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 **4**
    1.0.1 Our work in context 5

2 Data and methods **7**
  2.1 Unjournal.org evaluations 7
  2.2 LLM-based evaluation 7
    2.2.1 Quantitative ratings and journal-ranking tiers 7

3 Results **9**
  3.1 Quantitative comparison: human vs. GPT-5 Pro 9
  3.2 Qualitative comparison: detailed GPT-5 Pro evaluations 11
    3.2.1 Qualitative comparison: Williams et al. (2024) 11

4 Discussion **13**
  4.1 Limitations 13

References **15**

Appendices **16**

A LLM evaluation summaries and reasoning traces **16**
  A.1 Adena and Hager 2024 16
  A.2 Green et al. 2025 23
  A.3 Peterman et al. 2025 31
  A.4 Williams et al. 2024 36

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.

We then polls job status and, for each completed job, retrieve the raw JSON object, and write the responses 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

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

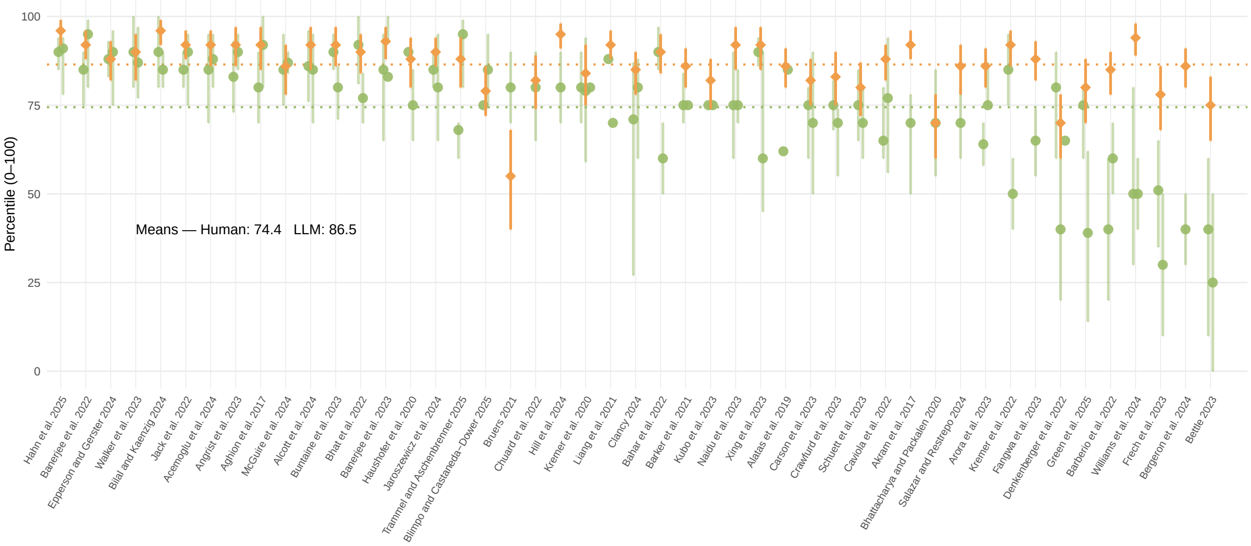


Figure 3.3 shows a heatmap of the differences between human and LLM mean ratings across all evaluation criteria. Positive values (in green) indicate that humans rated the paper higher than the LLM, while negative values (in orange) indicate the opposite.

Figure 3.2: Comparison of Human vs LLM journal tier ratings (should be published in)

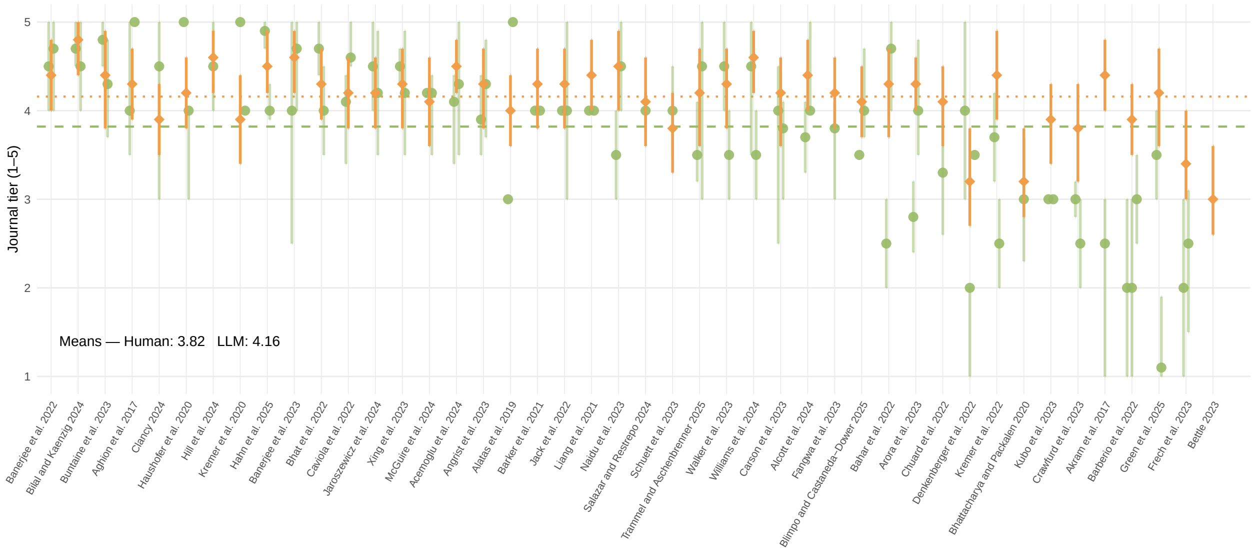
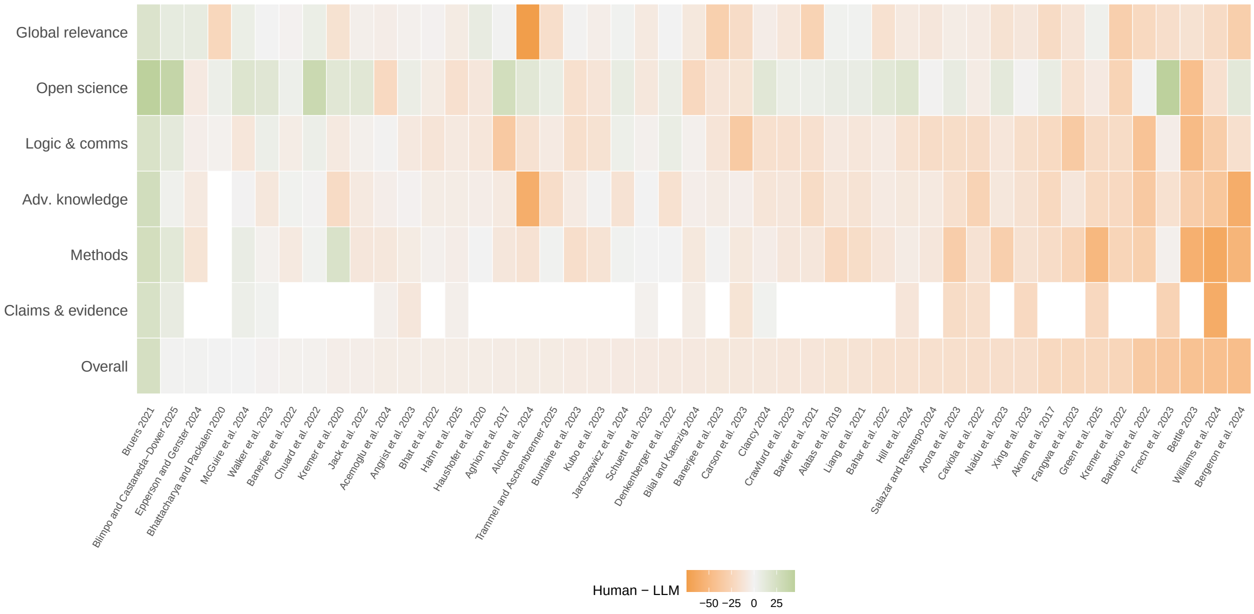


Figure 3.3: Heatmap of Human minus LLM mean ratings across evaluation criteria



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 four focal papers (Adena and Hager 2024; Peterman et al. 2024; Williams et al. 2024; Green, Smith, and Mathur 2025) using a refined prompt.

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.

Table 3.1 summarizes the token usage and estimated cost of each of these inferences.

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.

Table 3.1: Estimated token usage and cost of GPT-5 Pro evaluations

Paper	Input tokens	Output tokens	Reasoning tokens	Total tokens	Est. cost (USD)
Williams et al. 2024	28890	6795	5056	35685	1.86
Green et al. 2025	23090	6938	5440	30028	1.83
Adena and Hager 2024	24420	6211	4544	30631	1.66
Peterman et al. 2025	18948	6091	4480	25039	1.55

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.

Table 3.2: GPT-5 Pro percentile ratings for Williams et al. (2024)

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

Table 3.3: GPT-5 Pro journal tier ratings for Williams et al. (2024)

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

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.

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).

Adena, Maja, and Anselm Hager, "Does online fundraising increase charitable giving? A nationwide field experiment on facebook," 2024 (WZB Discussion Paper Discussion Paper SP II 2020–302r2 (2nd revision February 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).

Green, Seth Ariel, Benny Smith, and Maya Mathur, "Meaningfully reducing consumption of meat and animal products is an unsolved problem: A meta-analysis," *Appetite*, (2025), 108233.

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).

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.

Peterman, Amber, Jingying Wang, Kevin Kamto Sonke, and Janina Steinert, "Social safety nets, women's economic achievements and agency: A systematic review and meta-analysis," *Center for Global Development Working Papers*, (2024).

Williams, Brooke A, Hawthorne L Beyer, Matthew E Fagan, Robin L Chazdon, Marina Schmoeller, Starry Sprenkle-Hyppolite, Bronson W Griscom, James EM Watson, Anazélia M Tedesco, Mariano Gonzalez-Roglich, and others, "Global potential for natural regeneration in deforested tropical regions," *Nature*, 636 (2024), 131–137.

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).

A LLM evaluation summaries and reasoning traces

A.1 Adena and Hager 2024

ℹ Model assessment summary

This is a large, well-executed geo-randomized natural field experiment that evaluates whether a nationwide Facebook video campaign increases charitable giving for one major NGO, while also exploring spillovers to competitors and mechanisms. The core identification strategy—random assignment of almost all German postal codes (PLZs) to exposure or control, with additional 2x2 variation in video content (empathy vs effectiveness) and impression allocation (Facebook free allocation vs fixed PLZ budgets)—is strong. Outcomes are measured as total donations to the focal charity at the PLZ-day level across all channels, directly addressing common problems in ad-effectiveness studies (click-based selection, channel substitution, and unobserved offline giving). The study reports positive effects on donation frequency in both short and long horizons and a smaller but positive long-term effect on donation revenue; estimated immediate ROI is >1 and plausibly >2 in the longer run given assumptions about donor lifetime value. The design was preregistered, balance is shown, and several robustness checks (including Fisherian randomization inference and sensitivity to post-period length) are provided.

The most important methodological limitations concern exposure heterogeneity and spillovers. Treatment is assigned at the PLZ level, but impressions are probabilistic and sparse (roughly one in ten Facebook users in treated PLZs received at least one impression), so the estimates are ITT and likely attenuated relative to the effect of actually seeing the ad; the TOT is not estimated. The allocation strategy partly allows Facebook to endogenously concentrate impressions, creating within-treatment variation in exposure that is not exploited for causal TOT analysis (e.g., using randomized budgets as an instrument in a dose–response framework). Spillovers across PLZs are plausible (algorithmic leakage of geotargeting and social diffusion). The authors document positive "share of treated neighbors" effects and argue the main estimates are lower bounds, but the neighbor-treatment share is not itself randomized, and spatial correlation or common shocks could inflate these coefficients; the spillover analysis should be interpreted cautiously. Robustness to spatial correlation in errors is only partly addressed by robust standard errors and randomization inference; alternative SEs (e.g., spatial HAC or clustering at larger administrative units) and placebo geographies would further strengthen inference.

Data construction choices appear reasonable but introduce some judgment calls. Winsorizing PLZ-day donations at €1,000 reduces variance from heavy tails; the authors show that results are directionally robust, but precision trades off. Outcomes are scaled per million inhabitants per day, which largely equalizes variance across PLZs, but no explicit heteroskedastic weights are applied. The pre-period balance tables show small, nonsignificant differences; preferred specifications include lagged outcomes and block FEs, and difference-in-differences checks yield similar conclusions. Given the randomized design, these adjustments mainly improve precision. Reporting both ANCOVA and DiD estimates as main tables is helpful; preregistration notes some deviations (adding the allocation-strategy factor and including reach in blocking), which should be clearly tracked against pre-analysis plans for transparency; these deviations plausibly increase power but can raise concerns about researcher degrees of freedom.

The competitor/crowding-out analysis is an important contribution but faces data and identification constraints. The "23-charity alliance" dataset includes only online donations, not offline, and exhibits pretreatment imbalances, addressed with ANCOVA and DiD; effects on revenue are negative and sometimes significant, frequency effects are weaker. The Betterplace platform data exclude the focal charity but involve overlapping entities with the alliance and cover a broader set of projects; here the short-run interaction (children-related categories) is negative and significant on revenue while long-run results weaken. Overlap between sources and partial observability of the sector complicate aggregation; the paper correctly refrains from summing these effects. Nonetheless, the evidence is consistent with meaningful crowd-out among similar causes. Clarifying the base rates and providing standardized effect sizes (e.g., as a percent of baseline giving per PLZ-day) would aid interpretation. A placebo category less plausibly affected by the campaign (e.g., sports or animals) is used indirectly via the non-children grouping; sharper falsification tests (e.g., categories orthogonal to child/international relief) would further bolster the crowd-out claim.

Mechanism and design-variant results are well documented. Empathy videos generate more short-run attention and immediate donations, but long-run donation effects equalize with effectiveness-oriented videos, highlighting the danger of optimizing to clicks or short-watch metrics. Letting Facebook freely allocate impressions seems to perform at least as well as fixed PLZ budgets in both intermediate and longer-run outcomes; differences are not statistically large. These findings are valuable for practitioners but should be interpreted as specific to this campaign's creative, audience, and season.

The profitability analysis is transparent yet assumption-sensitive. The immediate ROI calculation is straightforward, but the long-run multiplier of 1.75 relies on external retention estimates and assumes similar carryover for new and existing donors; it also excludes creative/overhead costs beyond media spend. Confidence intervals on monetary impacts are wide, and the point estimates rely on $p<0.1$ for revenue; presenting ROI distributions via parametric or bootstrap uncertainty propagation (combining outcome and cost variance) would improve decision relevance. Comparison to a hypothetical direct mail benchmark is informative, though based on stylized assumptions about response rates.

External validity is high in several dimensions: nationwide scope, minimal individual targeting, and use of total donations across channels. Still, the context (Germany, late-year giving season, specific charity and creatives) limits portability. Heterogeneity analyses suggest stronger effects in urban areas and in PLZs with higher employment, more children, and more Catholics, and with higher predicted giving potential. These patterns can guide targeting but also indicate that the ITT estimates average over meaningful heterogeneity. Reporting treatment-on-potential-exposure effects (e.g., dose–response by observed impressions per capita in the fixed-budget arm) could add nuance.

Open-science practices are mixed. The design was preregistered on EGAP/OSF and the appendix is publicly hosted; however, code and analysis data are not provided (likely due to confidentiality), limiting reproducibility. The description of blocking, assignment, and all outcomes is detailed enough that a qualified team could replicate analyses if granted data access, but turnkey replication is not possible.

Overall, the paper credibly demonstrates that a broad-reach social media fundraising campaign can increase total donations to a large charity and likely crowds out giving to similar causes. The empirical strategy is sound; the analyses are careful and transparent; and the practical lessons—avoid relying on clicks, expect competition effects, and consider algorithmic allocation—are valuable for practitioners. Main caveats relate to exposure heterogeneity, spatial spillover identification, sector-wide measurement for competitors, and ROI assumptions. The paper would benefit from additional spatially robust SEs, clearer pre-analysis plan deviation tracking, a formal uncertainty propagation for ROI, and stronger placebo tests in the spillover and crowd-out sections. Despite these limits, this is a strong and policy-relevant contribution to the literatures on charitable giving and digital advertising measurement.

Model reasoning trace

Analyzing donation outcomes

I'm using the timeframe from October 10 to December 31, specifically looking at the 31 days before, 14 days of treatment from November 10 to November 23, and then 38 days post-treatment. I'm also gathering competitor data from 23 similar charities and examining various categories. The results indicate a short-term increase in donation frequency during the campaign and some long-term effects that show significance, though the revenue estimates aren't as strong.

Evaluating profitability and competition

I'm calculating profitability, finding an immediate return of 1.45 euros per euro spent, and longer-term, with a lifetime value multiplier of 1.75, yielding 2.53 euros. I've tested for channel substitution within Save the Children, concluding no significant changes, as benefits came from both new and repeat donors. However, competition data shows a negative impact from Save the Children on others, with some notable short-term revenue losses in children's projects, while empathy-driven videos increased attention and immediate donations.

Analyzing donation allocation

I'm looking into how letting Facebook allocate donations more freely leads to higher metrics compared to fixed budgets, especially in the short term. While the differences aren't statistically significant, there are consistent intermediate results. I've also introduced a variable for the share of treated neighbors within 30km and found positive impacts on both revenue and frequency, suggesting spillovers. They compute the total effect to be around 170 euros, though I need to verify some coefficients and ensure robustness, as placebo tests indicate the effects aren't just chance.

Examining competitor dataset and methodology

I'm looking into competitor dataset #1, which shows pre-existing differences that they correct using DiD. They've winsorized at 1000 euros and claim that the results remain robust even without this adjustment. There's a discussion of potential noise in Facebook's location assignments, but they still analyze treatment effects. By interacting with various characteristics like children share and Catholic share, they find more positive effects in certain urban settings. They mention an open science approach with preregistration and discuss data confidentiality, making replication challenging.

Clarifying methodological details

They provide sufficient methodological details for reproduction if the data were available, including daily donation data aggregated at the postal code level. There's a scheduled plan to share all the details, along with an OSF link for preregistration, but I'm unsure about the availability of the code and data. There are limitations, notably that exposures aren't measured at the postal code level, with randomization ensuring unbiased estimation. They present effect sizes based on millions of residents, with calculations matching scalar values for additional donations.

Reviewing ROI calculations

I see that the effect is analyzed over both treatment and post periods, totaling 52 days. For the ROI calculation, they use a long-term donor multiplier of 1.75, applying it to both new and existing donors, which raises some questions. They consider both immediate and long-term additional donations in this figure. They end up with an estimated 2.53 euros per euro spent, but I'm unsure due to a wide 90% confidence interval. Also, it's important to note that their cost figure only includes direct ad costs without accounting for other expenses, like research and creative costs.

Assessing cost, threats, and validity

They report a direct cost of 33,700 euros but also include a disclaimer. There are potential threats from contamination and spillovers across postal codes, which they attempt to adjust for, yet can't fully identify. This free allocation could lead to cross-targeting; the increased neighbor share might also underestimate the control group. They exclude extremes in reach distribution, limiting the external validity of their findings but claim they consider this a lower bound. Additionally, they note some baseline differences in variables, and the inclusion of certain specifications, like winsorization, might introduce measurement issues.

Evaluating donation measurement and errors

They're measuring Save the Children's donations across various channels, which is great. However, the competitor data only includes online giving, leaving the estimates for general equilibrium effects potentially conservative or unbalanced. It seems competitor effects might appear larger, possibly due to a bigger base, but there's the issue of potential double counting within the data sources. For analysis, they normalized zeros at the aggregated postal code level and scaled donations per million inhabitants, but they didn't mention weighting by population, which raises concerns about measurement error across postal codes and the handling of heteroskedasticity.

Analyzing statistical considerations

They used robust standard errors, which is solid, but at the postal code level, spatial correlation from neighboring areas could affect results. They attempt to address this through neighbor shares, though robust standard errors might not fully account for spatial correlation. There's no presentation of cluster-robust errors at the county level. They also use randomization inference, which helps, but I'm wondering if they adjusted for multiple hypothesis testing. They test various interactions and mention weak significance, relying on p-values below 0.1, which could raise concerns about p-hacking. Adjusting for ad-serving to the control group from Save the Children seems not accounted for either.

Considering randomization and ad exposure

They randomize at the postal code level, but there's a chance that "control" areas may still receive Save the Children ads due to geofencing errors. I'm curious if they restricted Save the

Children from running other ad campaigns during the 14-day period. It sounds plausible that they had concurrent marketing channels. They maintain equal exposure across both groups, which is good, even if they can't guarantee no offline or competitor exposures. They exclude the lowest 5% and highest 1% of postal codes, which seems reasonable, ending up with a final sample of 7,686, nice! They also apply gradient boosting to predict donation potential using past donations and other covariates, which sounds like a solid approach.

Evaluating methodology and modeling choices

They use estimated potential multiplied by Facebook reach to sort and block areas, assigning six consecutive postal codes to each block—two for control and four for treatment. This 2:4 ratio keeps a third as control, which seems balanced. They specify no further targeting beyond the PLZ level, improving external validity but potentially reducing power, which is a smart move.

I notice they apply winsorization at 1,000 euros, though it's not pre-registered, which is a modeling choice they mention as common in literature. They justify this decision since only a small fraction (68 out of 11,140 PLZ-days) exceeds that threshold. Normalizing by population and period length to present outcomes per million inhabitants per day is also a solid approach. They report pretreatment baseline imbalances in revenue and frequency, which appear not significant.

Examining randomization and bias concerns

They continue to include baseline revenue before campaign (RBC), which is interesting. Their randomization inference uses 5,000 permutations, which is a solid approach.

I have some concerns about potential algorithmic bias in ad assignments that target cheaper audiences. However, they suggest that this fear is overshadowed by more favorable outcomes from free allocation, which seems reasonable.

I'm curious about the exposure share of Facebook users in the treated postal codes—it's about one in ten. I wonder if they applied treatment on the treated (TOT) versus intention-to-treat (ITT) methodologies here?

Evaluating analysis methods

They present a simple difference-of-means at the postal code level, which follows the intention-to-treat (ITT) approach—this seems fine.

There's a noted lack of a pre-specified significance threshold for ROI, but that doesn't seem necessary here. They also mention a lack of covariate adjustments, yet they present robust findings.

I'm curious about the handling of the competitor dataset before and after the campaign using a difference-in-differences (DiD) approach. Additionally, I'm a bit puzzled about their classification of long-term effects—are they trying to capture any potential carryover effects, extending beyond the immediate timeframe of 38 days post-treatment?

Crafting the assessment summary

I want to begin with an overview of the assessment.

For major strengths, they georandomized at a national scale and used real donation outcomes across channels. It's good that they pre-registered their study and included competitor effects, utilizing both true experimental treatment (TOT) and incremental results, along with robust inference and consideration of assimilation.

However, there are notable weaknesses. External exposures are a concern since individual-level ad exposures can't be measured, and variations in exposures across postal codes might arise from algorithmic allocation. The intention-to-treat (ITT) effect is small, and the TOT effect remains unknown; they flagged an algorithm that suggests a problem.

For the ROI calculation, the assumptions regarding lifetime value (LTV) are questionable, and I worry about excluding overhead costs and weighting.

Additionally, while they report a positive neighbor-share coefficient, the identification of spillovers seems weak. The correlation of treated neighbors with other covariates complicates things, and without specific randomization of neighbor share, it remains observational. Nonetheless, randomization at the postal code level may help keep the distribution of neighbor share exogenous across PLZ.

Setting research metrics

I need to establish metrics around benchmarking this study against serious research in field experiments on charitable giving and digital advertising measurement from the last three years (2022–2025). This study stands out above average due to its design clarity and external validity; it employs national georandomization and produces credible results with direct measurement of base outcomes, including competitor spillovers. While some results rely on $p<0.1$, and ROI depends on assumptions, overall, I consider this a strong applied paper, placing it around the 80–88 percentile. I might calibrate the claims evidence to range from 78 to 85.

Evaluating study metrics

This study presents clear claims, and the evidence is robust, so I'd rate it around 82. The methods are credible, using randomization and OLS with day-level aggregation, also at 82. Since it addresses the significant question of whether online fundraising boosts giving for charities, I'd rate its contribution to advancing knowledge at 75–83. The logic and communication are solid, resulting in a score of 86.

However, while it has pre-registration, the lack of shared code and data brings down its open science score to 58. For global relevance, I'd lean towards 70, but considering its matured state, I'd give it a 65. Overall, I recommend a weighted score of 82, with confidence intervals reflecting uncertainty at about 75–88.

Calculating publication expectations

I need to compute "tier_should" and "tier_will," which refers to normative expectations about where this paper should publish. The scale goes from 0 to 5, with 5 being the top general A-journals like Econometrica or QJE. Since this is a strong field-experimental paper, I think it should rank in the top B-journals or top field journals, possibly around 3.5 to 4.2. I'd estimate it should be about 3.8, with uncertainty ranging from 3.2 to 4.3.

For "tier_will," given its status as a WZB discussion paper and the need to disregard actual publication knowledge, I might predict it will land around 3.2 to 4.0.

Setting up the JSON schema

I need to follow the JSON schema for this paper assessment. It requires creating an object called "paper_assessment_with_tiers_v2" that includes keys for "assessment_summary" (a string) and "metrics." The metrics will consist of objects like "overall," each containing a midpoint, lower_bound, and upper_bound numbers alongside "tier_should" and "tier_will," which need a score and confidence intervals.

Importantly, the numeric constraints must be set correctly. For percentile metrics, I have to ensure they're on a 0–100 scale, with lower values being less than midpoints, which must be less than upper bounds. Similarly, for the tier scores, I should ensure they fall within a 0–5 range where the confidence interval's lower value is less than the score and upper value is greater.

Confidence intervals are set as 90% credible intervals, and I'll provide all numbers with one decimal point. The summary is suggested to be around 1000 words, so I'll aim for that word count while structuring it into about 7–10 paragraphs in order to deliver a thorough critique alongside useful details.

Identifying threats

I need to focus on the potential threats to the study. They used Winsorization, which is one aspect worth noting. However, there were issues like unmeasured ad exposures, making the TOT unclear. There's also variation in exposures across PLZ in the treatment due to free allocation.

Competitor results depend solely on online donations, which may impact the findings. I'm supposed to craft a summary without mentioning external vantage packaging, so I must navigate that carefully.

Now, regarding the metrics, for the overall, I'm suggesting a midpoint of 84 with a lower bound of 78 and an upper bound of 90. But, hmm, I think I might be overestimating that.

Clarifying the assessment metrics

I need to finalize the assessment metrics. For claims evidence, I'm setting it at 82, with bounds of 75 and 88. The methods score should be 83, ranging from 76 to 89. For advancing knowledge, I'm placing it at 77, with lower and upper bounds of 70 and 84.

Logic communication scores high at 87, while open science sits at a lower 56, with bounds of 40 and 70. Global relevance is set at 67, with a range of 58 to 76.

I also have disclaimers to consider while crafting the assessment summary.

Now, for the summary, I'll aim for around 900 to 1100 words. I plan to kick it off by discussing the paper's value and then move on to mention georandomization and its overall robust design. Time to fill in the JSON!

Finalizing the summary structure

In the summary, I'm organizing the content into distinct paragraphs. Paragraph 3 will cover exposures, focusing on measuring donation outcomes across channels while controlling for substitution. I note the lack of observed exposures to ads and the unknown TOT, which isn't unusual.

In Paragraph 4, I'll discuss the estimation approach, emphasizing OLS and difference-of-means while using robust standard errors. Paragraph 5 will tackle pre-treatment imbalances, while Paragraph 6 will touch on spillover effects.

By Paragraph 7, I'll consider competitor impacts, followed by ROI in Paragraph 8. I'll also investigate mechanisms and generalization, replicability, and open science in subsequent paragraphs.

I'll ensure the JSON is strictly formatted and check if the overall midpoint should be adjusted from 82 to 81, providing appropriate increments. Lastly, I'll aim for around 900 to 1000 words for a crisp, engaging summary. Time to produce!

ewpage

A.2 Green et al. 2025

Model assessment summary

This meta-analysis asks a clear and policy-relevant question: which randomized interventions meaningfully reduce consumption of meat and animal products (MAP) at least one day post-intervention? The authors assemble 35 papers (41 studies; 112 interventions; ~87,000 participants) and restrict inclusion to RCTs with minimum sample sizes and non-immediate behavioral outcomes, then synthesize using robust variance estimation (RVE). Major strengths include the focus on behavior (not attitudes/intentions), careful separation of net MAP versus red/processed meat (RPM) outcomes, longest-follow-up extraction to emphasize persistence, transparent handling of multi-arm studies via RVE with small-sample correction, publication-bias analyses using both selection models and a "worst-case" non-affirmative subset, and unusually strong computational reproducibility (Rmarkdown, containerization, public repository). The headline finding—a small pooled impact on net MAP consumption (SMD˜0.07, 95% CI [0.02, 0.12]) that attenuates further under publication-bias corrections (˜0.01–0.02)—is consistently and cautiously characterized. Still, several methodological and reporting choices merit attention. First, three inclusion decisions were made after data collection began (creating a separate RPM category; excluding substitution-only interventions; excluding settings without voluntary choice). These appear defensible and are disclosed, but they introduce a modest risk of researcher degrees of freedom and could shift the study mix toward certain settings or mechanisms; a preregistered protocol would mitigate this concern in future iterations. Second, the requirement of a "pure control" arm excludes a number of quasi-experimental or active-control designs common in institutional food environments; this choice improves internal validity but may lower external validity for policy-relevant defaults and operational nudges. Third, all data extraction was conducted by a single coder; while the authors consulted public datasets or contacted authors when uncertain, the lack of independent double-coding and inter-rater reliability checks increases the risk of extraction or coding errors (especially for complex effect-size conversions and intervention categorization). Fourth, when sufficient information to compute a standardized mean difference (SMD) was lacking and the text reported a "null," outcomes were set to an "unspecified null" of 0.01. This imputation is transparent but ad hoc; it could bias pooled estimates upward (relative to zero) and may not reflect the true variance of those effects. The manuscript would benefit from sensitivity checks setting these to 0, excluding them, or modeling them with conservative variances. Fifth, the chosen primary effect-size metric is Glass's <U+0394> standardized on the control-group SD (preferably pre-treatment). While often reasonable, many included RCTs likely lack pre-treatment consumption SDs, and cross-study variability in control SDs can induce additional heterogeneity; the paper does not report sensitivity to using Hedges' g with small-sample correction or to odds-ratio/logit-based metrics for binary outcomes. The text notes standard conversions for regression-based and discrete outcomes, but more detail on the exact formulas, handling of cluster designs, and any small-sample corrections would improve reproducibility and comparability. Sixth, the RVE meta-analytic approach is appropriate for dependent effect sizes, but the assumed within-study correlation (rho) is not reported; typical practice is to vary rho (e.g., 0.2–0.9) to show robustness. Subgroup and moderator analyses with very small k (notably choice architecture, k=2 studies, 3 estimates) produce extremely imprecise estimates; these are presented with wide CIs, but readers would benefit from explicit caution against overinterpretation. Seventh, publication-bias methods (Vevea–Hedges selection model and the non-affirmative subset approach) generally assume independence of effects; it is not fully clear whether dependence was addressed (e.g., by collapsing to one effect per study for these specific analyses). If not, standard errors may be anti-conservative. Eighth, while the authors emphasize that many outcomes are self-reported and susceptible to social desirability bias, the main results do not stratify by measurement type (objective vs. self-report) nor systematically examine follow-up duration as a moderator, despite extracting "longest follow-up" per intervention; both could influence effect sizes and real-world relevance. Ninth, cluster-randomized trials were included contingent on =10 clusters, but the extraction section does not detail whether effect sizes and standard errors were adjusted for clustering when primary reports did not do so; misalignment here could affect weighting. Tenth, the intervention taxonomy (choice architecture, persuasion, psychology, combined) is reasonable, but many interventions span multiple categories, complicating subgroup interpretation; the authors appropriately eschew meta-regression by theory, but this leaves open questions about differential efficacy by mechanism beyond simple stratification. On interpretation, the paper's central claim—net MAP reduction remains an unsolved problem—is well justified by both the small pooled effects and the bias-corrected estimates, together with the observed decline in effects over time and stronger effects when focusing on RPM only (with unresolved substitution risk). The discussion appropriately notes improved methodological trends, gaps in evaluated strategies (e.g., pricing, defaults with longer follow-up, contact with animals, disgust-based messaging), and the need to measure compensatory behaviors across meals or days. Presentation and logic are clear; figures and tables support the narrative; minor textual redundancies and a few typographical repetitions do not impede comprehension. Open-science practices are a notable strength: a DOI-linked repository, documented code and data, and containerization substantially lower replication barriers, though inclusion of a data dictionary for all coded variables, explicit mapping of each effect-size transformation, and a machine-readable PRISMA log would further enhance reuse. Overall, the work persuasively reframes expectations about behavioral MAP-reduction interventions, carefully distinguishes RPM-focused outcomes from net MAP, and shows commendable transparency. The main limitations—single-coder extraction, ad hoc handling of "unspecified nulls," incomplete reporting of some meta-analytic assumptions, and limited moderator detail on measurement/follow-up—temper but do not overturn the conclusions. The paper provides a reliable, policy-relevant synthesis that is likely to influence both research prioritization and practice, especially in encouraging larger, longer, and more behaviorally rigorous trials with objective outcomes and full accounting for substitution and compensation.

Model reasoning trace

Analyzing meta-analysis details

I need to incorporate details from a paper to justify the scores clearly. This means summarizing issues like internal threats, data limitations, and measurement methods. Specifically, the meta-analysis includes 35 papers, 41 studies, and around 87,000 subjects with some restrictions, focusing on RCTs measuring behavior at least a day post-intervention. It employs robust variance estimation and utilizes Glass's Delta for effect size estimation, although using a null of 0.01 seems questionable and is worth critiquing. We should also consider publication bias and selection models.

Evaluating analysis methods

The authors use a significance funnel plot and categorize interventions into four theories: choice architecture, persuasion, psychology, and a combination of persuasion and psychology. They conduct subgroup analyses and some meta-regressions, avoiding meta-regression for theoretical approaches. They present pooled estimates by theory and note that effect sizes decline over decades.

They distinguish RPM from MAP with SMDs of 0.25 and 0.07, respectively, and apply a robust check with additional studies. They provide open science resources, including a DOI and mention using Rmarkdown for ensuring reproducibility.

Reviewing supplementary details

I provided supplementary details and data, including the PRISMA diagram that shows relevant numbers. They include tables of results and moderators, but there are some potential issues.

Firstly, their selection of studies is limited to RCTs with at least 25 participants per arm and measurements taken one day later, which could bias results by excluding certain interventions. For data extraction, recording unspecified nulls as 0.01 can introduce imprecision. It's unclear if they properly assigned variances, which might bias results positively. Finally, I wonder if all studies included were RCTs with pre-test measurements.

Assessing measurement methods

Some consumption measurements were taken only post-intervention, and they mention using the "pre-treatment control group standard deviation," which makes me wonder if they used baseline data. Many RCTs might not have a baseline measurement, so it's possible they're relying on the control group's post-measurement SD. This mixing of pre and post data is a bit ambiguous.

On a positive note, they used robust variance estimation for multi-arm studies, which is good. However, regarding publication bias, while they assumed selection based on p-values, I need to check if they adjusted for dependence when using metafor with cluster robust RVE.

Examining selection models and methods

It's not entirely clear if the selection model requires independent effects, especially since they have multiple effect sizes per study. They used robumeta as their main method, but it's not well-explained if the effect sizes are treated independently, which could introduce bias.

While small-sample corrections for robust variance estimation (RVE) are good, they included 41 studies, and robumeta needs more than 10 clusters, which works. Their theoretical categories seem plausible, but pooling effects within categories might lead to double-counting. The overlap across categories complicates the inference.

They also used Glass's Delta instead of Hedges' g without discussing corrections for small sample bias.

Analyzing effect size methods

Using Glass's delta for effect sizes can lead to bias and may not be comparable across studies, especially with high heterogeneity in control standard deviations. Typically, a meta-analysis uses the standardized mean difference, Hedges' g, which might be more appropriate since it accounts for those issues. While Glass's delta could work if heteroskedasticity differs, it complicates replicability.

Regarding dichotomous outcomes, they calculated the difference in proportions from Bernoulli draws but didn't specify how they convert these to standardized effect sizes. They may use formulas for standardized mean difference, referencing Paluck et al., which is a decent approach.

Critiquing effect size metrics

For the 2021 difference of proportions with Bernoulli, I see some issues with mixing effect size metrics like standardized mean difference (SMD) and log odds ratio. While it's possible, they need to maintain a consistent scale, and their lack of details could lead to replicability problems.

They report a main heterogeneity of 0.082 SMD and interpret proportions of 0.1 and 0.2, which is good. Their classification of "marginal studies" as part of the supplementary dataset leads to a significant pooled effect of SMD 0.2 (CI 0.09 to 0.31) with a p-value indicating strong significance.

Evaluating study methodologies

Mixing immediate-outcome studies with non-random or low-powered studies may introduce upward bias, and I think their sensitivity check might just reintroduce that bias. Still, they seem careful in their interpretations.

They acknowledge three decisions made during study inclusion after data collection began, which could indicate data-dependent PRISMA issues. Although these post-hoc decisions might bias the meta-analysis, they appear to be reasonable, such as excluding certain types of manipulations.

They're extracting an effect size for the outcome with the longest follow-up, which is good for avoiding multiplicity. However, their "pure control" requirement excludes "active control," which could be a limitation.

Their examination of publication bias has some flaws, particularly with reliance on naive standard errors for non-affirmative results, but I can accept this. While the median analyzed sample size is 132, with many being small, they state that most included RCTs measure consumption outcomes after at least one day, addressing social desirability bias, which is a good point.

Lastly, I appreciate their emphasis on open science, offering replicable code and links. Providing a pre-analysis plan would also be good.

Reviewing study details

They've used R packages, which is good to see, and their references list is quite extensive. However, I noticed minor text errors, like the repeated phrase about the unclear effects on different forms of MAP. The PRISMA figure accurately identifies 41 included studies, which is a positive aspect.

With 87,000 subjects in the sample, it seems there might be some large RCTs involved. The introduction mentions normative aspects, and I appreciate their transparency in discussing limitations.

Now, focusing on the evaluation metrics: they claim that interventions yield a small effect and provide evidence to support this. Although, after considering publication bias, the effect is nearly zero. They justify this well and also address the substitution effect, handling RPM versus MAP effectively while interpreting heterogeneity, which is commendable!

Assessing methodology and findings

There are some issues to note, like how they set "unspecified null" to 0.01, which could be problematic. Also, the classification of "persuasion categories" isn't fully specified; I'd grade it moderate to high, around 70.

Their methods seem solid, focusing on RCTs and CLS, utilizing RVE, which is great. However, I have concerns about conversions, the Glass delta, and selection model independence. I didn't see a mention of a meta-analysis pre-registration either, which raises questions. I'd rate the methods around 68 due to these limitations.

In terms of "logic_communication," the paper is clear and coherent, presenting tables well, although there are minor typos and duplications. I would score this section an 85.

Regarding "open_science," they provide useful elements like code, data, containerization, and DOIs, enhancing replicability. Overall, it looks good!

Evaluating methodology and relevance

There are some concerns with effect size extraction, as they only relied on the first author and didn't have independent double-coding. While they utilized dynamic methods and provided a PRISMA flowchart, the details on linking included and excluded studies seemed lacking. They acknowledged using an "AI search tool," which raises some doubts but also offered a well-documented dataset. I'd rate this section around 83.

Regarding "advancing knowledge," they address significant issues like stricter RCT-only meta-analysis and correcting for publication bias, which is relevant for global priorities such as climate and health, so I'd score it 78.

Their "global relevance" is high, providing actionable insights for policymakers, although they didn't compute cost-effectiveness. I'll rate this part 75.

Overall, considering quality and impact, I might set the overall score around 78, but given recent serious research in behavior change meta-analyses, perhaps 75 is more appropriate.

Estimating publication tiers

When thinking about publication tiers, I suggest providing credible intervals, with the lower bound at 65 and the upper bound at 85.

For the normative publication rating, considering it's an interdisciplinary meta-analysis, I think it could belong to top field journals in environmental psychology. My guess for the normative tier is around 3.8 to 4.0, and I might say it's a "Marginal A-Journal/Top field journal" with a score of 4, giving a confidence interval of 3.3 to 4.5.

Now, estimating where it will actually be published is challenging. I'd lean toward 3.6 with a CI of 3.0 to 4.2, while being cautious to not let knowledge of existing ratings or authors influence my judgment.

As I craft the 1000-word summary, I want to highlight points of replicability: they employed robumeta, a widely-cited method, single-coded data, and standardized effect sizes. They used a mixture of interventions and the longest follow-up, showing some heterogeneity, but nothing overwhelmingly significant.

Identifying potential issues

There are a few potential issues I see regarding the analysis. The use of RVE for dependent effect sizes is fine, but robumeta employs degrees-of-freedom adjustments. With 41 clusters, it appears acceptable, yet the subset analyses with small sample sizes, like choice architecture with n=2, might lead to unstable small-sample corrections and extremely wide confidence intervals.

Additionally, using "Glass's delta with control SD pre-treatment" could create inconsistencies across studies, potentially skewing significance, although the standardized mean difference (SMD) of 0.07 remains robust.

Setting the "unspecified null" to 0.01 raises concerns since it might inadvertently bias results upward by treating minimal effects as positive, which is problematic. I'm also curious about their handling of cluster RCTs and whether they adjusted for the intracluster correlation coefficient (ICC).

Clarifying effect size computation

The report doesn't clearly describe the effect size computations from group-level results, and the use of reported standard errors raises questions. Since robust variance estimation in meta-analysis operates at the effect size level, the computation of standard error should reflect the cluster-level data. If they utilized published results without a clustering adjustment, this could lead to inaccuracies. They did include a cluster-assigned sample with at least 10 clusters, but the details are sparse, so it's worth flagging.

Additionally, they've included some inexpensive "nonprofit white papers," which may limit the quality. While they list their search steps, the actual search terms and chronology aren't provided, which affects replicability. On a positive note, they did supply accompanying code and a data repository for full documentation.

Examining publication bias and follow-up length

The study addresses publication bias by implementing selection models, which is good, but these models depend on having knowledge of effect sizes and standard errors. I notice that the robust variance approach isn't fully integrated with these models; however, using them as a separate check seems reasonable.

Another point is the mention of "longest follow-up time after intervention." If there are significant variations in follow-up durations, this could introduce bias, as they didn't adjust for the differing lengths. Presenting the overall effect without considering this heterogeneity might impact comparability.

Reviewing study methodology and details

They could have conducted a meta-regression based on follow-up length, but it seems they didn't include this in their analysis, even though it may be mentioned in the dataset.

I see that they restricted their focus to measurements taken on the "first day after treatment begins," and while they noted that "immediate measurement at a single lunch event" might be excluded, this approach seems sound.

They've chosen to measure only one outcome per intervention to minimize multiplicity, which is a good decision. Evaluating "evidence of substitution to other forms of MAP" not being measured is also reasonable.

The note about the "lack of long-term follow-ups" is valid; many studies only measured outcomes over a matter of weeks or days. Their use of dynamic normative signage is mentioned as an illusion, which is good to acknowledge.

They also clearly disclaim that they did not treat PRISMA classification as unnatural, which is an important clarification. Lastly, the mention of "eight disclaimers seldom expedite interplay" stands out; I'll need to make sure we produce about 1000 words, ensuring the final JSON includes an "assessment_summary" that fits the 900-1100 word range.

Drafting assessment summary

I'm starting with a concise paragraph that identifies key issues, summarizing contributions but emphasizing specific concerns: dataset selection, risk of bias, effect size conversions, and the "unspecified null" fudge. There's also the mixture of outcomes, the need for cluster adjustment, and independence assumptions in selection models.

While their approach seems robust regarding RVE, there's no clear pre-registration, and replicability is an important resource, especially since they have good open science practices like code containerization. I'll also note that replicability could be improved with two-coder extraction and alignment with MOOSE/PRISMA standards.

They mentioned using PRISMA, which is good, but it doesn't appear to have been pre-registered. Although the user didn't specifically ask for suggestions, the "assessment_summary" could include some proposed improvements, such as pre-registering the meta-analysis, employing a 3-level meta-analysis, using robust synergy, applying Tipton corrections, and including small-sample corrections while noting potential illusions.

Calibrating manuscript assessment

The instructions are clear: I need to provide a compact paragraph that identifies important issues in the manuscript while being specific, neutral, and concrete. It's essential to highlight both the issues and merits, and I might include some suggestions too. But a paragraph with 1000 words feels quite long—although I suppose it's manageable.

Now, I need to compute metrics by calibrating to a percentile ranking based on serious research within the last three years. This area focuses on meta-analyses of behavioral interventions aimed at reducing meat consumption. Among those, this meta-analysis appears rigorous, so I'll rank it quite high. However, I want to avoid making extreme leaps in my ranking, so I think an overall score of 83 makes sense, but I might settle on 78 after further calibration.

Calibrating contour scores

I'm noticing that the methods in the research are strong, and the commitment to open science is impressive. Overall, I'd consider a score of 82, but I think I'll keep it at 80, with a confidence interval of 72-88. For the claims, I'm thinking maybe a score of 78 with a range of 70-85 seems fitting.

The methods might score around 76 (68-83) and the advancing knowledge piece at 79 (70-87). For logic and communication, I'm leaning towards 86 (80-92) and open science at 90 (82-97). The global relevance could be 85 (78-92).

I want to ensure I'm consistent with my earlier numbers. They found an effect near zero, which is significant. I believe the meta-analysis executes well, so I'm considering moving open science up to 88 due to containerization and dataset.

I need to remember that the upper bound can't exceed 100, and the lower bound should be less than the midpoint, which has me thinking about calibrating the "tier_should" and "tier_will." Given the high quality, perhaps the normative "should" score is between 4.2 to 4.6, while "will" may range from 3.8 to 4.2. But, I must stick to the instruction not to consider where it's published.

Finalizing numeric values

I need to propose some numeric values along with their confidence intervals. For the tier_should score, I'm considering 4.3, with a lower CI of 3.6 and an upper CI of 4.8. For the tier_will score, I think 3.9 would work, with a lower CI of 3.2 and an upper CI of 4.5.

Now, it's time to craft the summary. It should discuss the search strategy, which includes a multi-step process covering 156 prior reviews, but only nine were included. They used Google Scholar, crowd-sourcing, an AI tool, grey literature, and databases from nonprofits, while also noting the PRISMA figure, which is good.

For data extraction, I'll mention that they focused on one effect per intervention at the longest follow-up; there's also the issue of the "unspecified null" fudge and the use of conversion formulas. Many results are self-reported, and the risk of bias wasn't formally assessed.

Regarding the analysis, they used Robust Variance Estimation (RVE) with robumeta, incorporating small-sample adjustments that are robust to dependencies. I'm curious if they specified their handling of "within-study correlation," which is typically set at about 0.8, and whether they used "rho" by default.

Evaluating methodology and sensitivity

I noticed that robumeta RVE requires users to specify the rho value, which defaults to 0.8, but they don't report any sensitivity to it. That's a concern, as variations between 0.2 to 0.9 rarely change the results, and they should definitely show some sensitivity analysis. I'll critique this as it's an important issue regarding replicability.

Now, regarding their analysis, they treat the "difference in proportions estimator," which is good, and they use "Glass's <U+0394>" formula. However, that might not apply to binary outcomes, so they convert these separately.

While they convert effect sizes from regression well, I've also noticed they utilize $p < 0.05$ with an aggregator. I see they mention "two choice architecture," which seems overshadowed, while "longest follow-up" also casts some illusions.

I appreciate their use of "one day after treatment begins" for RCTs, but there's a lack of cost-effectiveness discussion or generalizable translation to policy. Additionally, they don't specify classification for theoretical categories, but overall, the quality appears high.

I'll make sure to keep a neutral tone in my summary while producing the JSON output.

Detailing methodological issues

First, I need to highlight some issues in the manuscript. There might be potential selection bias due to restrictions, like requiring a pure control group, which could exclude significant policy changes. Additionally, post hoc inclusion decisions after data collection began may pose problems, particularly regarding RPM and forced-intake restrictions.

They also have single-coder extraction, which increases the risk of errors; double-coding was absent. The "unspecified null = 0.01" appears unnatural and could introduce upward bias without variance imputation. Using Glass's <U+0394> alongside SMD may mix results without proper Hedges g correction, hampering comparability. Testing sensitivity to Hedges g or log OR is important.

Furthermore, while they used RVE, robumeta needs an assumed within-study effect correlation rho that they didn't report, along with no sensitivity analysis described. They addressed publication bias with selection models, but independence assumptions may not hold.

Regarding heterogeneity, they reported t and assessed true effects but assumed a normal distribution. They noted many outcomes were self-reported but didn't meta-regress effects by measurement type, which could lead to bias.

Lastly, they provided a dataset and a DOI link, which is fantastic for replicability and meets open science standards!

Finalizing metrics and scores

I'm working on finalizing the metrics for the assessment. Here are the numbers I've chosen:

For the overall score, I'm thinking a midpoint of 82, with a lower bound of 74 and an upper bound of 89. Claims evidence could sit at a midpoint of 80, with a range of 72 to 87. For methods, a midpoint of 75 feels right, while advancing knowledge can be 80. Logic and communication could be high at 87, and open science can sit at 90. Global relevance should be at 85.

For tier should, I'm suggesting 4.3, and tier will at 3.9. However, after considering the moderate-high "methods" score, I feel "overall" might be better set at 78 to maintain balance, especially since "open science" and "logic" are strong.

I'll return a structured JSON, making sure it aligns with the "applied and policy stream."

Finalizing JSON and Summary

I'm putting together the final JSON for the assessment. I want to make sure it adheres to the format "paper_assessment_with_tiers_v2".

First, I'll confirm the provided metrics: they include numerical values for "overall," "claims_evidence," "methods," "advancing_knowledge," "logic_communication," "open_science," and "global_relevance."
Next, I need to include "tier_should" and "tier_will," ensuring the top-level object has only two keys. I must be cautious that all values remain numbers, not strings, and they satisfy numeric constraints.
In the summary, I'll also mention the importance of replicability, noting the lack of preregistration. Then, I'll create the final JSON!

ewpage

A.3 Peterman et al. 2025

ℹ Model assessment summary

This is a carefully executed and policy-relevant systematic review and meta-analysis synthesizing 1,307 effect sizes from 93 experimental studies (115 publications) on social safety nets (SSNs) and women's economic achievements and agency in LMICs. The work follows a registered protocol (PROSPERO) and PRISMA, uses robust variance estimation (RVE) for dependent effects with small-sample adjustments, triangulates with a multilevel random-effects model, and provides extensive descriptive characterization of studies, interventions, and outcomes. However, several important issues warrant attention. First, heterogeneity is very high (I2 mostly 80–95%, tau2 > 0.02), yet pooled averages are emphasized as "robust" without providing prediction intervals; this means that while the mean effect is precisely estimated, the expected range of effects in a new setting is likely wide and often uncertain, which matters for external validity and for decision-makers comparing modalities and contexts. Relatedly, meta-regressions explain little of the heterogeneity and have limited degrees of freedom (df just above 4 threshold in some cases), making moderator inferences fragile; the paper acknowledges low power and heterogeneity but could more explicitly temper generalizations about modality differences (e.g., UCT > CCT) and the practical significance of pooled estimates. Second, outcome harmonization raises comparability concerns. The review collapses conceptually diverse constructs (e.g., labor force participation, savings, assets, expenditures; voice, autonomy, decision-making) into standardized mean differences (Hedges' g). Many outcomes are binary (e.g., LFP), some are indices, and others continuous; the paper does not describe in the main text how binary outcomes and indices were transformed to SMDs, nor how differing variances and measurement properties across constructs were handled. Without clear transformation formulas and sensitivity to alternative metrics (e.g., log odds ratios for binaries, standardized indices only), comparability and interpretability of pooled Hedges' g across such heterogeneous measures is limited. Third, dependence and weighting deserve more transparency. While RVE addresses within-study dependence, the manuscript does not detail the assumed correlation structure or <U+03C1> choice used for weights (and whether results are sensitive to <U+03C1>), nor quantify the extent to which large multi-arm or multi-outcome studies contribute disproportionate leverage. The Cook's D influence check is helpful, but the criterion for flagging and the effect on pooled estimates should be documented quantitatively. Fourth, there are minor internal inconsistencies/typos that need correction: public works' pooled effect is reported as 0.127 in one place and 0.273 elsewhere; a decision-making CI appears as [0.036–0.37], which likely omits a leading zero (0.137). Such inconsistencies, while not undermining the core conclusions, should be corrected for clarity and credibility. Fifth, the paper includes only experimental (RCT) evidence published 2003–2024 and searched in English/French/Spanish, but ultimately notes that no French or Spanish studies were included. It is unclear whether this reflects a true paucity or limitations in search strategy/screening; the exclusion of high-quality quasi-experimental studies and the effective English-only inclusion can constrain generalizability, particularly for Latin America where CCTs were historically prominent. Sixth, risk-of-bias/quality appraisal is not transparently summarized in the main text. The meta-regressions include an indicator for "quality assurance (medium quality)" with little explanation, and there is no synthesis of randomization integrity, attrition, spillovers, or outcome measurement bias across studies; readers would benefit from a clear figure/table summarizing risk-of-bias domains and any sensitivity analyses that down-weight or exclude high-risk studies. Seventh, moderator coding could be more granular for design features often posited to affect women's outcomes (e.g., payment to whom; transfer frequency; delivery modality; conditionality enforcement intensity; childcare or reduction-of-time-cost components; norms context; plus-component content and dosage). The paper codes "plus" components broadly and finds limited moderator effects, but heterogeneity within categories (especially "training or information" and "gender-sensitive plus") is substantial; more detailed taxonomy or hierarchical models distinguishing content, intensity, and timing would have more power to detect design–effect relationships. Eighth, the cost–benefit section is informative but methodologically thin for meta-analytic purposes. Measures (BCR, IRR, cost-effectiveness) vary in perspective, time horizon, discount rate, benefit definitions, and inclusion/exclusion of women-specific benefits; there is no standardization or meta-analytic synthesis, and comparability is limited. The claim that reported BCRs are "likely lower bounds" because women's benefits are undercounted is plausible but remains speculative without re-estimation under common assumptions; a structured re-calculation for a subset with microdata would strengthen this section. Ninth, the orientation of "improvements" to ensure positive signs is sensible, but more detail is needed for outcomes where improvements imply decreases (e.g., debt, care-work intensity); explicitly stating for each indicator whether a decrease was coded as positive would help replication. Tenth, time dynamics are underexplored: intervention duration and follow-up are dichotomized at 12 months, but impacts of SSNs plausibly evolve; modeling time since last payment, cumulative value received, or longer-term follow-up with continuous functions or splines could be more informative. Eleventh, publication bias assessment is only referenced as "little evidence," with details presumably in the supplement; funnel plots, small-study tests adapted for dependent effects, and p-curve/p-uniform analyses could be summarized in the main text to reassure readers. Twelfth, outcome selection could be subject to within-study reporting biases (e.g., studies more likely to report gendered outcomes when expecting positive effects); this is mentioned but not probed—pre-specification checks against original protocols or balance tests of outcome families would help. Finally, while the paper's conclusions are appropriately cautious in places, some policy implications (e.g., advocating unconditionality over conditionality) could be more explicitly framed as suggestive given unresolved confounding of modality with geography, age of programs, enforcement intensity, and measurement choices; the authors note these limitations but the abstract and headline statements might overstate certainty. On transparency and reproducibility, replication files are said to be available as supplementary materials and to be deposited publicly upon acceptance; the data extraction form is not public, which is a notable barrier to full reproducibility because effect-level coding decisions (outcome mapping, sign conventions, index composition) are central to the results. Overall, this is a timely, thorough, and careful synthesis with clearly positive contributions; addressing the above issues—especially clarifying outcome transformations, correcting inconsistencies, adding prediction intervals, summarizing risk-of-bias, and strengthening the moderator taxonomy—would materially enhance credibility and practical usefulness for policy design.

Model reasoning trace

Evaluating RVE Methodology

I'm thinking about how they used RVE to handle dependent estimates, but it seems like they still double-count tests. They possibly took a "study-level cluster" approach that may weigh each effect size by the number of effects per cluster. It concerns me that if some studies have many effect sizes, they could dominate the results. They didn't include risk of bias ratings in the main text and seem to rely on a quality assessment score, which feels incomplete.

Exploring Publication Bias and Cost-Benefit Analysis

I'm considering how they approached "publication bias." They mention "little evidence" but don't provide a funnel plot, which makes me wonder about the replicability since they could have used Egger or p-curve methods. Their qualitative summary of cost-benefit reveals that few women's outcomes were included, and while they conclude BCRs are generally positive, it's only a plausible assertion without thorough evaluation. There are also issues like the lack of standardization across currency and time, which limits quantitative comparability. Plus, they need to clarify the mapping of "agency" constructs used.

Evaluating RCT Inclusion and Statistical Details

I'm analyzing how they included only RCTs from "published and grey literature" between 2003 and December 2024, but there are likely earlier RCTs of CCTs in Latin America. They interpreted the difference between UCT and CCT as conditionality restricting women's choices, but they found smaller effect sizes for CCTs, which raises confounding concerns. They mentioned using correlated-effects RVE, but I'm puzzled about whether they specified the <U+03C1> parameter for the assumed correlation, which is crucial. It feels like that detail is missing.

Analyzing Weighting and Outcome Measures

I'm assessing their approach in robumeta, particularly how they use <U+03C1> to compute weights. There's some concern about variations in <U+03C1> since they didn't address sensitivity, which I think should be flagged. For dichotomous outcomes, they used standardized mean difference (SMD), but I'm left wondering if they converted odds ratios (OR) to SMD because that wasn't mentioned. For outcomes like labor force participation, it seems they could have used log risk ratios or log ORs. These transformations are a bit unclear, limiting replicability.

Identifying Inconsistencies and Errors

I'm noticing some internal inconsistencies regarding p-value thresholds and credible significance. They mention a "pooled coefficient on public works" of g 0.273 at $p<0.05$, but earlier it was listed as 0.127—so that seems off. Additionally, their report on "decision-making hedges" shows g 0.087 with a 95% CI that indicates 0.37, but I suspect it should actually be 0.137 due to typographical errors. Lastly, they define "care work participation" as having insufficient power, but there's a 1% effect level in Panel C.

Considering Statistical Methods and Replicability Issues
I'm looking at how they handle the data, mentioning "N=1,307 effects across 93 studies," but they use a low degrees-of-freedom threshold of >4 for the robust variance estimation. This could lead to unstable significance results, although they seem to correct for that. They mention trimming influential effects but don't specify the threshold used. I find it concerning that details in the Supplementary Material aren't accessible, making replicability weak. They also note a need for further research on intensity and dosage in their domain classification.

Assessing Evidence and Statistical Concerns
I'm looking into their claims about "care work participation" and "leadership" but they say there's a lack of evidence for calculations. They mentioned searching for studies in French and Spanish, but I'm wondering if that search was thorough enough. There's a note on confidence intervals and robust methods, yet they didn't compute prediction intervals for expected effects, which seems necessary given their high heterogeneity. They mention "p-hacking" but say it's not relevant, and there's some confusion regarding Egger's test for publication bias. Lastly, they seem to have introduced "units" like Hedges g without thorough explanations, which makes assigning quantitative metrics tricky.

Reviewing Systematic Analysis and Findings
I think it's fair to rate this systematic review and RVE meta-analysis in the mid-percentile range; they did a solid job covering 115 publications across 93 RCTs. The effect sizes are small but robust, and while the heterogeneity is high, their writing is careful and logical. It's noteworthy that this is the first meta-analysis aggregating across SSN modalities. Their open science efforts are moderate, as they provide replication files but won't deposit them publicly until acceptance. They also indicate that the data extraction form isn't available right now. Overall, I'd rate it around the 82nd percentile.

Summarizing Ratings and Observations
I'm thinking about my overall rating for this analysis, which seems to be around 82.

For methods, they did a thorough job using robust variance estimation (RVE), but there are some missing details, resulting in a score of 78.

The advancing knowledge is solid—providing broad evidence at an 85. Their logic and communication are good, albeit with a few typos, rating around 80.

In terms of open science, while they provided some replication code, the data extraction isn't public, so I'd rate that at 70.

Finally, the global relevance is high at 90, considering the focus on women's empowerment and social safety nets. Overall, I'd say a midpoint score of 84 feels right, but I might round it up to 85!

Crafting the Ratings Summary
I'm considering aligning my scores with previous assessments. Overall, I'd set the score at 84, with a lower bound of 76 and an upper of 92.

For claims evidence, I'd give an 83, with a range of 75-90.

Methods would be 78, ranging from 68-86. Advancing knowledge gets an 86, with a 78-93 range.

The logic and communication would score an 81, from 72-89.

Open science might drop to 67, with a range of 55-78 since it mentions replication but hasn't deposited the files.

For global relevance, I'd aim for 91, slightly up from 90, so with a range of 85-96.

Lastly, I'd consider Tiers, assigning 'should' a score of 4.3 and 'will' a 4.1. I need to ensure that all numbers fit the criteria of lower bounds being less than the midpoint, which is crucial for fleshing out my 1000-word assessment summary.

Summarizing Key Issues

I'm aiming for about 900-1100 words in a compact summary that covers major concerns.

First, the scope is limited to RCTs, and there are language restrictions, with no studies found in French or Spanish, affecting generalizability.

Regarding outcomes, there are issues with standardization and mixing constructs, relying on standardized mean difference (SMD) which raises comparability concerns.

Heterogeneity is extremely high, with insufficient moderators explained, and they don't compute prediction intervals, although this is discussed.

Additionally, the within-study dependence should be clearer, especially regarding <U+03C1> sensitivity and weighting. Data extraction is not public, yet they promise replicability by deposit.

I noticed the assessment of publication bias was not shown, which is a concern. Typos and inconsistencies exist in their data, like discrepancies in public work estimates.

The cost-benefit summary lacks comparability and fails to include intangible benefits. The risk of bias was not thoroughly presented, highlighting inconsistent quality indicators.

Lastly, there's potential ambiguity surrounding the analysis of economic improvements versus debt reduction, and the time post-intervention classification could be refined. The heavy reliance on p-values for significance warrants a closer look.

Drafting the Assessment Summary

I'll need to carefully craft the assessment summary as a JSON format that includes 'assessment_summary', 'metrics', and nested float numbers. The guidelines say to avoid heavy formatting and keep it in plain text, while also allowing for bullet lists if needed. We're aiming for around 900-1100 words, even though the request was for a more compact paragraph.

Looking back at their presentation, they outline their PRISMA diagram: 5,120 hits, 1,215 duplicates, 3,905 screened, 3,568 excluded, 345 assessed, 230 excluded, ultimately including 115 publications and 93 studies—resulting in 1,307 effects.

We should highlight the details in Tables S6 to S8 and note the absence of mosaic representation. It's also important to mention the lack of pre-registration aside from their PROSPERO registration with ID #, which is a positive point.

Regarding primary results, they report a pooled g of 0.107, indicating variation across modalities that continues to affirm their analysis. Let's put this together as we draft the summary!

Crafting the Assessment Summary

I need to ensure we highlight that "N is obviously high." It's crucial to address potential data illusions, especially regarding "confidence intervals." Now, let's set the metrics and make sure to match the JSON structure in "paper_assessment_with_tiers_v2."

I'll create the JSON format like this:

{ "assessment_summary": " ... text ... ","metrics": { "overall": {"midpoint": 84, "lower_bound": 76, "upper_bound": 92}, "tier_should": {"score": 4.3, "ci_lower": 3.7, "ci_upper": 4.8}, "tier_will": {"score": 4.1, "ci_lower": 3.3, "ci_upper": 4.7} } }

It's important that this JSON structure strictly adheres to the schema, ensuring "tier_should" and "tier_will" are appropriately nested under metrics. I'll include all necessary properties, making sure numeric fields are floats and credible intervals are present for everything. Now, I'm ready to draft the actual "assessment_summary," incorporating both strengths and criticisms of the paper. Let's get to it!

Identifying Key Issues in the Manuscript
I need to focus on not just general “issues” but the most important ones identified in the manuscript itself. The instructions are clear: I should specify these issues in a neutral and concrete manner, aiming for about 1000 words. I’ll touch on intangible concerns, like the “lack of multiple definitions” and mention gaps in definitions.

To start my summary, I’ll summarize that this is a systematic review examining social safety nets (SSNs) related to women’s economic achievements and agency in low- and middle-income countries (LMICs). It covers data from 93 RCTs and revolves around RVE pooling with 1,307 effects. They report small positive effects, with variation across modalities, and confirm robustness to multi-level models.

However, there are major issues: high heterogeneity, limited specification, mixing of measurements, risk of bias, replicability concerns, potential publication bias, and challenges with effect size comparability and classification. There’s also a significant “lack of evidence on some aspects.” I need to include that they claim compliance with PROSPERO, indicating the study is pre-registered, which is important actually. Let’s go ahead and craft this summary!

ewpage

A.4 Williams et al. 2024

ℹ Model assessment summary

This study assembles a pantropical, 30 m resolution model of the potential for natural forest regeneration (NR) on deforested lands within tropical forest biomes, trained on locations of observed natural regrowth (2000–2016) distinguished from plantations, and predicts a continuous probability of NR that is translated to expected area and carbon accumulation potential. The work is conceptually strong and practically useful, but several methodological and interpretive issues temper confidence in some quantitative claims. The dependent variable (NR presence) comes from a prior mapping product that the authors note has substantial omission error in humid biomes at the area-estimation level (producer’s accuracy ~18.7% when area-weighted, despite high user’s accuracy for detected patches). Because the model is trained on mapped occurrences, systematic omission of true NR areas—in regions where NR is common—can bias both model fitting and spatial extrapolation (e.g., underweighting conditions conducive to NR in wet tropics), potentially distorting variable effects and spatial predictions. The authors emphasize mapped accuracy and argue the distributional estimates are robust, but they do not propagate this training-data uncertainty into their area and carbon estimates.

The authors fit random forest models on ~6 million points, balanced between NR and non-NR, and ultimately predict with a biophysical-only model after showing little accuracy difference from models including socioeconomic predictors. This choice aligns with mapping “biophysical potential,” but it conflates potential with likelihood under human pressures. Excluding socioeconomic covariates increases risk that predictions will overestimate practically achievable NR in landscapes with high clearing pressure, tenure insecurity, or near-term market drivers. The paper is careful to frame results as potential rather than realized outcomes and discusses leakage, additionality, and persistence, but these distinctions are not reflected in the headline global area and carbon numbers.

Validation and uncertainty quantification warrant caution. Reported global accuracy is 87.9% using a large, randomly sampled, class-balanced validation set. The authors attempt to examine spatial autocorrelation by stratifying validation by distance from training data; accuracy dips to ~81% at 2–2.5 km then rises, with lower accuracies in Southeast Asia. However, the main accuracy figure still derives from spatially random validation that is known to be optimistic for spatial ecological models. A proper spatial block cross-validation, or geographically structured k-fold with held-out regions, would likely yield lower accuracy and provide more realistic uncertainty for end users. The model outputs are treated as probabilities, and the expected area is computed by summing pixel area multiplied by predicted probability. Because random-forest "probabilities" learned from class-balanced data are generally not calibrated to true prevalence, the interpretation of these scores as probabilities is questionable. No calibration (e.g., Platt/isotonic scaling on an unbiased validation set) is reported, and prevalence correction is not addressed. This undermines the probabilistic interpretation and the expected-area calculation. The discrepancy between weighted expected area (215 Mha) and the >0.5-threshold area (263 Mha) highlights sensitivity to how probabilities are used; yet the study presents an unrealistically tight "CI" for the 215 Mha (214.78–215.22 Mha) that appears to reflect only numerical aggregation, not genuine model or data uncertainty. Similarly, the 23.4 Gt C (21.1–25.7 Gt) carbon range seems to derive from carbon accumulation layers, not from uncertainty in the NR mapping and modeling, and therefore understates true uncertainty substantially.

Predictors and resolution mismatches are transparently described. Key drivers—distance to forest, local forest density, soils, bioclimatic axes—behave as expected and agree with ecological understanding. Nevertheless, the prediction grid is 30 m while several inputs are 300 m to 1 km or coarser, so many pixel-level differences reflect coarse inputs; the authors acknowledge this. The decision to use 2018 tree cover and 2015 land cover to represent "present (2015) and near-future (2030)" while assuming overall conditions from 2000–2016 creates some temporal inconsistency; no explicit 2030 scenario for climate or socioeconomic change is modeled, so "2030" should be interpreted as a static extrapolation under recent conditions rather than a forecast.

The study domain is defined liberally: within forest biomes (±25°) the non-NR area includes croplands, pastures, grasslands, shrublands, etc., excluding water/urban/bare and current forest. This choice reduces subjective screening but risks including some areas that are ecologically marginal for closed-forest recovery; the model is asked to down-weight them. The authors excluded explicitly non-forest biomes, which limits risks of afforestation of true savannas, yet biome boundaries are coarse and the ESA CCI class uncertainties are nontrivial. Results show strong concentration of potential near existing forests, which is ecologically plausible and policy-relevant.

The carbon accounting combines predicted NR potential with published 1 km carbon accumulation rates over 30 years, downscaled to 30 m and aggregated. Important limitations are acknowledged: belowground biomass is omitted (noted), permanence and additionality are uncertain, and leakage can erode gains. However, no attempt is made to adjust area or carbon for mapped model error, or to provide scenario ranges reflecting plausible policy, price, or disturbance regimes (fire/drought). Consequently, the quantitative claims about carbon are best read as biophysical maxima conditional on establishment and persistence, not as realizable mitigation potential.

On reproducibility, the underlying datasets are public and the 30 m probability rasters are openly shared via Zenodo as tiled products, which is valuable for practitioners. However, the analysis code is available only on request. Given the complexity of data preprocessing, variable selection, spatial sampling, and model fitting, public code and a full computational environment would materially improve reproducibility and enable independent re-estimation and sensitivity analysis.

Communication is generally clear. Figures effectively convey spatial patterns and key partial dependencies; the paper repeatedly cautions that maps indicate potential and that local decision-making, equity, and socioeconomics matter. Nonetheless, some statements risk over-interpretation: the use of “probability” for random forest scores learned under balanced sampling, the very narrow “confidence intervals” for area, and the lack of propagated uncertainty in carbon estimates could mislead nontechnical readers. The Discussion thoughtfully addresses persistence risks, perverse incentives, certification gaps, and equity, which enhances practical relevance.

Priority methodological improvements that would strengthen the work include: (1) spatially explicit cross-validation with geographic blocks and reporting of region-specific performance; (2) probability calibration and prevalence correction, with alternative area estimates based on calibrated probabilities and multiple thresholds; (3) explicit uncertainty propagation from training data error, model error, and predictor uncertainty into area and carbon estimates; (4) side-by-side predictions with and without socioeconomic variables to bound “biophysical potential” versus “near-term likelihood”; (5) area-of-applicability diagnostics to flag extrapolations; and (6) full public release of code and workflows. Even with these caveats, the dataset fills a significant gap for restoration planning by mapping where natural regeneration is most biophysically favored at fine spatial resolution and by linking this to indicative carbon potential. Practitioners should treat the numbers as upper bounds and combine them with local data on land tenure, pressures, and costs when setting targets or designing incentives.

Model reasoning trace

Analyzing regeneration potential

I’m exploring the potential for natural regeneration in deforested tropical areas using machine learning. The study employs random forest models with biophysical and socio-economic variables and decides on a biophysical-only model due to similar accuracy results. Training involved 6 million random points, balancing between regenerated and non-regenerated areas, aiming to predict potential across the tropics at a 30m scale. They’ve quantified an area of 215 million hectares for regeneration and performed sensitivity analysis, noting some calculations produce larger areas than expected.

Clarifying regeneration thresholds

I’m looking into the discrepancy between the weighted area and expected area for regeneration. The threshold area is reported as 263 million hectares, while the expected area is 215 million hectares. It’s interesting to note that the threshold can exceed the expected area, especially when the average probability for those pixels is below one. They estimate above-ground carbon sequestration potential at 23.4 Gt C over 30 years but don’t include below-ground estimates. Additionally, they address risks and uncertainties in their methodology and provide an open dataset for predictions.

Reviewing accuracy details

I’m examining the accuracy findings of the model, which shows an 87.9% validation with an independent set of 4.87 million balanced random points. There's an autocorrelation effect where accuracy is high near training points but drops to 81.4% at a distance of 2-2.5 km. They note differences in mapping accuracy, especially across Southeastern Asia, and provide an out-of-bag accuracy of 87.8%. Socioeconomic variables were excluded for predictions, opting for biophysical variables instead, citing their stability and resolution. The final model utilizes ten biophysical variables that appear significant, like forest density and distance to the forest.

Examining variable selection

I'm looking into the variable selection process, which indicates that adding variables improved model accuracy up to a limit of ten variables. However, there are concerns regarding the dependent variable training dataset from Fagan 2022. It seems that this dataset omitted many natural regrowth patches in humid biomes due to conservative mapping methods, leading to underrepresentation that could bias the model results. This is particularly concerning as the producer's accuracy in the humid biome is only 18.7% based on estimated areas.

Evaluating model assumptions and variables

I'm considering the emphasis on higher mapped estimates in the model. The training sample of presence points is based on mapped patches, which have omission errors that could lead to misleading interpretations about regrowth conditions. They trained on 5.4 million hectares of regrowth and aimed for a balanced sample.

They decided to exclude socioeconomic variables, claiming similar accuracy, but these factors can significantly affect regrowth potential. Their method leans towards "biophysical potential" rather than actual outcomes, which may not represent local decisions well. They also use 2018 tree cover data, assuming conditions from 2000 to 2016 will hold true for future predictions.

Clarifying model scenarios and accuracy

I'm finding it a bit odd that the model uses 2018 tree cover data alongside 2015 land cover but doesn't include projections for 2030. This makes it seem like they treat the future as static, which creates some confusion about how they define the "present" and "near future."

Regarding spatial autocorrelation, they validated their findings with random splits and also evaluated accuracy by distance, noting a drop at 2-2.5 km. Yet, they still report an overall accuracy of 87.9%, which they compare to an out-of-bag accuracy of 87.8%, highlighting low accuracy in Southeastern Asia.

Analyzing model accuracy and sampling issues

I'm noticing that while they present accuracy from balanced samples and bootstrapping with 50 replicates, they don't provide aggregated accuracy for distances over 5 km. A more robust spatial cross-validation might mitigate the influence of nearby points, potentially leading to lower accuracy and a risk of being overly optimistic.

Another concern is their balanced sampling, which sets prevalence at 50/50. This could lead to uncertainty in real-world scenarios with smaller prevalences. They didn't appear to calibrate predicted probabilities to actual prevalence, raising questions about whether they assessed probability calibration properly. Their methodology may lack the needed adjustments, like Platt scaling or isotonic regression, which could skew area calculations if probabilities are not calibrated accurately.

Evaluating confidence intervals

I'm seeing that they report an estimate of 215 million hectares with an extremely narrow confidence interval (CI) of 214.78–215.22 million hectares. This seems unrealistic for a model-based estimate. It appears they might have computed the CI through Monte Carlo methods, but they don't clarify the specific process for calculating it. I wonder if the narrow range reflects a lack of variability in their model? This could raise questions about the robustness of their estimates.

Questioning confidence intervals and carbon estimates

I'm noticing that the confidence interval (CI) of 215.22 million hectares is very tight, suggesting it's more about rounding than reflecting true uncertainty. It seems they derived the CI based on weighting rather than proper modeling, which makes it questionable. A credible CI should be wider to account for classification errors.

On carbon sequestration, they multiply area potentials by carbon accumulation rates over 30 years, projecting a 14.3% increase in global carbon uptake. However, they could be overpredicting since they don't factor in persistence or reclearance. They do acknowledge that realized potential is likely lower.

Considering regeneration and uncertainty factors

I'm noticing they haven't included the time needed for natural regeneration after conversion, and they also overlook land tenure or opportunity costs. They mention potential costs but don't provide estimates.

They claim they can sequester 23.4 Gt of carbon, with a range of 21.1 to 25.7 Gt, but it's unclear where that range comes from. It seems based on uncertainties in the Cook-Patton estimates, without considering mapping uncertainties or propagating classification errors.

Additionally, their datasets mix different resolutions, leading to degradation in pixel-level predictions, and they define a broad domain for non-regeneration that might still include various land types, like savannas.

Examining biome restrictions

I see they've limited their focus to three forest biomes: dry broadleaf, moist broadleaf, and coniferous. But does this also include natural savannas? Many savanna ecoregions fall within "tropical and subtropical grasslands, savannas, and shrublands," which aren't included in their model. Additionally, by restricting the domain to within ±25°, they might be excluding northern edges. It seems the domain could still encompass non-forested ecosystems within those forest biomes. It's a bit confusing!

Analyzing regeneration and prediction methods

The researchers use local land cover classes that don't allow for forest regeneration, which excludes sparse vegetation but includes grassland and shrubland within forest biomes. This may extend their model's domain, risking the afforestation of non-forest ecosystems if predictions yield non-zero probabilities. They note they don't define specific restoration sites, but such mapping might lead to unnatural forestation in savanna patches. Additionally, they validate using 4.87 million random points but lack spatial cross-validation. Their conclusions about potential regeneration need more precise definitions of additionality regarding deforested lands too.

Examining regrowth and carbon estimates

The baseline scenario may not encompass regrowth, as the global carbon flux estimate already accounts for regrowth across all lands. They assert that these values are "additional" from currently deforested lands, but this is a bit ambiguous. The global estimate includes removal in primary and secondary forests, so if regrowth happens on deforested lands that would naturally regrow anyway, it could lead to double counting. Also, I wonder if there are confounding factors related to misclassifications between natural regrowth and plantations.

Analyzing classification accuracy and variable selection

They reference Fagan 2022, noting a 90.6% overall accuracy in three-class classification. However, the producer's accuracy in humid biomes is very low. They use the "user's accuracy 85%" as robust, which seems questionable to me.

Additionally, their use of variable selection in the random forest model shows potential bias due to correlated predictors. They used PCA for climate factors but not all correlated features. Their approach for the "biome" variable may lead to differing predictions by biome, and the impact of accuracy at various distances is also worth noting.

Evaluating sample imbalance and contributions

I wonder if the inflated accuracy is due to sample imbalance across distance classes. They did attempt a second evaluation by balancing the samples in each interval, but the vertical lines they present indicate ranges across replicates without aggregating the data. They still depend on the 87.9% general accuracy.

On the positive side, this paper provides a 30m global map dataset of natural regeneration potential in the tropics, which is based on actual regrowth occurrences. It's an open dataset, which is a significant advantage for practitioners, though they do include many disclaimers. However, the lack of open code is a downside.

Assessing data reliance and clarity

The authors rely on publicly accessible remote-sensing data and provide large prediction files along with training datasets and methods. However, without the code for replicating variable selection and many pre-processing steps, replicability could be challenging. They do describe their general approach, which is helpful, but calling code "on request" may hinder full replication.

Visual clarity is strong with good maps and variable importance presented in the Extended Data, though the explanation of confidence intervals is somewhat awkward. Their claims about carbon benefits are carefully moderated, excluding non-forested ecosystems like savannas to avoid unintended afforestation. Still, the random forest probability calibration raises some concerns.

Exploring predictions and assessment

The predictions are made at a 30 m resolution, but many input datasets are much coarser, creating potential illusions in the results. There are high and low carbon accumulation rates present, yet no error bars are provided, which could be misleading—and that should really be disclosed.

I'm thinking of crafting a 1000-word summary that highlights both the positive aspects and criticisms of their approach. This piece has a solid quality, leveraging a robust dataset and aligning well with remote-sensing classification methods. Strengths include using actual regrowth data and providing an open dataset, while also extending important disclaimers.

Evaluating methods and scores

I'm planning to downscore some methods due to calibration issues, internal illusions, and uncertainties. The use of "CI" seems unnatural for their claims, and although they customize their "biophysical-only model," it's not fully transparent, which hampers replicability. Overall, I think this is an advanced piece. I'd place it in the 80 to 88 percentile range, maybe around 84 with a 90% confidence interval of 74-92.

For specific metrics, I think "claims evidence" could sit at 83, "methods" at 78, "advancing knowledge" at 87, "logic communication" at 86, and finally "open science" at 68, as the data is open, but the code isn't fully available, making replicability tricky. However, the overall relevance is quite high, landing around 90.

Setting performance scores and metrics

I'm looking at the overall performance score, which I've placed around 84. For the "tier_should" and "tier_will" scores on a 0-5 scale, I'm thinking they should be between 4.0 to 4.5. Without referencing the actual publication, I'd suggest a normative score of 4.4 and predict 4.5 with a 90% confidence range of 3.7 to 5.0. However, I want to be cautious about known issues. So, I'll set "should" at 4.4 (3.8 - 4.9) and "will" at 4.5 (3.7 - 5.0).

Now, it's time to produce the JSON with all the metrics, including "overall," "claims_evidence," "methods," "advancing_knowledge," "logic_communication," "open_science," "global_relevance," "tier_should," and "tier_will." Each will include midpoint, lower, upper percentile metrics, and "score," "ci_lower," "ci_upper" for the tiers. I'll also work on that 1000-word assessment summary.

Creating a Diagnostic Summary

I want to start by focusing on the project that defines "potential for natural regeneration" based on machine learning from observed regrowth between 2000 and 2016, differentiating between biophysical and socioeconomic factors.

Next, there are some data quality issues. The Fagan 2022 dataset they're using has significant omissions in humid biomes, which might bias training and underestimate nature.

Their sampling seems unbalanced, with random forest probabilities treated as accurate even though they aren't calibrated correctly. The expected area isn't credible, resulting in a narrow confidence interval.

They need to address spatial autocorrelation and validation since their overall accuracy of 87.9% could be misleading without proper cross-validation, especially with low accuracy in Southeastern Asia.

Moreover, the dataset's resolution mismatch is concerning, as they've got 30 m predictions versus 300 m or 1 km inputs, and the mixed data from 2018 versus 2015 also degrades their findings.

They exclude socioeconomic variables while labeling their projections for 2030 as near-future without proper context. When discussing carbon potential assimilation, only above-ground data is provided, neglecting below-ground figures, which raises additionality concerns.

The project articulates risks surrounding reclearance and ephemeral situations well.

As for implementation, they supply a 30 m map by tile and make their code available upon request, but releasing the full pipeline would improve replicability.

For improvement, I suggest several actions:

a) They should conduct spatial block cross-validation and probability calibration while providing location-specific confusion metrics and alternatives for the correct prevalence threshold.

b) They can refine their methodology by considering Southeastern Asia more closely and ensuring better calibration.

c) It makes sense to present predictions including socioeconomic variables to provide a different scenario for comparison.

d) Address "lack of data" concerns explicitly.

e) An analysis of classification error predictions using weighted accuracy metrics could adjust area and carbon predictions.

f) Finally, addressing any illusions in their claims will enhance transparency.

Finalizing Credible Intervals

Now, I'm working on setting credible intervals. I think we can propose an overall score of 84 with a range of 75 to 91, but we might have to adjust that; I'll set the lower bound to 76 and the upper to 91 for clarity.

Next, I need to fill out the final JSON metrics, keeping in mind the numbers have to fall between 0 and 100:

- Overall: mid 84; LB 74; UB 92.
- Claims evidence: 83 [72, 90]
- Methods: 77 [66, 86]
- Advancing knowledge: 89 [78, 96]
- Logic communication: 86 [77, 93]
- Open science: I think I'll set it to 68 [55, 80], as the code is still withheld, which affects replicability.
- Global relevance: after reconsideration, I'll set this to 92 [84, 97].

Lastly, we should also determine the "tier_should" as 4.4 [3.8, 4.8] and "tier_will" as 4.5 [3.7, 5.0]. But what about the applied and policy implications?

Finalizing the Assessment Summary

I'm putting together the final touches of the assessment summary. It looks like I have to address the applied and policy aspects in a friendly manner, while also noting that the reference group includes serious research in remote sensing and regeneration.

I need to draft approximately 900 words. The schema requires paragraphs instead of bullets, so I'll structure it accordingly. I should be cautious about the confidence intervals for the area of 214.78 to 215.22 million hectares; they seem unrealistically narrow. I'll provide calibration suggestions and ensure to include all necessary disclaimers.

Finally, I'll prepare the JSON string for "paper_assessment_with_tiers_v2" and make sure to craft a comprehensive "assessment_summary." Let's get this all implemented!

ewpage