

Table of Contents in Appendix

A	Extended: Methods	19
A.1	Subject category breakdown	19
A.2	Evaluation criteria for reviewing a paper	19
A.3	Evaluation criteria for reviewing a review	20
A.4	Processing official peer review files	20
A.5	AI reviewer configuration details	20
A.6	AI reviewer prompt	22
A.7	Domain scientist recruitment	25
A.8	Annotation guidelines	25
A.9	Inter-annotator agreement of the expert annotations	26
B	Extended: In which aspects are AI reviewers better or worse than human reviewers?	26
B.1	Item-level descriptive statistics	26
B.2	Complete pairwise paired comparisons	27
B.3	Generalized linear mixed-effects model robustness analysis	27
C	Extended: To what extent do AI reviewers overlap with human reviewers?	28
C.1	Detailed similarity breakdown	29
C.2	Similarity judge calibration and selection	29
C.3	Raw versus Rogan-Gladen-corrected comparison	31
D	Extended: What are the concrete strength and weaknesses of AI reviewers?	32
D.1	Detailed case-study examples	32
D.1.1	W1: Missing community / field norms	32
D.1.2	W2: Over-harsh, out-of-scope, or unrealistic demands	34
D.1.3	W3: Paper explicitly states X, AI says missing	36
D.1.4	W4: Redundancy across the three AI reviewers	38
D.1.5	W5: Vague, verbose, or without actionable recommendation	40
D.1.6	W8: Citing evidence that appeared after the preprint	42
D.1.7	S1: Statistical and methodological rigor	43
D.1.8	S2: Inspecting the submitted source code	45
D.1.9	S3: Domain-specific technical depth	48
D.2	Complete categorized expert comments	50
E	Tools for improving and using AI reviewers	71
E.1	PEERREVIEW BENCH: Can we automatically evaluate AI reviewers?	71
E.2	CMU PAPER REVIEWER: An open-source reviewer platform for pre-submission feedback	72
E.3	AI Meta-reviewers: Can AI agents review the reviews from AI reviewers?	74
E.4	AI meta-reviewer prompt	76
F	Recommendation for journal and conference organizers: Panel composition analysis	81
F.1	Methodology and metric definitions	81
F.2	Panel composition results	81

Broad Subject Area	Mid-level Category	Papers
Physical Sciences (38)	Physics	7
	Optics and photonics	7
	Engineering	6
	Astronomy and planetary science	5
	Chemistry	4
	Nanoscience and technology	4
	Energy science and technology	3
	Materials science	2
Biological Sciences (30)	Neuroscience	7
	Microbiology	5
	Ecology	3
	Computational biology and bioinformatics	3
	Drug discovery	2
	Genetics	2
	Biological techniques	1
	Biophysics	1
	Biotechnology	1
	Evolution	1
	Immunology	1
	Molecular biology	1
	Plant sciences	1
	Systems biology	1
Health Sciences (14)	Medical research	5
	Neurology	4
	Risk factors	3
	Biomarkers	1
	Oncology	1
Total		82

Table 8: Subject category breakdown of the 82-paper dataset.

A Extended: Methods

A.1 Subject category breakdown

Table 8 shows the full subject category breakdown of the 82 papers employed in our expert annotation study. This categorization is based on the *Nature Communications* subject taxonomy⁴.

A.2 Evaluation criteria for reviewing a paper

The six *Nature* peer-review evaluation criteria referenced in § 6.3 follow the *Nature* Portfolio peer review policy and are presented to each AI reviewer agent verbatim. They are also shown to each domain scientist as part of the annotation guidelines. They are defined as follows, ordered by priority.

- **Validity.** Does the manuscript have significant flaws which should prohibit its publication?
- **Conclusions.** Are the conclusions and data interpretation robust, valid, and reliable?
- **Originality and significance.** Are the results presented of immediate interest to many people in the field of study, and/or to people from several disciplines?
- **Data and methodology.** Is the reporting of data and methodology sufficiently detailed and transparent to enable reproducing the results?
- **Appropriate use of statistics and treatment of uncertainties.** Are all error bars defined in the corresponding figure legends, and are all statistical tests appropriate and the description of any error bars and probability values accurate?
- **Clarity and context.** Is the abstract clear and accessible? Are the abstract, introduction, and conclusions appropriate?

⁴<https://www.nature.com/ncomms/browse-subjects>

Criteria are ordered by priority: when the AI reviewer agent selects which criticisms to include in the final review (up to five items sorted from most to least significant), earlier criteria take precedence over later ones. For instance, a criticism based on *Validity* should be prioritized over one based on *Clarity and context*, all else equal. Domain scientists were not instructed to apply this priority when annotating, as their role is to evaluate each individual item on its own merits rather than to rank items.

A.3 Evaluation criteria for reviewing a review

The three-dimensional cascading rubric introduced in § 6.1 is operationalized through the following exact wording presented to each domain scientist in the annotation interface. While the criteria for reviewing a paper (§ A.2) operate at the level of a manuscript, these criteria operate at the level of an individual review item, with each subsequent dimension assessed only when the previous one is satisfied.

- **Correctness** (binary). “Is the main point of the criticism correct and clearly stated?” The annotator selects “Correct” if every aspect of the main point is correct and clearly stated, and “Not Correct” if there is any slight doubt, i.e., at least one aspect is incorrect or not clearly stated.
- **Significance** (ordinal, three-level; only if the item is marked Correct). “Does the main point of the criticism talk about a significant aspect of the paper that is constructive to enhance the paper rather than touching a minor issue?” The annotator selects one of three options:
 - *Significant*: an item that is insightful and helpful for improving the paper.
 - *Marginally Significant*: an item that is not directly helpful for improving the paper but is still worth remaining in the review, e.g., typos, stylistic issues, or suggestions to submit the paper to a different journal.
 - *Not Significant (very marginal issue)*: a very minor item that should not affect the acceptance of the paper and is better removed from the review.
- **Sufficiency of Evidence** (binary; only if the item is marked Correct and at least Marginally Significant). “Is the main point of the criticism well supported by evidence or concrete reasoning?” The annotator selects “Evidence is sufficient” or “Requires more evidence.”

Annotators may also add a free-form comment to any review item.

A.4 Processing official peer review files

We manually extract peer reviews from each paper’s publicly released Peer Review File, retaining only first-round reviewer comments directed at the same pre-review manuscript that AI reviewers evaluate; editor decision letters and author rebuttals are excluded. To maintain a consistent comparison between three AI reviewers and three human reviewers, we cap the number of human reviews at three by retaining the first three reviewers (Reviewer 1, 2, and 3) for papers with four or more reviewers. Each extracted review is then manually decomposed into **review items**, where each review item corresponds to a single atomic criticism directed at one aspect of the paper: bullet-pointed reviews are segmented at each bullet, and paragraph-based reviews are segmented at explicit enumeration markers (e.g., “First,” “Second,”).

A.5 AI reviewer configuration details

This subsection provides the full configuration of the three AI reviewer agents used in the expert annotation study (GPT-5.2, Claude Opus 4.5, and Gemini 3.0 Pro), including the agent framework, per-model hyperparameters, tool specifications, web search domain restrictions, and the full reviewer prompt.

Agent framework Each AI reviewer is implemented as an autonomous agent using the [OpenHands Software Agent SDK](#) [Wang et al., 2025]. The version used to generate the reviews for the expert annotation study described in § 2 is v1.1.0; the version used for the PEERREVIEW BENCH evaluation in § E.1 is v1.5.0. Both versions share the same core architecture; the later version incorporates minor improvements and additional tools that do not affect the reviewer prompt or the agent’s behavior on the tasks reported here.

Shared agent configuration All three agents share the same OpenHands agent settings, summarized in [Table 9](#). Each agent is equipped with three built-in OpenHands tools (`TerminalTool`, `FileEditorTool`, `TaskTrackerTool`) together with the Tavily web search tool described below. The agent uses the `LLMSummarizingCondenser` to manage long conversation histories: when the conversation exceeds a configured threshold, earlier turns are summarized by the same underlying language model while the first three turns (system prompt, task specification, and initial tool outputs) are preserved verbatim.

Parameter	Value
Framework	OpenHands SDK (v1.1.0 / v1.5.0)
Tools (built-in)	TerminalTool, FileEditorTool, TaskTrackerTool
Tools (external)	Tavily MCP web search
Max iterations	5,000
Max review items per paper	5 (sorted by significance)
Criteria preset	Nature (six evaluation criteria)
Condenser	LLMSummarizingCondenser (max_size=200, keep_first=3)
Per-API-call timeout	600 s
Prompt caching	enabled

Table 9: Shared OpenHands agent configuration for all three AI reviewer agents.

Shared LLM configuration All three agents are configured with identical LLM-level settings, passed to the SDK’s LLM class and forwarded to each provider via LiteLLM: `reasoning_effort = "high"`, `extended_thinking_budget = 200,000 tokens`, `temperature = 1.0`, `drop_params = True`, and `num_retries = 5`. The `drop_params = True` setting causes the SDK to silently drop any parameter that is not supported by a given provider: Anthropic models receive all three reasoning-related parameters, OpenAI (GPT) models receive `reasoning_effort` and `temperature`, and Gemini models receive `temperature` with `reasoning_effort` mapped to `thinking_level = "high"`. The temperature of 1.0 is required by Anthropic’s extended thinking mode and applied uniformly across all models for consistency.

Per-model specifications Context window and output length limits for each model are listed in Table 10. Each row reports the maximum input and output tokens used in our configuration, which equal each provider’s published maximum for the given model at the time we generated the AI reviewers for the expert annotation study (Dec 2025).

Model	Provider	Model identifier	Input tokens	Output tokens
GPT-5.2	Azure AI (OpenAI)	azure.ai/gpt-5.2	1,050,000	128,000
Claude Opus 4.5	Anthropic	anthropic/claude-opus-4-5	200,000	64,000
Gemini 3.0 Pro	Google	gemini/gemini-3.0-pro-preview	1,048,576	65,536

Table 10: Model identifier and context/output limits for each of the three AI reviewer agents used in the expert annotation study. All three models support multimodal (image) input.

Tool specifications Each agent has read and write access to the paper’s source files (preprint, supplementary materials, figures, and code) via the following four tools:

- **TerminalTool.** A bash-based terminal for executing shell commands, running code, and inspecting the filesystem.
- **FileEditorTool.** A file editor that supports reading files, creating new files, and applying targeted string replacements for writing the final review and any verification code.
- **TaskTrackerTool.** A lightweight task manager that lets the agent record, update, and mark off a plan of subtasks during the review.
- **Tavily web search.** An external web search tool provided via the [Tavily MCP server](#), used to retrieve relevant literature and external references during the review.

Domain restrictions on web search To prevent the AI reviewer from directly retrieving the published version of a benchmark paper or its existing peer review report, we exclude search results from four domains: `nature.com`, `researchsquare.com`, `springer.com`, and `springerlink.com`. This restriction is enforced at two levels. First, at the tool level, we apply a monkey-patch to the Tavily MCP tool executor that filters out any result whose URL matches one of the blacklisted domains before the result is returned to the agent. Second, at the prompt level, the reviewer prompt explicitly instructs the agent to avoid retrieving content from these domains. Both levels are required because an LLM agent may attempt to paraphrase or reconstruct blocked content, and because tool-level filtering alone does not prevent the agent from issuing queries that could be logged.

A.6 AI reviewer prompt

Figure 6, Figure 7, and Figure 8 reproduce the full reviewer prompt passed to each agent at the start of every (paper, model) run. The prompt is identical across all three models and across all 82 papers; only the placeholder [LINK TO THE PAPER] is replaced at runtime with the absolute path to the paper's source directory, and [MODEL NAME] with the identifier of the model being invoked.

You are a reviewer agent assessing the quality of a research paper.
You will be given the paper's content, images, and optionally its code and supplementary materials.
Your task is to write a review in markdown format, where your review must contain at most five items (from most significant to least significant). Each item represents an atomic criticism of the paper and points out a major issue.
If the paper contains no significant issues, then you can output zero items.

Principles guiding your review (ordered by importance)

1. Your review must be factually correct:
Your claims will be checked by domain experts. Any incorrect or unsupported criticism will undermine the credibility of your review. When uncertain, avoid speculation.
2. Your review must consist of only significant issues:
Only point out problems that meaningfully affect the paper's validity, soundness, methodology, claims, or reproducibility. Do not focus on minor or cosmetic issues. If you think there are less than five significant issues, then you should output less than five items (even zero items are allowed if there are no significant issues).
3. Your review must be concise and only criticize at most five major aspects with detailed evidence:
Each criticism must be supported with detailed evidence. Specifically, mention the contextual background of what the authors attempted to do, and why that was not sufficient when comparing to common practices in the field.

Rules for constructing each item

1. Each item consists of exactly two components: a claim and evidence.
2. The claim is the criticism itself. In the claim, you must clearly state:
 - a. What you are criticizing the paper for.
 - b. On which evaluation criterion or criteria the criticism is based.
 - c. Which component of the paper the criticism refers to.
3. The evidence must directly support the claim. You should quote:
 - a. Exact sentences from the main paper or supplementary materials.
 - b. Exact code blocks or functions from the paper's code.
 - c. Exact sentences from papers in the literature (hyperlinked and cited).
4. At the end of the review, include a citation list containing all literature references used in your evidence.
5. The review must not include an introduction, summary, or concluding remarks. It must contain at most five items and a citation list.
6. All output must be valid markdown.
7. You must separate each item with a blank line.
8. Try to avoid using what the paper listed in the "Limitations" or "Future work" section as your claim unless it is a significant issue.
9. The items should be sorted by their importance.
10. Use the format Item 1, Item 2, ..., with no fraction or denominator.

Figure 6: Reviewer prompt (Part 1 of 3): task description, principles, and rules for constructing each review item.

```

### Required structure and format of each item
Each item must be formatted exactly as follows:

## Item N: <short title summarizing the criticism>

#### Claim
* Main point of criticism: <State what you are criticizing the paper for>
* Evaluation criteria: <which evaluation criteria the criticism is based on>

#### Evidence
* Quote: <Exact sentence(s) 1 from the paper>
  * Comment: <Explanation of why this sentence is problematic>
* Quote: <Exact sentence(s) 2 from the paper>
  * Comment: <Explanation of why this sentence is problematic>
* Quote: <Exact code block 1 from the paper's code>
  * Comment: <Explanation of why this code block is problematic>
* Quote: <Exact sentence(s) from other papers [hyperlinked citation]>
  * Comment: <Explanation of how this contradicts the paper under review>

Each comment should be 5-7 sentences long (a single paragraph).
Insert two empty lines between each item to separate them.

### Required structure and format of the citation list

#### Citation List
[1] <citation 1> (hyperlinked to the retrieved literature)
[2] <citation 2> (hyperlinked to the retrieved literature)
[3] <citation 3> (hyperlinked to the retrieved literature)

There should be at least five citations in the citation list.

### Evaluation criteria (ordered by importance)
1. Validity: Does the manuscript have significant flaws which should
  prohibit its publication?
2. Conclusions: Are the conclusions and data interpretation robust,
  valid and reliable?
3. Originality and significance: Are the results presented of immediate
  interest to many people in the field of study, and/or to people from
  several disciplines?
4. Data and methodology: Is the reporting of data and methodology
  sufficiently detailed and transparent to enable reproducing the results?
5. Appropriate use of statistics and treatment of uncertainties: Are all
  error bars defined in the corresponding figure legends and are all
  statistical tests appropriate and the description of any error bars
  and probability values accurate?
6. Clarity and context: Is the abstract clear, accessible? Are abstract,
  introduction and conclusions appropriate?

Note that earlier evaluation criteria should be prioritized over later
ones when deciding the items in the review.

```

Figure 7: Reviewer prompt (Part 2 of 3): required output format for each review item and the citation list, and the six Nature evaluation criteria ordered by priority.

```

### TODO list for writing your review
- [ ] Read through the paper, supplementary files, and images; construct
  a potential list of items you will criticize.
- [ ] Read through the paper's code, check the functionality of each
  file, and attempt to execute the code if possible. You may
  implement additional code to validate the claims you make.
- [ ] Devise a list of search queries to find relevant literature.
- [ ] Retrieve relevant papers, read them, and update your list of
  criticisms.
- [ ] (Very Important) Iterate through your list and ensure each
  potential criticism is factually correct, significant, and
  eligible for inclusion.
- [ ] Write the review in markdown format and save it to the designated
  review file.

### Guidelines for opening the paper files
The directory to the paper you will be reviewing is [LINK TO THE PAPER].
The directory structure contains: the main paper in Markdown
(preprint.md), a JSON listing the images and their captions, an images
directory, an optional supplementary directory, and an optional code
directory.

### Guidelines for reading the paper's code
1. The code may include a README file that explains the purpose of the
  code and how to run it. Check it before trying to run the code.
2. If the code is not executable, try to resolve dependencies, download
  the necessary datasets, and run the code to validate your claims.
3. Do not try to run the code if it is non-executable or resource-prohibitive.

### Guidelines for retrieving literature
1. Do not iterate through all the papers included in the paper's
  references. Determine which papers are most relevant.
2. Be proactive and add search queries during the review process.
3. It is recommended not only to retrieve academic papers, but also
  blog posts, news articles, datasets, and code repositories.
4. Ensure you actually read what you retrieved.

### Tips
1. The paper's markdown may contain OCR errors. Do not assume the paper
  is incorrect solely because of OCR mistakes. Do not point out that
  the manuscript is incomplete due to formatting issues.
2. Image filenames are guaranteed to be figure1.png, figure2.png, etc.
  Do not point out broken or missing figure assets.
3. The code you are reviewing does not need to be perfect; focus on
  major issues such as non-reproducible experiments or mismatches
  with descriptions rather than minor issues.
4. When refining your review, ensure that all items are factually
  correct, significant, and mutually exclusive.

```

Figure 8: Reviewer prompt (Part 3 of 3): task workflow, guidelines for reading the paper and its code, guidelines for retrieving literature, and additional tips.

Item 3

In their previous work published in Adv. Mater. Interfaces 9, 2200776 (2022), the authors demonstrated tunable plasmonic resonances in graphene devices. The concept of controlling plasmonic resonance using a ferroelectric substrate appears to be quite similar. The primary advancement showcased in the current paper is the achievement of photodetection with the tunable responsive spectrum. However, regarding this as a minor enhancement in device functionality, its impact may not be substantial enough for publication in Nature Communications without the realization of scalable sensor arrays.

Evaluation

Is the main point of the criticism correct and clearly stated?

☐ Correct
☐ Not Correct (i.e., there is at least one incorrect component included)

(If main point of criticism was correct above,) Does the main point of the criticism talk about a significant aspect of the paper that is constructive to enhance the paper rather than touching a minor issue?

☐ Significant ☐ Marginally Significant
☐ Not Significant (Very marginal issue)

(If main point of criticism was correct and it was either Significant / Marginally Significant above,) Is the main point of the criticism well supported by evidence or concrete reasoning?

☐ Evidence is sufficient ☐ Requires more evidence

Are there any comments you would like to add about this criticism?

After reading the reviews of the paper "Type-printable photodetector arrays for multichannel meta-infrared imaging", which human reviewer do you think provided the best quality review overall?

☐ Reviewer 1 ☐ Reviewer 2 ☐ Reviewer 3

After reading the reviews of the paper "Type-printable photodetector arrays for multichannel meta-infrared imaging", which human reviewer do you think provided the worst quality review overall?

☐ Reviewer 1 ☐ Reviewer 2 ☐ Reviewer 3

Compared to the best human review, which AI-generated reviews do you think match or exceed that level of quality? (click all that apply)

☐ AI Reviewer 1 ☐ AI Reviewer 2 ☐ AI Reviewer 3

Compared to the worst human review, which AI-generated reviews do you think match or exceed that level of quality? (click all that apply)

☐ AI Reviewer 1 ☐ AI Reviewer 2 ☐ AI Reviewer 3

Did you notice anything in AI Reviewer 1's review that other reviewers (both AI and human) did not catch?

Figure 9: The annotation sheet presented to each domain scientist. **(Left) item-level annotation** for a single review item. The domain scientist judges its *correctness* (binary), *significance* (ordinal, three-level), and *sufficiency of evidence* (binary) along a cascading structure in which each dimension is only assessed when the previous one is satisfied, and may add an optional free-form comment. **(Right) paper-level overall survey** completed after all individual items are evaluated. The domain scientist selects the top-rated and lowest-rated human reviewer, identifies the AI reviewers whose overall review quality matches or exceeds each of these two human references via multi-select, and optionally notes, for each AI reviewer, anything that AI reviewer identified that others missed.

A.7 Domain scientist recruitment

We recruited the 45 domain scientists described in § 6.4 through a three-stage process conducted between September 2025 and April 2026. We first contacted candidate scientists via cold email describing the project and the expected time commitment. Candidates who expressed interest were then sent a list of 10 papers from our dataset and asked to identify papers they felt qualified to meta-review. Once a matching paper was confirmed, the candidate received the AI reviews (generated as described in § A.5) together with the extracted human review comments for that paper in the form of the annotation sheet described in § A.8 (see Figure 9), with at least one month allotted for completion. Annotation required approximately three hours per paper on average.

A.8 Annotation guidelines

This subsection reproduces the workflow-level annotation guidelines distributed to each domain scientist at the start of the annotation task. The item-level evaluation criteria themselves are in § A.3. Each annotator received these guidelines together with the custom-built PDF annotation sheet (see Figure 9) containing the paper’s human and AI reviews decomposed into review items. No training or calibration session was conducted before the annotation task; annotators were asked to contact the corresponding author if they had any questions about the guidelines.

Task description Each domain scientist is asked to check the quality of criticisms included in the human reviews and AI reviews of a paper they are qualified to meta-review. The resulting meta-reviews (reviews of reviews) are used for (i) comparing the quality of human reviews and AI reviews across multiple dimensions, and (ii) constructing PEERREVIEW BENCH, a benchmark for evaluating future generations of AI reviewer agents (§ E.1).

Review item structure Each point of criticism in a review is divided into atomic units called **review items**. For AI reviews, each item consists of a *Main Point*, *Evaluation Criteria*, and *Evidence* (see Figure 5 for an example). The concatenation of these three components is treated as a single item. For human reviews, which are written in a more free-form manner, the decomposition into items is performed manually by the authors as described in § 6.1, and each item is presented to the annotator as a single paragraph.

Item-level evaluation For each review item, the domain scientist provides three judgments (correctness, significance, sufficiency of evidence) structured as a cascading evaluation. The exact criteria, options, and wording presented to annotators are given in § A.3.

Dimension	Review type	<i>N</i>	% Agree.	Cohen’s κ	Gwet’s AC1
Correctness (binary)	Human	568	84.7%	0.12 (slight)	0.81 (alm. perf.)
	AI	340	87.6%	0.47 (moderate)	0.84 (alm. perf.)
	All	908	85.8%	0.28 (fair)	0.82 (alm. perf.)
Significance (3-way ord.)	Human	470	60.2%	0.34 (fair)	0.43 (moderate)
	AI	273	59.3%	0.22 (fair)	0.45 (moderate)
	All	743	59.9%	0.31 (fair)	0.44 (moderate)
Evidence (binary)	Human	353	83.9%	0.06 (slight)	0.81 (alm. perf.)
	AI	230	94.3%	0.29 (fair)	0.94 (alm. perf.)
	All	583	88.0%	0.12 (slight)	0.86 (alm. perf.)

Table 11: Full inter-annotator agreement breakdown across 908 doubly-annotated review items from 27 doubly-annotated papers, separated by reviewer type (human vs. AI). *N* decreases across dimensions due to the cascading annotation structure: significance is only annotated for correct items, and evidence sufficiency only for items that are correct and at least marginally significant. Interpretation labels follow Landis and Koch [1977]. *alm. perf.* = almost perfect agreement.

Paper-level overall survey After evaluating all individual review items, the domain scientist completes a paper-level survey comparing the reviewers, as summarized in § 6.4. The survey asks:

- “Which human reviewer do you think provided the best quality review overall?” (single-select among the paper’s human reviewers).
- “Which human reviewer do you think provided the worst quality review overall?” (single-select).
- “Compared to the best human review, which AI-generated reviews do you think match or exceed that level of quality?” (multi-select among the three AI reviewers).
- “Compared to the worst human review, which AI-generated reviews do you think match or exceed that level of quality?” (multi-select).
- For each AI reviewer: “Did you notice anything in AI Reviewer *k*’s review that other reviewers (both AI and human) did not catch?” (free-form).

Annotators may also add a free-form comment on the paper’s review history as a whole. The full annotation sheet and the code used to generate it are released as part of our public code release.

A.9 Inter-annotator agreement of the expert annotations

As shown in Table 11, comparing the human and AI subsets, Gwet’s AC1 is consistently high on correctness and sufficiency of evidence, and moderate on significance, across both subsets. Raw percent agreement is similar between human and AI items on correctness (84.7% vs. 87.6%) and significance (60.2% vs. 59.3%), but higher on AI items for evidence (94.3% vs. 83.9%). The pattern across dimensions mirrors the main-text result: judgments of correctness and evidence sufficiency are largely stable across annotators, while significance is the most subjective dimension.

B Extended: In which aspects are AI reviewers better or worse than human reviewers?

In this section, we provide the supplementary analyses referenced in § 2: item-level descriptive statistics across all reviewers and dimensions (§ B.1), complete pairwise paired comparisons across the five reviewer groups on each primary evaluation metric (§ B.2), and a generalized linear mixed-effects model (GLMM) robustness analysis with paper-level random intercepts (§ B.3).

B.1 Item-level descriptive statistics

Table 12 reports item-level counts and rates for each reviewer across the three evaluation dimensions (correctness, significance, evidence sufficiency). The main text (Table 1) reports paper-level means, which is the appropriate unit of analysis for the paired inferential tests. We include item-level rates here for completeness and for comparison with the GLMM analysis in § B.3, which directly models item-level outcomes with paper-level random intercepts. The item-level rates are close to but not identical to the paper-level means because papers contribute different numbers of items: item-level rates implicitly weight papers by their item count, while the main-text paper-level means weight each paper equally.

Reviewer	Correctness				Significance					Evidence			
	Rate [95% CI]	<i>n</i>	Breakdown		Mean [95% CI]	<i>n</i>	Breakdown			Rate [95% CI]	<i>n</i>	Breakdown	
			Corr.	Incorr.			Sig.	Marg.	Not			Suff.	Insuff.
Top-Rated Human	93.6 [92.0, 94.9]	1,139	1,066	73	1.38 [1.33, 1.42]	1,066	570	330	166	93.0 [91.1, 94.5]	900	837	63
Lowest-Rated Human	86.0 [83.4, 88.1]	833	716	117	1.06 [0.99, 1.12]	716	272	212	232	88.2 [85.0, 90.8]	484	427	57
GPT-5.2	86.9 [83.4, 89.7]	442	384	58	1.60 [1.54, 1.66]	384	260	94	30	97.5 [95.2, 98.7]	354	345	9
Claude Opus 4.5	84.6 [81.1, 87.6]	475	402	73	1.56 [1.50, 1.62]	402	261	106	35	97.8 [95.8, 98.9]	367	359	8
Gemini 3.0 Pro	82.6 [78.9, 85.8]	460	380	80	1.59 [1.53, 1.65]	380	249	105	26	91.8 [88.5, 94.2]	354	325	29

Table 12: **Item-level descriptive statistics across all evaluation dimensions.** Correctness and evidence rates use Wilson score 95% CIs; significance uses non-parametric bootstrap CIs (10,000 iterations). The cascading rubric (§ 6.1) explains the decreasing *n* across metrics: significance is annotated only on items rated Correct, evidence only on items rated Correct *and* at least Marginally Significant. Paper-level statistics and paired inferential comparisons are in Table 1.

Comparison	Correctness diff (Cohen’s <i>d</i>)	Significance diff (rank-biserial <i>r</i>)	Evidence Sufficiency diff (Cohen’s <i>d</i>)	<i>n</i> _{paired} range
Top-Rated H. vs Lowest-Rated H.	+13.2%*** (+0.54)	+0.13 (+0.26)	+2.7% (+0.11)	77–82
Top-Rated H. vs GPT-5.2	+6.1%* (+0.23)	−0.22*** (−0.49)	−5.0%* (−0.23)	79–81
Top-Rated H. vs Claude Opus 4.5	+8.6%** (+0.34)	−0.13* (−0.30)	−4.8%* (−0.24)	78–81
Top-Rated H. vs Gemini 3.0 Pro	+10.3%*** (+0.42)	−0.17** (−0.42)	+2.4% (+0.10)	80–82
Lowest-Rated H. vs GPT-5.2	−7.4% (−0.22)	−0.30*** (−0.56)	−5.3%* (−0.23)	74–81
Lowest-Rated H. vs Claude Opus 4.5	−4.8% (−0.13)	−0.24** (−0.43)	−7.1%* (−0.27)	73–81
Lowest-Rated H. vs Gemini 3.0 Pro	−2.9% (−0.09)	−0.27** (−0.46)	+0.3% (+0.01)	75–82
GPT-5.2 vs Claude Opus 4.5	+2.8% (+0.11)	+0.11* (+0.30)	+2.0% (+0.13)	75–80
GPT-5.2 vs Gemini 3.0 Pro	+4.5% (+0.18)	+0.05 (+0.15)	+7.0%** (+0.31)	77–81
Claude Opus 4.5 vs Gemini 3.0 Pro	+1.5% (+0.06)	−0.06 (−0.22)	+6.6%* (+0.24)	78–81

Table 13: **All pairwise paired comparisons across the five reviewer groups, on each of the three primary evaluation metrics.** Each row reports the paired difference A − B with effect size in parentheses (Cohen’s *d* for binary metrics, rank-biserial *r* for ordinal significance); positive = first-named reviewer higher. *n*_{paired} range reflects cascade-eligible paired-paper counts across the three metrics (highest for correctness; lower for significance and evidence due to additional cascade conditions). **p* < 0.05, ***p* < 0.01, ****p* < 0.001. Corresponding fully-positive comparisons are in Table 2.

B.2 Complete pairwise paired comparisons

Table 1 in the main text reports the six pairwise comparisons of each AI reviewer against the two human baselines. For completeness, Table 13 reports the full set of $\binom{5}{2} = 10$ pairwise comparisons across all five reviewer groups on each of the three primary metrics. Binary metrics (correctness, evidence sufficiency) use paired *t*-tests on per-paper rate differences with Cohen’s *d* as the effect size; the three-level ordinal significance score uses Wilcoxon signed-rank tests with rank-biserial correlation *r* as the effect size. Test statistics and degrees of freedom for all primary comparisons are reported in Table 14.

Three patterns in Table 13 complement the main-text findings. First, all three AI reviewers raise more significant issues than the Lowest-Rated Human at large effect sizes (*r* = −0.43 to −0.56), mirroring the Top-Rated Human comparison. Second, on correctness, the gap between the Lowest-Rated Human and each AI reviewer is directionally negative but smaller in magnitude than the Top-Rated Human gap, and does not reach significance for any AI reviewer. Third, among the three AI reviewers, GPT-5.2 has a modestly higher significance rate than Claude Opus 4.5 (*r* = +0.30, *p* < .05) and a higher evidence-sufficiency rate than Gemini 3.0 Pro (*d* = +0.31, *p* < .01); Claude Opus 4.5 and Gemini 3.0 Pro are indistinguishable on significance and on correctness.

B.3 Generalized linear mixed-effects model robustness analysis

As a robustness check against the paper-level paired analysis in § 2, we fit a generalized linear mixed-effects model (GLMM) that directly models item-level outcomes with a paper-level random intercept. This complements the paired analysis in two ways: (i) it explicitly models the hierarchical structure of the data (items nested within papers) rather than aggregating to the paper level, yielding greater statistical power when the within-paper variance is small; and (ii) it separates within-paper from between-paper variance via the intraclass correlation coefficient (ICC).

Comparison	Dimension	Test	n	Statistic	df
<i>Reference: Top-Rated Human (positive statistic = human higher)</i>					
vs GPT-5.2	Correctness	paired t	81	$t = +2.07$	80
vs GPT-5.2	Significance	Wilcoxon	79	$W^\ddagger$	—
vs GPT-5.2	Evidence	paired t	79	$t = -2.04$	78
vs Claude Opus 4.5	Correctness	paired t	81	$t = +3.06$	80
vs Claude Opus 4.5	Significance	Wilcoxon	79	$W^\ddagger$	—
vs Claude Opus 4.5	Evidence	paired t	78	$t = -2.12$	77
vs Gemini 3.0 Pro	Correctness	paired t	82	$t = +3.80$	81
vs Gemini 3.0 Pro	Significance	Wilcoxon	81	$W^\ddagger$	—
vs Gemini 3.0 Pro	Evidence	paired t	80	$t = +0.89$	79
<i>Reference: Lowest-Rated Human (positive statistic = human higher)</i>					
vs GPT-5.2	Correctness	paired t	81	$t = -1.98$	80
vs GPT-5.2	Significance	Wilcoxon	75	$W^\ddagger$	—
vs GPT-5.2	Evidence	paired t	74	$t = -1.98$	73
vs Claude Opus 4.5	Correctness	paired t	81	$t = -1.17$	80
vs Claude Opus 4.5	Significance	Wilcoxon	75	$W^\ddagger$	—
vs Claude Opus 4.5	Evidence	paired t	73	$t = -2.31$	72
vs Gemini 3.0 Pro	Correctness	paired t	82	$t = -0.81$	81
vs Gemini 3.0 Pro	Significance	Wilcoxon	77	$W^\ddagger$	—
vs Gemini 3.0 Pro	Evidence	paired t	75	$t = +0.09$	74

Table 14: **Test statistics and degrees of freedom for the primary human-vs-AI comparisons, complementing Table 13** (which reports the paired differences, effect sizes, and exact p -values). Two-sided paired t -tests for the binary dimensions, with $df = n_{\text{paired}} - 1$; two-sided Wilcoxon signed-rank tests for the ordinal significance dimension, for which no degrees of freedom apply. t statistics are computed as $d_z\sqrt{n}$ and should be regenerated from full-precision d_z so each reproduces its exact p . ‡ Signed-rank statistic W and normal-approximation z are in the released analysis output. Remaining AI-vs-AI pairwise statistics are available in the same output.

Model specification For each of the three primary metrics, we fit a Bayesian binomial generalized linear mixed model (BinomialBayesMixedGLM with variational Bayes) of the form:

$$\text{logit}(P(y_{ij} = 1)) = \beta_0 + \sum_k \beta_k \cdot \mathbb{I}[\text{reviewer}_{ij} = k] + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_u^2)$$

where y_{ij} is the item-level outcome for item j in paper i , $\mathbb{I}[\text{reviewer}_{ij} = k]$ is a reviewer indicator (dummy-coded against a chosen reference category), and u_i is the paper-level random intercept. For the ordinal significance score, we fit the same binomial GLMM at the “Significant” cut point ($y_{ij} = 1$ iff the item is rated at the highest significance level, $P(Y = 2)$), rather than a full cumulative-link mixed model, to keep the three metrics on a common log-odds scale and directly interpretable as the probability of passing the “Significant” bar conditional on being correct.

Interpretation Coefficients in Table 15 are on the log-odds scale; a positive coefficient means the row reviewer has higher log-odds of the outcome than the reference category. Probability-scale changes at the reference can be computed as $\sigma(\text{intercept} + \text{coef}) - \sigma(\text{intercept})$, where σ is the logistic function. The two halves of the table refit the same model with a different reference category (Top-Rated or Lowest-Rated Human) so that all pairwise contrasts to either human baseline can be read directly off a single row.

Concordance with paper-level paired analysis The GLMM reaches the same directional conclusions as the paper-level paired analysis in § 2 on all three dimensions: every AI reviewer has significantly lower correctness and significantly higher significance than the Top-Rated Human, and GPT-5.2 and Claude Opus 4.5 have significantly higher evidence sufficiency while Gemini 3.0 Pro is indistinguishable. Several contrasts that are borderline in the paired analysis reach stronger significance in the GLMM (e.g., Top-Rated vs GPT-5.2 correctness: paired $p = .046$, GLMM $p < .001$), which is consistent with the GLMM’s greater statistical power from modeling items directly rather than averaging to the paper level. The paper-level ICC of 0.25 to 0.29 indicates that roughly one quarter of the latent log-odds variance sits between papers, justifying the random-intercept specification and, equivalently, the paired structure of the main-text analysis.

C Extended: To what extent do AI reviewers overlap with human reviewers?

This appendix collects the supplementary analyses referenced in § 3: the full 4-category similarity breakdown across all pair types with within-reviewer and between-reviewer splits (§ C.1), the similarity judge calibration and

	Reference: Top-Rated Human			Reference: Lowest-Rated Human		
	Correctness	Significance	Evidence	Correctness	Significance	Evidence
Intercept	+2.860*** (0.057)	+0.228*** (0.041)	+3.050*** (0.086)	+1.925*** (0.057)	−0.251*** (0.041)	+2.706*** (0.086)
Top-Rated Human	—	—	—	+0.968*** (0.126)	+0.479*** (0.067)	+0.356** (0.137)
Lowest-Rated Human	−0.931*** (0.106)	−0.488*** (0.083)	−0.344* (0.151)	—	—	—
GPT-5.2	−0.639*** (0.149)	+0.684*** (0.117)	+1.173*** (0.339)	+0.295* (0.149)	+1.163*** (0.117)	+1.508*** (0.338)
Claude Opus 4.5	−0.846*** (0.136)	+0.506*** (0.113)	+1.314*** (0.358)	+0.090 (0.136)	+0.985*** (0.113)	+1.648*** (0.356)
Gemini 3.0 Pro	−1.033*** (0.132)	+0.518*** (0.116)	−0.092 (0.204)	−0.098 (0.132)	+0.996*** (0.116)	+0.250 (0.204)
ICC (paper)	0.253	0.272	0.289	0.253	0.272	0.290
Observations	3,349	2,948	2,459	3,349	2,948	2,459
Papers	82	82	82	82	82	82

Table 15: **Mixed-effects logistic regression confirms that AI reviewers raise more significant issues but are less factually correct than the Top-Rated Human reviewer.** Coefficients on the log-odds scale from a Bayesian binomial GLMM (BinomialBayesMixedGLM, variational Bayes, paper random intercept); for significance, fit at the highest cut point ($P(Y = 2)$). The two halves refit the same model with a different reference category so that all pairwise contrasts to either human baseline read off a single row. Probability change at the reference: $\sigma(\text{intercept} + \text{coef}) - \sigma(\text{intercept})$, $\sigma = \text{logistic}$. SE in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. ICC (paper) = $\sigma_u^2 / (\sigma_u^2 + \pi^2/3)$, the share of latent variance attributable to between-paper differences.

selection details (§ C.2), and a side-by-side comparison of raw versus Rogan-Gladen-corrected prevalences (§ C.3).

C.1 Detailed similarity breakdown

Table 16 reports the full 4-category distribution of cross-reviewer review-item pairs for each pair type (Human-Human, AI-AI, and Human-AI), with within-reviewer and between-reviewer splits for the two same-group pair types. Figure 2 (left panel) in the main text visualizes the “diff reviewer”, “diff model”, and H-A rows of this table; the extended table adds the same-reviewer and same-model rows, which capture the diversity of items raised by a single reviewer within one paper rather than the agreement between different reviewers.

Two observations complement the main-text findings. First, the within-reviewer similarity rates (same-reviewer H-H: 6.0%; same-model A-A: 13.5%) are higher than the corresponding between-reviewer rates (diff-reviewer H-H: 3.4%; diff-model A-A: 20.9%) *only* for human reviewers; for AI reviewers, different models of the same paper converge on the same criticism more often than two items written by the same model in the same review. This is consistent with the view that a single AI reviewer deliberately diversifies its own items across different parts of a paper, while different AI models, each independently identifying the paper’s top concerns, end up surfacing many of the same targets. Second, the “same target, different criticism” category (topical co-occurrence without agreement on what is wrong) is roughly stable across pair types at the between-reviewer level (H-H: 8.4%, A-A: 20.2%, H-A: 14.8%) and is responsible for a substantial share of the H-A difference from the H-H baseline, suggesting that AI reviewers often flag the same parts of a paper that human reviewers do, even when their specific criticisms do not align at the threshold used in the main text.

C.2 Similarity judge calibration and selection

To select an automated similarity judge for the full-scale analysis, we compared candidate models on a small, manually-annotated calibration set, then characterized the chosen model’s error rates for use in the Rogan-Gladen prevalence correction. This subsection describes each step in turn.

Comparison of candidate judges We compared two families of automated similarity judges on a 164-pair similarity calibration set (described in the next paragraph). The first family is **embedding models** that compute cosine similarity between item embeddings (OpenAI’s `text-embedding-3-large` or Gemini’s `gemini-embedding-001`), with the binary similarity threshold tuned on the similarity calibration set. The second family is **LLM-as-judge models** that prompt a frontier language model with both review items, the paper’s text, relevant images, and any supplementary materials, to produce a 4-way label using extended-thinking reasoning. The LLM judges include

Pair type	<i>N</i>	Same tgt., same crit., same evid.	Same tgt., same crit., diff evid.	Same tgt., diff crit.	Different tgt.	<i>P</i> (similar)
H–H (all)	34,931	0.4 [0.0, 0.7]	4.1 [0.0, 8.2]	8.4 [6.9, 10.5]	87.1 [81.4, 91.8]	4.5 [0.0, 8.9]
same reviewer	15,030	0.5 [0.1, 0.9]	5.5 [0.9, 9.7]	8.4 [6.9, 10.5]	85.7 [79.8, 90.9]	6.0 [0.9, 10.5]
diff reviewer	19,901	0.3 [0.0, 0.6]	3.1 [0.0, 7.1]	8.4 [6.9, 10.8]	88.3 [82.6, 92.3]	3.4 [0.0, 7.7]
A–A (all)	6,053	2.2 [1.6, 2.8]	16.6 [12.3, 20.6]	20.7 [18.3, 23.3]	60.5 [56.4, 64.8]	18.8 [14.0, 23.2]
same model	1,724	0.1 [0.0, 0.3]	13.4 [8.2, 17.9]	22.0 [19.0, 25.2]	64.6 [60.0, 69.6]	13.5 [8.3, 18.1]
diff model	4,329	3.1 [2.3, 3.9]	17.9 [13.8, 21.8]	20.2 [17.7, 22.7]	58.9 [54.9, 63.0]	20.9 [16.2, 25.4]
H–A	23,642	0.3 [0.0, 0.5]	4.9 [0.2, 8.5]	14.8 [13.3, 16.5]	80.1 [76.2, 84.4]	5.1 [0.3, 9.0]

Table 16: **Review-item-pair similarity across pair types.** $P(\text{similar})$ is the sum of the first two columns (same target, same criticism, regardless of evidence). Within H–H and A–A, rows are further split by whether both items come from the same reviewer (or the same AI model within a paper): “same reviewer”/“same model” rows capture within-reviewer pairs; “diff reviewer”/“diff model” rows capture between-reviewer pairs. All pairs are within the same paper. All values are Rogan-Gladen-corrected; 95% CIs from 10,000-iteration paper-level bootstrap with resampled sensitivity/specificity.

Model	Agreement with Experts on “Similarity”	
	Binary [95% CI]	4-way [95% CI]
<i>Embedding models</i>		
text-embedding-3-large	77.4 [70.5, 83.2]	—
gemini-embedding-001	80.5 [73.8, 85.8]	—
<i>LLM-as-judge models</i>		
GPT-5.4	92.7 [87.6, 95.8]	83.5 [77.1, 88.4]
Gemini 3.1 Pro	92.7 [87.6, 95.8]	81.1 [74.4, 86.4]
Claude Sonnet 4.6	92.7 [87.6, 95.8]	80.5 [73.8, 85.8]
Claude Opus 4.7	92.7 [87.6, 95.8]	79.9 [73.1, 85.3]
Gemini 3 Flash	92.1 [86.9, 95.3]	79.9 [73.1, 85.3]
GPT-5.4-mini	89.0 [83.3, 92.9]	80.5 [73.8, 85.8]

Table 17: **Agreement with expert decisions on the 164-pair similarity calibration set.** Binary judgment: same target and same main criticism vs. not. 4-way judgment: exact match with the full taxonomy. 95% Wilson score CIs in brackets.

616 six recent frontier models from three families: GPT-5.4 and GPT-5.4-mini from OpenAI, Claude Opus 4.7 and
617 Claude Sonnet 4.6 from Anthropic, and Gemini 3.1 Pro and Gemini 3 Flash from Google. As shown in Table 17,
618 embedding baselines reach 77.4% and 80.5% binary accuracy, while the six LLM judges cluster tightly at 89 to
619 93% binary accuracy and 80 to 84% 4-way accuracy, with no systematic gap between families. For the full-scale
620 overlap analysis, we use GPT-5.4 (92.7% binary accuracy, 83.5% 4-way accuracy), which achieves the highest
621 4-way accuracy among candidate similarity judges.

622 **Similarity calibration set construction** The 164-pair similarity calibration set is drawn from domain scientists’
623 paper-level survey responses to the free-form question “Did you notice anything in AI Reviewer k ’s review that other
624 reviewers did not catch?”. Specifically, some domain scientists explicitly listed specific AI review items as unique to
625 that reviewer (*i.e.*, having no counterpart in any other reviewer’s review of the paper), which provided ground-truth
626 not-similar pairs; we then identified candidate similar pairs by manually matching AI review items to items from
627 other reviewers of the same paper. The 94 *not similar* pairs were sampled from annotator-flagged unique items and
628 manually subdivided into *different target* (67 pairs) or *same target, different criticism* (27 pairs); the 70 *similar*
629 pairs were sampled from AI items not flagged as unique, manually matched to items from other reviewers, and
630 subdivided into *convergent conclusion* (48 pairs) or *near-paraphrase* (22 pairs). The resulting 164-pair similarity
631 calibration set covers 32 papers and 266 unique review items, with 85 AI-AI pairs and 79 AI-human pairs. Because
632 its labels derive from the same pool of annotations used throughout our study, we treat it as a calibration set for our
633 automated judge rather than an independent benchmark.

634 **Confusion matrix of the chosen judge** Table 18 reports GPT-5.4’s binary similar / not-similar predictions on the
635 164-pair similarity calibration set, from which we derive the judge’s sensitivity (87.1%) and specificity (96.8%);
636 these two values are the inputs to the Rogan-Gladen prevalence correction below.

637 **Rogan-Gladen prevalence correction** Given an apparent prevalence p of similar pairs among classifier predictions
638 on a large set, the corrected estimate of the true prevalence π is:

$$\pi = \frac{p + \text{Spec} - 1}{\text{Sens} + \text{Spec} - 1} = \frac{p - 0.032}{0.840}. \quad (1)$$

	Predicted Similar	Predicted Not Similar	Total
True Similar	61 (TP)	9 (FN)	70
True Not Similar	3 (FP)	91 (TN)	94
Total	64	100	164

Table 18: Confusion matrix of GPT-5.4 on the binary similar/not-similar classification task over the 164-pair similarity calibration set. Sensitivity = $61/70 = 0.871$; specificity = $91/94 = 0.968$.

Pair type	Category	<i>N</i>	Raw (%)	RG-corrected (%)	95% CI
H-H (diff reviewer)	Same target, same crit., same evid.	19,901	0.5	0.3	[0.0, 0.6]
	Same target, same crit., diff evid.		5.5	3.1	[0.0, 7.1]
	Same target, diff criticism		8.1	8.4	[6.9, 10.8]
	Different target		85.8	88.3	[82.6, 92.3]
H-A	Same target, same crit., same evid.	23,642	0.4	0.3	[0.0, 0.5]
	Same target, same crit., diff evid.		7.1	4.9	[0.2, 8.5]
	Same target, diff criticism		14.4	14.8	[13.3, 16.5]
	Different target		78.1	80.1	[76.2, 84.5]
A-A (diff model)	Same target, same crit., same evid.	4,329	3.0	3.1	[2.3, 3.9]
	Same target, same crit., diff evid.		17.7	17.9	[13.7, 21.8]
	Same target, diff criticism		20.2	20.2	[17.7, 22.7]
	Different target		59.0	58.9	[54.8, 63.1]

Table 19: **Raw versus Rogan-Gladen-corrected similarity prevalences for Figure 2 (left).** The correction adjusts for the GPT-5.4 similarity judge’s binary misclassification rate (sensitivity = 87.1%, specificity = 96.8%, on the 164-pair similarity calibration set). It is applied to the binary similar/not-similar boundary; within each binary group, the 4-way split is maintained proportionally from the raw data. 95% CIs are cluster-bootstrap (10,000 paper-level resamples) with sensitivity and specificity themselves resampled from Binomial distributions per iteration to propagate calibration uncertainty.

The correction is valid when $p \geq 1 - \text{Spec} = 0.032$; when apparent prevalences fall below this threshold we clip the corrected estimate to zero.

Cluster-bootstrap confidence intervals To construct 95% confidence intervals for each Rogan-Gladen-corrected prevalence, we use a cluster bootstrap at the paper level. In each of 10,000 bootstrap iterations, we (i) resample 82 papers with replacement, (ii) pool all cross-reviewer pairs of the relevant type from the resampled papers to form a new apparent prevalence p^* , (iii) resample the similarity calibration set’s sensitivity and specificity from binomials with the observed parameters ($\text{Sens} \sim \text{Binomial}(70, 0.871)/70$ and $\text{Spec} \sim \text{Binomial}(94, 0.968)/94$), and (iv) apply the Rogan-Gladen formula to p^* , Sens^* , Spec^* . The 2.5th and 97.5th percentiles of the resulting distribution give the 95% confidence interval. This procedure propagates both the within-paper classifier error (via sensitivity/specificity resampling) and the cross-paper variance (via paper-level resampling).

Limitations The Rogan-Gladen correction assumes that the judge’s sensitivity and specificity are constant across the full set of 65,704 pairs. Since the similarity calibration set does not include Human-Human pairs (it was constructed around AI reviews), we assume that the judge’s error rates transfer to Human-Human comparisons; this is a plausible but untested assumption.

C.3 Raw versus Rogan-Gladen-corrected comparison

All percentages in Figure 2 (left) and Table 16 are adjusted for the GPT-5.4 similarity judge’s classification error using the Rogan-Gladen prevalence correction. Table 19 reports the raw (uncorrected) prevalences alongside the corrected values for the three between-reviewer pair types that appear in the left panel of Figure 2, so that readers can assess the magnitude and direction of the correction.

The correction shifts the same-target-same-criticism categories downward (reflecting the judge’s false-positive rate on the similarity calibration set) and shifts the not-similar categories upward, consistent with the judge’s estimated sensitivity of 87.1% and specificity of 96.8%. The magnitude of the shift is largest for the H-H and H-A pair types, where the raw similar-pair rates are low (around 5 to 8%) and are therefore more sensitive to the asymmetry between the judge’s false-positive and false-negative rates; for A-A pairs, where raw similar rates are already substantial (around 20%), the correction changes the estimate by less than one percentage point. The main-text conclusions are qualitatively unchanged by whether the raw or corrected values are used, but we report the corrected values throughout for consistency with the calibration-set sensitivity/specificity estimates.

D Extended: What are the concrete strength and weaknesses of AI reviewers?

This appendix contains the complete qualitative record behind the strength/weakness analysis introduced in § 4. § D.1 presents 24 detailed case-study examples spanning the nine most informative categories, each shown with the original review item, the expert annotator's commentary, and our discussion. § D.2 lists every one of the 442 substantive expert comments behind the S/W classification, organized by category in longtable form, so that every claim in the main text and every example in § D.1 is traceable back to the underlying data.

D.1 Detailed case-study examples

This subsection contains the complete set of expert-annotated strength and weakness examples for the categories introduced in § 4. Three of these examples (W1 Example #1, W3 Example #2, S2 Example #1) are also shown in the main text; they are reproduced here so that the subsection is self-contained as a complete atlas of all 24 examples. The examples are organized by category in the same numbering used in Figure 3, and each example preserves the structure of the main-text examples (review-item box with claim, criticism, evidence, and AI commentary; expert-comment box; and a short discussion paragraph).

D.1.1 W1: Missing community / field norms

The three examples below illustrate the severity-miscalibration pattern in three different subfields (particle physics, bacterial community modeling, agricultural ecology). Example #1 is also shown in the main text (??) and reproduced here for completeness.

AI Reviewer Weakness #1: Missing community / field norms ($n = 54$)

The AI reviewer flags something as a methodological gap that is in fact accepted or expected practice within the subfield of the paper. The content of the critique is typically accurate against a discipline-neutral standard (for reproducibility, sample size, out-of-distribution validation, and so on), but the severity is miscalibrated because the reviewer lacks the accumulated knowledge of what the field treats as normal.

Example #1 for W1 — Particle physics paper, reviewer: GPT-5.2

Claim: “Analysis choices are not documented enough for independent reproduction”

- **Main point of criticism:** The manuscript's analysis choices (simulation-fixed fit shapes, external calibration inputs, and pseudoexperiment-based uncertainties) are not documented in enough detail, and are not accompanied by reusable released artifacts, to enable independent reproduction or reanalysis of the key result.
- **Evaluation criteria:** Data and methodology; Reproducibility

Evidence:

Quote 1 (from main text): “The signal shape is modelled as a combination of a Gaussian and two Crystal Ball functions, all sharing the same peak position. The parameters of the signal function are determined from simulated events and are fixed in the fit to data, except for the Gaussian width, the average width of the Crystal Ball functions, and the peak position, which accounts for imperfections in the simulation.”

| **Comment:** The extracted yields, and therefore the measured asymmetry, depend on simulation-derived shape parameters that are fixed rather than floated in the data fit. For an external reader, reproducing the measurement requires access not only to the data but also to the exact simulation configuration, the derived fixed parameter values, and the fitting implementation. The manuscript does not tabulate these fixed parameters or provide a machine-readable specification of the full likelihood model, which would make independent reproduction infeasible even in principle. Providing the parameter values and a compact statistical model description (or a public likelihood artifact) would significantly strengthen reproducibility.

Expert comment. It is true that these information are not documented in the paper, however these are kept internally at CERN and the authors are not expected to share any data or simulation when publishing the paper for the full reproduction of the results. This is done in exceptional cases where relevant work has been done to provide data for example to the theory community. This comment for this particular paper is irrelevant.

The AI reviewer's request is reasonable by generic open-science standards: tabulated fit parameters, a machine-readable likelihood, and enough metadata for an external reanalysis. None of that is factually incorrect, and in some subfields it would be a legitimate revision request. But in LHCb physics at Nature-level submission, the relevant calibration and fit-configuration artifacts are maintained and used internally by the collaboration rather than published alongside the paper, and a specialist reviewer knows this by default. The critique is therefore not wrong in

content so much as miscalibrated in severity: what the AI reads as a reproducibility gap is the community’s normal publication boundary.

971
972

Example #2 for W1 — Bacterial community modeling paper, reviewer: GPT-5.2

Claim: “The ‘universal collapse’ claim is not established beyond the studied system”

- **Main point of criticism:** The paper makes an overgeneralized claim that invasion probability universally collapses onto survival fraction across conditions, but the main theoretical support relies on an exchangeable generalized Lotka–Volterra setup that can mechanically enforce such a collapse, and the experiments do not establish universality beyond the studied library and protocol.
- **Evaluation criteria:** Validity; Conclusions

Evidence:

Quote 1 (from main text): “Despite these conflicting diversity-invasibility relationships, after scaling richness with species pool size to get the survival fraction, we found that all communities collapsed to a universal line in which the invasion probability is approximately equal to the survival fraction (Fig. 4b).”

| **Comment:** The sentence asserts a universal relationship, but the paper’s own phrasing indicates that the collapse is obtained after a specific rescaling (richness divided by pool size), not from an independently validated mechanism. A collapse can occur for multiple reasons, including shared definitions and exchangeability assumptions, without implying universality across ecological contexts. The manuscript does not provide out-of-distribution validation across different taxa, interaction types, or invasion protocols. Because later conclusions extrapolate this relationship to natural communities, the evidentiary standard should be higher than a within-system collapse. The paper would need clearer limits on scope, or additional tests demonstrating that the relationship persists under qualitatively different invasion settings.

Expert comment. The main point of the critique is correct. But to the bacterial modeling community this paper does not read that it claims a universal law. The AI reviewer is correct with the request for clearer limits on scope and additional tests demonstrating the relationship persists under qualitatively different invasion settings. It is harsh, but correct.

This example is a cleaner dissociation of content from framing than the main-text example. The expert explicitly endorses the AI’s methodological point (it is harsh, but correct) and agrees that the scope should be tightened. What the expert disputes is the AI’s reading of the paper as a universality claim in the first place: to someone familiar with how the bacterial modeling literature uses the word “universal,” the paper is making a bounded empirical statement rather than an unconditional physical law. The AI reviewer, operating on the text alone, does not have the community-level prior that calibrates this reading. The resulting review item is a correct technical request bundled with an overly strong interpretation of what the authors had actually claimed.

973
974
975
976
977
978
979

Example #3 for W1 — Agricultural ecology paper, reviewer: Claude Opus 4.5

Claim: “Weakly predictive models are used to support strong global projections”

- **Main point of criticism:** The models explaining biodiversity responses to yield have marginal pseudo- R^2 values ranging from 0.05 to 0.36, and the authors acknowledge that their “models do not fully fit model assumptions,” yet these weakly predictive models are used to project biodiversity changes globally, generating estimates of species losses up to 89% that should be interpreted with substantial uncertainty.
- **Evaluation criteria:** Validity; Appropriate use of statistics

Evidence:

Quote 1 (from main text): “The variance explained by fixed effects (marginal pseudo- R^2) for all models varied between 0.05 and 0.36, whereas the variance explained by fixed and random effects together (conditional pseudo- R^2) varied between 0.52 and 0.98.”

| **Comment:** A marginal R^2 of 0.05 means that the fixed effects (including yield and its interactions) explain only 5% of the variance in biodiversity outcomes, with the remainder explained by random effects such as study-level variation and spatial blocks. While low marginal R^2 is common in PREDICTS-based analyses, using such models to project global biodiversity changes generates predictions where most of the variance is driven by factors not captured in the projections, since random effects cannot be extrapolated beyond the sampled studies. The projection of 89.1% maximum species loss relies on these weakly predictive relationships.

33

Expert comment. I think the AI does not understand how hard ecology is! R^2 of 0.36 is impressive!

The AI reviewer applies thresholds for “weakly predictive” that are common in a controlled experimental setting but are very far from the noise floor of observational ecology on cross-study datasets. The expert’s single-sentence reaction captures this neatly: by agricultural-ecology standards, a marginal R^2 of 0.36 on a variance-dominated response is an unusually strong fixed-effect signal, not an alarming one. The AI’s numeric description of the situation is accurate (the R^2 values are what they are, and the authors do acknowledge assumption violations); what goes wrong is the implicit reference class the AI is comparing against, which is not the reference class the field actually uses. The same numerical result therefore supports opposite severity assessments depending on the reviewer’s field prior.

D.1.2 W2: Over-harsh, out-of-scope, or unrealistic demands

AI Reviewer Weakness #2: Over-harsh, out-of-scope, or unrealistic demands ($n = 46$)

The AI reviewer raises a criticism that is technically valid in isolation but falls outside the scope of the paper or asks for additional experiments that the authors cannot realistically provide. The claim itself is usually not wrong; what is missing is the reviewer’s calibration of severity to what the paper is actually claiming and what is feasible to change in a revision.

Example #1 for W2 — Bacterial community modeling paper, reviewer: GPT-5.2

Claim: “Continued dispersal of the invader contaminates the invasion-success measurement”

- **Main point of criticism:** The paper’s core experimental outcome measures treat invasion success and priority effects as emergent community properties, but the protocol includes continued dispersal and immigration of the invader and residents after invasion, so “invasion success” is not cleanly separated from sustained propagule pressure and may not reflect intrinsic community resistance or true priority effects.
- **Evaluation criteria:** Validity; Conclusions

Evidence:

Quote 1 (from main text): “After six days of culturing, each community was exposed to an invader species (Fig. 1a) and we continued to culture the communities for another 6 days with dispersal of all species on each dilution cycle (Fig. 1a, b).”

| **Comment:** The invader is not a one-time introduction; instead, it experiences ongoing immigration throughout the invasion assay, which is a central driver of establishment in invasion ecology. Under such conditions, the measured “invasion probability” can reflect the imposed propagule pressure regime as much as community resistance mechanisms. The paper interprets differences between regimes as emergent outcomes of resident interactions, rather than as outcomes shaped by continued introductions. The manuscript does not present a parallel experiment with a single invader pulse to show that the main conclusions are robust to this distinction.

Expert comment. This is too harsh a critique. The authors do not claim that their result is universal, although the AI would have had a good point if they did. This critique is somewhat along the lines of Human Reviewer 2.

The AI reviewer is correctly describing an experimental design choice (continued dispersal of the invader) and correctly identifying that it complicates the interpretation of “invasion success” as an intrinsic community property. The critique is not factually wrong. What the expert objects to is the severity of the request: the AI asks for a parallel experiment with a single-pulse invader, which would roughly double the wet-lab workload of the study, on the basis of a conclusion (universal intrinsic invasion resistance) that the paper in fact does not claim. The expert notes that a human reviewer raised a similar concern but calibrated to what the paper actually argues rather than to the strongest possible reading of it.

Example #2 for W2 — Neuroimaging of psychiatric disorders paper, reviewer: Claude Opus 4.5

Claim: “Healthy-derived predictors cannot be validly applied to patient populations”

- **Main point of criticism:** All molecular predictors (gene expression, receptor densities, metabolism, synapse density, myelination) and connectivity predictors are derived exclusively from healthy adult participants, yet these are used to predict and explain cortical abnormalities in patient populations where these features may be substantially altered by disease processes.
- **Evaluation criteria:** Validity; Appropriate use of data

Evidence:

Quote 1 (from main text): “Gene expression data was collected by the Allen Human Brain Atlas as described in Hawrylycz et al.”

| **Comment:** The Allen Human Brain Atlas is derived from postmortem brain tissue of donors without known neurological or psychiatric disorders. However, the disorders studied in this paper (schizophrenia, depression, epilepsy, Parkinson’s disease, and so on) are known to involve altered gene expression patterns in the brain. Using healthy-derived gene expression to predict patient abnormalities assumes that the spatial pattern of gene expression is preserved in disease, which may not be valid for many of the studied conditions.

997

Expert comment. This critique is factually accurate but it highlights a limitation that the authors cannot realistically defend against or resolve by modifying their data, making it a non-constructive critique.

The expert explicitly agrees with the AI on content: the Allen Human Brain Atlas is indeed healthy-donor data, and the studied disorders do indeed alter the relevant spatial patterns. What makes the critique fall under W2 is not factual error but that there is no actionable path forward: a patient-derived atlas of gene expression across the relevant disorders simply does not exist, and producing one is orders of magnitude beyond the scope of any single neuroimaging paper. A human reviewer typically recognises this asymmetry and reframes the observation as a limitation to be acknowledged in the discussion rather than as a correctness concern. The AI does not perform this reframing and instead flags it as a validity issue, which escalates the severity beyond what the paper can plausibly respond to.

998
999
1000
1001
1002
1003
1004
1005

Example #3 for W2 — Near-field optical spectroscopy paper, reviewer: GPT-5.2

Claim: “Fitting procedure for the central quantitative outputs is not fully reproducible”

- **Main point of criticism:** The paper does not provide sufficient information and runnable artifacts to reproduce its central quantitative outputs (complex dielectric function, radiative rate/lifetime, damping rate) derived from fitting s-SNOM spectra with the point dipole model.
- **Evaluation criteria:** Data and methodology; Reproducibility

Evidence:

Quote 1 (from main text): “As shown in Figs. 2c,d, the point dipole model with Eq. 1 as the input dielectric function provides an adequate fit to the data (see SOM for details on fitting).”

| **Comment:** The paper’s core quantitative claims rely on a fitting procedure whose details are deferred to the SOM, but the fitting workflow (objective function, constraints, priors, initialization, treatment of background, and handling of parameter correlations) is not described in the main text. Without those details, an independent reader cannot reproduce the reported parameter values or assess whether the fits are unique and robust to alternative model choices. The extracted parameters are sensitive to normalization, tip-sample coupling, and baseline assumptions in s-SNOM analyses, so providing a full fitting recipe with sensitivity analysis, or shared scripts that deterministically regenerate the fit results, is standard for quantitative near-field spectroscopy.

Expert comment. This is such a harsh comment for an experimental study. The authors provided all the datasets. The measurements were taken several times and averages were reported. The real world measurements are not perfect or noise free. What else could authors do?

The expert’s reaction here is the clearest version of the W2 pattern: the reviewer has essentially made a conceptually correct reproducibility point (fitting details matter, sensitivity analyses are good practice), but the request implies a standard of experimental reproducibility that is not achievable for the measurement in question. s-SNOM data are intrinsically noisy, the reported values already come from averaging multiple runs, and the “fitting

1006
1007
1008
1009

recipe with sensitivity analysis and deterministic scripts” the AI asks for is closer to what a computational paper provides than what an experimental physics group can realistically deliver. The expert’s rhetorical “what else could authors do?” captures the asymmetry: the critique gives the authors no room to respond because the thing being asked for is not a thing that exists in their workflow.

D.1.3 W3: Paper explicitly states X, AI says missing

The three examples below illustrate the reading-failure pattern in three distinct variants: (i) Example #1, a structural variant where the AI fails to recognize information present in an unlabeled section; (ii) Example #2 (also shown in main text), a direct-omission variant where the AI claims absent a procedure that is explicitly described; and (iii) Example #3, a partial-coverage variant where the AI underweights cross-section evidence.

AI Reviewer Weakness #3: Paper explicitly states X, AI says missing (*n* = 37)

The AI reviewer asserts that a piece of information is absent from the manuscript when in fact the paper provides it, whether in a different section, the supplementary material, or elsewhere in the same passage. The critique itself would often be valid if the information were truly missing, but the factual premise is wrong. This happens because of limited long context management: Through out writing a review, an AI reviewer must read multiple files (context from the main text, supplementary materials, code, information retrieved from the web) and when it exceeds its context length limit, it summarizes the context, resulting in losing information it should hold to process information and write a coherent review.

Example #1 for W3 — 11.7T MRI methodology paper, reviewer: Gemini 3.0 Pro

Claim: “The imaging sequences and reconstruction methods are not described”

- **Main point of criticism:** The manuscript fails to provide a detailed description of the imaging sequences and reconstruction methods used to acquire the “first in vivo images,” violating reproducibility standards.
- **Evaluation criteria:** Data and methodology; Reproducibility

Evidence:

Quote 1 (from main text): “We tackled this problem by deploying parallel transmission hardware and software, first developed by the laboratory at 7T and further optimized for 11.7T (Methods).”

| **Comment:** The “Methods” section referenced here is missing from the manuscript. The only methods sections present are “Genotoxic tests,” “Behavioral tests,” and “Balance tests,” leaving the core imaging methodology undefined.

Expert comment. The AI made factual errors. The sequence parameters were reported, and the methods section starts with this. It is true that the MRI methodology does not have a subheading, which is why the AI got confused.

The AI reviewer looked for a section explicitly labelled “Methods” or with an imaging-related subheading, did not find one, and concluded that the imaging methodology was absent. The expert’s correction is precise: the imaging methodology is in the manuscript, and the sequence parameters are reported, but the section that contains them does not carry a recognisable heading, which is what misled the AI. This is the characteristic failure mode of W3 at the structural level: the AI indexes a paper by headings and fails to recover information that is present in the running text. A human reviewer scans for content rather than labels and would not have missed this.

Example #2 for W3 — Future air-quality projections paper, reviewer: Claude Opus 4.5

Claim: “Known CMIP6 PM2.5 bias is not corrected for”

- **Main point of criticism:** The paper does not adequately discuss or correct for the known systematic underestimation of PM_{2.5} concentrations by CMIP6 models, particularly in highly polluted regions like China, which could bias the DAPP projections.
- **Evaluation criteria:** Data and methodology; Validity

Evidence:

Quote 1 (from main text): “We calculated future PM_{2.5} concentration in China based on empirical formulas and future pollutant concentration data.”

- | **Comment:** The paper mentions using CMIP6 pollutant concentration data but does not discuss bias correction. Multiple studies have documented that CMIP6 models systematically underestimate surface PM_{2.5} concentrations, particularly in regions with high pollution levels like East Asia, which could lead to underestimation of future DAPP if not addressed.

1026

Expert comment. This is factually incorrect. The paper explicitly addresses this issue. Lines 489–496 describe a calibration procedure (Equation 5) that adjusts modeled PM_{2.5} concentrations against observed baseline period data (2012–2017).

The AI reviewer is not missing a subheading; it is missing an entire calibration procedure (Equation 5, with its own methodological description spanning seven lines) that is already in the manuscript and addresses exactly the bias the AI is raising. The expert’s reply cites the line numbers because the content is unambiguously present. What makes this critique fall under W3 rather than a more fundamental validity concern is that the underlying methodological point is valid in general (CMIP6 does underestimate surface PM_{2.5} in East Asia), but the accusation that the paper ignores it is factually wrong. A reviewer who reads the full Methods section will see the bias correction; the AI apparently did not.

1027
1028
1029
1030
1031
1032
1033

Example #3 for W3 — Structural variant discovery paper, reviewer: GPT-5.2

Claim: “Validation is not independent of training data”

- **Main point of criticism:** The paper’s conclusion that SurVIndel2 yields a substantially more complete 1000 Genomes CNV catalogue with “high validation rate” is insufficiently supported because the validation is reported only “in samples with long reads” that overlap the HGSVC2 cohort used to train and benchmark SurVIndel2’s filtering model, making the validation non-independent and potentially optimistic.
- **Evaluation criteria:** Validity; Appropriate use of statistics

Evidence:

Quote 1 (from main text): “The machine learning model of SurVIndel2 was trained using the 34 samples in HGSVC2 sampled by the NYGC. For this reason, when testing them, we use a leave-one-out strategy.”

- | **Comment:** Leave-one-out within the same 34-sample cohort reduces direct per-sample label leakage, but it does not provide an independent external validation of the learned filtering behavior. The training and evaluation data share the same sequencing center, platform, library preparation, and mapping pipeline characteristics, which can create homogeneity and risk overfitting. When the same cohort is used to compute “validation rates” for catalogue-level claims, the evaluation can become circular, and reported validation is measured against resources tied to the training distribution. A stronger design would validate the 1000 Genomes discoveries on long-read samples not used in model training or tuning, ideally sequenced and processed independently.

Expert comment. Another valid point that is somewhat offset by the benchmarking on other organisms.

This example shows a subtler variant of W3: the information the AI treats as absent is not entirely absent but only partial. The AI notes, correctly, that validation within the HGSVC2 cohort shares sequencing center and pipeline with the training data and therefore is not a fully independent test. The expert acknowledges the point but observes that the paper does benchmark SurVIndel2 on other organisms (non-human genomes) elsewhere, which the AI did not mention and which provides exactly the kind of out-of-distribution evaluation the AI says is missing. The critique is therefore half-right: the human-genome validation is indeed not independent, but the paper’s overall validation strategy is stronger than the AI represents because it extends beyond human data. This kind of partial coverage is common in W3 cases where the AI reads one section closely but does not integrate it with evidence

1034
1035
1036
1037
1038
1039
1040
1041

elsewhere in the manuscript.

D.1.4 W4: Redundancy across the three AI reviewers

AI Reviewer Weakness #4: Redundancy across the three AI reviewers (n = 28)

The three AI reviewers assigned to the same paper tend to converge on substantially overlapping criticisms, so a second or third AI reviewer often adds little information that the first has not already raised. This weakness is inherent to the AI panel as a whole rather than to any single review item; it becomes visible only when the reviews for the same paper are read side by side. For this reason, each example in this category shows all three AI reviewers’ overlapping review items together. Please note that we look into this phenomenon in more detail at § 3.

Example #1-a for W4 — Machine-learning interatomic potentials paper, reviewer: GPT-5.2

Claim: “The 1000× data-efficiency claim may be inflated by train/test correlation”

- **Main point of criticism:** The headline conclusions about extreme data efficiency and superiority over prior ML interatomic potentials are not robust because key comparisons (notably the “1000× fewer data” DeepMD comparison) are evaluated on highly correlated MD trajectory frames with train/test splits that can leak near-duplicate configurations.
- **Evaluation criteria:** Validity; Appropriate use of statistics

Evidence:

Quote 1 (from abstract): “NequIP outperforms existing models with up to three orders of magnitude fewer training data, challenging the widely held belief that deep neural networks require massive training sets.”

| *Comment:* This is the central, strongest quantitative claim in the abstract, and it is comparative (“outperforms existing models”), so it requires an evaluation protocol that rules out confounders. In MD-generated datasets, adjacent frames can be extremely similar, so randomly sampled training frames can have near-neighbors in the test set even when they are formally disjoint. If such leakage occurs, test errors can substantially underestimate true generalization error. The paper does not establish that its orders-of-magnitude comparisons are robust to these well-known correlation effects.

Example #1-b for W4 — Machine-learning interatomic potentials paper, reviewer: Claude Opus 4.5

Claim: “The NequIP vs DeepMD comparison is not a valid data-efficiency test”

- **Main point of criticism:** The comparison between NequIP (133 training structures) and DeepMD (133,500 training structures) in Table III does not constitute a valid data-efficiency comparison because it compares different architectures at vastly different operating points rather than showing learning curves for both methods.
- **Evaluation criteria:** Validity; Appropriate use of statistics

Evidence:

Quote 1 (from main text): “A DeepMD NN-IP model was previously trained for water and ice using a joint training set containing 133,500 reference calculations of these four systems. To assess data efficiency of the NequIP architecture, we similarly train a model jointly on all four parts of the data set, but using only 133 structures for training, i.e. 1000x fewer data.”

| *Comment:* The comparison shows that NequIP achieves better force RMSE with 133 structures than DeepMD does with 133,500 structures, but this does not demonstrate data efficiency in the proper sense. Data efficiency would require showing that NequIP consistently achieves lower errors than DeepMD at the same training set size, or equivalently, that NequIP reaches a target accuracy with fewer data points. The comparison as presented could simply mean that NequIP is a fundamentally better architecture regardless of data size, rather than being more data-efficient.

Example #1-c for W4 — Machine-learning interatomic potentials paper, reviewer: Gemini 3.0 Pro

Claim: “The NequIP vs DeepMD comparison lacks a direct control experiment”

- **Main point of criticism:** The claim that NequIP requires “orders of magnitude fewer training data” is not rigorously supported because the comparison with DeepMD is unfair (comparing NequIP trained on 133 samples against DeepMD trained on 133,500 samples) and lacks a direct control experiment in which DeepMD is trained on the same 133 samples as NequIP.
- **Evaluation criteria:** Validity; Data and methodology

Evidence:

Quote 1 (from main text): “Table III shows the comparison of the predictive force accuracy of NequIP trained on the 133 structures vs DeepMD trained on 133,500 structures. We find that with 1000x fewer training data, NequIP significantly outperforms DeepMD on all four parts of the data set.”

| **Comment:** Table III compares NequIP (133 samples) with DeepMD (133,500 samples). This comparison is flawed because it does not show how DeepMD performs on 133 samples. It is possible that DeepMD would perform poorly, but without that data point, the “orders of magnitude” claim is an inference based on the assumption that DeepMD needs 133,500 samples to reach that performance, which is not proven. A fair comparison would involve training DeepMD (and other baselines) on the same 133 samples to demonstrate the specific advantage of NequIP in the low-data regime.

1045

Expert comment. To a large extent, all three AI reviewers raise very similar points. Thus, the overall review would not cover many substantial points a human reviewer raised.

All three reviewers converge on the same observation about the NequIP vs. DeepMD comparison (different sample counts, no shared operating point), differing only in phrasing and emphasis. GPT-5.2 frames it as train/test correlation, Claude Opus as missing learning curves, and Gemini as a missing control experiment, but the underlying concern is identical. The expert annotator’s comment captures the cost: panel-level redundancy means a multi-AI review covers fewer distinct concerns than a multi-human review of the same paper.

1046

1047

1048

1049

1050

Example #2-a for W4 — Compound temperature-COVID mortality paper, reviewer: GPT-5.2

Claim: “The ‘compound mortality’ calculation violates its own independence assumption”

- **Main point of criticism:** The paper’s central “compound mortality impact” calculation adds temperature-attributable deaths and COVID-19 deaths under an explicit independence and additivity assumption, even though the manuscript itself presents evidence and prior literature that the two hazards can interact and share causal pathways; this affects the core results on heatwaves and cold snaps and the interpretation of “compound” burdens.
- **Evaluation criteria:** Validity; Conclusions

Evidence:

Quote 1 (from main text): “These temperature-related deaths are theoretically independent of COVID-19 deaths because they are calculated from DLNMs that describe the relationships between daily mean temperature and daily all-cause mortality, after COVID-19 mortality has been removed (see Methods).”

| **Comment:** Removing deaths with COVID-19 on the death certificate changes the outcome series, but it does not establish causal independence between temperature and pandemic conditions. COVID-19 can change population vulnerability, healthcare access, and reporting or diagnostic patterns, all of which can modify the temperature–mortality relationship even if COVID-coded deaths are excluded. The claim of theoretical independence is therefore stronger than what the described preprocessing supports, and because the paper’s main “compound” totals depend on this premise, any violation can materially bias compound magnitudes.

Example #2-b for W4 — Compound temperature-COVID mortality paper, reviewer: Gemini 3.0 Pro

1051

Claim: “Removing COVID deaths does not make the remaining mortality pandemic-independent”

- **Main point of criticism:** The paper incorrectly assumes that removing COVID-19 deaths from the all-cause mortality series makes the remaining temperature-related mortality “theoretically independent” of the pandemic, ignoring the significant indirect effects of the pandemic (healthcare strain, behavioral changes) that likely altered the temperature–mortality relationship for non-COVID causes.
- **Evaluation criteria:** Validity; Appropriate use of statistics

Evidence:

Quote 1 (from main text): “These temperature-related deaths are theoretically independent of COVID-19 deaths because they are calculated from DLNMs that describe the relationships between daily mean temperature and daily all-cause mortality, after COVID-19 mortality has been removed (see Methods).”

| **Comment:** This claim of independence is methodologically unsound. The non-COVID mortality during the pandemic was not independent of the pandemic context; it was influenced by factors such as overwhelmed healthcare systems, delayed treatments for chronic conditions, and lockdowns. Recent literature has shown that the risk of heat-related mortality was significantly elevated during the pandemic compared to pre-pandemic periods, even for non-COVID causes, due to these synergistic effects. By using a model trained primarily on pre-pandemic data to estimate mortality in 2020–2022, the study fails to capture this altered vulnerability.

Example #2-c for W4 — Compound temperature-COVID mortality paper, reviewer: Claude Opus 4.5

Claim: “Vulnerable populations overlap, so additive compound estimates can be biased”

- **Main point of criticism:** The paper assumes that temperature-related deaths and COVID-19 deaths are independent and simply adds them together to estimate compound mortality impacts. However, vulnerable populations for heat and cold mortality and COVID-19 mortality substantially overlap (elderly, those with chronic conditions), which violates the independence assumption and may lead to either double-counting or underestimation of compound effects.
- **Evaluation criteria:** Validity; Conclusions

Evidence:

Quote 1 (from main text): “By adding temperature-related deaths (with COVID-19 deaths removed prior to calculation) and COVID-19 deaths together to estimate the compound mortality impact, we have assumed that they are independent of each other.”

| **Comment:** The paper explicitly acknowledges this assumption but does not adequately address its implications for the validity of the compound mortality estimates. Simply removing COVID-19 deaths from the mortality data before fitting the temperature–mortality model does not account for potential effect modification, where COVID-19 may have amplified temperature–mortality relationships through shared vulnerabilities, behavioral changes, or healthcare system strain. Recent studies using interaction models have found that heat-related mortality was significantly elevated during COVID-19 periods compared to pre-pandemic periods, suggesting the relationship is not simply additive but involves effect modification.

Expert comment. AI Reviewer 1 (Items 2 and 3) identified both the violated independence assumption and the misleading “compound” terminology. AI Reviewer 2 (Item 1) restated it with additional references. This reviewer articulates it more formally, mentioning joint models, attributable fraction partitioning, and counterfactual decomposition, but the core insight is identical.

This second example sharpens what the side-by-side view can reveal. GPT-5.2 and Gemini both quote the exact same sentence from the Methods (“theoretically independent of COVID-19 deaths”), while Claude Opus quotes the paper’s own acknowledgment of the additive assumption two paragraphs later; all three then argue that removing COVID-coded deaths does not make the remaining mortality pandemic-independent. The divergence is in emphasis rather than substance: Gemini adds references to heat–COVID interaction literature, Claude adds quantitative estimates (RR 1.07 vs 1.55), GPT frames the issue in terms of formal statistical decomposition. Each addition is useful, but they are elaborations of a single observation that any one reviewer could have made on their own.

D.1.5 W5: Vague, verbose, or without actionable recommendation

AI Reviewer Weakness #5: Vague, verbose, or without actionable recommendation (n = 24)

The AI reviewer identifies a concern at length but does not translate it into a concrete change the authors could make, either because the review is too long and diffuse to pinpoint the main issue, or because the critique ends without any actionable revision request. Both patterns leave the authors and the editor unsure what would satisfy the reviewer.

Example #1 for W5 — 11.7T MRI methodology paper, reviewer: GPT-5.2

Claim: “Safety conclusions at 11.7T are drawn from an underpowered protocol”

- **Main point of criticism:** The paper draws overly strong conclusions about human safety and tolerability at 11.7T from a small, short-term, endpoint-limited protocol, which weakens the validity of the safety claims made in the abstract and main text.
- **Evaluation criteria:** Validity; Appropriate use of statistics

Evidence:

Quote 1 (from main text): “To assure the safety of human imaging at such high field strength, we performed physiological, vestibular, behavioral and genotoxicity measurements on the volunteers. The data shows no evidence of adverse effects.”

| **Comment:** “No evidence of adverse effects” is a non-equivalence statement that depends strongly on sample size, endpoint sensitivity, and follow-up duration. The described measures target a subset of plausible acute and subacute effects, but do not address detection of rare events or delayed outcomes, which is central when translating a first-in-human ultra-high-field exposure into a generalised safety assertion. The paper does not report an a priori safety endpoint hierarchy, minimally important differences, or any power or sensitivity analysis for the bioeffects outcomes, so readers cannot assess what magnitude of harm the study could have detected. As written, the conclusion materially overreaches what the reported study design can support.

Expert comment. Too verbose. I am not a statistics expert, but this is a paper about MR methodology. The first comment should not be six paragraphs about biological effects.

The AI reviewer’s underlying statistical point is defensible: a small, short-term safety study cannot support an unqualified “safe” claim, and the paper should scope its safety language accordingly. But the expert’s complaint is not with the content, it is with the volume and placement. The review opens with six paragraphs on biological effects in a paper whose primary contribution is magnetic-resonance engineering, and the statistical critique is spread across those six paragraphs rather than stated once crisply. A shorter review that said “soften the safety language and report a sensitivity analysis” would have delivered the same point with a tenth of the text and would have made the actionable request unambiguous. W5 is most commonly this kind of length-to-signal imbalance: the AI reviewer knows where the weakness is but buries it in elaboration.

1061
1062
1063
1064
1065
1066
1067
1068

Example #2 for W5 — Agricultural ecology paper, reviewer: Gemini 3.0 Pro

Claim: “Intensification projection is inconsistent with the landscape-scale framing”

- **Main point of criticism:** The projection of biodiversity impacts for intensification is methodologically inconsistent with the expansion scenario and contradicts the authors’ claim that their models capture landscape-scale impacts. The intensification projection weights the biodiversity change by the crop’s relative area fraction, treating the impact as local to the crop’s footprint, whereas the models (based on generic “Cropland” data and landscape-scale yields) and the expansion scenario imply landscape-wide effects. This likely underestimates the biodiversity cost of intensification.
- **Evaluation criteria:** Validity; Conclusions

Evidence:

Quote 1 (from main text): “Therefore, our models capture landscape-scale but not local-scale impacts of crop-specific management.”

| **Comment:** The authors explicitly state that their models capture landscape-scale impacts, implying that the yield variable (measured at approximately 10 km resolution) influences biodiversity across the landscape, or at least across all cropland in the landscape, and not just within the specific crop’s fields.

Expert comment. The review does not provide any suggestions for how to address this criticism.

This example captures the other end of the W5 spectrum: rather than being buried in verbose elaboration, the critique is structurally incomplete. The AI correctly identifies a genuine methodological inconsistency (landscape-scale models applied with local-scale area weighting), and the claim is precise enough that a human reviewer could work with it. What is missing is any guidance on what the authors should do about it. Should the intensification projection be rerun without area weighting? Should the paper reframe its impact claims to local scale only? Should a sensitivity analysis bracket the two interpretations? The AI flags the problem and stops, leaving the authors with no path to revision. Human reviewers typically close this loop even when the suggestion is a single sentence.

D.1.6 W8: Citing evidence that appeared after the preprint

AI Reviewer Weakness #8: Citing evidence that appeared after the preprint ($n = 9$)

The AI reviewer's criticism depends on work that was published after the preprint under review, which means the critique uses information that was not available to the authors at the time of submission. This is an artifact of our evaluation setup, where AI reviewers see already-published papers rather than genuinely concurrent preprints; a deployed AI reviewer would not have access to post-submission evidence.

Example #1 for W8 — Neural-network wavefunctions paper, reviewer: Claude Opus 4.5

Claim: "The claimed HEG superiority is contradicted by subsequent analysis"

- **Main point of criticism:** The paper claims that the neural-network results for the homogeneous electron gas (HEG) "outperform many traditional ab initio methods" and achieve "error of less than 1%" using BF-DMC as reference. However, subsequent independent analysis in the literature demonstrates that the DeepSolid method (referred to as "LiNet") does not actually improve upon backflow DMC results across the full density regime for $N = 54$ electrons, undermining the claimed superiority of the method for this benchmark system.
- **Evaluation criteria:** Validity; Conclusions

Evidence:

Quote 1 (from main text): "Overall, our neural network performs very well, with an error of less than 1% in a wide range of density."

| **Comment:** This claim presents the correlation error relative to BF-DMC, but the framing suggests the neural network method is achieving higher accuracy than traditional methods. A subsequent study by Pescia et al. (2024) in Physical Review B directly compared DeepSolid against BF-DMC for $N = 54$ electrons and found that it does not improve upon BF-DMC energies over the whole density regime, which contradicts the paper's implication of superior performance.

Expert comment. While the claims are actually fine, I marked it as not correct because the paper they use to support these claims appeared two years after the presented one. The field was evolving very fast, and in 2022 the presented results could be considered as good. The only thing one could move on with is the claim of improving upon the state of the art; the authors should clarify what exactly they mean by it.

The AI reviewer's critique relies on a Pescia et al. (2024) analysis that did not exist at the time of the 2022 submission. At submission, the authors' claim was defensible against the then-current state of the art, and a reviewer from that period could not have raised this objection. The expert's response separates two things: the narrow issue of whether the "improves upon state of the art" wording should be tightened (a small editorial fix that is legitimate), and the larger issue of whether the method's performance is genuinely inferior (which only became clear in later work). W8 cases like this are methodologically informative despite being artifacts of our setup: they show that the AI reviewer tends to integrate all accessible literature without reasoning about the reviewing timeline, which a deployed system could mitigate by constraining its search to content dated before submission.

Example #2 for W8 — Polymer property prediction paper, reviewer: Claude Opus 4.5

Claim: “Frozen polyBERT fingerprints deviate from standard fine-tuning practice”

- **Main point of criticism:** The paper uses frozen polyBERT fingerprints (without fine-tuning the transformer weights) for property prediction, which deviates from established practices in chemical language modeling where fine-tuning typically improves downstream task performance.
- **Evaluation criteria:** Data and methodology; Validity

Evidence:

Quote 1 (from main text): “polyBERT fingerprints (dashed arrow) are the averages over the token dimension (sentence average) of the last Transformer encoder.”

| **Comment:** The polyBERT fingerprints are extracted as fixed representations from the pre-trained model, with no indication that the transformer weights are updated during property prediction. This means the 600-dimensional fingerprint vectors are computed once and remain static, while only the downstream multitask neural network is trained. This approach may not fully leverage the power of transfer learning, as the pre-trained representations cannot adapt to capture property-specific patterns that may be critical for accurate prediction.

1086

Expert comment. This needs to be seen in historical context. PolyBERT was very early.

This example shows the softer version of the W8 pattern. The AI reviewer applies what is now “established practice” in chemical language modeling (fine-tune rather than freeze the transformer backbone) to a paper that was one of the earliest transformer-based polymer representation efforts. At the time of submission, pre-trained chemical transformers were themselves a recent development, and using their representations as static fingerprints was a reasonable initial design choice; the fine-tuning literature the AI implicitly invokes matured afterwards. The expert’s single-sentence correction is all that is needed: the critique is not wrong by current standards but is applied out of historical context. As with Example #1, the underlying failure mode is the AI reviewer treating its entire training corpus as simultaneously available, without reasoning about when the manuscript was written.

1087

1088

1089

1090

1091

1092

1093

1094

D.1.7 S1: Statistical and methodological rigor

1095

AI Reviewer Strength #1: Statistical and methodological rigor ($n = 45$)

The AI reviewer identifies a genuine statistical or methodological issue that neither the human reviewers nor, in several cases, the authors themselves had addressed. These include violations of independence assumptions, inappropriate significance tests, inadequate validation splits, and missing uncertainty quantification. The expert annotators treat these as legitimate additional scrutiny that a thorough review would have surfaced.

Example #1 for S1 — Protein-protein interaction prediction paper, reviewer: GPT-5.2

1096

Claim: “Reported SHS27K gains may not reflect generalization due to evaluation protocol”

- **Main point of criticism:** The paper’s reported SHS27K protein-protein interaction (PPI) gains are not a valid estimate of generalization because the PPI evaluation lacks a validation split, and the released code uses the test split for model selection and early stopping, which directly undermines the paper’s state-of-the-art performance claims.
- **Evaluation criteria:** Appropriate use of statistics; Validity

Evidence:

Quote 1 (from main text): “MUSE outperformed all existing models including single-scale (DrugVQA and TAG-PPI), and multi-view methods (GNN-PPI and HIGH-PPI). Our model showed substantial improvements over the strongest baseline HIGH-PPI, with an increase of 8.41% in the BFS split, 13.04% in the DFS split, and 7.69% in the Random split.”

| **Comment:** This passage is the paper’s central empirical claim for multi-scale interaction prediction on SHS27K, asserting large improvements over the strongest baseline across three splits. Such claims require a strictly held-out test set so that the reported numbers reflect generalization rather than adaptation to the evaluation labels. Because the magnitude of the claimed gains (for example, 13.04% on DFS) is used to justify the overall superiority of the EM framework, any test-set contamination directly affects the paper’s conclusions. In common ML practice, test labels must not influence checkpoint selection, early stopping, threshold tuning, or graph augmentation decisions.

Expert comment. The PPI trainer has no validation split (only train/test) and uses test metrics for model selection (misleadingly named `best_valid_f1`).

The AI reviewer identifies a subtle but consequential methodological flaw: without a validation split, any hyperparameter choice or early-stopping criterion based on test metrics tunes the model to the test set rather than measuring held-out generalization. The expert’s confirmation is especially striking because the flaw is corroborated at the code level: the variable used for model selection is named `best_valid_f1`, suggesting a validation split was intended, but the actual implementation uses test metrics. This is exactly the kind of failure mode a superficial review might miss, since the paper’s text describes standard protocols and the issue only surfaces when the code is read. The AI reviewer is not simply applying ML best practice in the abstract; it is reading the evaluation loop carefully enough to notice the mismatch between described and implemented protocol.

Example #2 for S1 — Compound temperature-COVID mortality paper, reviewer: Claude Opus 4.5

Claim: “K-S tests on a handful of extreme events may be underpowered”

- **Main point of criticism:** The paper uses two-sample Kolmogorov–Smirnov tests to conclude that temperature-related mortality distributions during COVID-19 years differ significantly from non-COVID years. However, the number of extreme temperature events is limited (approximately 10–22 heatwaves and 8–24 cold snaps across the periods), which may affect the statistical power and reliability of these conclusions.
- **Evaluation criteria:** Appropriate use of statistics; Validity

Evidence:

Quote 1 (from main text): “Two-sample Kolmogorov–Smirnov tests confirm that the COVID-19 event distributions are significantly different from the non-COVID-19 distributions at the 5% significance level. These results suggest that COVID-19 may have impacted temperature-related mortality during extreme weather events.”

| **Comment:** The conclusion that COVID-19 “may have impacted” temperature-related mortality is drawn from K-S tests applied to limited samples. From the code and figure descriptions, the comparison involves approximately 12 heatwaves in 2016–2019 versus 10 heatwaves in 2020–2022, and similar numbers for cold snaps. With such small sample sizes, the K-S test may have limited power to detect true differences or may be sensitive to individual outliers, making the statistical significance potentially unstable.

Expert comment. Even if the overall sample size were adequate, treating correlated regional observations within the same event as independent violates the K-S test assumptions. This is a genuinely useful statistical critique that neither the first reviewer in this set nor the human reviewers identified.

This example captures a characteristic S1 pattern: the AI reviewer raises a valid statistical concern (small-sample

K-S tests are underpowered), and the expert not only agrees but extends the critique with an even sharper point that the AI did not make (independence violations from clustered regional observations). Neither the other AI reviewer on this paper nor the human reviewers raised either issue. The expert’s explicit framing is informative: “a genuinely useful statistical critique that neither the first reviewer in this set nor the human reviewers identified.” Statistical rigor at this level of detail is a recurring blind spot in human peer review for domain-specific empirical papers, and the AI panel is reliably filling it.

Example #3 for S1 — Meta-optic imaging paper, reviewer: GPT-5.2

Claim: “In-the-wild imaging claims rely on single captures without uncertainty reporting”

- **Main point of criticism:** The paper’s real-world imaging conclusions are not robust because the “in-the-wild” comparisons are based on a semi-blind deconvolution that jointly estimates the image, PSF, and fixed-pattern noise from a single capture per scene, with primarily qualitative assessment and limited reporting of constraints and hyperparameters, making it difficult to rule out reconstruction artifacts or non-uniqueness.
- **Evaluation criteria:** Conclusions; Appropriate use of statistics and treatment of uncertainties

Evidence:

Quote 1 (from main text): “In each case, we captured a single image using a FLIR A65 sensor, and performed a numerical deconvolution (details in the Methods and supplementary information).”

| **Comment:** A single-capture demonstration provides limited evidence for robustness, especially when the post-processing includes solving an inverse problem that can be sensitive to initialization and regularization. In imaging, reporting multiple trials or at least sensitivity to hyperparameters is a common way to separate repeatable improvements from incidental reconstructions. Because the scenes are uncontrolled, sensor noise, temperature gradients, and alignment can vary substantially, and a one-off result can be cherry-picked unintentionally. The key qualitative comparisons therefore depend on the deconvolution pipeline, not only on the optical PSF itself.

Expert comment. I agree with the AI: “a great deal” does not mean anything from a scientific point of view. They should have quantified their finding.

This example broadens S1 beyond strictly statistical critiques to include quantification and uncertainty reporting more generally. The AI reviewer notices that the paper relies on subjective language (“a great deal,” “clearly better”) for its real-world imaging claims and correctly pushes back that such comparisons should be quantitative when the underlying pipeline includes an inverse problem. The expert’s endorsement is unambiguous: subjective sharpness judgments on single reconstructions are not a scientific standard for demonstrating that a meta-optic outperforms a baseline. Here again, the AI reviewer surfaces a rigor gap that is easy to miss on first read because the paper’s overall presentation looks polished, and only a careful reader distinguishes the qualitative claims from what the data actually support.

D.1.8 S2: Inspecting the submitted source code

The three examples below illustrate the code-reading capability of an agentic AI reviewer across three different settings: (i) Example #1 (also shown in main text), an internal-consistency check between the manuscript’s mathematics and the implemented code; (ii) Example #2, a data-leakage discovery where the manuscript text alone could not reveal the issue; and (iii) Example #3, the most consequential S2 catch in our dataset, where the implemented code contradicts a central reproducibility claim.

AI Reviewer Strength #2: Inspecting the submitted source code ($n = 28$)

The AI reviewer opens the submitted source code, reads it, and uses what it finds to support or refute specific claims in the manuscript. In several cases this surfaces concrete implementation bugs, data leakage, or mismatches between described and implemented methodology that the manuscript text alone does not reveal.

Example #1 for S2 — Neural-network wavefunctions paper, reviewer: GPT-5.2

Claim: “The wavefunction ansatz is inconsistent between the main text and the algorithm”

- **Main point of criticism:** The manuscript’s formal description of the solid wavefunction ansatz is internally inconsistent between the main-text expression and the provided algorithm, making it difficult to verify correctness (spin structure and determinant factorization) and undermining reproducibility of the central methodological contribution.
- **Evaluation criteria:** Data and methodology; Reproducibility

Evidence:

Quote 1 (from main text): “ $\Psi(\mathbf{r}) = \text{Det}_1[e^{i\mathbf{k}\cdot\mathbf{r}_1} u_{mol}^\uparrow(d)] \text{Det}_1[e^{i\mathbf{k}\cdot\mathbf{r}_1} u_{mol}^\downarrow(d)]$.”

| **Comment:** This equation indicates a product of two determinants associated with spin-up and spin-down channels. However, the notation is ambiguous (both are labeled Det_1), and it does not clearly specify whether the determinants are block-diagonalized by spin or full determinants over all electrons. In periodic systems with complex-valued orbitals, these details matter for the correctness of antisymmetry constraints and the implementation of phase factors.

Quote 2 (from submitted source code):

```
def logdet_matmul(xs, w=None):  
    # Combines determinants in log-domain.  
    # xs: FermiNet orbitals in each determinant.  
    # Either of length 1 with shape  
    #     (ndet, nelectron, nelectron) # full_det=True  
    # or length 2 with shapes  
    #     (ndet, nalpha, nalpha), (ndet, nbeta, nbeta)  
    # (full_det=False, determinants are factorised  
    # into block-diagonals for each spin channel).
```

| **Comment:** The released implementation explicitly distinguishes two cases: a single full determinant over all electrons (`full_det=True`) or a product of spin-channel determinants when factorized (`full_det=False`). This illustrates that the determinant factorization and spin handling are subtle and explicitly parameterized in code. When contrasted with the manuscript’s Algorithm 1 line that multiplies two identical determinants, it reinforces that the paper’s written description does not clearly convey which mode was used. Because determinant structure can change expressivity, optimization behavior, and numerical stability, readers cannot reproduce or compare results without a clear specification in the main text.

Expert comment. I find it quite impressive that this reviewer actually looks at the code to understand whether some parts of the paper are not properly discussed. I am personally not doing it on a regular basis as it would consume a substantial amount of time.

This is the clearest expert endorsement of S2 behavior in our dataset: the annotator, a domain specialist, explicitly acknowledges that cross-checking the main-text mathematics against the provided code is something human reviewers typically do not do because it is too time-consuming. The AI reviewer caught a real ambiguity: the main text describes two spin-separated determinants, but the algorithm and code use notation consistent with two identical determinants, leaving readers unable to tell which formulation was actually executed. For a paper whose central contribution is the wavefunction ansatz, this is a genuine reproducibility issue. The expert does not confirm which version is correct; the point is that the reader cannot tell, and the AI reviewer is the only reviewer in the entire panel who checked.

Example #2 for S2 — Protein-protein interaction prediction paper, reviewer: Gemini 3.0 Pro

Claim: “The pseudo-interaction augmentation leaks test edges into training”

- **Main point of criticism:** The paper’s methodology contains a severe data leakage issue where predicted interactions from the test set are added to the molecular network graph, which is then used to predict the existence of those same interactions.
- **Evaluation criteria:** Validity; Data and methodology

Evidence:

Quote 1 (from main text): “To alleviate this incompleteness, we adopted the pseudo-likelihood learning in our variational expectation-maximization framework, augmenting the molecular network graph $\mathcal{N}$ with pseudo interactions predicted by the atomic structure scale model GNN_{ϕ} .”

| **Comment:** The paper describes this augmentation as a feature, but fails to mention that the “pseudo interactions” are drawn from the test set. In a valid inductive or transductive link prediction setting, the existence of the test edges should not be revealed in the input graph structure used for message passing, as it trivializes the prediction task.

Quote 2 (from submitted source code):

```
# In trainer_ppi.py
def gnn_inference(self, gnn_model, protein_dataset):
    # Uses test_edges to generate predictions
    edges = self.test_edges[...]
    pseudo_preds = gnn_model(batch)
    return pseudo_edges

def train(self, ...):
    if gnn_model is not None:
        # Adds predicted test edges back to the graph
        pseudo_edges = self.gnn_inference(
            gnn_model, protein_dataset)
        self.graph.edge_index = torch.cat(
            [self.graph.edge_index,
             pseudo_edges.to(self.device)], dim=1)
        # Evaluates on the same test edges using
        # the augmented graph
        valid_loss, test_metrics = self.evaluate(
            self.graph, self.test_edges, self.test_label)
```

| **Comment:** The code explicitly uses `self.test_edges` to generate pseudo-labels in `gnn_inference`, and the resulting `pseudo_edges` are then concatenated to `self.graph.edge_index` inside `train`. The molecular network model is subsequently evaluated on the same `test_edges` using this augmented graph, which means the network’s input graph structure already contains the edges it is being asked to predict. The leakage that the manuscript’s text does not disclose is unambiguous at the code level.

1138

Expert comment. Code verification fully confirms the data leakage claim. The DPI link trainer (`trainer_dpi.py`) also has a bug using test metrics for model selection despite having a validation set.

This example combines S1 and S2: the AI reviewer raises a methodological concern from reading the paper (data leakage in the graph augmentation), and the expert’s verification at the code level confirms the concern and extends it with an additional bug the AI did not catch (using test metrics for model selection). What makes it an S2 example rather than strictly an S1 one is that the leakage claim cannot be resolved from the manuscript alone; it is only by tracing where the “pseudo interactions” come from in the implementation that the issue becomes visible. Both the AI’s catch and the expert’s extension require reading the code, and this is a paper where the AI panel’s code-reading behavior directly uncovered a flaw that neither the manuscript’s text nor a reader’s prior expectations would have surfaced.

1139
1140
1141
1142
1143
1144
1145
1146

Example #3 for S2 — Wireless health-monitoring patch paper, reviewer: Gemini 3.0 Pro

1147

Claim: “Claimed 800 Hz sampling is contradicted by the implemented code”

- **Main point of criticism:** The paper claims a sampling frequency of 800 Hz and analyzes frequencies up to 400 Hz, but the provided code limits the sampling rate to approximately 2 Hz, making the reported results impossible to reproduce with the described setup.
- **Evaluation criteria:** Validity; Reproducibility

Evidence:

Quote 1 (from main text): “The high-sensitivity accelerometer ADXL-345 with a sampling frequency of up to 800 Hz in the patch allows successful continuous monitoring, with a wide frequency spectrum from 0 to 400 Hz.”

| **Comment:** The authors claim to capture high-frequency signals up to 400 Hz, which by the Nyquist theorem requires a sampling rate of at least 800 Hz. The paper presents this sampling rate as a feature of the wireless patch and uses it to support the frequency-spectrum plots shown in the results.

Quote 2 (from submitted source code):

```
// In code/sketch_may06a.ino (Arduino)
void setup() {
    Serial.begin(9600); // baud rate caps throughput
    // ...
}

void loop() {
    // Read accelerometer and transmit
    delay(500); // forces sampling rate ~2 Hz
}
```

| **Comment:** The Arduino sketch includes a 500 ms delay in the main loop, which forces the sampling rate to approximately 2 Hz: three orders of magnitude below the claimed 800 Hz, and insufficient for capturing anything in the reported 0–400 Hz band. In addition, the serial communication is configured at 9600 baud, which by itself limits the throughput to well under 100 Hz even without the delay. Either setting alone contradicts the 800 Hz claim; together, they show that the wireless pipeline as submitted cannot produce the frequency content the paper reports.

Expert comment. A human reviewer would not open up the raw code to find this out. If this code is correct, basically the core data reported in this paper is wrong. The authors in the code said “It appears that delay is needed in order not to clog the port,” meaning it is not possible to wirelessly transmit high-frequency information. If this is true, Figure 4 would mean the authors measured in a wired manner and falsely claimed that their system is wireless, which is one key aspect of this paper.

This is the most consequential S2 catch in our dataset. The AI reviewer noticed a $400\times$ discrepancy between the paper’s claimed sampling rate (800 Hz) and the implemented rate (approximately 2 Hz), from which it correctly inferred that the reported frequency spectrum could not have been produced by the wireless setup described in the manuscript. The expert’s follow-up reading of the code goes further still: an inline code comment from the authors themselves explicitly states that a delay was added “in order not to clog the port,” which implies that the wireless channel could not carry high-frequency data at all. If this interpretation holds, the paper’s wireless claim, not merely a reproducibility statistic, is incorrect at the level of the headline result. This is the category of finding that human peer review systematically misses: the source of the error is not in what the paper says but in what the implementation does, and only a reviewer willing to read the code can see the mismatch.

D.1.9 S3: Domain-specific technical depth

AI Reviewer Strength #3: Domain-specific technical depth ($n = 27$)

The AI reviewer raises a subfield-specific technical point that is correct, nontrivial, and often pivots on a piece of specialized knowledge the authors could reasonably have expected a domain reviewer to bring. These cases show that the AI reviewer is not limited to generic methodological critique and can engage with the technical content of a paper.

Example #1 for S3 — Multimode fiber image transmission paper, reviewer: GPT-5.2

Claim: “The paper’s scope claim overstates what the method actually recovers”

- **Main point of criticism:** The paper’s main claims overstate the scope of what is demonstrated, because the described experimental setup and learning objective target binary amplitude pattern recovery rather than arbitrary optical fields (for example, complex field transmission) or direct high-resolution video transmission, which weakens the robustness of the paper’s conclusions.
- **Evaluation criteria:** Conclusions; Validity

Evidence:

Quote 1 (from abstract): “In this paper, we present a self-supervised dynamic learning approach that achieves long-term, high-fidelity transmission of arbitrary optical fields through unstabilized MMFs.”

| **Comment:** The abstract asserts transmission of “arbitrary optical fields,” but the experimental setup and learning objective target binary amplitude patterns rather than the full complex-valued optical field. Amplitude-only recovery is a substantially easier problem than complex-field retrieval through a multimode fiber, and the two are not interchangeable for the applications the paper’s framing suggests. The method’s performance on binary amplitude should not be extrapolated to arbitrary optical fields without direct evidence.

Expert comment. This is a good catch. Retrieving the complex field (with the imaginary part) is very ill-posed, and to the best of my knowledge there is no method to do that well with multimode fiber propagation. Arbitrary optical fields is for sure not within the scope of the paper, and the method has nothing to do with the complex field. The authors state the relation as amplitude-to-amplitude, so this part is accurate, but it is not enough to extrapolate this huge leap.

The AI reviewer’s critique pivots on a distinction that a specialist in multimode-fiber imaging would be expected to make: “arbitrary optical field” is a phrase with a specific technical meaning (full complex-valued field, including phase) that is substantively harder to recover than the amplitude-only setting the paper actually demonstrates. The expert not only confirms this distinction but adds the stronger claim that no existing method, without additional information or measurements, handles complex-field recovery through a multimode fiber well. This means the paper’s framing is not just imprecise but reaches beyond what its setup and objective can support. This is S3 at its sharpest: the AI is applying a subfield-specific reading of a scoping word that a generalist reviewer would pass.

1159
1160
1161
1162
1163
1164
1165

Example #2 for S3 — Late-stage pharmaceutical modification paper, reviewer: GPT-5.2

Claim: “Stereocontrol is claimed but diastereomeric ratios are not reported”

- **Main point of criticism:** The paper claims “stereo-controlled” late-stage modification of pharmaceuticals but does not report stereochemical outcomes (diastereomeric ratios, enantiomeric ratios or excess, or absolute-configuration assignment for newly formed stereocenters) for the late-stage modification examples, making the stereocontrol conclusion unsupported.
- **Evaluation criteria:** Validity; Conclusions

Evidence:

Quote 1 (from main text): “The stereo-controlled late-stage modification of some complicated pharmaceuticals indicated the versatility of this protocol.”

| **Comment:** This sentence asserts stereocontrol as an achieved feature and uses it as evidence of method versatility. Stereocontrol is a quantitative claim: it requires reporting which stereoisomer(s) form and in what proportions, and ideally how the stereochemistry was assigned. The paper’s main text does not provide diastereomeric ratios, enantiomeric excess or ratios, or chiral analytical data for the late-stage modification products. In late-stage settings, where complex substrates often contain stereocenters and reactive intermediates can be planar (for example, cationic), stereochemical outcomes are not predictable without measurement.

Expert comment. Since the substrate used for the synthesis of compound 4v contains a stereogenic center, the issue of diastereoselectivity is important. Presumably, a mixture of diastereomers was formed in compound 4v. It was not mentioned by the human reviewers.

The AI reviewer raises exactly the objection a trained organic chemist would raise on first reading: the word “stereo-controlled” is a technical commitment that requires diastereomeric or enantiomeric ratio data to back it up, and that data is absent for the late-stage modification examples. The expert confirms the point with a specific

1166
1167
1168

substrate (compound 4v) and notes that none of the human reviewers flagged this despite its centrality to the paper's versatility claim. This is the pattern that makes S3 a meaningful category distinct from S1: not generic statistical rigor, but the ability to recognise that a particular word in a particular context is a claim that demands particular kinds of evidence which the paper does not provide.

Example #3 for S3 — s-SNOM spectroscopy of TMD monolayers paper, reviewer: Gemini 3.0 Pro

Claim: “The Point Dipole Model is insufficient for quantitative analysis of atomically thin films”

- **Main point of criticism:** The authors use the Point Dipole Model (PDM) to extract the complex dielectric function of transition-metal dichalcogenide monolayers, which is insufficient for quantitative analysis of s-SNOM data on atomically thin materials.
- **Evaluation criteria:** Data and methodology; Validity

Evidence:

Quote 1 (from main text): “We choose the point dipole model to interpret the data as it is well documented to capture the response of atomically thin samples laid on thick substrates.”

| **Comment:** The Point Dipole Model treats the tip as a simple dipole, ignoring the extended geometry of the tip and the complex field distribution in the tip-sample gap. While PDM can qualitatively reproduce material contrast, it is widely recognized as insufficient for quantitative extraction of optical constants, especially for layered systems where the near-field interaction is sensitive to the vertical field distribution. More advanced models such as the Finite Dipole Model or the Lightning Rod Model are required for accurate quantitative analysis.

Expert comment. I agree with the AI reviewer that the PDM with their implementation is not the most accurate; they should have used the layered-medium Green's functions and take care of the inhomogeneous background. However, this doesn't mean that the main findings of these researchers were wrong. A more accurate numerical model would generate results with smaller error limits, that's it.

The AI reviewer identifies a modeling choice that is standard in qualitative s-SNOM work but inadequate for the quantitative analysis the paper attempts, and proposes specific alternative models (Finite Dipole, Lightning Rod) that a specialist would recommend. The expert agrees with the modeling critique but softens the implication: using a more accurate tip-sample model would tighten the error bars on the extracted dielectric function, not invalidate the paper's main findings. S3 examples often have this structure where the AI is correct about the technical point but neutral on severity; the expert then calibrates the severity to what the paper's central claims actually require. The valuable contribution of the AI reviewer in this case is not to escalate the concern but to surface a specific, actionable improvement (switch to layered-medium Green's functions) that a generalist reviewer would not have suggested.

D.2 Complete categorized expert comments

This subsection lists every expert annotator comment that participated in the S/W classification of AI reviews. The full corpus is 442 substantive comments plus 121 paper-level descriptive comments. Comments labeled as being about a human reviewer, or carrying explicit *item-number* references instead of free-form prose, are handled by a separate artifact and are not listed here. Within each category, comments are sorted first by source (item-level before paper-level), then by paper id and reviewer. Each row is citable via its paper id, reviewer, and item number or paper-level slot number, so any comment can be traced back on our HuggingFace dataset.

Category summary

1188

Code	Name	n
W1	Missing community / field norms	54
W2	Over-harsh / out-of-scope / unrealistic	46
W3	Paper explicitly states X, AI says missing	37
W4	Redundancy across the 3 AI reviewers	28
W5	Vague / verbose / no actionable recommendation	24
W6	Trivial / nitpicking	16
W7	Technical term-of-art confusion	13
W8	Cites evidence from after the preprint	9
W9	Over-inflates small code/text inconsistencies	9
W10	Criticizes what authors already flagged as a limitation	6
W11	AI misreads a figure or caption	6
W12	AI misquotes or fabricates a verbatim quote	5
W13	AI misses supplementary content	3
W14	Ignores authors' own cited prior work	2
W15	AI misreads a table	1
W16	Cannot analyze figures — only text	1
W_unspecified	Residual: AI judged Not Correct without specific reason	10
S1	Statistical / methodology rigor	45
S2	Code reading	28
S3	Specialized niche field catch	27
S4	Internal consistency across sections	15
S5	Reproducibility / dependency failures	10
S6	Big-picture / counter-narrative synthesis	7
S_generic	Residual: AI judged Correct without specific reason	40
Total		442

1189

W1: Missing community / field norms — n = 54

1190

Citation	Expert comment
P2 · Claude 4.5 · item 4 · primary	I believe, it would make sense if this AI reviewer states what is necessary from the authors to support their claims. Yes, it is an issue that some calculations are not converging but it happens in computations/quantum chemistry.
P2 · GPT-5.2 · item 2 · primary	While this point is essentially valid, I find it quite arguable. None AL approach provides reasonable error bounds and none ML applied to physical sciences realistically has any. AL literally helps finding new data points to improve the accuracy of models without acquiring excessive amount of them. It is typically tested empirically. Therefore, asking the authors to comment on the issue is fine but making this argument too string is unfair.
P3 · Claude 4.5 · item 4 · primary	It is well known that 1D corrections are not working well and that is the reason why researchers using better approximations like RPMD and Instanton theory.
P8 · GPT-5.2 · item 1 · primary	SPlot is a common technique in particle physics and widely used in many publications.
P9 · GPT-5.2 · item 2 · primary	The comment is correct from statistical point of view, but the likelihood test cannot be done here due to how the parameters of interest are obtained.
P9 · GPT-5.2 · item 5 · primary	It is true that these things are not documented in the paper, however they are intenally at CERN. It is not a practice to publish these as part of Nature papers so this comment is irrelevant.
P9 · Gemini 3.0 · item 2 · primary	Full amplitude analysis is subject to a different measurement cannot be done in a CP violation measurement.
P13 · Gemini 3.0 · item 4 · primary	Although the criticism is valid, this method has been used in many recent papers. Such maps cannot be obtained easily from many individuals, does the average maps from a sample sample are currently methodologically acceptable.
P15 · Claude 4.5 · item 2 · primary	It's not possible to obtain intracranial data from healthy controls - it won't pass the ethical committee protocol. Unless it's a very specific population of quadriplegic patients who are getting an intracranial implant but that is extremely rare.
P15 · Claude 4.5 · item 3 · primary	Some of this data might not be feasible to be obtained.
P16 · GPT-5.2 · item 5 · primary	Sometimes authors release the code after publication, it could be the case for this preprint as well.
P17 · GPT-5.2 · item 2 · primary	Although the pointing out is quite correct but I feel that it neglects that in reality the observational data is quite limited and a comprehensive data cannot be available for now.

Citation	Expert comment
P18 · Claude 4.5 · item 2 · primary	The point is correct but neglecting the fact that obtaining enough observation data is always very difficult in reality. The reviewer should show some possible alternative observational data to be analyzed if proposing this claim.
P18 · Gemini 3.0 · item 3 · primary	Neglecting the difficulty of obtaining observational data by satellites.
P18 · Gemini 3.0 · item 4 · primary	Unfortunately, still common for physics community.
P19 · GPT-5.2 · item 4 · primary	This comment neglects the fact that obtaining sufficient amount of observational data is impossible for astrophysical community.
P20 · Claude 4.5 · item 4 · primary	The AI reviewer seems to lack context that it is reasonable to assume that dust attenuation can not be that high so early in the Universe from physical reasons. The AI reviewer could have provided more evidence that this can actually be a concern at these redshifts either via citations or further arguments. Also, it is reasonable to remove templates that are older than the universe, and it's unphysical to assume otherwise.
P20 · Claude 4.5 · item 5 · primary	It's a valid point, but again the context lacks — the preprint compares with at that time state-of-the-art measurements and the purpose is to prove the point that these brand-new observations are in striking contrast with the current best estimates with the selection function that was possible pre-JWST.
P21 · GPT-5.2 · item 1 · primary	Valid point, but the conclusion can be tentatively excepted given the evidence in the preprint and the early/exploratory nature of the work.
P22 · GPT-5.2 · item 2 · primary	While this point is valid, it can be acceptable to report such speculative caluculations, while clearly stating the caveats of basing this on only three objects — which would have been accepted by a human reviewer with enough context of the field.
P28 · Claude 4.5 · item 1 · primary	I did not do the statistical analysis the AI is claiming, but the statistical analysis in the paper is generally accepted in the community doing this type of research.
P28 · GPT-5.2 · item 1 · primary	The main point of the critique is correct. But to the bacterial modeling community this paper does not read that it claims a universal law. However, the AI reviewer is correct with the statement "The paper would need clearer limits on scope (e.g., "within these assembled microcosms and this dispersal/invasion protocol") or additional tests demonstrating the relationship persists under qualitatively different invasion settings." It is harsh, but correct request.
P35 · GPT-5.2 · item 2 · primary	this is a benchmark limitation shared by all methods evaluated on BioSNAP, not a flaw unique to this paper's methodology.
P37 · Claude 4.5 · item 4 · secondary	The authors might have not given much thought about clinical significance, but that doesn't seem like a critical problem if we treat this paper as an ML paper than a clinical paper.
P37 · GPT-5.2 · item 2 · secondary	This AI Reviewer seems to focus on statistical rigor in research (maybe because it's following the Nature's review guideline), but I don't think the ML community these days care so much about statistical rigor any more.
P41 · Claude 4.5 · item 4 · primary	I think the AI does understand how hard ecology is! R2 of 0.36 is impressive!
P43 · Claude 4.5 · item 2 · primary	Privacy comes up often in these reviews. I think that it is a technical aspect that is overrepresented in the literature, but less concerning in practice than what the reviewers (AI and humans) make it seem. Especially since we are training a model on skull-stripped data.
P50 · Claude 4.5 · item 2 · primary	Again, if this was a clinician-targeted paper, it would be important but the novelty of this paper lies in the system/engineering so normal subject is sufficient to demonstrate proof-of-concept.
P50 · GPT-5.2 · item 1 · primary	These are important points but typically we understand that this paper is 'biomedical' field paper and not 'medicine' paper, meaning accuracy in clinical term is less emphasized since the novelty of this paper lies in engineering, not a clinical application.
P52 · Claude 4.5 · item 2 · primary	This paper novelty lies in engineering/system, not a fully-deployed medical study.
P53 · Gemini 3.0 · item 1 · primary	this comment should account for the rest of the tools available and also all the data available and how difficult is to actually obtain all taxonomy :) so the comment lack prior knowledge that an expert in the filed would know by reading other studies or having used the MSMS tools.
P54 · Claude 4.5 · item 3 · primary	even though the criticism is on point, a human reviewer would have accounted for how difficult analysis and case studies are and that it is ok to publish as such. however yes, the authors could be careful not to overstate the application. but the potential is big and if tried in other settings than the validity will hold more.
P54 · Gemini 3.0 · item 5 · primary	it is allowed to do such an estimation, but the programs are accounting for all of these when doing discovery and allignment. AI is probably not trained to account for this knowledge.
P55 · Gemini 3.0 · item 3 · primary	the point 3 about the sample size, the AI is correct in what the literature suggest, but it all depends on the circumstances of the study and in this study the patient size number and the way they have been chosen is quite valuable enough.
P59 · Claude 4.5 · item 1 · secondary	This critique applies to virtually all in vitro toxicology studies that inform policy and in my opinion this is more of a theoretical caveat.

Citation	Expert comment
P61 · Gemini 3.0 · item 1 · secondary	While the concern is technically valid, it applies to virtually every study using CMIP6 for subnational health impact assessment coarse-resolution climate model output is a universal constraint. The authors partially mitigate this through the baseline calibration (Equation 5).
P62 · Claude 4.5 · item 2 · primary	The comment is true but at this proof of concept level, there are other main problems while one can speculate that, in practice, this might not be that critical thanks to prior knowledge on samples.
P62 · GPT-5.2 · item 4 · primary	I did not personally check the repository and the code, but the missing data could be retrieved from the authors by request. In pinciple, AI reviewer is right about the point that the paper should contain all the details and if something is missing this is a concern. However, in practice, providing everything without missing out a single point is not usually happening and this paper is already above average in terms of description of methods and details for reproducibility.
P62 · Gemini 3.0 · item 2 · primary	The comment is true but at this proof of concept level, there are other main problems while one can speculate that, in practice, this might not be that critical thanks to prior knowledge on samples.
P62 · Gemini 3.0 · item 3 · primary	I did not personally check the repository and the code, but the missing data could be retrieved from the authors by request.
P63 · GPT-5.2 · item 4 · primary	It is true that all the details to reproduce the results are not provided. However, the methods in the paper are well-known and there is sufficient depiction to reproduce the results, which will not be exactly the same with the paper since all the details are not provided but it is conceivable that the results would be fairly close.
P65 · Claude 4.5 · item 2 · primary	Modified Strehl formulation is very common in metamaterial research.
P65 · Gemini 3.0 · item 3 · primary	I don't think AI is correct. Despite having a few issues, the authors' design is one of the state-of-the-art and comparing its performance against a base design makes sense.
P67 · Claude 4.5 · item 2 · primary	This a correct statement but here the purpose is a proof of a concept.
P67 · Gemini 3.0 · item 4 · primary	The method has been used for nano-fabrication since early 2000. This might not be novel enough for AI, but it sure is for humans!
P76 · Claude 4.5 · item 3 · primary	These xenograft models are typically used to assess efficacy rather than on-target, off-tumor toxicity
P78 · Claude 4.5 · item 2 · primary	Faradaic efficiency is not highly weighed in electro-organic synthesis. AI reviewer is too harsh on this matter considering the accepted standard of the field.
P78 · Claude 4.5 · item 3 · primary	This RPC process is assumed in this field. No need to show the machanism of the second oxidation. The AI reviewer is too harsh on this matter.
P53 · Claude 4.5 · paper-level slot 2 · primary	Yes. the AI reviewers in general do not account to a prior knowledge of the filed so that they can make a fair point when criticizing the tool in terms of accuracy and application. Human reviewer, if chosen as a good expert in the filed, would have applied this knowledge when making the comments. This applies to all AI.
P53 · GPT-5.2 · paper-level slot 1 · primary	Yes. the AI reviewers in general do not account to a prior knowledge of the filed so that they can make a fair point when criticizing the tool in terms of accuracy and application. Human reviewer, if chosen as a good expert in the filed, would have applied this knowledge when making the comments. This applies to all AI. AI does not pay attention to Figure commenting when it comes to introducing information in the main paper but lacks information from the supplementary files. So more connection to what info can be written in the main part and taken from supplementary is something that AI misses to account for.
P53 · Gemini 3.0 · paper-level slot 3 · primary	Yes. the AI reviewers in general do not account to a prior knowledge of the filed so that they can make a fair point when criticizing the tool in terms of accuracy and application. Human reviewer, if chosen as a good expert in the filed, would have applied this knowledge when making the comments. This applies to all AI.
P54 · Claude 4.5 · paper-level slot 3 · primary	AI-3 was too detailed on the statistics, the numbers of analysis which is something that the nanopore sequencing account by itself and it is not something this paper is focused on. Human reviewers that are expert in genomics would have prior knowledge about this and would usually not comment on these very specific statistics. For example the percentage of accuracy for genome alignment etc. usually you do not find a 100 % match due to the way the softwares work.
P54 · GPT-5.2 · paper-level slot 1 · primary	In general the AI lacks prior knowledge about what genomics community accepts as a golden standard of matching or gene alignment for some of the trained softwares. Human reviewer would know this and would not be critic about it. The AI also comments on the probable application and lacks the knowledge that even case studies of such degree can be very useful when published.
P78 · Claude 4.5 · paper-level slot 1 · primary	AI reviewer 1 was too harsh on Faradaic economy, which is a concept that is not that highly weighed yet in synthetic organic electrochemistry.

Citation	Expert comment
P1 · Claude 4.5 · item 3 · primary	Overall, I agree with this comment. However, I would not claim that what the authors do is completely wrong. One may simply ask for doing that additional set of experiments.
P2 · Claude 4.5 · item 2 · primary	However, requesting CBS may be too much.
P4 · Gemini 3.0 · item 1 · primary	It is an architectural choice and it is valid to have advantages by using it.
P6 · Claude 4.5 · item 1 · primary	This is the least interesting aspect of the paper. None of the 3 human reviewers even bothered with it, because the paper is using safety as an excuse to not talk about the MR engineering. The AI may be right (I am not a statistics expert), but it is a marginal issue.
P6 · Claude 4.5 · item 2 · primary	The entire safety portion of the paper is just filler to justify not talking about the MR engineering. This is partially the paper’s fault, but the human reviewers were not fooled by the misdirection.
P6 · Claude 4.5 · item 4 · primary	Again irrelevant. The safety aspects belong in another paper and the duration of the follow-up is very field-specific. While it is true that the Grant et al. study looked at longer time horizons, I doubt that a follow-up of 1 week vs. 2-4 weeks makes any difference.
P6 · GPT-5.2 · item 1 · secondary	not focus of the paper
P8 · Claude 4.5 · item 2 · primary	This is not relevant and cannot be covered in this paper.
P8 · GPT-5.2 · item 4 · primary	It’s true that the assumption is made, however these inputs are not subject to a measurement in this paper.
P28 · Claude 4.5 · item 3 · primary	The statement of AI is correct, however fitting models against experimental data is beyond the scope of the paper.
P28 · Claude 4.5 · item 4 · primary	The statements are correct, but too harsh and are beyond the scope of the paper.
P28 · Claude 4.5 · item 5 · primary	The statement is correct, but a threshold must be chosen so the comment, although technically correct, is too harsh for a review.
P28 · GPT-5.2 · item 2 · primary	This is too harsh critique. The authors do not claim that their result is universal, although the AI would have had a good point if they did. This critique is somewhat along the lines of Human Reviewer 2.
P28 · GPT-5.2 · item 3 · primary	The AI is finding problems with external packages and software. This is too much, I did not go down the rabbit hole to evaluate it.
P28 · Gemini 3.0 · item 3 · primary	The AI here is in principle correct. However, this critique is too harsh. From the context it is clear that the scope of the paper is somewhat limited. This critique is similar to Human Reviewer 2.
P29 · Claude 4.5 · item 1 · primary	This is technically correct, but outside the scope of the paper.
P29 · Claude 4.5 · item 4 · primary	Correct, but out of the scope of the paper.
P30 · Claude 4.5 · item 2 · primary	Out of the scope of the paper.
P30 · Gemini 3.0 · item 5 · primary	Out of the scope of the paper.
P31 · Claude 4.5 · item 4 · primary	Too harsh critique.
P37 · Claude 4.5 · item 2 · secondary	The first two quote-comment pairs are logically valid, but I don’t think they are enough to reject the paper. However, the third pair is more critical, as one of the core claims (i.e. the importance of date embedding) is negated by the empirical result.
P42 · Claude 4.5 · item 3 · primary	Technically the AI is correct, but I think much like the other paper I reviewed, the AI does not understand the complexity of studying ecosystems and so has unrealistic expectations of the resultant classification performance. The signal in the models is strong even with the moderate classification performance.
P42 · GPT-5.2 · item 3 · primary	While technically correct again, I think the AI does not understand the complexity and challenges of studying ecosystems; however, there is some causal language in the paper that could be revised.
P43 · Claude 4.5 · item 1 · primary	The interesting part about the paper is more about the model itself rather than the federation.
P46 · Claude 4.5 · item 2 · primary	Yes, but what can you do?!
P46 · Claude 4.5 · item 4 · primary	Yes, but it is a valid and good comparison!
P46 · GPT-5.2 · item 3 · primary	It is correct that this is the case, but PG cannot be effectively implemented on the GPU, whereas polyBERT can. A comment in the text may be good, but to me this is not a big deal.
P49 · Claude 4.5 · item 2 · secondary	This critic is factually accurate but it highlights a limitation that the authors cannot realistically defend against or resolve by modifying their data, making it a non-constructive critique.
P52 · Claude 4.5 · item 4 · primary	Given this is academic paper, it’s difficult to ask authors to provide full biocompatibility data which would take too long.

Citation	Expert comment
P54 · Claude 4.5 · item 1 · primary	Bit too rigid criticism however, it does point out when authors overstate their point trying to sell their paper.
P59 · Claude 4.5 · item 4 · secondary	The dose response non-linearity for total PM 2.5 mass has been characterized, but there is no established dose-response function for toxic potency-adjusted PM2.5. This is a future research direction, not something the authors could realistically incorporate. Also citation 5, the link is wrong.
P59 · GPT-5.2 · item 4 · secondary	Applies an unrealistically stringent reproducibility standard. This standard of reproducibility, while ideal, is rarely met in chemical transport modeling studies published in top journals.
P59 · Gemini 3.0 · item 3 · secondary	reviewer seems to recognize that this is not a methodological flaw - "it does not necessarily invalidate the relative ranking" but still frames it as a significant concern. It can be a valuable addition to the discussions but its not affecting the conclusions.
P62 · Claude 4.5 · item 3 · primary	The comment is true but one can easily make a counter-argument of these not being a huge problem in controlled experiments.
P65 · Claude 4.5 · item 1 · primary	AI got this wrong. Comparing simulation and experimental results is not fair.
P65 · Gemini 3.0 · item 5 · primary	It looks like this was a badly trained AI platform. What could the authors do if their imaging takes place at certain wavelength values other than averaging the discrete outcome? This was my first study on Human vs AI reviewers. It is interesting to see that one of three AI platforms can come up with a totally different review than the other two. Clearly its training was done with different sources than the other two. LTSM was an issue 5-10 years ago. The AI reviewer III shows me that it might be still an issue for some AI platforms. There is a more significant alignment between AI reviewers I and II. They saw some important points which were not caught by human reviewers.
P66 · GPT-5.2 · item 1 · primary	This is such a harsh comment for an experimental study. The authors provided all the datasets. The measurements were taken several times and averages were reported. The real world measurements are not perfect or noise free. What else could authors do?
P66 · GPT-5.2 · item 2 · primary	I again do not agree with the AI reviewer. The wavelength is close to 800 nm and the tip is 20-25 nm wide, which is 1/30th - 1/40th of the wavelength. We can treat it as a point dipole. The AI reviewer's comment is too harsh!
P67 · GPT-5.2 · item 4 · primary	I understand AI's request, which is fair. But this shouldn't undermine the authors' contribution
P68 · Gemini 3.0 · item 1 · primary	This is technically correct, but morphological traits are very correlated, which makes the underlying dimensionality much lower than the number of coordinates. So, I am two ways with regards to this comment. On one way, the criticism is correct. On the other, it is likely overstated.
P72 · Claude 4.5 · item 5 · primary	DeBreak is more computationally efficient than one of the three standard callers it is compared to (pbsv), so I think this criticism is overblown.
P72 · Gemini 3.0 · item 4 · primary	One could always compare to more tools, but I probably would not have asked the authors to include an additional tool unless there was a specific reason for including it. e.g. one of the human reviewers' requested a comparison to PBHoney and had a good reason for doing so.
P77 · Claude 4.5 · item 5 · secondary	A definitive answer to this question may not be readily available. Nevertheless, it would be valuable if the authors could provide insight into the different reactivities of various metals.
P6 · Claude 4.5 · paper-level slot 1 · primary	AI Reviewer 1 focused too much on the safety aspects of the paper. 70% of the review was about safety, whereas the paper is primarily about MRI methodology. I am sure that some of the statistics observations were original, but they were irrelevant in my opinion.
P8 · Claude 4.5 · paper-level slot 1 · primary	Human review 2 is not available in the checklist. I noticed that there was one question in the AI reviewer that was suggesting to provide a measurements of some parameters which is out of the scope of the analysis done there and also some trivial questions on plot that are not applicable. That would never be asked by a human reviewer. In the AI reviewer 2 there were some trivial questions that should not be asked in Nature paper
P28 · GPT-5.2 · paper-level slot 1 · primary	Yes, the AI reviewer 1 was talking about the equations and math, but those comments are relevant for external software, not authored in this paper.

W3: Paper explicitly states X, AI says missing — n = 37

1192

Citation	Expert comment
P1 · Gemini 3.0 · item 1 · primary	The authors write explicitly that they employ finite-size correction.
P3 · Claude 4.5 · item 1 · primary	I guess one can criticize it but authors have already provided a reasonable evaluation.
P3 · Claude 4.5 · item 3 · primary	They give values they got in their experiments which is appropriate for this case.
P4 · GPT-5.2 · item 3 · primary	The authors provide the code and it does not matter whether it is on GitHub or GitLab.
P4 · Gemini 3.0 · item 5 · primary	The code is available.

Citation	Expert comment
P5 · Claude 4.5 · item 4 · primary	Formally, this criticism is correct, but it can be omitted because the authors provide molecular dynamics simulations for other benchmark systems.
P6 · Gemini 3.0 · item 1 · primary	The AI made factual errors. The sequence parameters were reported, and the methods section starts with this. It is true that the MRI methodology does not have a subheading, which is why the AI got confused.
P8 · Claude 4.5 · item 1 · primary	The comment make sense however the authors consider more complicated model as a systematic study.
P9 · GPT-5.2 · item 4 · primary	The comment is too vague of what is insufficient in the paper since there is some discussion on the resonant part.
P12 · Gemini 3.0 · item 3 · secondary	The concern about J1 being a non-detection is valid, but the Lorentz factor is derived including J2.
P17 · Gemini 3.0 · item 1 · primary	The pointing out on how to define classification during non-equilibrium process is appropriate but it neglects that the paper considers a response to linear perturbation and a near equilibrium initial profiles of observation and PIC simulations, leading to inadequately strong criticism to the authors' claim. The wording seems common on ML community unfriendly review comment but less common in physics community review.
P18 · GPT-5.2 · item 2 · primary	AI confused by a confusing authors' explanation, mixing 1D analyses with 2D reconnection case without explicitly clarifying. Just suggesting clarifying the explanation on that part is more appropriate as a review comment.
P25 · Gemini 3.0 · item 3 · primary	I was doubting between correct and not correct. What the criticism reffers to are instantaneous correlations, due to the wave there are lagged correlations.
P28 · Gemini 3.0 · item 1 · primary	The paper claims usage of a Matlab method, and does not explicitly state the equations mentioned in the review here. This is clearly a fail of the AI reviewer.
P30 · Claude 4.5 · item 1 · primary	It is clear from the context that only QS systems are included.
P30 · GPT-5.2 · item 1 · primary	It is clear from the context that QS systems only are considered.
P31 · Gemini 3.0 · item 1 · primary	Correct, but explained in the paper that this is inevitable.
P34 · Gemini 3.0 · item 1 · primary	The test should be done ~3000 times consistently with the paper, single run shouldn't be interpreted like this.
P37 · GPT-5.2 · item 4 · primary	the model architecture for finetuning is provided (ICDBartForSequenceBinaryClassification in icdmodelbart.py) but the training script is not.
P43 · Claude 4.5 · item 3 · primary	This comment disregards the existence of out-of-sample sites
P44 · Claude 4.5 · item 2 · primary	It's failing to take into account the reduced set, which naturally leads to higher p-values.
P50 · Gemini 3.0 · item 3 · primary	There is actually 'methods' text in the manuscript. It probably wasn't properly inputted into AI.
P55 · Claude 4.5 · item 2 · primary	the age is defined as above 18 so there is a need for more info to make such a comment
P59 · Gemini 3.0 · item 1 · secondary	This critic assumed "zero-out" meant removing all emissions (PM, SO ₂) from each sector but in the paper authors clearly mentioned that they zero out only primary PM 2.5 and they acknowledge this in the limitations section as well (i.e ignoring secondary aerosol contribution).
P61 · Claude 4.5 · item 5 · secondary	this is factually incorrect, the paper explicitly addresses this issue. Lines 489–496 describe a calibration procedure (Equation 5) that adjusts modeled PM _{2.5} concentrations against observed baseline period data (2012–2017).
P61 · Gemini 3.0 · item 4 · secondary	The paper's title, abstract, and methods explicitly scope the study to PM _{2.5} . The reviewer's argument that the paper's broader framing around "air pollution-related deaths" and SDGs implicitly requires ozone consideration is a stretch no single study is obligated to cover all pollutants, and the PM _{2.5} focus is clearly stated throughout.
P61 · Gemini 3.0 · item 5 · secondary	The paper actually uses the GBD2019 MR-BRT model (line 470), not the older IER function, and explicitly discusses the GEMM comparison (lines 547–550), acknowledging that GEMM yields higher estimates.
P62 · GPT-5.2 · item 2 · secondary	Marked it maginal, because important metrics are provided in their experimental context.
P66 · Claude 4.5 · item 2 · primary	The authors mention repetition of measurements and using averaged results in the end. I don't think this is a fair comment.
P66 · Claude 4.5 · item 3 · primary	The authors mention that their resolution is limited by the width of their tip and their experimental results in a good agreement with theoretical results (10 nm). There is no issue here!
P67 · GPT-5.2 · item 3 · primary	The zip file I downloaded has all the *.mat files
P67 · GPT-5.2 · item 3 · secondary	In the code bundle that was provided to me, the .mat files were empty. I believe this is what the AI reviewer is referring to. However, I could find the actual .mat files on GitHub.
P67 · Gemini 3.0 · item 3 · primary	They wrote "to simulate the process."

Citation	Expert comment
P67 · Gemini 3.0 · item 5 · primary	They already showed field enhancement far from the contacts.
P74 · Claude 4.5 · item 3 · primary	The point is valid but overblown given that the authors also assess accuracy on non-human genomes.
P74 · GPT-5.2 · item 2 · primary	Another valid point that is somewhat offset by the benchmarking on other organisms.
P77 · Claude 4.5 · item 2 · primary	The substrates described in this paper do not have this issue.

W4: Redundancy across the 3 AI reviewers — n = 28

1193

Citation	Expert comment
P5 · Claude 4.5 · item 2 · primary	To a large extent, all three AI reviewers raise very similar points. Thus, the overall review would not cover many substantial points human reviewer raised.
P42 · GPT-5.2 · item 1 · primary	Same comment as AI 2 item 2.
P46 · Gemini 3.0 · item 1 · primary	Same as point 1 of other reviewer.
P59 · GPT-5.2 · item 2 · secondary	This item largely duplicates AI Review #1 Item 1. Despite being more elaborately argued, it doesn't add a fundamentally new dimension. The authors should temper their language from "health-oriented" to something like "toxicity-informed" and explicitly acknowledge that TPAE is an exploratory index rather than a calibrated health impact function is easy to implement.
P59 · GPT-5.2 · item 3 · secondary	Essentially the same critique as AI Review #2 Item 1. It actually creates a conceptual tension: secondary sulfate/nitrate are widely considered less toxic, so including them would increase each sector's attributable concentration while simultaneously lowering the average toxicity per unit mass. The reviewer asserts this "can materially change" rankings but provides no quantitative basis for that claim.
P60 · Claude 4.5 · item 1 · secondary	This is a well-constructed critique but largely redundant with the AI reviewer1's Items 2 and 3. The main added value is the specific Ali et al. (2025) reference demonstrating synergistic effects with quantified RRs, which provides stronger empirical backing than the first reviewer offered for the same point.
P60 · Claude 4.5 · item 2 · secondary	This is essentially the same critique as the AI reviewer 1 Item 2, which already identified that removing COVID-19 deaths doesn't make the remaining mortality independent of pandemic context, and that the model trained on primarily pre-pandemic data fails to capture altered vulnerability.
P60 · GPT-5.2 · item 1 · secondary	AI Reviewer 1 (Items 2 and 3) identified both the violated independence assumption and the misleading "compound" terminology. AI Reviewer 2 (Item 1) restated it with additional references. This reviewer articulates it more formally mentioning joint models, attributable fraction partitioning, and counterfactual decomposition but the core insight is identical
P60 · GPT-5.2 · item 2 · secondary	This is a thorough and well-referenced elaboration of a point other reviewers already made more concisely. The added value is the explicit connection to the Vicedo-Cabrera projection framework, which helps explain why the authors may have made this choice (importing a projection-style baseline into a retrospective analysis) and why it's inappropriate here.
P61 · Gemini 3.0 · item 2 · secondary	This item is essentially a less developed version of what AI Review #2 Item 2 covers more comprehensively.
P62 · Claude 4.5 · item 3 · secondary	Aligned with Item 7 of Human Reviewer 3.
P62 · GPT-5.2 · item 1 · secondary	Aligned with Item 7 of Human Reviewer 3, Item 3 of AI Reviewer 1, and Item 1 of AI Reviewer 2. But not addressing other important challenges in in vivo settings like Human Reviewer 3.
P62 · GPT-5.2 · item 3 · primary	Again the same comment about NMF rank that is common in all AI reviews.
P62 · GPT-5.2 · item 3 · secondary	Aligned with Item 2 of AI Reviewer 1 and Item 2 of AI Reviewer 2.
P62 · GPT-5.2 · item 4 · secondary	Aligned with Item 3 of AI reviewer 2. Marked marginal by the same reason.
P62 · Gemini 3.0 · item 1 · secondary	Aligned with Item 7 of Human Reviewer 3, and Item 3 of AI Reviewer 1. But not addressing other important challenges in in vivo settings like Human Reviewer 3.
P62 · Gemini 3.0 · item 2 · secondary	Aligned with Item 2 of AI Reviewer 1.
P72 · GPT-5.2 · item 3 · primary	A fair point, echoing a comment from one of the human reviewers.
P72 · GPT-5.2 · item 4 · primary	Echoes a point by one of the human reviewers that the authors could emphasize that their benchmarking results do not encompass all SV types handled by DeBreak.

Citation	Expert comment
P4 · Claude 4.5 · paper-level slot 3 · primary	I found the point on the comparison with prior work important. Other points are rather similar to other AI reviewers.
P22 · Claude 4.5 · paper-level slot 3 · primary	In general, they pick up on the same major points
P22 · GPT-5.2 · paper-level slot 1 · primary	In general, they pick up on the same major points
P22 · Gemini 3.0 · paper-level slot 2 · primary	In general, they pick up on the same major points
P62 · Claude 4.5 · paper-level slot 1 · primary	None. In this paper's specific case, all human reviewers were exceptional and their combined feedback covered every possible aspect. All AI reviewers to some extent failed to give feedback that could possibly improve the manuscript while keeping the intricacies and difficulties of optics+bio experimental work landscape.
P62 · GPT-5.2 · paper-level slot 3 · primary	None. In this paper's specific case, all human reviewers were exceptional and their combined feedback covered every possible aspect. All AI reviewers to some extent failed to give feedback that could possibly improve the manuscript while keeping the intricacies and difficulties of optics+bio experimental work landscape.
P62 · Gemini 3.0 · paper-level slot 2 · primary	None. In this paper's specific case, all human reviewers were exceptional and their combined feedback covered every possible aspect. All AI reviewers to some extent failed to give feedback that could possibly improve the manuscript while keeping the intricacies and difficulties of optics+bio experimental work landscape.
P67 · Claude 4.5 · paper-level slot 3 · secondary	The review of this AI is similar to the review of AI reviewer 1 but I find that the AI reviewer 1 commented on matters that are more important first.
P72 · GPT-5.2 · paper-level slot 3 · primary	No. This AI Reviewer was the only one to complain about the assembly-based callset being treated as ground-truth, but at least one other reviewer was clearly aware that this was not the authors' true intention—they were just using it as a decently accurate set that might be worth comparing alignment-based methods to.

W5: Vague / verbose / no actionable recommendation — n = 24

Citation	Expert comment
P1 · Claude 4.5 · item 2 · primary	The valid point is that authors refer to ref. 21 as providing the experimental value. However, ref. 21 takes the corresponding reference value from a different work. This could be requested to correct. But, in general, the comment/critique by this reviewer is rather pointless.
P5 · GPT-5.2 · item 1 · primary	While I agree with the points raised by this AI reviewer, I found them somewhat too lengthy and repetitive.
P5 · GPT-5.2 · item 4 · primary	Similar to the previous item, this AI reviewer tends to overemphasize each quotation and comment. It would be sufficient to state that repeating the simulation with different starting points and random seeds is necessary to estimate the variance and thus draw more reliable conclusions. This AI reviewer also appears to misuse the term of "uncertainty".
P6 · GPT-5.2 · item 1 · primary	Too verbose. I am not a statistics expert, but this is a paper about MR methodology. The first comment should not be six paragraphs about biological effects.
P6 · GPT-5.2 · item 2 · primary	Too verbose. The previous two AI reviewers did a much better job pinpointing the main issues.
P17 · Gemini 3.0 · item 2 · primary	The pointing out is basically correct but needs more comments. Maybe a more appropriate review comment on this point is to suggest to take more effort on validating why the authors used MMS observation data with reconnection to compare 1D non-reconnection current sheet PIC simulations. Basically, it criticized too much.
P22 · GPT-5.2 · item 1 · primary	Thorough and well articulated, but perhaps unnecessarily long.
P22 · GPT-5.2 · item 3 · primary	Too long.
P29 · Claude 4.5 · item 2 · primary	This critique, technically correct, is confusing the significance of some of the results.
P29 · GPT-5.2 · item 2 · primary	There are some correct elements here, but it is stated in a super confusing way.
P29 · GPT-5.2 · item 3 · primary	This has an element of truth, but evidence is not provided well, and the statement is convoluted.
P29 · Gemini 3.0 · item 3 · primary	This might be correct but it is stated in a convoluted way.
P36 · Gemini 3.0 · item 3 · primary	The evidence consists only of paper quotes restated with the reviewer's interpretation, without external citations or concrete reasoning to support the reverse causality claim
P41 · Claude 4.5 · item 1 · primary	As with other reviews, it is unclear what the recommendation for addressing this criticism is.
P41 · Claude 4.5 · item 3 · primary	Similar to AI review #1, this comment is true and yet the signal in the results is still strong, it's unclear from the comment whether this is a fundamental criticism or something that requires addressing in the discussion.

Citation	Expert comment
P41 · GPT-5.2 · item 1 · primary	As with other reviews, it is unclear what the recommendation for addressing this criticism is.
P41 · Gemini 3.0 · item 1 · primary	The review does not provide any suggestions for how to address this criticism.
P41 · Gemini 3.0 · item 2 · primary	It is unclear from the criticism how this issue results in bias per se, which is a much greater concern than “noise”, given the strong signal in the results of the paper.
P49 · GPT-5.2 · item 5 · secondary	While this AI reviewer (and others) is excellent at pointing out flaws, it fails to provide constructive alternatives or actionable suggestions on how the authors should address these criticisms to improve the paper.
P60 · GPT-5.2 · item 4 · secondary	The point is valid but the level of detail suggests the reviewer may be optimizing for thoroughness over insight.
P77 · GPT-5.2 · item 4 · secondary	I could not understand the reviewer’s main point.
P78 · Gemini 3.0 · item 2 · secondary	Although the point mentioned by the reviewer is valid, more evidence (references) should be given.
P5 · GPT-5.2 · paper-level slot 2 · primary	I found the items themselves to be reasonable, but the underlying argumentation was diffuse and overcrowded with statements that were largely irrelevant. Overall, as noted in the checklist, all reviewers raised very similar points. However, the first reviewer provided the strongest arguments, whereas this reviewer’s arguments were the weakest.
P6 · GPT-5.2 · paper-level slot 3 · primary	This reviewer hallucinated the most and it spit out word salads. I am sure there were some original insights in there (the review was 5 times longer than the others), but they were clouded by verbosity

W6: Trivial / nitpicking — n = 16

1195

Citation	Expert comment
P3 · GPT-5.2 · item 3 · primary	I believe it is just a yet another result the authors present which is essentially disconnected from the main narrative.
P17 · GPT-5.2 · item 3 · primary	It seems in the end criticizing a typo in the caption of Figure 3, $1000 \omega^{-1}_{-1-ci} \rightarrow 100 \omega^{-1}_{-1-ci}$ but the comment overly criticizing. Just pointing out a typo is enough.
P25 · Claude 4.5 · item 1 · primary	more of a technicality.
P26 · GPT-5.2 · item 4 · primary	seems like a technicality to me.
P42 · Gemini 3.0 · item 3 · primary	Pretty sure that sentence is just a typo, not a misunderstanding and validity risk.
P46 · GPT-5.2 · item 2 · primary	Do not think this is particular relevant. Sure, there may be badly predicted values outside the range...
P53 · Gemini 3.0 · item 2 · primary	It is ok observation but definitely does not limit the study and the tools. it just makes it specific.
P59 · Claude 4.5 · item 2 · secondary	This pairs with Human reviewer 3 critique about acidity preservation however here the critique is more of a “the authors should discuss this limitation” point than a fatal flaw.
P62 · Gemini 3.0 · item 1 · primary	The comment is valid but the paper presents this as a grand vision rather than something it achieves, that is why this comment is marginally significant rather than significant.
P66 · Claude 4.5 · item 1 · primary	When we fit data with a Lorentzian using MATLAB for example, MATLAB gives us the best fit and range of values which leads to, let’s say, R2 value of 0.99. The point raised by the AI reviewer is minor with respect to the main message of the manuscript.
P67 · Claude 4.5 · item 3 · primary	This is a very minor issue. In fact, sometimes I personally forget to include the statistical analysis of ML outcomes.
P68 · Claude 4.5 · item 1 · primary	This is true, but unlikely to change any major conclusion
P68 · Gemini 3.0 · item 5 · primary	These are indeed speculative assertions that can be easily dealt with.
P72 · Claude 4.5 · paper-level slot 1 · primary	AI reviewer 1 noted that the GIAB ground-truth set contained only SVs within the “high-confidence” regions. I consider to be a relatively minor point that could be addressed by emphasizing that benchmarking results may not generalize to more difficult/repetitive parts of the genome.
P72 · Gemini 3.0 · paper-level slot 2 · primary	An unused argument in the code, which is helpful but about as minor as a criticism can be.
P73 · Gemini 3.0 · paper-level slot 1 · primary	AI Reviewer 1 was the only reviewer to point out that there are newer methods for detecting microsat mutations, but I felt that this was a minor point.

W7: Technical term-of-art confusion — n = 13

1196

Citation	Expert comment
P25 · Gemini 3.0 · item 1 · primary	On first sight it seems like a good comment, but AI and TW are established as two different aspects of neuronal dynamics, the first is local, the second is spatially extended.
P28 · Gemini 3.0 · item 4 · primary	There is a misunderstanding of what the term "universal" means here. "Universality" is mere a physics term that means the data collaps to a single curve when scaled. It does not mean it is universally valid.
P42 · Gemini 3.0 · item 2 · primary	This seems a bit of a terminology/scope issue, not a validity issue.
P54 · Gemini 3.0 · item 3 · primary	AI should be aware of the connotation that is accepted by the genomics community for the use of terminologies such is real time.
P65 · Gemini 3.0 · item 2 · primary	INR might take a few seconds for training, but once the training is complete, then it is extremely fast during inference.
P66 · Gemini 3.0 · item 3 · primary	Here the resolution depends on the width of the tip, not the excitation wavelength.
P67 · Claude 4.5 · item 4 · primary	I think AI got this totally wrong. Since 6 photodetectors have transmission peaks at different at different wavelengths, they can provide a much richer training data which eventually leads to a more accurate edge detection.
P67 · Gemini 3.0 · item 1 · primary	SCA acts like a monochromatic filter, MCA is more like an array filter.
P67 · Gemini 3.0 · item 2 · primary	SCA output is more like the average of 6 photodiodes of the same type. MCA output is more like an output of 6 different type photodiodes, which can distinguish the edges more accurately.
P72 · GPT-5.2 · item 1 · primary	This comment is in line with one of the human reviewer's criticisms about the lack of methodological detail included in the paper. I don't get the AI reviewer's complaint about requiring a reference genome though, given that DeBreak is an read alignment-based method for detecting and locating SVs.
P77 · Claude 4.5 · item 2 · secondary	In most of the substrates, there are no stereogenic centers to consider. If the reviewer is referring to the newly formed stereogenic centers arising from the C–C bond cleavage (Table 3), this is a valid point; however, these centers are not related to the electrocyclization step.
P77 · Gemini 3.0 · item 3 · secondary	While the terminology appears to be somewhat misleading, both C–H and C–C bond cleavages would not take place in the absence of the metal carbene.
P2 · Gemini 3.0 · paper-level slot 3 · primary	The issue with the size-extensivity of the proposed approach is a very important one. However, it should not be an argument for rejecting this work. I believe, the authors should simply thoroughly discuss this limitation. Overall, I found all AI reviewer very good but quite critical and rather not differentiating issues stemming from quantum chemistry and machine learning. I think, the issues from the former should be discussed in the paper, but not be a major issue when evaluating the proposed machine learning approach.

W8: Cites evidence from after the preprint — n = 9

Citation	Expert comment
P1 · Claude 4.5 · item 1 · primary	While claims are actually fine, I marked it as not correct as the paper they use to support these claims appeared two years after the presented one. The field was evolving very fast and in 2022 the presented results could be considered as good. The only thing one could move on with is the claims of improving upon the state-of-the-art. The authors should clarify what exactly they mean with it.
P5 · Claude 4.5 · item 1 · primary	Funnily, this AI reviewer references in the last quote the same paper published later in 2022.
P20 · Claude 4.5 · item 1 · primary	This point is significant and valid, however it lacks context and chronology. While it is true that spectroscopic confirmation is crucial, photometric evidence can still be compelling, especially given the fact that this discovery is reported very early after JWST started operations. Given this context, reporting such a potentially far-reaching photometric discovery is acceptable. Additionally, the AI reviewer provides evidence from papers from 2025 which is invalid.
P20 · Claude 4.5 · item 2 · primary	Again, while this point is valid and significant, the evidence is based on knowledge obtained in the future related to when this preprint was reviewed. Although a human reviewer should have asked the authors to comment on the AGN contamination, the significance would have been less because at that time we didn't have the knowledge of this class of objects.
P20 · Gemini 3.0 · item 1 · primary	Again chronology is not preserved and the cited evidence is after this preprint
P21 · Claude 4.5 · item 1 · primary	The comments cite a paper published 2 years after this preprint which is not valid.
P46 · Claude 4.5 · item 3 · primary	This needs to be seen in historical context. PolyBERT was very early.
P20 · Claude 4.5 · paper-level slot 1 · primary	Yes, but the evidence was based on knowledge obtained several years after the publication of this preprint.

Citation	Expert comment
P20 · Gemini 3.0 · paper-level slot 3 · primary	Yes, but the evidence was based on knowledge obtained several years after the publication of this preprint.

W9: Over-inflates small code/text inconsistencies — n = 9

1198

Citation	Expert comment
P12 · Claude 4.5 · item 2 · secondary	The two Doppler estimates are not shown to be in serious conflict, because the variability estimate has no quoted uncertainty.
P34 · Gemini 3.0 · item 2 · primary	This one could be an unit error in the code, over-interpreted by AI-reviewer.
P34 · Gemini 3.0 · item 4 · primary	This could be also the unit error in the code.
P42 · GPT-5.2 · item 2 · primary	I think the AI is overstating the implications of small deviations between code and text that occur over the lifetime of a scientific paper.
P42 · GPT-5.2 · item 5 · primary	Same comment as item 2.
P46 · GPT-5.2 · item 4 · primary	This is most likely not a real problem and solved upon final submission.
P54 · Claude 4.5 · item 5 · primary	Quote: "Functional annotation of the assembled plasmid genomes using BugSeq revealed that KPC-2 and KPC-14 were both located on IncN plasmids, which showed 99.7% identity according to sequence alignments"-overstatement by AI and overanalysis
P59 · Claude 4.5 · item 3 · secondary	This feels very much AI generated critique that identifies a real asymmetry in methodology but fails to evaluate whether it matters quantitatively. the cement sector's contribution to total TPAE is so small that even substantial provincial variation in cement toxicity would not meaningfully alter the conclusions.
P59 · GPT-5.2 · item 1 · secondary	The within-sector variation is modest compared to the ~5 fold between-sector differences. So while the two aggregation methods would yield different absolute TU values, the sector ranking is very unlikely to change. However, the reviewer is correct that the authors don't demonstrate this, a sensitivity analysis would be needed to confirm.

W10: Criticizes what authors already flagged as a limitation — n = 6

1199

Citation	Expert comment
P3 · Claude 4.5 · item 2 · primary	The authors state explicitly that it is a proof-of-concept study. Transferability claims are somewhat unrelated to the main narrative but it does not make the reviewer's claim correct.
P46 · GPT-5.2 · item 1 · primary	It is a problem that is also already mentioned by the authors. It is up to the reviewer to judge if that invalidates everything. I would not think so.
P60 · Gemini 3.0 · item 2 · secondary	The core criticism is valid but should be calibrated against what the paper actually discusses later. The deeper methodological question whether to fit separate models for the COVID period is something the authors themselves flag as future work, constrained by insufficient data (only 2 years of COVID-period observations for regional-level estimation).
P60 · Gemini 3.0 · item 3 · secondary	This is a well-articulated critique, but it partly overlaps with Item 2 and with what the authors themselves discuss. The paper actually acknowledges this directly in the text quoted by the reviewer and in the discussion around Figure 4, where the authors present evidence suggesting the two hazards do interact and explicitly state they assumed independence.
P62 · Gemini 3.0 · item 3 · secondary	The algorithm is essentially NMF and authors stated that they used the existing NMF package already available. The way to use NMF in fiber photometry context is new, so I think this comment, which is regarding code implementation, is marginal.
P62 · Gemini 3.0 · item 4 · secondary	The key point of this study is not in NMF, nor the authors does not claim NMF as their developed method. The way to use NMF in fiber photometry context is new, so I think this comment is marginal.

W11: AI misreads a figure or caption — n = 6

1200

Citation	Expert comment
P3 · Gemini 3.0 · item 1 · primary	For example, Figure S10 does not show barriers of 40-50 kJ/mol. Furthermore, Figure S13 shows every barriers of a comparable height to previous studies. Finally, the chosen DFT may be a valid critique but the results should be interpreted within this setting if it is not dramatically wrong.
P19 · Gemini 3.0 · item 3 · primary	Maybe neglecting Figure 4.

Citation	Expert comment
P35 · Claude 4.5 · item 3 · primary	review misidentifies the optimal threshold as $t=0.4$ when Figure 5c shows the peak at $t=0.3$.
P35 · Claude 4.5 · item 4 · primary	error bars are present in figure2c (small gray lines on each bar)
P66 · GPT-5.2 · item 4 · primary	I do not agree with the AI reviewer again. The uniformity in Fig3. d and e removes the questions about the defects. Since these were mechanical exfoliated samples, strain criticism is also not relevant.
P77 · Claude 4.5 · item 3 · secondary	The reviewer may have misunderstood Figure 2 in the main text.

W12: AI misquotes or fabricates a verbatim quote — n = 5

Citation	Expert comment
P29 · GPT-5.2 · item 1 · primary	I doubt AI run the model.
P49 · Gemini 3.0 · item 1 · secondary	Instead of quoting the author's exact words, this AI reviewer incorrectly quoted the author and then claimed that the author's claims were incorrect. It's bit surprising that AI reviewers can make such mistakes, and I think we need to be aware that they can make such simple mistakes.
P65 · Gemini 3.0 · item 4 · primary	Again, AI made a mistake by comparing the sentences taken from different parts of the manuscript talking about different parts of the process.
P77 · GPT-5.2 · item 3 · secondary	I could not find CuTe in both main text and supplementary information.
P77 · Gemini 3.0 · item 1 · secondary	I could not observe the discrepancies pointed out by the reviewer.

W13: AI misses supplementary content — n = 3

Citation	Expert comment
P6 · Gemini 3.0 · item 1 · secondary	very detailed in the SI: The RF and DB 0 π eld maps were acquired for each subject for the pulse design and shimming using 2D interferometric turbo-FLASH [30] at 5 mm isotropic resolution and 3D multi-echo ($TE = 1.6/3.5/6$ ms) GRE acquisitions at 2.5 mm isotropic resolution respectively. Sequence parameters for the T2-weighted variable π ip angle turbo spin-echo acquisition were: resolution = $0.55 \times 0.55 \times 0.55$ mm ³ , $TR = 6$ s, $TA = 13$ min, GRAPPA = 3×2 , $TE = 301$ ms, matrix = $400 \times 400 \times 320$, bandwidth = 250 Hz/pixel.
P37 · GPT-5.2 · item 1 · secondary	The first comment is correct. The second comment is correct based on what is written in the main text. But if you refer to the Supplementary Figure 3, then we can see the number 281 in the main text is a typo. The actual number is 69281. It seems the AI Reviewer did not check the supplementary material.
P53 · Claude 4.5 · item 1 · primary	In supplementary

W14: Ignores authors' own cited prior work — n = 2

Citation	Expert comment
P28 · Claude 4.5 · item 2 · primary	The authors make the claim about correlation of nutrient concentration and interaction strength based on their previous work. This critique is similar to the one of the Human Reviewers.
P28 · Gemini 3.0 · item 2 · primary	The authors make this claim based on their previous work. Similar to one of the human reviewers, the AI reviewer did not read those cited papers.

W15: AI misreads a table — n = 1

Citation	Expert comment
P32 · Claude 4.5 · item 2 · primary	The reviewer's assessment seems to be based on an over-interpretation of Table 1, where a typographical error in the values may have led to an incorrect conclusion.

W16: Cannot analyze figures — only text — n = 1

1205

Citation	Expert comment
P6 · Gemini 3.0 · item 5 · primary	This could be argued better, but to do this the AI would need to be able to analyze the images. It went by text only, did not download the figures from Figshare, so it gave boilerplate feedback.

W_unspecified: Residual: AI judged Not Correct without specific reason — n = 10

1206

Citation	Expert comment
P28 · Gemini 3.0 · item 5 · primary	I don't think this is correct.
P29 · Gemini 3.0 · item 1 · primary	Complete fail here.
P30 · GPT-5.2 · item 4 · primary	I am not sure where this code is.
P34 · Gemini 3.0 · item 1 · secondary	I did not verify this in detail by running the Python code, and I assume the authors did not make this mistake. However, if this is indeed the case, it would be a significant issue.
P3 · Claude 4.5 · paper-level slot 2 · primary	I found this reviewer rather bad.
P3 · Gemini 3.0 · paper-level slot 3 · primary	I found this reviewer rather bad.
P9 · GPT-5.2 · paper-level slot 1 · primary	Human review 2 was not provided in the checklist. For the last question here I added AI reviewer 3 however it's a bit borderline, I think the 3 AI reviews were not very good.
P77 · Gemini 3.0 · paper-level slot 1 · primary	AI Reviewer 1 offered the weakest assessment among all reviewers.
P79 · GPT-5.2 · paper-level slot 1 · primary	AI reviewer 1 does not seem to have a fundamental understanding of the key discovery of the manuscript.
P80 · GPT-5.2 · paper-level slot 1 · primary	it misses the main point of the paper.

S1: Statistical / methodology rigor — n = 45

1207

Citation	Expert comment
P7 · Gemini 3.0 · item 2 · primary	Here in general the problem is that the fit is of bad quality
P35 · GPT-5.2 · item 1 · primary	The PPI trainer has no validation split (only train/test), uses test metrics for model selection (misleadingly named best_valid_f1)
P60 · Claude 4.5 · item 3 · secondary	Even if the overall sample size were adequate, treating correlated regional observations within the same event as independent violates the K-S test assumptions. This is a genuinely useful statistical critique that neither the first reviewer in this set nor the human reviewers identified.
P60 · Gemini 3.0 · item 1 · secondary	This is an outstanding technical critique. It's worth noting that the attributable AF itself is independent of the baseline, so the relative risk estimates are unaffected but the absolute death counts, which are the paper's primary output, scale directly with deathexp.
P60 · Gemini 3.0 · item 5 · secondary	The 1.7–3.8x fold increases for heatwaves and 2.1–3.5x for cold snaps are presented as key results demonstrating the severity of compound crises, but as the reviewer notes, these numbers are mathematically inevitable once COVID-19 deaths exist. This connects to Item 3's point about the paper not truly measuring compound/interactive effects.
P61 · Claude 4.5 · item 3 · secondary	this is logically paired with Item 2 and together they form a coherent critique: the provincial methodology is underspecified (Item 2) and unvalidated (Item 3). It's possible that the Pearson correlation in Figure S3 does include some provincial-level comparisons but again since its not explicit, authors needs to work on the presentation.
P61 · Claude 4.5 · item 4 · secondary	The authors should clarify: did they combine bounds simultaneously (which would overestimate uncertainty for independent sources), use Monte Carlo sampling, or apply some other approach? This is a fair ask for methodological transparency even if the practical impact on conclusions is small.
P61 · GPT-5.2 · item 2 · secondary	The paper would benefit from a supplementary table of fitted parameters, a description of the ARIMA specification procedure, and explicit statement of the TMREL distribution used.
P61 · GPT-5.2 · item 3 · secondary	The reviewer correctly distinguishes between association (correlation) and agreement (accuracy in magnitude), citing the Bland-Altman framework which is the standard methodological reference for exactly this issue

Citation	Expert comment
P61 · GPT-5.2 · item 4 · secondary	This critique pairs well with Review 1's Item 4 (incomplete uncertainty propagation) but is far more precise. The critique doesn't necessarily invalidate the central qualitative findings (aging as dominant driver), but it does undermine the quantitative confidence statements that the paper emphasizes throughout.
P63 · Gemini 3.0 · item 4 · primary	The power criticism is very correct and very significant. On the time, comments are correct but the mentioned challenges can be overcome. However the energy part is the fundamental criticism that should be enough to reject the paper.
P65 · GPT-5.2 · item 4 · primary	I agree with the AI: "a great deal" does not mean anything from a scientific point of view. They should have quantified their finding.
P68 · Gemini 3.0 · item 2 · primary	Given that there are categorical PLS options, this is a reasonable criticism.
P73 · GPT-5.2 · item 1 · primary	These methods do appear to be missing, and could be consequential.
P74 · Claude 4.5 · item 2 · primary	I agree that data leakage is possible if the same regions are included in both training and test sets. A safer approach would be to split the genome up into segments that are partitioned into the training and test sets.
P3 · GPT-5.2 · paper-level slot 1 · primary	This AI reviewer recognized ambiguity in the description of methods used for certain calculations and asks for the clarification for the computed coverages.
P4 · Gemini 3.0 · paper-level slot 2 · primary	Essentially, I find again great that also this reviewer questions if random splitting is a good strategy, similar to the first AI reviewer.
P5 · Gemini 3.0 · paper-level slot 1 · primary	I found this AI reviewer quite good. They were very constructive and concise. Their arguments are close to those I would make. Similar to the third reviewer, the most significant contribution, in my opinion, was pointing out the lack of a proper baseline comparison and the absence of variance estimates in the computed errors of the predicted values.
P14 · Claude 4.5 · paper-level slot 1 · secondary	Yes, Potential confounding of circadian rhythm detection by ongoing therapeutic stimulation.
P15 · Claude 4.5 · paper-level slot 2 · primary	Human reviewer only consider the relation with the data analysis issues.
P15 · Claude 4.5 · paper-level slot 2 · secondary	Absence of multiple comparisons
P15 · GPT-5.2 · paper-level slot 1 · primary	[non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin].
P15 · Gemini 3.0 · paper-level slot 3 · primary	Yes, data imputation methodological problem and inconsistencies; multiple comparisons absence.
P16 · Claude 4.5 · paper-level slot 3 · primary	Lack of analyses on EEG microstructures; data imputation problem
P24 · GPT-5.2 · paper-level slot 3 · primary	granger causality group level significance but not individual level
P35 · Claude 4.5 · paper-level slot 3 · primary	catching the normative atlas issue missed by the humans. nice catching the lack of spurious correlations null model.
P35 · GPT-5.2 · paper-level slot 2 · primary	The utilization rate metric mathematical error and the suspicious standard deviation entry in Table S2 were unique observations.
P37 · GPT-5.2 · paper-level slot 3 · secondary	The evaluation protocol violation (missing validation split, model selection on test set) provides a complementary angle to AI Review 1's data leakage finding. The BioSNAP benchmark limitation and code reproducibility issues were unique to AI Reviewer 2.
P42 · Gemini 3.0 · paper-level slot 2 · primary	It was the only reviewer who noticed: - Class imbalance in the dataset affects AUPRC, which can undermine the SOTA claim of the paper. - The submitted code does not accurately calculate AUROC
P47 · Claude 4.5 · paper-level slot 3 · primary	AI reviewer 2 found the collinearity of habitat fragmentation and amount and had a good suggestion for sensitivity analysis.
P47 · Gemini 3.0 · paper-level slot 1 · primary	AI #3 uniquely pointed out that the entropy equation is mathematically incorrect.
P49 · Claude 4.5 · paper-level slot 1 · secondary	AI #1 was the only reviewer who raised a critical question regarding the intra-hamming distance, and argued that its magnitude is too large to support stable cryptographic key generation.
P50 · GPT-5.2 · paper-level slot 1 · primary	AI reviewer 1 pointed out a statistical flaw in the distance-dependent cross-validation method. Specifically, test regions adjacent to the boundary of the 75% training set and the 25% test set still share spatial autocorrelation, which can potentially inflate the model's performance estimates.
P50 · GPT-5.2 · paper-level slot 1 · secondary	Yes, AI reviewer 1 found out about SNR which all other reviewers failed to catch.
P52 · GPT-5.2 · paper-level slot 3 · primary	AI Reviewer 1's review included details such as the SNR that was not in other reviews, human or otherwise.
P52 · GPT-5.2 · paper-level slot 3 · secondary	Yes. Issue of experimental setting on "control".
	confounding factors related to contralateral effects of the drug that was used unilaterally, with the contralateral eye as a control. Also, this reviewer mentioned the code that was used in the paper.

Citation	Expert comment
P55 · Claude 4.5 · paper-level slot 3 · primary	AI reviewer 3 overestimated the age of the population and had similar comments as the other AI reviewers. it did cough more statistics and model flaws than all the human comments. same applies for AI 1 and AI2.
P55 · Gemini 3.0 · paper-level slot 1 · primary	MS methodology in for example catching the imputation in the data analysis. this was not caught by the human reviewer.
P67 · Claude 4.5 · paper-level slot 3 · primary	AIR 3 caught the lack of statistical analysis, which was not caught by other reviewers. AIR2 and AIR3 did a very good job in terms finding the different resolution used in SCA and MCA, which I hope was fixed later by the authors.
P68 · Gemini 3.0 · paper-level slot 2 · primary	The ordinal data issue was raised by AI reviewer 2 and not by others.
P79 · Claude 4.5 · paper-level slot 2 · primary	AI reviewers 2 and 3 both made critical assessment regarding the causality
P79 · Gemini 3.0 · paper-level slot 3 · primary	AI reviewers 2 and 3 both made critical assessment regarding the causality. AI Reviewer 3 was more in depth than AI reviewer 2.
P80 · Claude 4.5 · paper-level slot 3 · secondary	Yes. AI Reviewer 3 gave reasonable comments on the simulation setting, theoretical assumption, and uncertainty in real data analysis that other reviewers did not catch.
P85 · GPT-5.2 · paper-level slot 1 · primary	Library preparation batch effects
P85 · Gemini 3.0 · paper-level slot 3 · primary	AI Reviewer 3 was very focused on data analysis (computational) procedures and almost all comments were distinct from the other reviewers (human and AI).

S2: Code reading — n = 28

1208

Citation	Expert comment
P1 · GPT-5.2 · item 4 · primary	I find it quite impressive that this reviewer actually looks at the code to understand whether some parts of the paper are not properly discussed. I am personally not doing it on a regular basis as it would consume a substantial amount of time.
P6 · Gemini 3.0 · item 4 · primary	Impressive that the AI looked at the code and realized there is nothing about the MRI reconstruction. It missed that some of the links are broken even for the ANT-R task.
P35 · Gemini 3.0 · item 1 · primary	Code verification fully confirms the data leakage claim. The DPI link trainer (trainer_dpi.py) also has a bug using test metrics for model selection despite having a validation set.
P37 · Claude 4.5 · item 3 · secondary	I checked the validity of this item using Claude Code.
P37 · GPT-5.2 · item 3 · secondary	I checked the validity of this item using Claude Code.
P37 · GPT-5.2 · item 4 · secondary	I checked the validity of this item using Claude Code.
P37 · Gemini 3.0 · item 2 · secondary	I didn't manually go through the code to verify the claim. Instead, I asked Claude Opus 4.6 to verify the claim and it said the claim was correct.
P49 · GPT-5.2 · item 2 · secondary	Unlike human reviewers, this reviewer actually visited the GitHub open source code repository the authors listed in their papers, dissected and analyzed the Python scripts to identify problems in the analysis.
P50 · Gemini 3.0 · item 2 · primary	Human reviewer would not open up the raw code to find this out - if this code is correct, basically the core data reported in this paper is wrong. Authors in the code said "It appears that delay is needed in order not to clog the port" meaning it's not possible to wirelessly transmit high frequency information. If this is true, Figure 4, for example, would mean the authors measured in wired manner and falsely claimed that their system is wireless (which is one key aspect of this paper).
P67 · Claude 4.5 · item 1 · primary	Same as the AI reviewer-2, this reviewer also caught the same resolution issue. Maybe this is one of the a few advantages over the human reviewers. AI is good at reviewing the codes and finding bugs.
P67 · GPT-5.2 · item 2 · primary	I didn't examine their code, but it looks like it wasn't a fair comparison. I hope the authors later fixed this.
P74 · Gemini 3.0 · item 1 · primary	The AI reviewer seems to be correct here, although I did not go through the entirety of the code so I cannot be certain that this check was not done elsewhere.
P1 · GPT-5.2 · paper-level slot 3 · primary	I think this reviewer also more clearly identified the systematic issue in the evaluation of the finite-size correction. I was also impressed by how this reviewer examined the code, identifying inconsistencies in the mathematical formulations.
P15 · GPT-5.2 · paper-level slot 1 · secondary	[non-Latin][non-Latin] [non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin]. [non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin][non-Latin][non-Latin][non-Latin].

Citation	Expert comment
P23 · Claude 4.5 · paper-level slot 3 · primary	Good job in checking the code and finding the discrepancies.
P23 · GPT-5.2 · paper-level slot 2 · primary	Good job in checking the code and finding the discrepancies.
P24 · Gemini 3.0 · paper-level slot 2 · primary	nice checking the code and catching the normative atlas issue missed by the humans.
P25 · GPT-5.2 · paper-level slot 3 · primary	Thoroughly checking the code and whether it can reproduce the results.
P26 · GPT-5.2 · paper-level slot 2 · primary	good job on noticing and calling up the circularity in the work. Alsoo very useful for checking the code.
P27 · Claude 4.5 · paper-level slot 2 · primary	human reviewers, especially 1, were very good. AI only exceeded them in checking the code.
P27 · GPT-5.2 · paper-level slot 1 · primary	human reviewers, especially 1, were very good. AI only exceeded them in checking the code.
P27 · Gemini 3.0 · paper-level slot 3 · primary	human reviewers, especially 1, were very good. AI only exceeded them in checking the code.
P35 · Gemini 3.0 · paper-level slot 1 · primary	The graph structure data leakage mechanism and the swapped EM step naming in code were unique. The co-training critique overlaps with others but adds unique code evidence.
P49 · GPT-5.2 · paper-level slot 2 · secondary	AI Reviewer 2 went beyond merely reading the text of the manuscript and providing logical critiques. It delved directly into the actual Python code scripts, data arrays, and references of the original datasets provided by the authors, accurately pinpointing critical discrepancies between what was claimed in the paper (methods) and what was actually executed (code or data). - False description of the cross-validation code: The manuscript claims to have performed 1,000 random iterations for distance-dependent cross-validation, but AI Reviewer 2 analyzed the provided Python code and revealed that it actually executed only 68 deterministic splits corresponding to the 68 brain regions. - Mismatch in the number of predictors: The methodology explicitly states that nine global connectome predictors were used, but AI Reviewer 2 found that only seven predictors were actually enumerated in the text and implemented in the code. - Age range error in ENIGMA data: Contrary to the authors' assertion that all cortical abnormality maps were collected from "adult patients," the cited ENIGMA ASD dataset actually comprises participants aged 2 to 64 years, highlighting that unaddressed developmental lifespan effects could heavily distort the results.
P50 · Gemini 3.0 · paper-level slot 2 · primary	Yes, AI Reviewer 2 found a problem that can be raised from raw code - which if true would flip all the claims this paper made on their capability of wireless system.
P60 · Gemini 3.0 · paper-level slot 1 · secondary	This reviewer went directly into the code to uncover a concrete computational bias that no other reviewer could have found from reading the manuscript alone. It also identified opposing directions of bias, harvesting potentially overestimating temperature deaths while pandemic context underestimates them. It questioned that compound term was used only in wordings but not analytically.
P67 · Gemini 3.0 · paper-level slot 1 · secondary	I find the review of AI reviewer 1 to be the best among all the human and AI reviews. One thing I would like to mention when comparing to human reviews is that this AI looked thoroughly into the code to find inconsistencies.
P74 · Gemini 3.0 · paper-level slot 3 · primary	AI Reviewer 3 dug into the code and noticed the possibility that "hidden split reads" are not actually being detected in the same way as described in the paper.

S3: Specialized niche field catch — n = 27

Citation	Expert comment
P61 · GPT-5.2 · item 1 · secondary	This critique substantially improves upon Human Reviewer 1's Item 20 (which asked about the formula's basis). If nitrate data were unavailable from the CMIP6 models used, the authors should have stated this as a limitation and discussed its implications for their estimates.
P64 · GPT-5.2 · item 4 · primary	this is a good catch. retrieving the complex field (with the imaginary part) is very ill-posed and to the best of my knowledge there is no method to do that well with multimode fiber propagation. arbitrary optical fields is for sure not within the scope of the paper and the method has nothing to do with the complex field. the authors state the relation as amplitude-to-amplitude, so this part is accurate but for sure not enough to extrapolate this huge leap.
P66 · GPT-5.2 · item 3 · primary	It is true that the bright-versus-dark exciton interpretation is not adequately supported and we do not know the exact illumination angle. But the authors claim a bright exciton dominance, so it should be close to a normal excitation. We should remember that the rapid spatial variation of the field at the tip apex is also required for dark excitons. If they were moving their tip slowly, then why would they have a serious dark excitons??
P66 · Gemini 3.0 · item 1 · primary	I agree with the AI reviewer that the PDM with their implementation is not the most accurate (they should have used the layered medium Green's functions and take care of the inhomogeneous background). However, this doesn't mean that the main findings of these researchers were wrong. A more accurate numerical model would generate results with smaller error limits, that's it.

Citation	Expert comment
P73 · GPT-5.2 · item 3 · primary	I don't quite follow the argument about the GWAS results here, but the broader point about homopolymer errors is valid.
P77 · GPT-5.2 · item 2 · primary	Since the substrate used for the synthesis of compound 4v contains stereogenic center, the issue of diastereoselectivity is important. Presumably, mixture of diastereomers were formed in compound 4v. It was not mentioned by the human reviewers.
P77 · Gemini 3.0 · item 2 · secondary	Although the formation of a seven-membered ring via an 8π electrocyclicization is relatively rare, such transformations have been reported. Nevertheless, I agree with the reviewer that the manuscript would be more impactful if additional experimental or computational evidence were provided to support the proposed mechanism.
P1 · Claude 4.5 · paper-level slot 2 · primary	I think this reviewer more clearly identified the systematic issue in the evaluation of the finite-size correction.
P2 · Claude 4.5 · paper-level slot 1 · primary	I found AI reviewer quite good. This identified important issues such as the lack of statistical sampling or the use of the minimal basis set, which may weaken the presented results. Also pointing out to a rather unfair comparison to GAP potentials appears to be raised by this AI reviewer but not the human ones, which I found very good. Overall, I found every single point raised by this AI reviewer as important and not necessarily found them in the human reviewer.
P2 · GPT-5.2 · paper-level slot 2 · primary	This reviewer raises an issue concerning how close the provided results are really to the reference on the Zundel-cation benchmark. I found it quite important for the scope of this paper. And it appears not to be raised by the human reviewer. Computational cost is another point raised by this AI reviewer I found particularly important. Finally, this AI reviewer discusses issues with the code which I find very good, as a human reviewer I check the code only if I am very suspicious about the results and could have overlooked it.
P7 · Claude 4.5 · paper-level slot 2 · primary	Yes there was a mention how systematic uncertainties were reported in Table I that was not spotted by other reviewers.
P15 · Gemini 3.0 · paper-level slot 3 · secondary	[non-Latin][non-Latin] [non-Latin][non-Latin] [non-Latin][non-Latin](local sleep disruption)[non-Latin] circadian clock [non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin][non-Latin] [non-Latin][non-Latin]
P23 · Gemini 3.0 · paper-level slot 1 · primary	It did significantly better for the depth of the comments, regarding the other 2 AI reviewers, but also made good point for some of the items, that were missed by the human reviewers (frontal alpha in propofol).
P24 · Claude 4.5 · paper-level slot 1 · primary	nice catching the normative atlas issue missed by the humans.
P52 · Gemini 3.0 · paper-level slot 2 · primary	Yes. Raised issue of biosafety of gold dissolution.
P52 · Gemini 3.0 · paper-level slot 2 · secondary	dissolution of gold for drug release can be toxic - needs testing
P73 · Claude 4.5 · paper-level slot 2 · primary	AI Reviewer 2 was the only one to mention the potential confounding effect of population structure on the GWAS results.
P73 · GPT-5.2 · paper-level slot 3 · primary	AI Reviewer 3 was the only one to raise a concern about the authors' extrapolation to estimate genome-wide mutation rates per microsatellite, although I consider this to be a relatively minor concern.
P74 · GPT-5.2 · paper-level slot 2 · primary	AI Reviewer 2 noted that the non-human benchmark data were limited to a single SV caller and thus could have many errors.
P77 · Claude 4.5 · paper-level slot 2 · primary	Yes, AI reviewer 2 pointed out an important point regarding the difference of Ag catalysis and Cu catalysis.
P77 · GPT-5.2 · paper-level slot 3 · primary	AI reviewer 3 nicely brought up the issue of diastereoselectivity in the product formation which was not mentioned by the human reviewers.
P81 · Gemini 3.0 · paper-level slot 1 · primary	- Only reviewer to question the use of the term "LED," given that the device operates via gate oxide breakdown rather than conventional carrier injection. - Uniquely criticized the use of spatial intensity (mW/cm^2) as misleading, since the absolute EQE is extremely low compared to other Si emitters.
P81 · Gemini 3.0 · paper-level slot 1 · secondary	AI Reviewer 1 was the only reviewer to question about the name "LED". It focused on how the light emission mechanism is different from traditional LEDs while other reviewers simply accepted the term. Second, while other reviewers also raised EQE concerns, it uniquely criticized the author's choice of the benchmark itself pointing out that the spatial intensity argument is misleading since the EQE is very small compared to other works.
P83 · Gemini 3.0 · paper-level slot 1 · primary	- Only review to flag that 50 bending cycles at 5 mm radius is insufficient to justify the "ultraflexible".
P83 · Gemini 3.0 · paper-level slot 1 · secondary	AI reviewer 1 focused on the mechanical durability compared to the standards in the field.
P84 · GPT-5.2 · paper-level slot 2 · primary	- Only review to catch EM simulations used air/copper properties instead of tissue-equivalent properties. - Mentioned about insufficient strain sensor calibration based on fractional shortening metric.
P84 · GPT-5.2 · paper-level slot 2 · secondary	AI reviewer 2 points out the insufficient calibration on the strain sensors, which makes cardiac contractility not convincing.

S4: Internal consistency across sections — n = 15

Citation	Expert comment
P1 · Gemini 3.0 · item 2 · primary	The point is that the goal may be to get closer to this accuracy and not to outperform it but at a reduced computational cost. The real critique here is that the authors don't clearly demonstrate how more and whether their approach is more computationally efficient.
P37 · Claude 4.5 · item 1 · secondary	I wholeheartedly agree with this comment. This alone is enough to reject the paper, as the core claim doesn't match the empirical results.
P50 · Claude 4.5 · item 3 · primary	Mistakes in writing - they switched the numbers. Numbers are correctly shown in Fig. 4f.
P60 · Claude 4.5 · item 4 · secondary	This is a factual correction, as also noted by human reviewer 1 who caught a different labelling error in Item 4, this paper have slight proof reading issues.
P60 · GPT-5.2 · item 3 · secondary	AI Reviewer 1's Item 5 noted the comparison is logically trivial (adding COVID deaths guarantees an increase); this reviewer shows that even setting that aside, the comparison is technically flawed in its construction. The inconsistent event definitions across periods is a particularly important point that no other reviewer raised.
P61 · GPT-5.2 · item 5 · secondary	This is one of the most damaging critiques across all reviews because it challenges the paper's internal consistency rather than requiring external knowledge or judgment. A reader can verify every claim in this item solely from the manuscript.
P65 · GPT-5.2 · item 2 · primary	This was one of my concerns regarding the original manuscript. They design for the 8-12 um range, but they their laser covers the 8.23 - 10.93 [non-Latin]m range.
P37 · Claude 4.5 · paper-level slot 2 · secondary	AI Reviewer 2 was the best in terms of analyzing what the authors claimed, and if the experiment design/results support the claims.
P41 · Claude 4.5 · paper-level slot 3 · primary	AI Reviewer 3 caught the temporal mismatch issue.
P41 · Gemini 3.0 · paper-level slot 1 · primary	AI Reviewer 1 caught the inconsistent projection methodology.
P50 · Claude 4.5 · paper-level slot 3 · primary	Yes, AI Reviewer 3 found about mistakes in writing where authors reported numbers in wrong order and thus physiologically impossible number was reported.
P50 · Claude 4.5 · paper-level slot 3 · secondary	AI Reviewer 3 correctly pointed out a discrepancy in the reported heart rate value - which was physiologically impossible for a healthy adult.
P67 · GPT-5.2 · paper-level slot 2 · primary	Inconsistency in the number of epochs.
P83 · GPT-5.2 · paper-level slot 3 · primary	- Uniquely caught that the abstract claims "no passivation" while Figure 2f's caption explicitly states the ultraflexible device used "parylene encapsulation"
P83 · GPT-5.2 · paper-level slot 3 · secondary	AI reviewer 3 was the only reviewer to question about the encapsulation contradiction of "no passivation" and "parylene encapsuation".

S5: Reproducibility / dependency failures — n = 10

Citation	Expert comment
P54 · GPT-5.2 · item 4 · primary	significant point well supported that applies for reproducibility
P66 · Gemini 3.0 · item 2 · primary	They do not provide the exact measurement results in their code, but one could extract the average data from the provided figures.
P85 · GPT-5.2 · item 5 · primary	It is likely that upon further review the journal would require the deposition of these materials.
P13 · Claude 4.5 · paper-level slot 2 · primary	Commented on the code availability; demographic mismatch
P13 · GPT-5.2 · paper-level slot 3 · primary	Commented on the code availability; demographic mismatch
P13 · Gemini 3.0 · paper-level slot 1 · primary	Commented on the code availability
P42 · GPT-5.2 · paper-level slot 3 · primary	AI reviewer 3 mostly found reproducibility issues, which while valid, likely just require tidying and are highly unlikely to reflect scientific validity.
P47 · GPT-5.2 · paper-level slot 2 · primary	AI #2 identified the reproducibility failures in the experimental pipeline such as missing parameter specifications and modified external tools.
P63 · GPT-5.2 · paper-level slot 3 · primary	Reproducibility item has not been raised by other AI and human reviewers.
P68 · GPT-5.2 · paper-level slot 3 · primary	The code reproducibility issue was raised by AI reviewer 3 and not by others.

S6: Big-picture / counter-narrative synthesis — n = 7

Citation	Expert comment
P60 · Gemini 3.0 · item 4 · secondary	This is one of the most insightful critiques across all reviews because it runs counter to the paper's narrative rather than supporting the claim that compound impacts are underestimated, it suggests temperature-related mortality during the pandemic may actually be overestimated due to a less vulnerable surviving population.
P85 · GPT-5.2 · item 2 · primary	Another way of phrasing this is that the authors over-interpret the connection between chromosome organization and metabolic state.
P6 · Gemini 3.0 · paper-level slot 2 · primary	No, but this is primarily because it did not get bogged down in details. It was very good at seeing the big picture and pinpointing the biggest problems.
P26 · Claude 4.5 · paper-level slot 1 · primary	this provided the most nuance in the comments, though all 3 AI did good job on noticing and calling up the circularity in the work, which was missed even by the first reviewer which was quite good.
P26 · Gemini 3.0 · paper-level slot 3 · primary	good job on noticing and calling up the circularity in the work
P37 · Gemini 3.0 · paper-level slot 1 · secondary	AI Reviewer 1 was the only reviewer who correctly noticed: 1. The training objective was not as novel as the authors claimed. 2. Insufficient comparison with state-of-the-art baselines. While other reviewers also expressed concerns about using connectome and molecular maps derived from healthy adults to explain the damage patterns of brain disorders, this AI reviewer 3 went a step further, demonstrating a deeper understanding of disorder-specific pathology. The logic is highly compelling: because neurodevelopmental disorders such as ASD and ADHD emerge at an early stage of life, unlike other brain conditions, their brain architecture fundamentally deviates even more from the normative connectomes and molecular maps of a healthy adult. Therefore, since these developmental disorders are the most likely to deviate from the normative baseline, it would significantly strengthen the study if the examples could obtain patient-specific data at least for these specific conditions to validate whether their observations still hold true.
P49 · Gemini 3.0 · paper-level slot 3 · secondary	

S-generic: Residual: AI judged Correct without specific reason — n = 40

1213

Citation	Expert comment
P1 · Gemini 3.0 · item 3 · primary	I would also willing to see how their method performs with a larger cell.
P6 · Claude 4.5 · item 3 · primary	This is well-argued and correct!
P6 · Gemini 3.0 · item 2 · primary	Good point and well-argued!
P7 · Gemini 3.0 · item 5 · primary	Yes this was mentioned previously as well, the paper should be rephrased a bit on this point, however the result is still valuable
P9 · GPT-5.2 · item 3 · primary	What is being measured in Lb -> RR is not enough for an observation therefore the text is fine.
P24 · GPT-5.2 · item 1 · primary	it is worth discussing it, I wouldn't say it is correct or incorrect.
P25 · Gemini 3.0 · item 2 · primary	More references are needed. Roberts et al Nat Commun 2019 for example.
P29 · Claude 4.5 · item 3 · primary	This is a good point.
P29 · Claude 4.5 · item 5 · primary	Very good point. very well stated.
P29 · Gemini 3.0 · item 4 · primary	This is quite good critique.
P31 · Gemini 3.0 · item 2 · primary	Great point.
P31 · Gemini 3.0 · item 4 · primary	Great point.
P37 · Gemini 3.0 · item 3 · secondary	I wrote the paper "Doctor AI", and it is true what the AI reviewer is claiming here.
P42 · Claude 4.5 · item 4 · primary	The AI is correct, while difficult to address methodologically, this is something that should be addressed in the discussion.
P49 · GPT-5.2 · item 1 · secondary	By examining the materials used by authors more thoroughly than human reviewers, this reviewer was able to point out very important points.
P53 · GPT-5.2 · item 1 · primary	very well supported observation
P54 · Claude 4.5 · item 4 · primary	it catches detailed text that the authors should consider rewriting for clarity.
P55 · Claude 4.5 · item 4 · primary	it is OK

Citation	Expert comment
P61 · Claude 4.5 · item 1 · secondary	This is a major issue which none of the Human reviewers pointed out but at the same time they all did emphasize more clarity on this figure especially.
P61 · Gemini 3.0 · item 3 · secondary	This connects to Human Reviewer 1 Item 13 about Wang et al.: multiple prior studies did multi-factor analysis, and the paper should engage with them honestly rather than dismissing the entire prior literature.
P62 · Claude 4.5 · item 1 · secondary	Addressing the same key point as Human reviewer #2 and #3, but the distinction between ex vivo and in vivo experiments were not detailed as in the comments from the Human reviewers.
P64 · Gemini 3.0 · item 2 · primary	Good catch as AI reviewer 2
P65 · Claude 4.5 · item 3 · primary	I agree. A comparison of numerical vs. experimental Strehl ratios would be useful.
P65 · Claude 4.5 · item 5 · primary	I am not sure how the authors could quantify the effect of LPA
P68 · Claude 4.5 · item 3 · primary	Agree and also raised by the human reviewer 3
P68 · Claude 4.5 · item 4 · primary	This is an important point, especially when considering sexual dimorphism, geographic variation, etc.
P68 · Gemini 3.0 · item 4 · primary	Some of the circularity is broken by the fact that analogs are defined based on multiple characteristics (not only inner ear). But the criticism is still relevant.
P73 · Claude 4.5 · item 4 · primary	Again, I don't think that mutations in this gene must necessarily increase cancer risk to be candidates for slight increases in microsat mutation rates, but this would be worth discussing as it does weaken the case for this mutation somewhat.
P74 · Claude 4.5 · item 4 · primary	I agree that this would have been useful to include.
P77 · Claude 4.5 · item 5 · primary	This is a good point that was missed by the human reviewers.
P77 · GPT-5.2 · item 1 · primary	The proposed 8-pi electrocyclization is assumed.
P77 · GPT-5.2 · item 2 · secondary	The term "stereo-controlled" appears to be misleading.
P85 · GPT-5.2 · item 1 · primary	While there are other samples C4 and C7 that are late time points in which data were generated with the same library kit, this is a valid concern.
P85 · GPT-5.2 · item 4 · primary	I am not an expert on the analysis of these data, but this critique seems valid to me.
P85 · Gemini 3.0 · item 1 · primary	I am not an expert in this analysis procedure, but this critique seems reasonable.
P85 · Gemini 3.0 · item 2 · primary	I am not an expert in this analysis procedure, but this critique seems reasonable.
P85 · Gemini 3.0 · item 3 · primary	I am not an expert in this analysis procedure, but this critique seems reasonable.
P5 · Claude 4.5 · paper-level slot 3 · primary	Similar to my comments regarding the second AI reviewer, though the third AI reviewer provides stronger arguments.
P35 · GPT-5.2 · paper-level slot 2 · secondary	Yes, All the Items.
P67 · GPT-5.2 · paper-level slot 2 · secondary	I would say that the review of this AI is better than all the human reviews but not as good as the other AI reviews. Overall I find that the AI reviews are very good at judging the technical matters of the paper and human reviews have points that can only come from years of experience although not as thorough as the AI reviews.

E Tools for improving and using AI reviewers

E.1 PEERREVIEW BENCH: Can we automatically evaluate AI reviewers?

Motivation The analyses in the preceding sections rest on a 469-hour expert annotation study, which is costly to repeat for every new AI reviewer model. We therefore investigate whether we can automatically evaluate the quality of an AI review without relying on a human expert. This enables evaluating future generation of AI reviewers to keep track of progress.

Evaluation protocol. For each paper p , let $R(p) = \{r_1, \dots, r_{n_p}\}$ denote the *rubric*, the set of fully-positive human review items associated with p , and let $G(p) = \{g_1, \dots, g_{m_p}\}$ denote the items produced by a candidate AI reviewer on p . “Fully positive” is defined exactly as in our expert-annotation study (§ 2): both domain-scientist annotators rated the item correct, significant, and evidence-sufficient. We compute three per-paper metrics:

$$\text{Recall}(p) = \frac{|\{r \in R(p) : \exists g \in G(p), \text{match}(g, r)\}|}{|R(p)|}, \quad (2)$$

$$\text{Precision}(p) = \frac{|\{g \in G(p) : \text{MetaReviewer}(g, p) = \text{fully positive}\}|}{|G(p)|}, \quad (3)$$

$$\text{F1}(p) = \frac{2 \cdot \text{Precision}(p) \cdot \text{Recall}(p)}{\text{Precision}(p) + \text{Recall}(p)}, \quad (4)$$

and report each metric averaged across the benchmark’s papers. Here, $\text{match}(g, r)$ is the GPT-5.4 similarity judge of § 3 at the “at least same target and criticism” threshold, and $\text{MetaReviewer}(g, p)$ is the AI meta-reviewer described in § E.3, run on p with g supplied as the candidate item. The candidate AI reviewer runs under the same constraints as the AI reviewers in our expert-annotation study (up to five review items per paper by default; full access to the preprint, supplementary materials, and source code). Per-paper averaging treats each paper as one independent observation and avoids letting papers with larger rubrics dominate the score.

Construction. PEERREVIEW BENCH comprises 78 of the 82 papers in our expert-annotation study. The remaining four are excluded because $R(p) = \emptyset$: their three human reviewers jointly produced zero items that both annotators labeled correct, significant, and evidence-sufficient, leaving no fully-positive content from which to construct a rubric. For each included paper, the rubric $R(p)$ is the union of fully-positive human review items pooled across the paper’s three human reviewers. Rubric sizes $|R(p)|$ span 1 to 36 items per paper, with a mean of 10 and a median of 8, reflecting substantial variation in how much high-quality content human reviewers produced. We deliberately do not deduplicate overlapping rubric items across a paper’s three human reviewers: the analysis in § 3 established that overlap between independent human reviewers is low, and when two or three human reviewers do converge on the same concern, the convergence is itself a signal that the concern is important. Leaving duplicates in means that an AI reviewer covering a repeatedly-raised concern receives correspondingly more credit.

Baselines In addition to the three models we asked human experts to evaluate in § 2, we further add nine additional models which uses the same framework to produce the reviews. These models include Claude-Opus-4.7, Claude-Sonnet-4.6, GPT-5.4, GPT-5.4-mini, Gemini-3.1-Pro-Preview, Gemini-3-Flash, DeepSeek-V4-Pro, Kimi-K2.6, and Qwen3.6-Plus. In particular, Claude-Opus-4.7, GPT-5.4, and Gemini-3.1-Pro-Preview are models that were released after Claude-Opus-4.5, GPT-5.2, and Gemini-3.0-Pro (the models we evaluated in § 2).

Main results. Table 44 in the main text reports recall, precision, and F1 across twelve backbone models. Claude-Opus-4.5 leads at F1 = 50.89, with Claude-Opus-4.7 a fraction of a point behind at F1 = 50.46; the two Opus models are essentially tied within the resolution of the benchmark. A second tier follows at F1 in the high-40s: DeepSeek-V4-Pro (F1 48.52), GPT-5.2 (F1 47.37). The remaining nine models cluster between F1 35 and 45, with Gemini-3.1-Pro-Preview an outlier at F1 22.57, driven primarily by a collapse in recall to 13.92%.

Different model families specialize in different aspects of AI reviewing. The Claude-Opus models occupy a balanced operating point on the precision-recall plane (precision 71-76%, recall 38-39%) and as a result lead in F1 despite not maximizing either axis. The GPT-5 family (excluding the mini variant) sits at the high-precision, low-recall corner: GPT-5.4 attains the benchmark’s highest precision at 93.81% while raising only 26.55% of the human rubric; GPT-5.2 follows the same shape with 88.92% precision and 32.28% recall. The Gemini-3.0 family and several other models (Qwen3.6-Plus, Gemini-3-Flash) sit at the opposite corner with lower precision (53-60%) and similar or higher recall (32-38%). DeepSeek-V4-Pro and Kimi-K2.6 sit closer to the Claude-Opus balance. No single backbone dominates on both axes, and the choice of backbone is therefore a choice about which kind of error to tolerate: GPT-family models produce few items but most are well-formed, Gemini-family and similar models cast a wider net at lower per-item quality, and the Claude-Opus models split the difference.

AI Reviewer backbone LLM	PeerReview Bench			
	Precision	Recall	F1 score	# Review items
Claude-Opus-4.7	71.47	39.00	<u>50.46</u>	4.73
Claude-Opus-4.5	75.49	<u>38.39</u>	50.89	4.38
Claude-Sonnet-4.6	82.03	29.74	43.65	4.42
GPT-5.4	93.81	26.55	41.38	3.60
GPT-5.2	<u>88.92</u>	32.28	47.37	4.12
GPT-5.4-mini	86.59	22.55	35.78	3.04
Gemini-3.1-Pro-Preview	59.55	13.92	22.57	3.31
Gemini-3.0-Pro-Preview	53.35	37.65	44.14	4.23
Gemini-3-Flash	60.46	31.99	41.84	4.86
DeepSeek-V4-Pro	76.75	35.47	48.52	4.59
Kimi-K2.6	74.38	28.76	41.48	4.15
Qwen3.6-Plus	56.67	36.27	44.23	4.62

Table 44: **PEERREVIEW BENCH results.** Recall is the fraction of correct, significant, and well-evidence human review items matched by review items from the AI reviewer; precision is the fraction of generated review items that are correct, significant, and well-evidence; F1 is the harmonic mean. All metrics are computed at the paper level and averaged across 78 papers. Best scores are **bolded**, second best are underlined.

Newer models are not consistently better. A pattern that runs against the implicit assumption of the benchmark is visible across all three families. Within Anthropic, Claude-Opus-4.5 narrowly outperforms the newer Claude-Opus-4.7 in F1 (50.89 vs 50.46, a 0.43-point gap). Within OpenAI, GPT-5.2 outperforms GPT-5.4 by nearly six points (47.37 vs 41.38), with the newer model trading recall (32.28 to 26.55) for precision (88.92 to 93.81). Within Google, Gemini-3.0-Pro-Preview outperforms Gemini-3.1-Pro-Preview by more than 21 points (44.14 vs 22.57), driven almost entirely by a collapse in recall (37.65% to 13.92%). For model developers who want their models to be useful as AI reviewers for scientific peer review, PEERREVIEW BENCH provides a concrete testbed on which targeted improvements can be measured.

E.2 CMU PAPER REVIEWER: An open-source reviewer platform for pre-submission feedback

We release the CMU PAPER REVIEWER⁵, an open-source platform for authors, students, and researchers who want detailed feedback on a manuscript before submission, built on the pipeline employed in our expert annotation study.

Implementation. The platform uses the OpenHands framework [Wang et al., 2025] and the same configuration of tools (terminal, file editor, Tavily web search, Mistral OCR) used in our expert-annotation study, so its behavior matches the AI reviewer that produced the items analyzed in § 2.

Mitigations for AI reviewer weakness categories. Building on the qualitative analysis in § 4, the platform implements concrete mitigations for the most consequential weakness categories observed in our study.

- *To address vague, verbose, or non-actionable critiques*, every review item is paired with a concrete action: a proposed manuscript edit (which paragraph to rewrite or where to insert new content) or, when the author submits source code, a generated patch with a runnable bash script and README that adds a missing experiment or fixes a code-manuscript mismatch.
- *To address over-harsh, out-of-scope, or unrealistic demands*, the platform offers two features. First, when the author opts in, the platform grounds review items in concerns the manuscript itself acknowledges as limitations, and annotates each item with whether it overlaps a stated limitation. Second, authors can challenge any review item through an interactive debate: a pop-up opens in which the AI defends valid concerns but acknowledges items as misplaced when the author’s counter-argument is sound, so authors gain context rather than receiving a single static round of feedback.
- *To address citing evidence that appeared after the preprint*, the platform allows authors to set a publication date (or auto-detects it from the web), annotates each cited reference with whether it predates or postdates the manuscript, and optionally forbids post-preprint citations entirely.

Comparison with other AI reviewer platforms We compare the CMU PAPER REVIEWER against two free public services on the 78 papers from PEERREVIEW BENCH using the same precision, recall, and F1 metrics: the Stanford Agentic Reviewer⁶, where each bullet point under “Weaknesses” is treated as a single review item, and

⁵<https://prometheus-eval.github.io/cmu-paper-reviewer/>

⁶<https://paperreview.ai/>

AI Reviewer platforms	LLM Backbone	# Max review items	PeerReview Bench			
			Precision	Recall	F1	# Review items
Stanford Agentic Reviewer	N/A	✗	59.84	45.43	<u>51.65</u>	11.08
OpenAIRreview	Claude-Opus-4.7	✗	57.57	40.98	47.88	18.64
CMU PAPER REVIEWER (OURS)	Claude-Opus-4.7	5	71.47	39.00	50.46	4.73
	GPT-5.4	5	<u>93.81</u>	26.55	41.38	3.60
	GPT-5.4	15	95.46	<u>42.32</u>	58.64	7.35

Table 45: **Comparison of publicly available AI reviewer platforms on PeerReview Bench (78 papers).** For each platform, we evaluate the review items it produces on **PeerReview Bench**, and report Precision, Recall, and F1 score. For Stanford Agentic Reviewer, each bullet point under the “Weaknesses” section is treated as a single review item; for OpenAIRreview, each feedback card is treated as a single review item.

OpenAIRreview⁷, where each feedback card is treated as a single review item. Table 45 reports the results.

The CMU PAPER REVIEWER dominates on F1 across configurations. The GPT-5.4 backbone with a 15-item cap reaches the highest F1 on the benchmark at 58.64, against 51.65 for the Stanford Agentic Reviewer and 47.88 for OpenAIRreview. The same backbone with a 5-item cap leads on precision (93.81%) at the cost of recall (26.55%), reflecting the high-precision-low-recall character of the GPT-5 family observed across Table 44. The Claude-Opus-4.7 backbone with 5 items reaches an F1 of 50.46 with a more balanced operating point (precision 71.47%, recall 39.00%) and a much smaller absolute item count than the competing platforms (4.73 items per paper, against 11.08 and 18.64).

Raising the item cap improves coverage without sacrificing precision. A non-trivial observation is that tripling the per-paper item cap (from 5 to 15) for the GPT-5.4 backbone does not triple the number of items the agent emits, but only roughly doubles it (3.60 to 7.35 items per paper on average). The agent’s internal selection prioritizes fully-positive items rather than filling the budget, so additional items only appear when the agent is confident they meet the three-axis bar; precision is actually higher with the larger cap (95.46% vs 93.81%), and recall climbs sharply (26.55% to 42.32%). This is the only configuration in our comparison that simultaneously achieves the benchmark’s highest precision and approaches the recall of the Stanford Agentic Reviewer. We leave to future work the question of the optimal item-generation strategy for AI reviewers: whether to prioritise quality at the cost of coverage (as our 5-item cap does), to expand coverage without sacrificing precision (as the 15-item cap appears to do), or to seek a different operating point altogether may depend on the use case and on the reviewer’s intended audience.

Availability and intended use The platform is publicly accessible at cmu-paper-reviewer.com; users can generate up to three reviews per day for free, or register their own API key for unlimited use. The implementation code is released as open source so any institution can deploy its own instance. We emphasize that the CMU PAPER REVIEWER is intended as an assistive tool for pre-submission feedback, and **should not be used at conferences or journals that prohibit AI reviewers in their official review process**; the platform is not a substitute for a sanctioned human peer-review workflow. Conference and journal organizers interested in officially incorporating the CMU PAPER REVIEWER into their review workflow are welcome to contact the authors.

⁷<https://openaireview.org/>

E.3 AI Meta-reviewers: Can AI agents review the reviews from AI reviewers?

Motivation. In § E.1, we investigate whether AI agents can serve as meta-reviewers, performing the same three-axis judgment of correctness, significance, and evidence sufficiency that domain scientists did in our study. To understand if this precision meta-judgment is accurate, we use our expert annotations as a calibration set (which we refer to as ‘meta-review calibration set’).

Construction. The meta-review calibration set is built from the 27 papers in our expert-annotation study for which two domain scientists independently annotated every review item. This yields 908 dual-annotated review items (568 human, 340 AI) on which both annotators provided three-axis judgments (correctness, significance, evidence sufficiency) under the cascade protocol described in § 6. The meta-review calibration set supports two settings.

The *primary setting* is per-axis judgment, in which the AI meta-reviewer produces its own correctness, significance, and evidence labels for each item, exactly as the human meta-reviewers in our study did, and is scored against a single expert annotator. The *secondary setting* additionally asks the AI meta-reviewer to predict how the two expert annotators would jointly judge the item, which we encode as a ten-class label that captures both the cascade outcome and inter-annotator agreement: classes 1–7 cover the cases where the two annotators agree on correctness and significance and either agree or disagree on evidence; class 8 covers correctness agreement with significance disagreement; class 9 is agreed-on-incorrect; class 10 is correctness disagreement. The ten-class distribution is shown in Table 46, and the top-three classes are “both correct, both significant, but disagree on significance” (32.8%), “both correct, both significant, evidence sufficient” (30.5%), and “disagree on correctness” (14.2%).

Evaluation protocol. The meta-review calibration set evaluates an AI meta-reviewer on two paired tasks. In the *primary setting*, the AI meta-reviewer produces its own three-axis judgment (correctness, significance, evidence sufficiency) for each item, exactly as the human meta-reviewers in our study did, and is scored against a single expert annotator using percent agreement as the main metric. Per-axis agreement is reported for correctness (computed on the 779 items where both annotators agree on correctness), significance (445 items where both annotators agree through significance), and evidence (357 items where both agree through evidence sufficiency). To assess whether AI–human agreement matches the empirical ceiling set by inter-human agreement, we additionally report Gwet’s AC1 for the Human–Human, Human–AI, and AI–AI pairwise comparisons; AC1 is robust to the prevalence-induced kappa paradox that affects Cohen’s κ on these heavily skewed label distributions. In the *secondary setting*, the AI meta-reviewer additionally predicts how the two expert annotators would jointly judge the item, expressed as one of the ten classes defined above; we report ten-class accuracy on all 908 items.

AI meta-reviewer setup. A single review item often anchors its claim in a specific figure, source-code file, or supplementary section, so the meta-reviewer needs the same access to the paper that a human meta-reviewer (the expert scientist in our annotation study) would. We implement the meta-reviewer as an agent on the OpenHands framework with terminal, file-editor, and web-search tools that let it open the preprint markdown, view referenced figures, read submitted source code, and look up specialist terminology. The AI meta-reviewer is blocked from accessing the paper’s published version (via a domain blacklist) so it cannot retrieve actual referee reports. We evaluate three frontier models as the AI meta-reviewer backbone: Claude-Opus-4.7, GPT-5.4, and Gemini-3.1-Pro. The full prompt, including the axis definitions and the ten-class decision procedure, is in § E.4.

AI meta-reviewers closely approach the human inter-annotator agreement on all three axes. Table 47 reports the per-axis accuracy of the three meta-reviewers in the primary setting, scored against the primary annotator. Claude-Opus-4.7 reaches 87.9% correctness, slightly above the 85.8% primary–secondary baseline; GPT-5.4 and

Statistic	Value
Papers	27
Total review items	908
Human items	568
AI items	340
<i>Primary setting</i> ground truth labels (cascade applied)	
(Axis 1) Correct	743
(Axis 1) Not Correct	36
(Axis 2) Significant	299
(Axis 2) Marginally Significant	91
(Axis 2) Not Significant	55
(Axis 3) Evidence Sufficient	351
(Axis 3) Evidence Not Sufficient	6
<i>Secondary setting</i> ground truth labels	
1. Correct + Sig. + Evi. Sufficient	277 (30.5%)
2. Correct + Sig. + Evi. Not Suff.	2 (0.2%)
3. Correct + Sig. + Evi. Disagree	20 (2.2%)
4. Correct + Marg. Sig. + Evi. Sufficient	74 (8.1%)
5. Correct + Marg. Sig. + Evi. Not Suff.	4 (0.4%)
6. Correct + Marg. Sig. + Evi. Disagree	13 (1.4%)
7. Correct + Not Significant	55 (6.1%)
8. Correct + Sig. Disagree	298 (32.8%)
9. Not Correct	36 (4.0%)
10. Disagree on Correctness	129 (14.2%)

Table 46: **Meta-review calibration set statistics.** Each of 908 review items from 27 dual-annotated papers carries a 10-class ground truth label encoding both the cascade outcome (correctness → significance → evidence) and inter-annotator agreement. “Sig.” = both significant; “marg.” = both marginally significant.

Meta-Reviewer	ALL REVIEWS			HUMAN REVIEWS			AI REVIEWS			SELF REVIEW		
	Corr.	Sig.	Evid.	Corr.	Sig.	Evid.	Corr.	Sig.	Evid.	Corr.	Sig.	Evid.
RANDOM BASELINE	50.0%	33.3%	50.0%	50.0%	33.3%	50.0%	50.0%	33.3%	50.0%	50.0%	33.3%	50.0%
HUMAN (SECONDARY ANNOTATOR)	85.8%	59.9%	88.0%	84.7%	60.2%	83.9%	87.6%	59.3%	94.3%	—	—	—
CLAUDE-OPUS-4.7	87.9%	56.7%	85.6%	87.7%	55.8%	79.4%	88.2%	58.2%	94.9%	85.7%	55.4%	97.6%
GPT-5.4	82.0%	56.9%	85.3%	80.0%	55.3%	80.7%	85.3%	59.4%	91.8%	90.0%	52.5%	93.1%
GEMINI-3.1-PRO	81.5%	54.0%	87.4%	79.4%	51.8%	82.6%	84.7%	57.3%	94.2%	82.9%	65.5%	90.9%

Table 47: **Meta-review calibration set results: per-axis accuracy (primary setting).** Each row is a meta-reviewer model; columns show per-axis accuracy across four review subsets: all items, human review items only, AI review items only, and “self review” where the meta-reviewer judges items from its own model family (e.g., GPT-5.4 judging GPT-5.2 reviews). Random baseline assumes uniform prediction over label categories (2 for correctness/evidence, 3 for significance). The human baseline treats the secondary expert annotator as the “meta-reviewer” and the primary as ground truth, representing the ceiling of human inter-annotator agreement.

Comparison	Correctness		Significance		Evidence		N
	Agr.	AC1	Agr.	AC1	Agr.	AC1	
<i>Human–Human</i>							
Primary vs Secondary	85.8%	0.82	59.9%	0.44	88.0%	0.86	908
<i>Human (Primary) vs AI Meta-Reviewer</i>							
vs Claude Opus 4.7	87.9%	0.86	56.7%	0.43	85.6%	0.83	908
vs GPT-5.4	82.0%	0.78	56.9%	0.44	85.3%	0.83	908
vs Gemini 3.1 Pro	81.5%	0.77	54.0%	0.44	87.4%	0.86	908
<i>Human (Secondary) vs AI Meta-Reviewer</i>							
vs Claude Opus 4.7	85.3%	0.82	62.3%	0.52	92.1%	0.91	908
vs GPT-5.4	82.4%	0.77	64.5%	0.55	91.1%	0.90	908
vs Gemini 3.1 Pro	82.9%	0.78	61.6%	0.54	93.3%	0.93	908
<i>AI vs AI</i>							
GPT-5.4 vs Gemini 3.1 Pro	88.8%	0.86	85.3%	0.83	91.2%	0.91	908
GPT-5.4 vs Claude Opus 4.7	91.0%	0.89	89.3%	0.88	90.9%	0.90	908
Gemini 3.1 Pro vs Claude Opus 4.7	92.0%	0.91	86.1%	0.84	93.3%	0.93	908

Table 48: **Pairwise agreement on the 908-item meta-review calibration set.** Each cell reports percent agreement and Gwet’s AC1 between two judges. Human–Human: primary vs secondary expert annotator. Human–AI: each AI meta-reviewer vs each human annotator. AI–AI: pairs of AI meta-reviewers. AC1 is robust to class imbalance and avoids the “kappa paradox” where high agreement produces low κ when one category dominates. AI–AI agreement substantially exceeds Human–AI agreement, especially on significance (AC1 = 0.83–0.88 AI–AI vs 0.43–0.55 Human–AI), indicating that AI meta-reviewers converge on a shared judgment style.

Gemini-3.1-Pro reach 82.0% and 81.5%, a few points below. On significance, the three AI meta-reviewers reach 56.7%, 56.9%, and 54.0%, against a baseline of 59.9%; on evidence, they reach 85.6%, 85.3%, and 87.4%, against a baseline of 88.0%. Notably, significance is the hardest axis for both humans and AI meta-reviewers, and the modest 59.9% primary–secondary baseline reflects this directly: domain scientists themselves often disagree on whether a given critique materially improves a paper, so the upper bound for any meta-reviewer on this axis is correspondingly limited. Against this backdrop, the AI meta-reviewers land in the same regime as the humans, which is the empirical basis for using an AI meta-reviewer as the precision judge in PEERREVIEW BENCH.

Self-review enhancement bias is uneven across axes. When the GPT-5.4 meta-reviewer judges items written by GPT-5.2 reviewers (its same-family “self review” setting), per-axis accuracy is 90.0% / 52.5% / 93.1%, compared to 85.3% / 59.4% / 91.8% on AI items overall: correctness and evidence rise, but significance falls. Claude-Opus-4.7 (85.7% / 55.4% / 97.6% vs. 88.2% / 58.2% / 94.9%) shows the opposite asymmetry, with evidence rising but correctness and significance falling. Gemini-3.1-Pro (82.9% / 65.5% / 90.9% vs. 84.7% / 57.3% / 94.2%) shows yet another pattern, with significance rising but correctness and evidence falling. The self-enhancement bias well-known for LLM judges [Wataoka et al., 2024] is therefore present in AI meta-reviewers, but does not act uniformly across axes; on every model, at least one axis goes down rather than up under same-family judgment. The implication for PEERREVIEW BENCH is that precision scores derived from a single meta-reviewer should be read with caution when the candidate AI reviewer is from the same family as the meta-reviewer.

AI meta-reviewers agree with each other more than with humans. For a finer-grained view of the agreement structure, Table 48 reports pairwise agreement among all meta-reviewer pairs (Human–Human, Human–AI, AI–AI). The three AI meta-reviewers converge on a shared judgment style that is markedly more consistent than any

individual AI’s agreement with a human. On significance, the three pairwise AI–AI comparisons reach AC1 between 0.83 and 0.88, while AI–human AC1 ranges from 0.43 to 0.55 – only modestly above the 0.44 human–human baseline. The same pattern holds on evidence (AI–AI 0.90–0.93 vs AI–human 0.83–0.93) and, less dramatically, on correctness (AI–AI 0.86–0.91 vs AI–human 0.77–0.86). The three AI meta-reviewers therefore assign similar significance and evidence labels to similar items even when those labels differ from what a human expert would assign, suggesting the AI meta-reviewers have settled on a shared standard that does not fully match the human one. We do not have direct evidence on whether this shared standard is closer to or further from a hypothetical “true” standard than the human one.

Predicting where experts will disagree is a much harder task.

Per-axis accuracy is a natural stress test on each individual judgment, but the ten-class label adds a stricter requirement: the meta-reviewer must predict not only what the cascade outcome is, but also where in the cascade two human experts would and would not converge. Table 49 shows this is a substantially harder task. Claude-Opus-4.7 reaches 44.3% ten-class accuracy across all items, well ahead of GPT-5.4 (30.1%) and Gemini-3.1-Pro (29.0%) and far above the 10% random baseline, but still well short of perfect prediction of the joint cascade-and-agreement structure. The gap between Claude-Opus-4.7 and the other two frontier models is roughly fourteen percentage points, considerably larger than the gap on per-axis accuracy. We do not include a human ten-class baseline because the human annotators were not asked to predict the other annotator’s labels; a “human ceiling” is therefore not directly defined for this metric.

Meta-Reviewer	ALL	HUMAN	AI	SELF
RANDOM BASELINE	10.0%	10.0%	10.0%	10.0%
CLAUDE-OPUS-4.7	44.3%	40.0%	51.2%	47.9%
GPT-5.4	30.1%	31.2%	28.2%	29.1%
GEMINI-3.1-PRO	29.0%	26.1%	33.8%	39.6%

Table 49: **Meta-review calibration set results: ten-class accuracy (secondary setting).** The 10-class label encodes both the annotation cascade outcome (correctness → significance → evidence) and inter-annotator agreement/disagreement. Columns show accuracy across four review subsets: all items, human-written items, AI-written items, and “self review” where the meta-reviewer judges items from its own model family.

E.4 AI meta-reviewer prompt

The prompt for the AI meta-reviewer agent is below across Figure 10, Figure 11, Figure 12, and Figure 13. The same prompt is used for all three frontier-model backbones (Claude-Opus-4.7, GPT-5.4, Gemini-3.1-Pro). Placeholders of the form {paper_preprint_dir}, {paper_reviews_dir}, {reviewer_list}, {paper_id}, and {output_file} are replaced per paper at runtime.

You are a meta-reviewer agent for scientific papers. You will be given a paper's source files on disk AND reconstructed reviews written by multiple reviewers (human and AI). For EVERY review item from EVERY reviewer, you must:

1. Judge the item along three axes: correctness, significance, and evidence sufficiency (your own meta-review judgment).
2. Predict how two independent expert meta-reviewers would jointly judge the item, expressed as one of 10 collapsed class labels that encode both the cascade outcome and inter-expert agreement.

You are NOT writing a new review. You are judging existing review items by verifying their claims against the paper.

Paper location

The paper's source files are at: {paper_preprint_dir}

The directory structure is:

```
preprint/  
| preprint.md          (main paper, markdown)  
| images_list.json     (list of figure images with captions)  
| images/              (figure files referenced by the paper)  
| supplementary/        (optional supplementary materials)  
| code/                (optional source code)
```

Reviews to meta-review

The reconstructed reviews are at: {paper_reviews_dir}

Each file is one reviewer's full review with all their items.

The reviewers are: {reviewer_list}

Principles (ordered by importance)

1. Correctness -- judge the main point, not peripheral details.
If the reviewer's core concern is valid even though one specific supporting claim is inaccurate, the item is still Correct. Only mark Not Correct when the main point itself is wrong.
2. Significance -- the bar is "would this improve the paper?"
Any criticism that would genuinely help the paper if addressed is Significant -- it does NOT need to threaten the paper's validity. Missing statistics, undefined figure annotations, unreported methodological details, internal inconsistencies between text and figures, and missing ablations are typically Significant. Only pure stylistic preferences and trivial typos are Marginally Significant.
3. Evidence -- verifiability, not exhaustiveness. If a meta-reviewer can verify the reviewer's claim from what the reviewer wrote plus the paper, the evidence is Sufficient. When the reviewer's criticism IS that something is absent, identifying the specific absence IS the evidence. Reserve Requires More for cases where the meta-reviewer cannot even locate what part of the paper the reviewer is talking about.

Figure 10: Meta-reviewer prompt (Part 1 of 4): role, paper context, and the three principles that establish the bar for each axis.

```

### Decision procedure (follow for EACH item)

Part A -- Your initial judgment (axis labels)
Step 1 (Understand). Read the item. What is the main point?
Step 2 (Correctness). Verify the MAIN POINT against the paper.
  - Core concern valid even if a peripheral detail is off?
    -> "Correct". Continue to Step 3.
  - Main point itself factually wrong? -> "Not Correct".
    Set significance and evidence to null.
Step 3 (Significance). Would addressing this criticism genuinely
  improve the paper?
  - Would fixing it make the paper substantively better?
    -> "Significant"
  - Minor presentation or stylistic issue that does not affect
    substance? -> "Marginally Significant"
  - Trivial enough that the review would be better without it?
    -> "Not Significant". Set evidence to null.
Step 4 (Evidence). Can a meta-reviewer verify the claim from the
  review item + the paper without substantial extra work?
  - Yes -> "Sufficient"
  - No -> "Requires More"

Part B -- Devil's advocate + expert disagreement prediction
Expert disagreement is common in meta-review -- it is not a rare
edge case. For each axis, you must argue the other side before
deciding whether experts would agree.

Step 5 (Correctness -- devil's advocate).
  - If "Not Correct" in Step 2: argue the reviewer's defense.
    Could a charitable expert still find the core concern valid?
    If plausible -> predict "disagree_on_correctness" (and you
    may revise Step 2 to "Correct" if the defense convinces you).
  - If "Correct" in Step 2: argue the prosecution. Could a
    skeptical expert find the claim factually wrong? If
    plausible -> predict "disagree_on_correctness".
  - If neither side is plausible -> experts agree on correctness.

Step 6 (Significance -- devil's advocate). Only if Step 5 = agree
  on Correct.
  - If "Significant": argue for downgrading. Could an expert see
    this as merely Marginally Significant?
  - If "Marginally Significant": argue for upgrading. Could an
    expert see this as genuinely improving the paper?
  - If counter-argument plausible -> predict
    "correct_disagree_on_significance".

IMPORTANT: Correctness disagreement means the factual truth of
the claim is ambiguous. Significance disagreement means the
fact is clear but its importance is debatable. Do not mislabel
correctness uncertainty as significance uncertainty.

Step 7 (Evidence boundary). Only if Step 6 = agree on >= Marginal.
  Is evidence clearly sufficient or clearly lacking, or
  borderline?
  - Firm -> experts agree. Pick the matching agreement class.
  - Borderline -> "correct_significant_disagree_on_evidence" or
    "correct_marginal_disagree_on_evidence".

```

Figure 11: Meta-reviewer prompt (Part 2 of 4): the per-item decision procedure, with Part A producing the agent's own three-axis judgment and Part B applying a devil's-advocate check that determines whether the agent should predict expert agreement or disagreement on each axis.

```

### Consistency constraint (critical)
Your prediction_of_expert_judgments MUST be consistent with your
FINAL axis labels (after any Part B revisions):
- If final correctness = "Correct", the prediction must be
  a "correct_*" class OR "disagree_on_correctness". Never
  "incorrect".
- If final correctness = "Not Correct", the prediction must be
  "incorrect" OR "disagree_on_correctness". Never a "correct_*"
  class.
- Similar logic applies downstream: significance and evidence
  labels constrain which agreement/disagreement classes are valid.

### TODO list for your workflow
- [ ] Read preprint.md thoroughly. Understand the paper's main
  contribution, methodology, and key claims.
- [ ] Skim images_list.json and open any figures/tables that the
  reviewers mention by name. Verify visual claims.
- [ ] Read the code/ directory if any reviewer references specific
  files or functions. Check the README first, then look at the
  referenced files. If the code is executable and not resource-
  prohibitive, try running it to verify code-related claims.
- [ ] For each reviewer's review file, read every item carefully.
- [ ] For each item, follow Part A then Part B. Reasoning must
  include: (1) axis judgment + why, (2) devil's advocate
  argument for correctness (and significance if applicable)
  and whether plausible, (3) the resulting prediction.
- [ ] (Very important) Review your judgments for consistency.
  Are you applying the same bar for "Significant" across all
  reviewers? Iterate if needed.
- [ ] Write the final prediction JSON to: {output_file}

### Output format (STRICT)
Write a single JSON file to {output_file} with this exact shape:

{
  "paper_id": {paper_id},
  "reviewers": [
    {
      "reviewer_id": "<reviewer name>",
      "items": [
        {
          "item_number": 1,
          "reasoning": "...",
          "correctness": "Correct",
          "significance": "Significant",
          "evidence": "Sufficient",
          "prediction_of_expert_judgments":
            "correct_significant_sufficient"
        }
      ]
    }
  ]
}

Rules:
- You MUST include ALL reviewers and ALL items -- no exceptions.
- Each item must have: item_number, reasoning, correctness,
  significance, evidence, prediction_of_expert_judgments.
- "reasoning" must include: (1) axis judgment + why, (2) devil's
  advocate argument and whether plausible, (3) expert prediction.
  Aim for 4-6 sentences.
- Use exact label strings: "Correct" / "Not Correct";
  "Significant" / "Marginally Significant" / "Not Significant";
  "Sufficient" / "Requires More". Use null for cascade-skipped
  fields.

```

Figure 12: Meta-reviewer prompt (Part 3 of 4): consistency constraint linking the predicted ten-class label to the agent's own per-axis labels, the workflow checklist, and the strict JSON output format.

CRITICAL: Verification before finishing
After writing the JSON file, you MUST perform these checks before stopping. Do NOT skip this step.

1. Read the file back and verify it is valid JSON (no syntax errors, no trailing commas, no truncated content).
2. Count reviewers: number of reviewer entries must match number of .md files in the reviews/ directory.
3. Count items per reviewer: number of item entries must match the number of "## Item" sections in that reviewer's .md file.
4. Check label strings: every correctness value must be exactly "Correct" or "Not Correct"; every prediction must be one of the 10 valid strings (listed below).
5. Check consistency: prediction must agree with axis labels.
6. If any check fails, fix the file and re-verify.

Only after all checks pass, print:
"Verification complete. All [N] reviewers and [M] total items included. Prediction written to {output_file}"
Then stop.

Filesystem boundaries

- READ from {paper_preprint_dir} and {paper_reviews_dir}. These are the paper's source files. Do not modify anything there.
- WRITE only to {output_file}. Do not create any other files.
- Do not navigate to parent or sibling directories of the paper.
- Do NOT access nature.com, researchsquare.com, springer.com, or springerlink.com -- these host the published versions of benchmark papers and may contain reviewer comments. Judge items solely from the paper files provided on disk.

The 10 valid prediction_of_expert_judgments labels

correct_significant_sufficient	(class 1)
correct_significant_requires_more	(class 2)
correct_significant_disagree_on_evidence	(class 3)
correct_marginal_sufficient	(class 4)
correct_marginal_requires_more	(class 5)
correct_marginal_disagree_on_evidence	(class 6)
correct_not_significant	(class 7)
correct_disagree_on_significance	(class 8)
incorrect	(class 9)
disagree_on_correctness	(class 10)

Tips

1. The paper's markdown may contain OCR errors. Do not penalize reviewers for pointing out things that are actually OCR artifacts; infer content from context.
2. Image links may be broken. Figures are stored at preprint/images/figure1.png, etc.; open images_list.json to see captions.
3. Do not try to read every file in code/ -- focus on the files that reviewers explicitly reference.
4. Apply the same significance bar consistently across all reviewers. Do not be lenient on one reviewer and strict on another.
5. Your judgment must be independent of who wrote the review. Do not infer reviewer identity (human/AI) from writing style.

Figure 13: Meta-reviewer prompt (Part 4 of 4): verification checklist before finishing, filesystem and access boundaries (including the domain blocklist that prevents the agent from retrieving the published version of the paper), the ten valid prediction labels, and additional tips for handling OCR artifacts and applying a uniform bar across reviewers.

F Recommendation for journal and conference organizers: Panel composition analysis

Motivation This appendix documents the panel composition simulation to indicate what journal and conference organizers should consider and prioritize when deploying AI reviewers. We describe the four panel configurations, the seven review-item metrics used to evaluate each panel, the main results, three configurations indexed to organizer priorities, caveats on generalization, and the 95% confidence intervals.

F.1 Methodology and metric definitions

Panel construction. We construct four panel compositions by varying the mix of human and AI reviewers: 3 humans (3H), 2 humans plus 1 AI (2H+1AI), 1 human plus 2 AIs (1H+2AI), and 3 AIs (3AI). The simulation runs over the 53 papers in our expert-annotation study with three human reviews and three AI reviews available. For mixed panels, we average over all possible compositions: 2H+1AI averages over all $\binom{3}{2} \times 3 = 9$ combinations of two humans and one of the three AI reviewers; 1H+2AI averages over all $3 \times \binom{3}{2} = 9$ combinations of one human and two of the three AIs. We additionally evaluate every panel composition with and without an AI meta-reviewer filter (GPT-5.4) that removes items the meta-reviewer judges not fully positive.

Metric definitions. Table 50 reports seven per-paper metrics for each panel composition, organized by the stakeholder concern each metric targets:

- **# Total Items:** how many review items the panel produces per paper; what the author receives for their submission.
- **# Unique Items:** items that have no similar counterpart from other reviewers in the panel (similarity ordinal < 2).
- **# Not Fully Pos. Items:** items that fail at least one of correctness, significance, or evidence sufficiency; this corresponds to the burden that authors or editors must triage to surface useful items.
- **% Fully Pos. + Unique / # Unique Items:** the fraction of distinctive contributions that are fully positive, i.e., how often a reviewer’s distinct item turns out to be useful for the authors.
- **# Fully Pos. + Unique Items:** items that are simultaneously fully positive (correct, significant, and evidence-sufficient by expert annotation) and unique within the panel. This is the bottom-line metric for what authors actually receive, since the remaining items are either redundant, incorrect, not significant, or insufficiently evidenced.
- **% Fully Pos. + Unique / # Total Items:** the fraction of all items shown that are useful to authors, i.e., the panel’s overall yield.
- **Noise per Gem:** # Not Fully Pos. / # Fully Pos. + Unique, the number of items the author reads to encounter one item containing useful feedback.

F.2 Panel composition results

Substituting a second human with another AI begins to cost useful content. Among unfiltered panels, 2H+1AI does not lose to 3 humans on any metric in Table 50: same fully-positive-and-unique count (3.9 vs 3.9), fewer total items (21.4 vs 25.8), fewer not-fully-positive items (11.5 vs 14.6), higher yield (18.2% vs 15.1%), higher quality of unique items (35.8% vs 33.9%), and lower noise per gem (2.95 vs 3.74). Replacing a second human with an AI reviewer (1H+2AI) gives a smaller, not-yet-significant drop in fully-positive-and-unique items (3.5 per paper, CI [2.9, 4.1]). Only at the all-AI panel does the count drop sharply and significantly to 1.8 (CI [1.4, 2.2], non-overlapping with all three other compositions), driven by a collapse in unique coverage: AI reviewers overlap with each other much more than human reviewers do, so a panel of three AIs produces only 3.1 distinct issues per paper, against 11.5 for three humans. Adding more than one AI to a panel therefore begins to erode the useful content authors receive, and replacing all three humans erodes it sharply.

The AI meta-reviewer filter raises every panel’s efficiency at the cost of absolute volume. Filtering each panel’s items through the AI meta-reviewer (§ E.3) raises the yield of every panel composition (e.g., 2H+1AI from 18.2% to 20.0%, 1H+2AI from 20.5% to 20.6%) and lowers noise per gem (e.g., 2H+1AI from 2.95 to 2.14). The trade-off is that the same filter also reduces the absolute count of items that are simultaneously fully positive and unique: from 3.9 to 1.9 for three humans (a 51% drop, CIs do not overlap), from 3.9 to 2.2 for 2H+1AI, from 3.5 to 2.1 for 1H+2AI, and from 1.8 to 1.2 for the all-AI panel. The meta-reviewer removes some genuinely useful items along with the not-fully-positive items it is designed to filter, a consequence of the meta-reviewer’s leniency on significance documented in ?? For deployments that prioritise editor or author triage time over absolute volume, the filter is a reasonable addition; for deployments that prioritise the volume of useful feedback delivered to authors, it loses more than it removes.

Panel Composition			Volume	Diversity	Quality		Diversity & Quality	Triage Efficiency	
# Human Reviewers	# AI Reviewers	AI Meta-Rev. Filtering	# Total Items	# Unique Items (↑)	# Not Fully Pos. Items (↓)	% Fully Pos. + Unique / # Unique Items (↑)	# Fully Pos. + Unique Items (↑)	% Fully Pos. + Unique / # Total Items (↑)	Noise per Gem (↓)
3	0	✗	25.8	11.5	14.6	33.9%	3.9	15.1%	3.74
2	1	✗	21.4	10.9	11.5	35.8%	3.9	18.2%	2.95
1	2	✗	17.1	8.5	8.4	41.2%	3.5	20.5%	2.40
0	3	✗	12.7	3.1	5.2	58.1%	1.8	14.2%	2.89
3	0	✓	11.8	4.0	5.3	47.5%	1.9	16.1%	2.79
2	1	✓	11.0	4.6	4.7	47.8%	2.2	20.0%	2.14
1	2	✓	10.2	4.1	4.1	51.2%	2.1	20.6%	1.95
0	3	✓	9.4	1.9	3.5	63.2%	1.2	12.8%	2.92

Table 50: **Reviewer panel composition analysis.** Mean per-paper values across 53 papers with complete data (3 human reviewers, 3 AI reviewers, and AI meta-reviewer predictions for every item). A *unique* item has no similar counterpart from any other reviewer in the panel (similarity ordinal < 2); a *fully positive* item is correct, significant, and evidence-sufficient by expert annotation. “2H+1AI” averages over all $\binom{3}{2} \times 3 = 9$ panels; “1H+2AI” over all $3 \times \binom{3}{2} = 9$ panels. Rows 5–8 apply an AI meta-reviewer (GPT-5.4) filter that removes items the meta-reviewer judges not fully positive. Noise per Gem is # Not Fully Pos. / # Fully Pos. + Unique, the filler an editor or author reads to surface each useful item. Bold values mark the single best performer across all eight panel configurations on each metric (# Total Items is descriptive and not bolded; # Fully Pos. + Unique has a tie between the 3-Human and 2H+1AI configurations). 95% confidence intervals on the count columns are reported in Table 51 (§ F.2).

Panel Composition			Volume	Diversity	Quality	Diversity & Quality
# Human Reviewers	# AI Reviewers	AI Meta-Rev. Filtering	# Total Items	# Unique Items	# Not Fully Pos. Items	# Fully Pos. + Unique Items
3	0	✗	25.8 [21.8, 30.2]	11.5 [9.6, 13.6]	14.6 [11.3, 18.5]	3.9 [3.0, 4.8]
2	1	✗	21.4 [18.7, 24.4]	10.9 [9.3, 12.6]	11.5 [9.2, 14.1]	3.9 [3.2, 4.6]
1	2	✗	17.1 [15.7, 18.6]	8.5 [7.2, 9.8]	8.4 [7.0, 9.8]	3.5 [2.9, 4.1]
0	3	✗	12.7 [12.3, 13.0]	3.1 [2.6, 3.5]	5.2 [4.2, 6.3]	1.8 [1.4, 2.2]
3	0	✓	11.8 [9.7, 14.1]	4.0 [3.3, 4.8]	5.3 [4.0, 6.8]	1.9 [1.5, 2.4]
2	1	✓	11.0 [9.5, 12.6]	4.6 [4.0, 5.3]	4.7 [3.7, 5.8]	2.2 [1.8, 2.6]
1	2	✓	10.2 [9.3, 11.1]	4.1 [3.4, 4.7]	4.1 [3.3, 4.9]	2.1 [1.7, 2.5]
0	3	✓	9.4 [8.7, 10.0]	1.9 [1.5, 2.3]	3.5 [2.8, 4.3]	1.2 [0.9, 1.5]

Table 51: **Panel composition count metrics with 95% confidence intervals.** Brackets show paper-level bootstrap 95% CIs (10,000 resamples) for each of the four count metrics in Table 50. Metric definitions, panel-composition methodology, and combinatorial averaging conventions are given in § F.1.

All-AI panels with filtering are selective, not productive. The strength of the all-AI configuration is selection. AI reviewers tend to converge on a small set of well-known issues, so most of what they produce overlaps; the items where one AI breaks from the pack to raise something distinct tend to be narrow, well-targeted critiques that are easy to verify, which makes them more likely to be rated correct, significant, and evidence-sufficient. The meta-reviewer filter sharpens this selectivity by removing the items it judges not fully positive. The result is a panel whose distinctive contributions are unusually high quality and whose absolute reviewer-side noise is the lowest of any configuration, but at the cost of producing only one or two pieces of useful, non-redundant feedback per paper. This configuration is appropriate when an editor or author wants a small, vetted shortlist of high-confidence critiques rather than broad coverage of a manuscript.

Confidence intervals on count metrics Table 51 reports paper-level bootstrap 95% confidence intervals (10,000 resamples) for the four count metrics in Table 50.

Three configurations indexed to organizer priorities The findings above support three configuration-priority pairings that an organizer can pick from depending on what their venue most wants to optimize. Each is summarized below.

Recommendation 1 (2H+1AI): same useful feedback as today’s 3-human panel, with less filler
Both 3H and 2H+1AI deliver the highest amount of useful feedback among all panel configurations, producing 3.9 items per paper that are simultaneously fully positive (correct, significant, and well-evidenced) and unique to that reviewer, with heavily overlapping 95% CIs in Table 51. The 2H+1AI configuration is preferable in practice because it preserves the same amount of useful feedback while reducing the total number of items by 17% (21.4 vs 25.8) and the number of not-fully-positive items by 21% (11.5 vs 14.6), so authors and editors read less filler to reach the same useful content. Given the rising volume of paper submissions and the increasing difficulty of recruiting qualified human reviewers [Chen et al., 2025], 2H+1AI is the natural recommendation for venues that want to preserve today’s review experience while reducing reviewer-side burden.
Recommendation 2 (1H+2AI + meta-reviewer filter): minimize editor and author triage time
The 1H+2AI panel with the meta-reviewer filter achieves the highest fraction of items that are simultaneously fully positive and unique within the panel (20.6% of all items shown, against 15.1% for the 3-human baseline, a 36% relative improvement) and the lowest noise per gem (1.95 against 3.74, a 48% reduction). The trade-off is volume: this configuration delivers 2.1 fully-positive-and-unique items per paper against 3.9 for the 3-human baseline, so authors receive less useful content overall but spend less time reading filler to reach each useful item. This configuration suits venues where editor or author triage time per paper is the binding constraint.
Recommendation 3 (3AI + meta-reviewer filter): high-confidence shortlist instead of full reviews
The all-AI panel with the meta-reviewer filter produces the fewest not-fully-positive items per paper (3.5 against 14.6 for the 3-human baseline) and the highest fraction of distinctive contributions that are fully positive (63.2% of unique items against 33.9% for three humans, a 71% relative improvement). When this panel raises a distinctive item, it is fully positive nearly twice as often as when a human panel does. The trade-off is volume across the board: the all-AI filtered panel produces only 1.9 unique items and 1.2 fully-positive-and-unique items per paper. This configuration suits editorial pre-screening, desk-review aids, and other contexts where a small, vetted shortlist of high-confidence critiques is preferable to broad coverage of a manuscript.

Caveats Two caveats temper these recommendations. First, our entire study draws from one Nature paper submission pool, where reviewers are typically from broad disciplines and review papers tend to address a non-specialist audience; conferences with narrower technical scope, single-discipline journals, or venues with very short review timelines may exhibit different dynamics. Second, the per-AI five-item cap and the choice of three frontier AI reviewers in our study are specific to our experimental setup; a larger per-AI item budget would change the absolute counts, and frontier-model behavior will continue to evolve. The qualitative pattern (humans contribute most of the unique coverage, AIs contribute fluent and well-evidenced items but overlap with each other, AI reviewers’ distinctive contributions are unusually high quality when they appear, and the meta-reviewer filter improves efficiency at the cost of removing some genuinely high-value items) is what we expect to generalize; specific numbers should be re-validated for other venues or future AI reviewer generations.