knowledge and practice", "Logic and communication", "Open, collaborative, replicable science", and "Relevance to global priorities, usefulness for practitioners"

[3] "a normative judgment about 'how well the research should publish' " and "a prediction about where the research will be published"

**methods**, **logic & communication**, **open science**, **global relevance**, and an **overall** assessment. Ratings are interpreted as percentiles relative to serious recent work in the same area. For each metric, evaluators are asked for the midpoint of their beliefs and their 90% credible interval, to communicate their uncertainty. For the journal rankings measure, we ask both "what journal ranking tier should this work be published in? (0.0-5.0)" and "what journal ranking tier will this work be published in? (0.0-5.0)", with some further explanation.The full prompt can be seen in the code below – essentially copied from the Unjournal's guidelines page.

We captured the versions of each paper that was evaluated by The Unjournal's human evaluators, downloading from the links provided in The Unjournal's Coda database.

We evaluate each paper by passing the PDF directly to the model and requiring a strict, machine-readable JSON output. This keeps the assessment tied to the document the authors wrote. Direct ingestion preserves tables, figures, equations, and sectioning, which ad-hoc text scraping can mangle. It also avoids silent trimming or segmentation choices that would bias what the model sees.

We enforce a JSON Schema for the results. The model must return one object for each of the named criteria including a midpoint rating and a 90% interval for each rating. This guarantees that every paper is scored on the same fields with the same types and bounds. It makes the analysis reproducible and comparisons clean.

We request credible intervals (as we do for human evaluators) to allow the model to communicate its uncertainty rather than suggest false precision; these can also be incorporated into our metrics, penalizing a model's inaccuracy more when it's stated with high confidence.

Relying on GPT-5 Pro, we use a single-step call with a reasoning model that supports file input. One step avoids hand-offs and summary loss from a separate "ingestion" stage. The model reads the whole PDF and produces the JSON defined above. We do not retrieve external sources or cross-paper material for these scores; the evaluation is anchored in the manuscript itself.

The Python pipeline uploads each PDF once and caches the returned file id keyed by path, size, and modification time. We submit one background job per PDF to the OpenAI Responses API with "high" reasoning effort and server-side JSON-Schema enforcement. Submissions record the response id, model id, file id, status, and timestamps.

A separate script polls job status and, for each completed job, retrieves the raw response, extracts the first balanced top-level JSON object, and writes both the raw response and parsed outputs to disk.

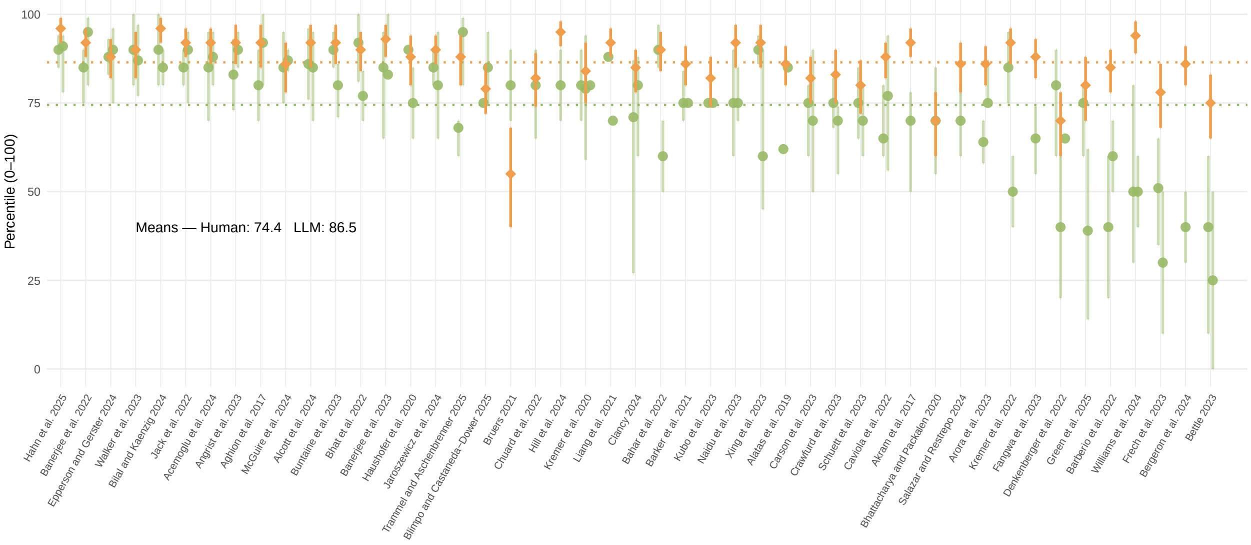
# 3 Results

Here we present preliminary results, starting with a comparison of the LLM-generated quantitative ratings (model: `gpt-5-pro`, see the(previous section) with human evaluations across the Unjournal's criteria.

## 3.1 Quantitative comparison: human vs. GPT-5 Pro (initial run)

We first use the earlier GPT-5 Pro evaluation run that covered all papers in our Unjournal sample with a simpler JSON-schema prompt. Figure 3.1 shows the overall percentile ratings from this initial run, averaged across human evaluators and compared to the LLM's "overall" scores for each paper.

Figure 3.1: Comparison of Human vs LLM overall percentile ratings

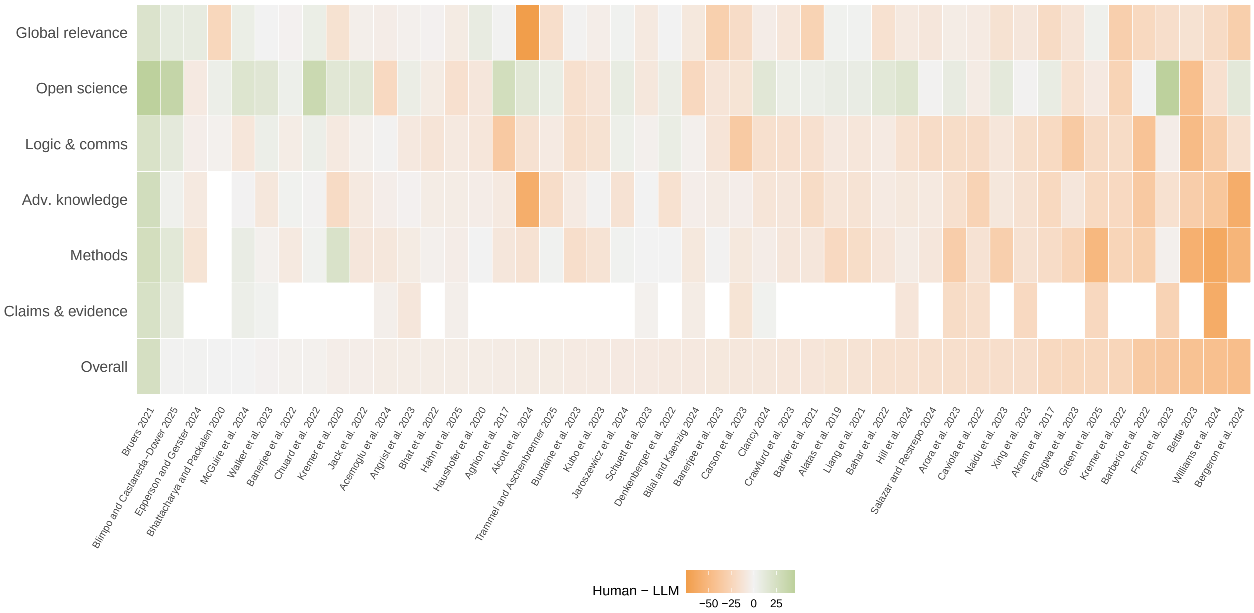


**?@fig-heatmap-human-minus-llmshows** a heatmap of the differences between human and LLM mean ratings across all evaluation criteria. Positive values (in green) indicate that humans rated the paper higher than the LLM, while negative values (in orange) indicate the opposite.

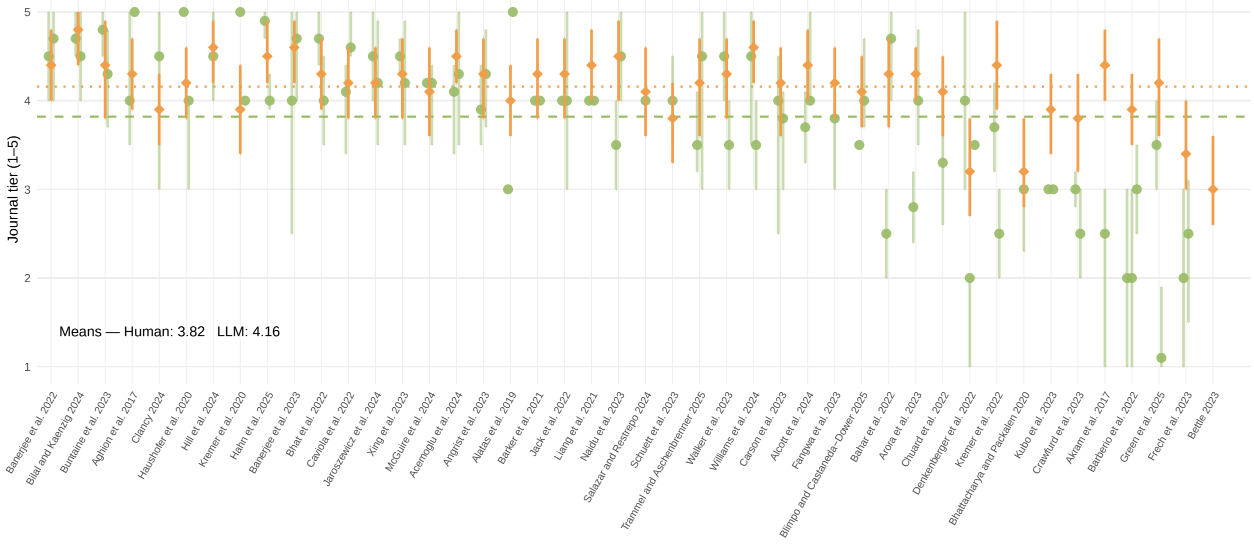
## 3.2 Qualitative comparison: detailed GPT-5 Pro evaluations

To understand what GPT-5 Pro is actually responding to, we re-ran the model on five focal papers (Adena and Hager 2024; Kudymowa et al. 2023; Peterman et al. 2025; Green et al. 2025; Williams et al. 2024) using a refined prompt.

## Figure 3.2: Heatmap of Human minus LLM mean ratings across evaluation criteria



## Figure 3.3: Comparison of Human vs LLM journal tier ratings (should be published in)



This second run keeps the same quantitative metrics but additionally requires a diagnostic summary of about 1,000 words and high-effort reasoning, with the full reasoning trace returned by the "thinking" model. For each paper we can therefore inspect:

- the LLM's quantitative scores and journal-tier predictions,
- the hidden reasoning steps used to arrive at those scores, and
- the token usage and approximate API cost of the evaluation.

We start by examining the Williams et al. (2024) evaluation in detail and then show the analogous summaries for the other four focal papers. In the next step we will juxtapose these LLM assessments with the human evaluators' written reports.

### 3.2.1 Qualitative comparison: Williams et al. (2024)

In the refined run, GPT-5 Pro reads about 28,704 input tokens and produces 6,327 visible output tokens plus 5,120 reasoning tokens.
At current API prices this evaluation costs roughly $1.80.

The table below shows the model's percentile ratings and 90% credible intervals for the Unjournal criteria.

| Criterion | Midpoint | Lower 90% | Upper 90% |
|---|---|---|---|
| Overall assessment | 86 | 86 | 86 |
| Claims & evidence | 78 | 78 | 78 |
| Methods | 74 | 74 | 74 |
| Advancing knowledge and practice | 92 | 92 | 92 |
| Logic and communication | 84 | 84 | 84 |
| Open, collaborative, replicable science | 63 | 63 | 63 |
| Relevance to global priorities | 94 | 94 | 94 |

| Measure | Score | Lower 90% | Upper 90% |
|---|---|---|---|
| Deserved journal tier (should) | 4.4 | 4.4 | 4.4 |
| Predicted journal tier (will) | 4.8 | 4.8 | 4.8 |

Model reasoning trace – Williams et al. 2024

**Analyzing model limitations**
I see another limitation in the model's assumptions. They assumed stationarity from 2000-2016 when predicting for 2015 to 2030. They included present conditions like 2018 tree cover but didn't factor in future climate or socio-economic changes. By excluding socio-economic predictors, they reduced potential model improvements slightly, but this could bias predictions in areas where socio-economic factors play a significant role, potentially leading to unrealistic outcomes in "biophysical potential." They clarify it's about modeling potential, not feasibility.
**Examining dataset limitations**
I notice another issue with the primary regrowth dataset used, which came from Global

Forest Watch to map regrowth over 5 m. In humid tropics, the producer's accuracy is low at just 18.7% because of omission errors, although user accuracy is high at 85.1%. The detection biases may skew the training sample and affect model calibration. Their conservative approach to mapping regrowth likely favors more accessible and larger areas, and many small regrowth patches are excluded. It's unclear if the training sample is representative across different biomes without proper stratification.

**Clarifying domain definitions**

I need to define the domain used in the study. They included all parts of forest biomes within ±25° but excluded water, bare land, urban areas, and so on, while purposely opting for a liberal domain. They used 2018 tree cover to exclude forests but may misclassify certain savannas or drylands as part of forest biomes. While they mention not including savannas, their definition could still capture mosaic-savanna zones. Additionally, there's a scale mismatch since predictions at a finer 30 m resolution are computed alongside coarser predictors, which could impact accuracy.

**Assessing the weighted area approach**

I see another issue with the weighted area approach, where the probability of regrowth times pixel area is calculated. However, the random forest classification is trained on presence/absence from 2000-2016, and the output probabilities aren't properly calibrated. Calibration techniques like Platt scaling or isotonic regression are necessary to interpret results as probabilities. They also didn't present AUC, only accuracy, which may mislead since balanced sampling artificially sets prevalence to 0.5. Without proper probability calibration, the estimate of expected area could be questionable and not reflect actual conditions.

**Reviewing definitions and calibration**

They used specific definitions, claiming that "the continuous potential" can be seen as the probability of natural regeneration per pixel, with the weighted-area value representing the expected area for regeneration. However, they didn't perform any calibration on this. Their accuracy and out-of-bag accuracy are reported as 0.878, but they failed to discuss probability calibration, which is important. Regarding carbon sequestration potentials, they calculated totals and ranges from previous datasets, but it's unclear where the range comes from—possibly from Monte Carlo methods or carbon dataset quartile variation, yet they didn't show how they computed uncertainty.

**Identifying dataset uncertainties**

The dataset presents uncertainties by grid cell, but their method multiplies 30-year totals without mentioning how uncertainty is handled. The "range" they provide likely comes from minimum and maximum values across these uncertainties, which is unclear. They need to clarify how uncertainty propagates and address independence and omitted below-ground carbon, since they reference adding 22–28% from literature. Additionally, they didn't consider factors like degradation baseline, leakage, or permanence. While they frame the "biophysical potential," the abstract could mislead policy implications with its stated above-ground carbon sequestration potential of 23.4 Gt C over 30 years.

**Checking carbon sequestration claims**

I noticed they mentioned that "new carbon sinks could increase global sequestration by 14.3% per year" but also stated "mitigating current pantropical forest carbon losses by 90.5% per year," which might be inconsistent. They claim assisted natural regeneration can sequester significant carbon but didn't show how they calculated this. It seems to derive from 23.4 Gt over 30 years, equating to 0.78 Gt C per year. With a net carbon loss of about 0.86 Gt C per year, I need to verify Baccini 2017's values, which reported a net carbon source of 1.1 Gt C per year. My memory on this is a bit fuzzy!

**Evaluating carbon loss mitigation claims**

They state that "mitigating current pantropical forest carbon losses by approximately 90.5% per year," referencing Baccini, raises some questions. If regeneration can offset about 0.8 Gt per year, that suggests total losses of around 0.9 Gt per year. While this might be plausible, there are logical leaps involved since their potential benefits might not actually be realized. It's a bit tricky to interpret, yet they present it confidently as potential. I need to consider these assumptions carefully!

**Analyzing dataset issues**

They wisely exclude plantation regrowth from their analysis, but it seems their training dataset might still contain misclassified plantations. They reported a previous algorithm with 90.6% accuracy; however, an area-based producer accuracy of just 18.7% in humid tropics suggests significant under-detection. This non-random classification error could undermine generalizability. While they argue for the robustness of the detected patches, the validation accuracy of 87.9% relies on potentially biased training dataset presence points. Field validation is unlikely, and poor accuracy in Southeastern Asia raises another red flag.

**Evaluating modeling assumptions**

The modeling approach assumes that predicted 'potential' is valid from 2015 to 2030 without considering changes, which is questionable since the climate is constantly shifting. Additionally, the "distance to forest" will likely change as regrowth occurs, yet they didn't model this dynamic feedback. They mentioned underestimating impacts like positive feedback and the exclusion of savanna/dryland. Climate change could also hinder regrowth, leading to ambiguity. They present a figure showing "available for restoration" area by country, but they don't clarify how this was calculated, leaving some mystery around the methodology used.

**Analyzing definitions and reproducibility**

They define the study's domain as all tropical forest biomes, excluding areas like water, urban, and existing forests, while including factors like 'distance to water' as covariates. It seems that "available for restoration" refers to the total area of this domain, but their disclaimers indicate some harvest/regrowth classifications might be included. Although they mention using a forestry dataset, they didn't clearly define "available for restoration." Additionally, requesting code via email isn't ideal; it should be publicly accessible. While they use accessible datasets, the lack of code hampers reproducibility, and the confidence intervals aren't credible.

**Evaluating accuracy and methodology**

They mention the model overcame potential overfitting with an out-of-bag (OOB) accuracy of 87.8%. They used a balanced sample and omitted socioeconomic features due to similar accuracy, but I wonder if the limited role of socioeconomic data in predicting presence versus absence, as indicated by remote sensing, could affect results. They also disclaim that the maps are only one input, which is good. Since this is published in Nature, it seems high quality, but I believe we should look beyond that. We need to propose credible metrics.

**Ranking evaluation criteria**

Based on the content, I would rank high—maybe around the 85th percentile overall.

For claims and evidence, they articulate strong claims and provide robust mapping and a predictive model. However, there are issues like reliance on one training dataset with major omissions and unrealistic confidence intervals. I'd place this at about 75.

For methods, they've implemented random forest modeling and accounted for autocorrelation, yet there's heavy reliance on coarse data. This leads me to perhaps a 70.

Regarding advancing knowledge, their maps on natural regeneration potential are valuable, globally relevant for restoration planning, so I'd mark this as a 90.

On logic and communication, while the writing is clear and disclaimers are present, there are leaps in concluding about carbon losses and some misuses of "CI." I'd score this around 82.

For open science, while the data is available on Zenodo and the input data is open, the code is only available on request, which limits replicability. I'd give a score of 65.

Finally, in terms of global relevance, this work ties in well with global priorities and quantifies information by country, so I'd rate this a solid 92.

**Assigning tier ratings**

I need to assign "tier_should" and "tier_will" ratings from 0 to 5. Based on quality, I think "should" deserves a high score, perhaps around 4.5 or 4.7, due to some identified issues. For "will," since it's already published in Nature, I'd predict around 4.6 to 4.8. However, they want continuous numbers, so I'll set "should" at 4.4 and "will" at 4.7.

If I consider possible underestimation due to method-related issues, perhaps "should" could be 4.2 or 4.4, while "will" could be 4.8 because of strong interest. Still, the dataset not being open-coded is a consideration, but it's published in Nature, so I want to remain fair.

Now, I need to craft a summary, aiming for around 1000 words. I'll start by summarizing the approach and identifying major issues.

**Highlighting dataset issues**

I'm noticing that the training dataset has biases, especially regarding natural regrowth mapping, which shows significant omissions in humid biomes. This can lead to a non-random sample of presences, and the model might learn from incomplete detections, distorting variable relationships and potentially mispredicting areas.

Next, regarding probability calibration and the "expected area" method, they treat random forest outputs as probabilities, but without proper calibration and baseline prevalence, the resulting estimates of 215 million hectares could be inaccurate. They present highly precise confidence intervals that don't seem credible because they didn't account for uncertainty across inputs.

Lastly, concerning spatial autocorrelation, while they tried to evaluate it, accuracy tends to decline with distance, so their optimism might not be justified. They didn't apply block cross-validation and reported overall accuracy, avoiding metrics like AUC. The balanced classification gives an unrealistic sense of accuracy, and there are no confusion matrices or biome-specific metrics presented beyond overall accuracy.

**Addressing scale and domain issues**

I see that there's a scale mismatch with the 30 m predictions derived from 300 m to 1 km predictors, which could cause smoothing and artifacts. They've acknowledged this with visual evidence in Fig. 1, suggesting it could impact local planning as they consider these results to be preliminary.

The domain definition appears overly broad, including areas where natural regeneration may be undesirable—like non-forest ecosystems, productive farmland, or communities—which could potentially mislead their "available for restoration" metrics.

Lastly, they assume conditions from 2000 to 2016 without modeling climate change or socioeconomic factors, which could really undermine their 2030 predictions. Also, the carbon accumulation overlay from the Cook-Patton dataset has a scale mismatch issue, with 1 km data being downscaled to 30 m without considering edges and heterogeneity.

**Evaluating carbon accounting and code accessibility**

They don't initially account for below-ground carbon, indicating an additional 22–28% should be included, and they don't consider the risk of reclearance. Their discussions on leakage and additionality seem inadequate since their yearly rates are used to support claims of significant offsetting of losses, which likely exaggerates realized mitigation.

While they exclude plantations, algorithmic misclassification might still exist, and they haven't incorporated a classification probability step.

Regarding code accessibility, it's not publicly available, making replication difficult. Although they provide predicted maps on Zenodo, the training and modeling code needing to be requested further hampers replicability. Additionally, using 2018 tree cover for distance predictions while training on 2000 cover could lead to variable shifts, potentially degrading the predictions since they haven't cross-validated changes.

**Observing assumptions and suggestions for improvement**

The weighted area assumption sums across all pixels to determine "expected area," but it overlooks spatial dependence and positive feedback, which are mentioned but not integrated. The country-level summarization is good, although it uses administrative boundaries without addressing ecoregion differences, and there's a lack of error budget for these country-level sums.

They note that predictions may degrade in Southeastern Asia, but don't confirm this for later years. For improvements, I should suggest block cross-validation, probability calibration via isotonic regression, and incorporating socioeconomic variables.

Overall, I thought the score should be around 85, but I'm reconsidering an 86.

**Finalizing ranking and summary considerations**

I'm setting the overall score to 86, considering it's intended to reflect percentiles within [0, 100]. For research in remote sensing global mapping, I'd lean toward an 88, but due to certain issues, I might adjust this to 82. However, advancing knowledge could be rated higher at 90, leading me to average around 86, so choosing 87 seems fair.

I also need to fill out the JSON carefully, ensuring that the assessment summary is a concise paragraph of about 800-1100 words and omitting any author names or external knowledge outside the provided PDF. The confidence intervals will match the midpoint of 86. For claims and evidence, I'm thinking a score of 79?

**Setting scores and producing JSON details**

I've decided to set the score of 78 for overall assessment, with specifics for other categories: 73 for methods, 92 for advancing knowledge, 84 for logic communication, 62 for open science, and 94 for global relevance. For tier should, I'll use a score of 4.4, and for tier will, I'll set it at 4.8, all while keeping consistency in mind with overall scores at 86, advancing at 92, and global relevance at 94.

The JSON needs to include an assessment summary of around 1,000 words and metrics for each score calculated with midpoint, lower, and upper bounds. I'll also avoid naming authors or their institutions, but I can reference previous analyses without using specific names or citation numbers.

**Crafting the summary points**

I'm going to paraphrase the training dataset description, ensuring I accurately reference the analysis of natural regrowth from 2000 to 2016 without including citation numbers. I should say, "The study builds on an existing pantropical remote sensing analysis that mapped natural regrowth and distinguished it from plantations."

Next, I want to finalize the summary by outlining key points. For data, I'll mention potential mapping across tropical forests at a 30 m resolution, calculating weighted areas for 215 Mha, and noting variations across ecoregions.

For methods, I'll detail the balanced random forest approach, variable selection, the training dataset derived from regrowth mapping, and the overall accuracy. I'll also highlight potential issues like omissions in the training dataset and domain definition relative to thresholds.

**Assessing model limitations**

I need to go over some aspects of the model's predictive capabilities. First, I see that there's a lack of probability calibration in the balanced training; the predicted probabilities aren't calibrated, which might lead to bias in the weighted area approach and unrealistic confidence intervals.

Next, there's the issue of carbon overlay and uncertainty propagation; downscaling seems to be done naively. The coarser covariate resolution could degrade the 30 m predictions, causing blockiness. Socio-economic variables are omitted in terms of feasibility versus potential, and there's a failure to integrate dynamic processes like climate change impacts.

I should also note that accuracy in Southeastern Asia is poor, and the code's availability is limited, impacting replicability due to the heavy dataset.

Lastly, I'll mention that the additional risk of reclearance isn't accounted for in the carbon measurements. I need to finalize this while including suggestions and discussing uncertainties qualitatively as we maintain strict JSON formatting.

Diagnostic assessment – Williams et al. 2024

This paper delivers a high-resolution, pantropical map of where deforested lands have biophysical potential to undergo natural forest regeneration, and it aggregates this to country and biome summaries and an overlay-based carbon removal potential. The approach is transparent in intent—predicting potential, not feasibility—but several methodological choices and data limitations affect inference strength and precision. The dependent variable (presence/absence of natural regeneration during 2000–2016) is derived from a prior global remote-sensing classification that conservatively emphasized omission errors in humid regions and required a minimum patch size; area-based producer accuracy for natural regrowth in the humid biome was low while user accuracy for detected patches was high. This means the positive training class is a non-random subset of true regrowth, with systematic under-detection likely concentrated in particular forest types, geographies, and patch sizes. The predictive model then inherits these detection and sampling biases, potentially distorting learned relationships and shifting mapped potential away from contexts where regrowth was common but poorly detected. Model fitting uses random forests on millions of stratified points, with a sensible variable-selection procedure; predictions are ultimately based on biophysical predictors alone because a combined socioecological model yielded very similar accuracy. Omitting socioeconomic predictors makes the maps more temporally stable and widely available, but it also reframes outputs as "biophysical potential" and risks overpredicting potential in places where human pressure or governance constraints are binding. Validation emphasizes overall accuracy around 88% using a balanced validation set; however, accuracy declines at moderate distances from training locations, indicating residual spatial autocorrelation and some optimism in headline accuracy. The use of accuracy rather than threshold-free metrics, and the balancing of classes (prevalence set to 50%) during training and validation, further complicates interpretation of both accuracy and the calibration of predicted probabilities. The paper interprets the continuous model output as a probability and converts this to "expected area" by multiplying by pixel area and summing, yielding 215 Mha. Yet the calibration of random-forest scores to true probabilities is not established; without calibration (e.g., isotonic or Platt scaling) under the real-world prevalence, the expected-area identity may be biased. The extraordinarily narrow "confidence intervals" reported for area (±0.22 Mha at global scale) are not credible statistical uncertainty; they reflect computational or rounding artifacts, not propagation of uncertainty from training data errors, cross-validated prediction error, input-layer uncertainty, or probability calibration. A threshold (>0.5) sensitivity shows a substantially larger area, highlighting dependence on interpretation choices. Many predictors are coarse (250 m–1 km) while predictions are at 30 m; this scale mismatch introduces blockiness and may overstate local precision. The domain of "available for restoration" is defined liberally by excluding only forests, water, urban, bare, and forestry areas; while appropriate for modeling, its presentation alongside country totals risks being misread as areas that are socially or legally available for reforestation. Temporal assumptions also matter: the model projects to 2015/2030 by assuming that 2000–2016 relationships hold and by updating some forest-context layers with 2018 data; it does not incorporate climate change trajectories, shifting fire regimes, or evolving socioeconomics that strongly affect regeneration and permanence. Carbon estimates are calculated by downscaling a 1 km accumulation dataset to 30 m, multiplying by modeled potential, and summing; the stated range lacks a clearly defined uncertainty model and appears not to propagate classification, calibration, or permanence risks. The paper is commendably explicit that realized sequestration will be lower due to leakage, reclearance, and policy constraints, yet some statements comparing potential annual removals to current pantropical losses could be interpreted too strongly given these omitted risks and uncertainties. Spatial performance is uneven, with lower accuracy in parts of Southeast Asia; country and biome summaries lack uncertainty decomposition, which would be useful for decision makers. On replicability, large gridded outputs are publicly archived and inputs are open, but code is available only upon request; combined with the volume of preprocessing, this raises barriers to full reproduction, code review, and extension. Strengths include a clearly articulated objective focused on actionable planning, a very large training sample, careful variable selection, explicit treatment of autocorrelation effects (even if not using full spatial block cross-validation), and extensive caveats about use and governance. The resulting layers will be highly valuable as a first-pass screen for assisted natural regeneration opportunities and for national restoration dialogues. The main improvements needed relate to uncertainty characterization (probability calibration, spatially explicit error propagation, and country-level error budgets), spatial cross-validation that reduces spatial leakage, clearer separation of statistical precision from conceptual ranges, dynamic scenario analysis including plausible climate and human-pressure changes, and full open release of code. Overall, this is an influential, practically useful mapping contribution with limitations that mainly concern probabilistic interpretation, uncertainty, and temporal and socioecological external validity.

| Paper | Input tokens | Output tokens | Reasoning tokens | Total tokens | Est. cost (USD) |
|---|---|---|---|---|---|
| Peterman et al. 2025 | 18762 | 7617 | 6208 | 26379 | 1.94 |

| Paper | Input tokens | Output tokens | Reasoning tokens | Total tokens | Est. cost (USD) |
|---|---|---|---|---|---|
| Adena and Hager 2024 | 24234 | 7019 | 5312 | 31253 | 1.84 |
| Williams et al. 2024 | 28704 | 6327 | 5120 | 35031 | 1.80 |
| Kudymowa et al. 2023 | 65096 | 4147 | 2432 | 69243 | 1.77 |
| Green et al. 2025 | 22904 | 5884 | 3904 | 28788 | 1.52 |

# 4 Discussion

## 4.1 Limitations

*Sample size and representativeness:* We evaluated on only ~40–50 papers, all in the social sciences and specifically chosen by The Unjournal for evaluation (which means they were considered high-impact or interesting). This is not a random sample of research literature. The papers also skew toward empirical and policy-relevant topics. The AI's performance and alignment might differ in other fields (e.g., pure theory, biology) or on less polished papers.

*Human agreement as a moving target:* The Unjournal human evaluations themselves are not a single ground truth. As evidence of this, we note substantial variability between reviewers.

*Potential AI knowledge contamination:* We attempted to prevent giving the AI any information about the human evaluations, but we cannot be 100% sure that the model's training data didn't include some fragment of these papers, related discussions, or even The Unournal evaluations. We will be able to exclude this for the evaluations of *future* Unjournal evaluations.

Model limitations and "alignment" issues: While powerful, is not a domain expert with judgment honed by years of experience. It might be overly influenced by how a paper is written (fluency) or by irrelevant sections. It also tends to avoid extremely harsh language or low scores unless there is a clear reason, due to its alignment training to be helpful/polite – this could explain the general score inflation we observed. The model might fail to catch subtle methodological flaws that a field expert would notice, or conversely it might "hallucinate" a concern that isn't actually a problem. Without ground truth about a paper's actual quality, we used human consensus as a proxy; if the humans overlooked something, the AI could appear to "disagree" but possibly be pointing to a real issue.

(There is also evidence, e.g. Pataranutaporn et al. (2025) that LLMs show biases towards more prestigious author names, institutions, and towards male prestigious authors. We will provide further evidence on this in the next iterations, de-identifying the work under LLM evaluation.)

Scoring calibration: The AI was prompted to use the 0–100 percentile scale, but calibrating that is hard. Humans likely had some calibration from guidelines or community norms (e.g. perhaps very few papers should get above 90). The AI might have been more liberal in using the high end of the scale (hence higher means). In future, a different prompt or examples could calibrate it to match the distribution of human scores more closely. We also only took one run from the AI for each paper; LLM outputs can have randomness, so a different run might vary slightly. (To do: aggregate across multiple runs.)

Small differences and rounding: Our analysis treated the AI's numeric outputs at face value. Small differences (e.g. AI 85 vs human 82) might not be meaningful in practice – both indicate a similar qualitative assessment ("very good"). Some of our metrics (like kappa) penalize any difference, even if minor. Thus, the "low agreement" statistics might sound worse than the reality where in many cases AI and humans were only off by a few points. We intend to analyze the distribution of absolute differences: a large portion might be within say ±5 points which could be considered essentially agreement in practice. The credible intervals add another layer: sometimes an AI's score fell outside a human's interval, but overlapping intervals could still mean they agree within uncertainty. We did observe that AI's intervals were often narrower than humans' (LLM tended to be confident, giving ~10-point spreads, whereas some human evaluators gave 20-point or left some intervals blank), which is another aspect of calibration.

⚠ Planned updates and extensions

Also see internal tasks/issues in Coda

- Related work
- Slides
- Extended evaluation:
  - Journal ranking tiers and predictions
  - Claim identification
  - Qualitative assessments and full evaluations
  - Comparing evaluations across fields/areas
- Improved workflow:
  - Improve PDF ingestion
  - System prompt optimization
  - Alternative models
  - Extend set of papers
- Aggregating multiple LLM runs
- Anonymization
- Evaluation of papers for prospective (uncontaminated) evaluation
- More grounded information theoretic metrics and robust statistical tests.

# References

We used R version 4.4.2 (R Core Team 2024) and the following R packages: ggforce v. 0.4.2 (Pedersen 2024), ggrepel v. 0.9.6 (Slowikowski 2024), glue v. 1.8.0 (Hester and Bryan 2024), here v. 1.0.1 (Müller 2020), janitor v. 2.2.0 (Firke 2023), kableExtra v. 1.4.0 (Zhu 2024), knitr v. 1.49 (Xie 2014, 2015, 2024), rmarkdown v. 2.29 (Xie, Allaire, and Grolemund 2018; Xie, Dervieux, and Riederer 2020; Allaire et al. 2024), scales v. 1.3.0 (Wickham, Pedersen, and Seidel 2023), tidyselect v. 1.2.1 (Henry and Wickham 2024), tidyverse v. 2.0.0 (Wickham et al. 2019).

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).

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone, “rmarkdown: Dynamic documents for r,” (2024).

Eger, Steffen, Yong Cao, Jennifer D’Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller, “Transforming science with large language models: A survey on AI-assisted scientific discovery, experimentation, content generation, and evaluation,” *arXiv preprint arXiv:2505.05151*, (2025).

Firke, Sam, “janitor: Simple tools for examining and cleaning dirty data,” (2023).

Henry, Lionel, and Hadley Wickham, “tidyselect: Select from a set of strings,” (2024).

Hester, Jim, and Jennifer Bryan, “glue: Interpreted string literals,” (2024).

Luo, Ziming, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du, “LLM4SR: A survey on large language models for scientific research,” *arXiv preprint arXiv:2501.04306*, (2025).

Müller, Kirill, “here: A simpler way to find your files,” (2020).

Pataranutaporn, Pat, Nattavudh Powdthavee, Chayapatr Achiwaranguprok, and Pattie Maes, “Can AI solve the peer review crisis? A large scale cross model experiment of LLMs’ performance and biases in evaluating over 1000 economics papers,” 2025.

Pedersen, Thomas Lin, “ggforce: Accelerating ‘ggplot2’,” (2024).

R Core Team, “R: A language and environment for statistical computing,” (Vienna, Austria, R Foundation for Statistical Computing, 2024).

Slowikowski, Kamil, “ggrepel: Automatically position non-overlapping text labels with ‘ggplot2’,” (2024).

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani, “Welcome to the tidyverse,” *Journal of Open Source Software*, 4 (2019), 1686.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel, “scales: Scale functions for visualization,” (2023).

Xie, Yihui, “knitr: A comprehensive tool for reproducible research in R,” in *Implementing reproducible computational research*, Victoria Stodden, Friedrich Leisch, and Roger D. Peng, eds. (Chapman; Hall/CRC, 2014).

——, “Dynamic documents with R and knitr,” (Boca Raton, Florida, Chapman; Hall/CRC, 2015).

——, “knitr: A general-purpose package for dynamic report generation in r,” (2024).

Xie, Yihui, J. J. Allaire, and Garrett Grolemund, “R markdown: The definitive guide,” (Boca Raton, Florida, Chapman; Hall/CRC, 2018).

Xie, Yihui, Christophe Dervieux, and Emily Riederer, “R markdown cookbook,” (Boca Raton, Florida, Chapman; Hall/CRC, 2020).

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).

Zhang, Yaohui, Haijing Zhang, Wenlong Ji, Tianyu Hua, Nick Haber, Hancheng Cao, and Weixin Liang, “From replication to redesign: Exploring pairwise comparisons for LLM-based peer review,” *arXiv preprint arXiv:2506.11343*, (2025).

Zhu, Hao, “kableExtra: Construct complex table with ‘kable’ and pipe syntax,” (2024).