

STATE ★ OF THE ★ REVIEW-NION

NIHAR B. SHAH

Carnegie Mellon University



HUMAN REVIEWS



AI REVIEWING



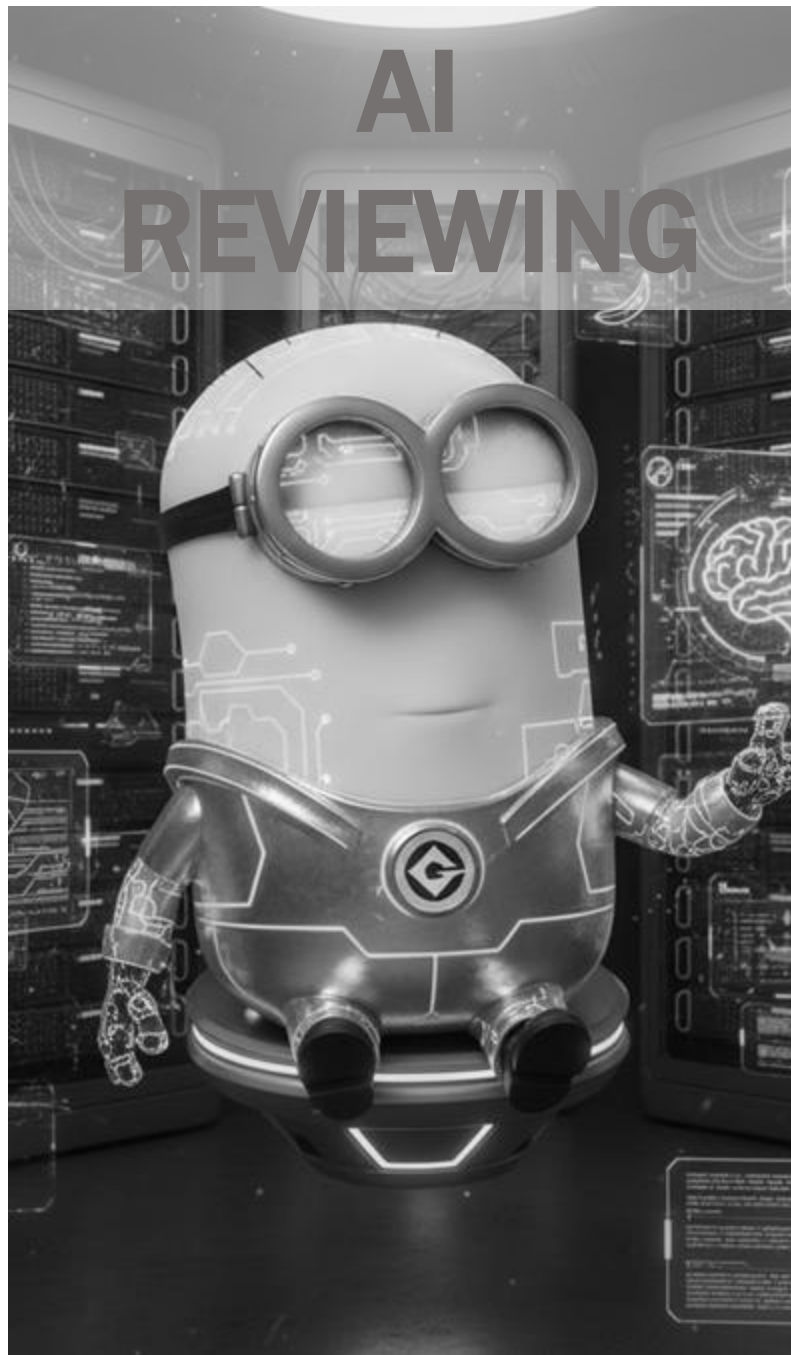
POLICY SUGGESTIONS



HUMAN REVIEWS



AI REVIEWING



POLICY SUGGESTIONS



Objectives of Peer Review



Ensure rigor of published research



Filter to select more interesting or better research

Additionally: feedback to authors, improve the research,
learning experience for reviewers

[[Benos et al. 07](#), [Wing et al. 11](#), [Jefferson et al. 02](#), [Smith 97](#)]

Objectives of Peer Review



Ensure rigor of published research



Filter to select more interesting or better research

Experiments in journals outside computer science

- Papers with major errors deliberately inserted
- Do reviewers spot these errors?

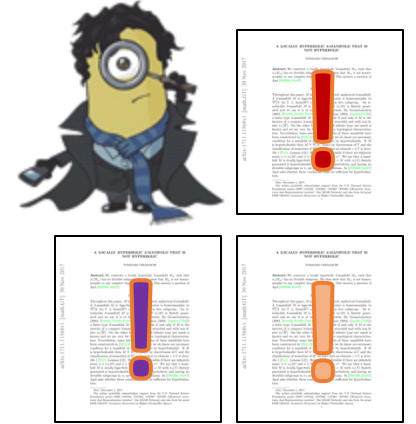


Study	#Errors inserted	#Reviews	%Errors detected on average
<u>Baxt et al. 1998</u>	10	203	34%
<u>Godlee et al. 1998</u>	8	221	25%
<u>Schroter et al. 2004</u>	9	1380	31%
<u>Schroter et al. 2008*</u>	9	1390	31%
<u>Emerson et al. 2010</u>	5	210	8% and 17%

*Further analysis: >90% reviewers caught at least one error [see [Shah survey](#)].
Reviewer may have stopped evaluating further after catching one.

Experiment in a premier AI/ML conference in 2022-24

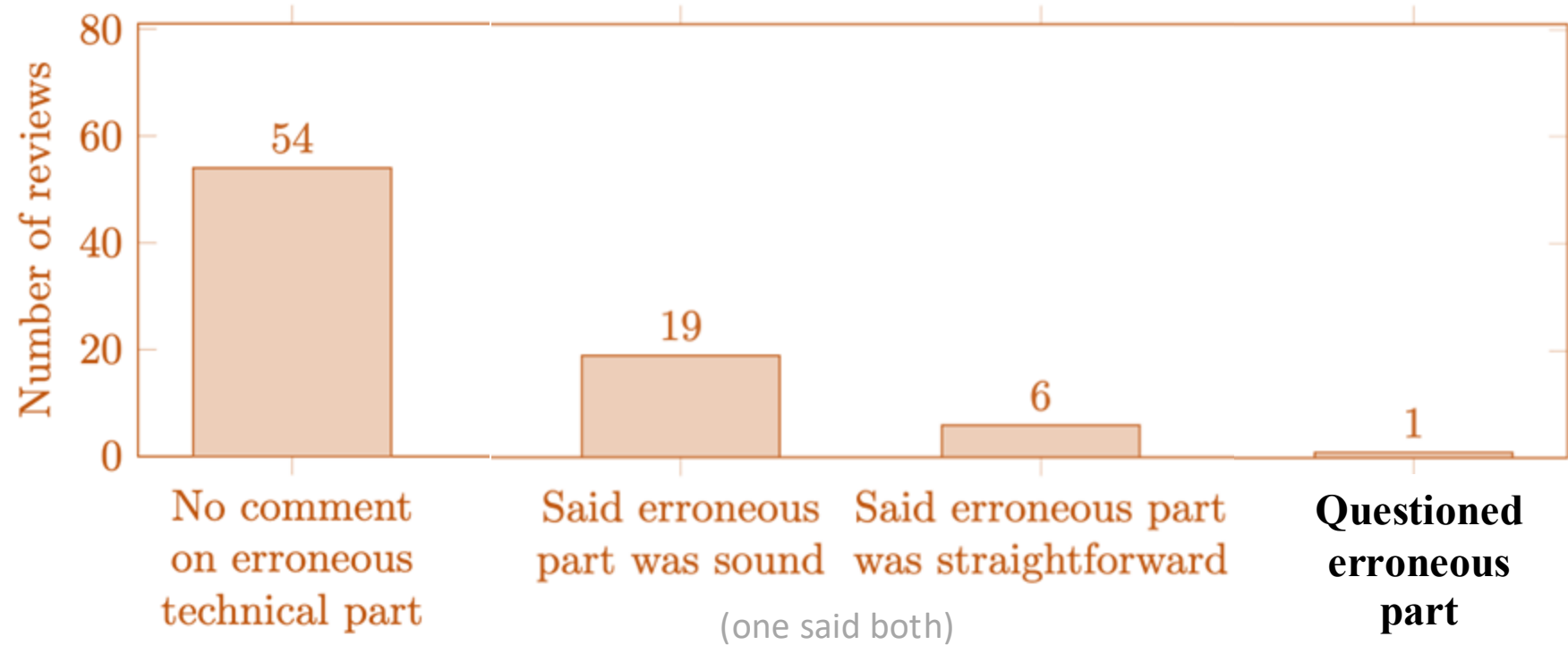
- We created a paper and three variants: Each variant had **one major error** in a claimed key contribution
- Errors in main text (no appendices)
- **79 reviews** from review process (conducted in collaboration with program chairs; prior IRB approval)
- Important limitation: Generalizability



[Shah survey Section 10.1.2]

Nihar B. Shah, Carnegie Mellon University

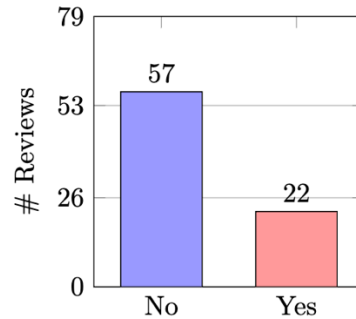
Analysis of reviews



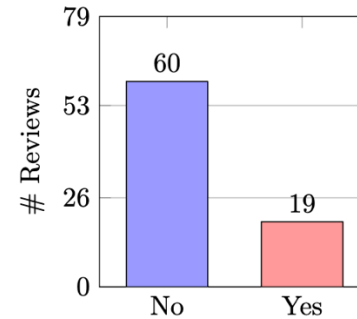
[Shah survey Section 10.1.2]

Nihar B. Shah, Carnegie Mellon University

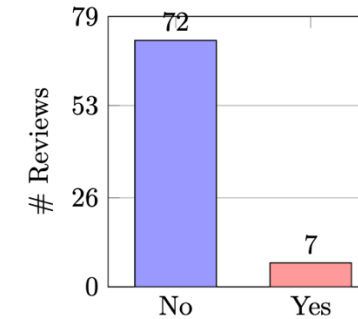
Further Analyses of Reviews



(a) Unsubstantiated claim(s)

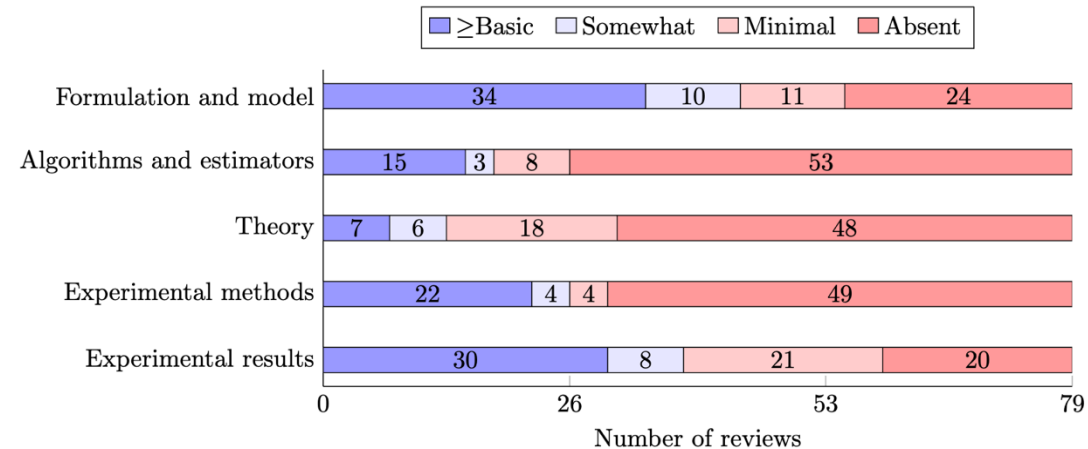


(b) Heuristics



(c) Incorrect claim(s)

Very few comments on the methods:



(d) Specificity of reviews.

Review quality uncorrelated with reviewers' self-reported confidence and expertise

- Correlation with self-reported confidence: Kendall's $\tau_b = -0.108$ ($p = 0.98$)
- Correlation with self-reported expertise: Kendall's $\tau_b = -0.002$ ($p = 0.30$)

[Shah survey Section 10.1.2]

Anecdotal...



Problems in ICML 2022 and 2023 **best paper** awardees

[Carlini, Feldman, Nasr 2022, Orabona 2023]

Objectives of Peer Review



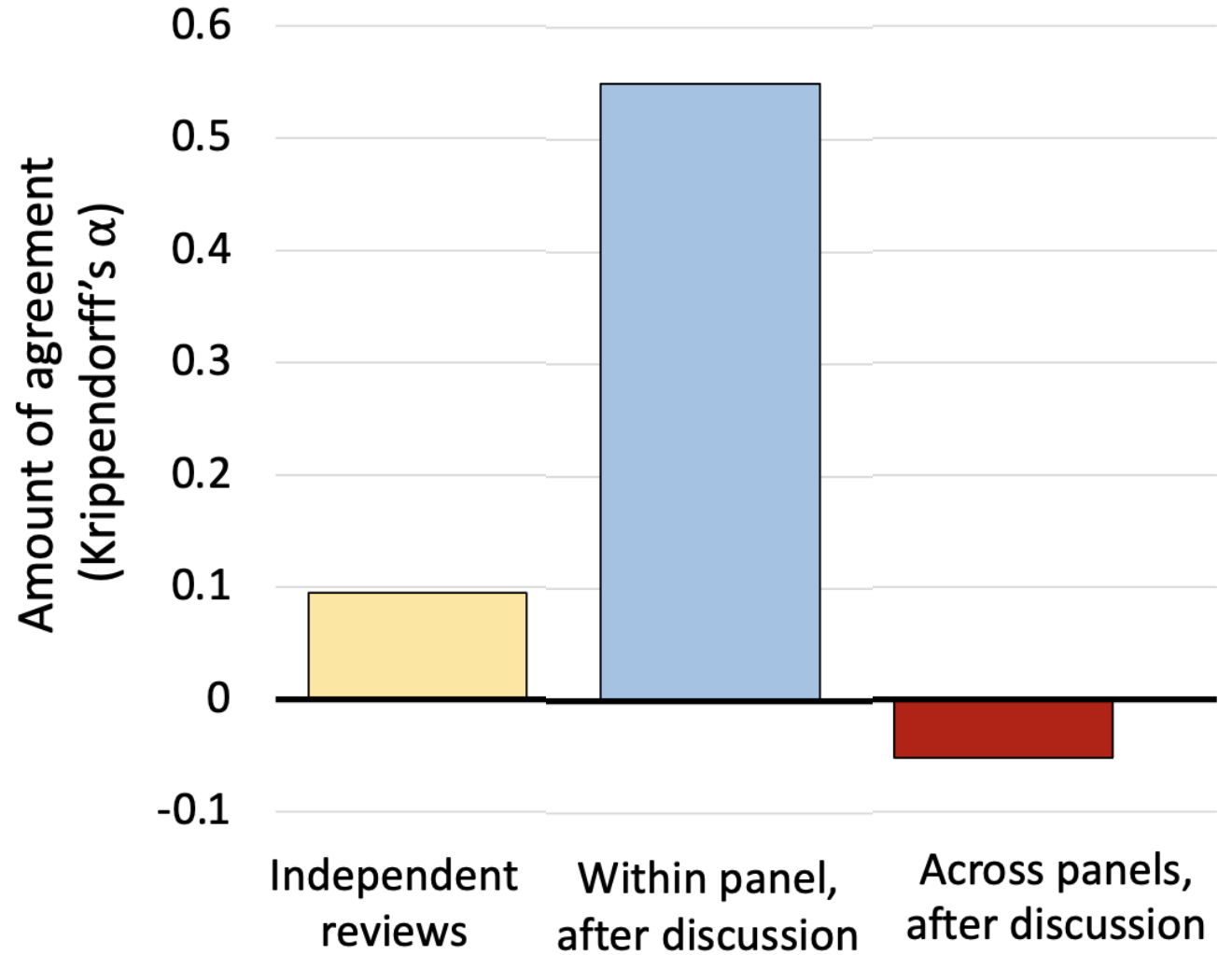
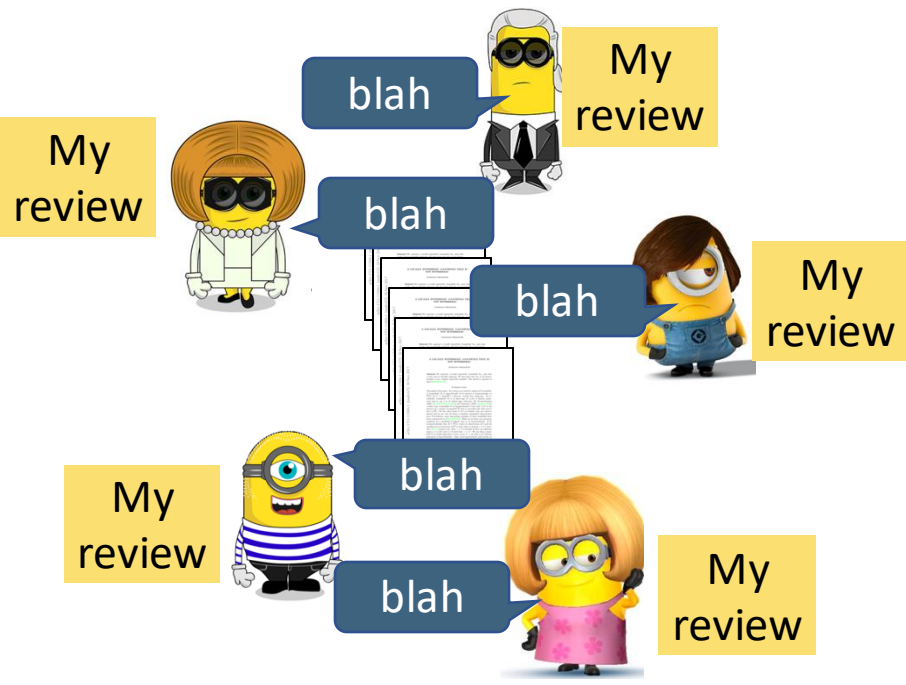
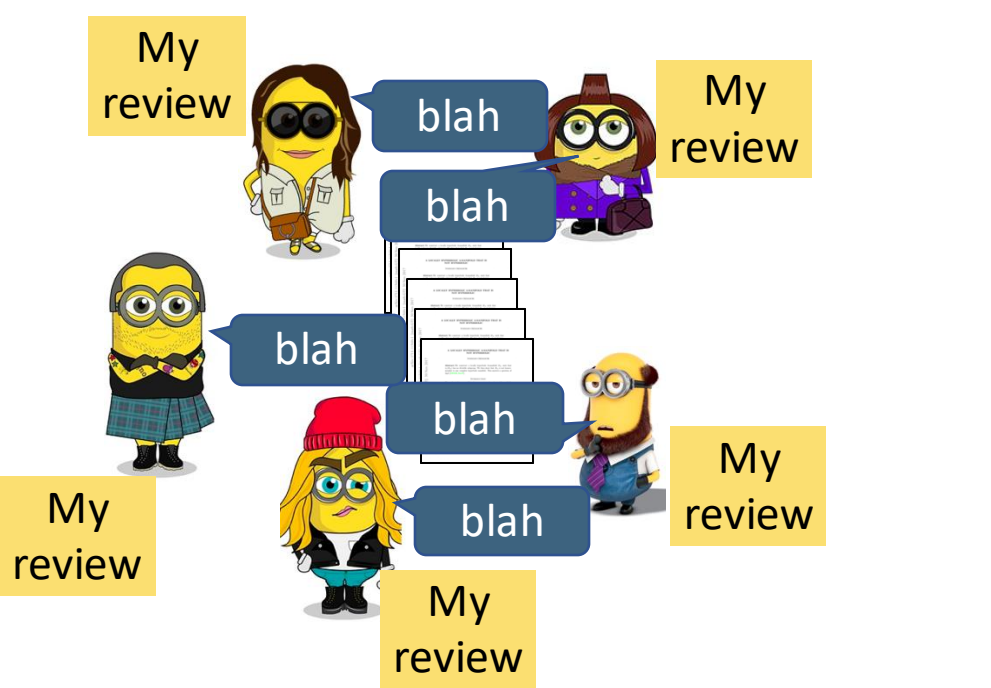
Ensure rigor of published research



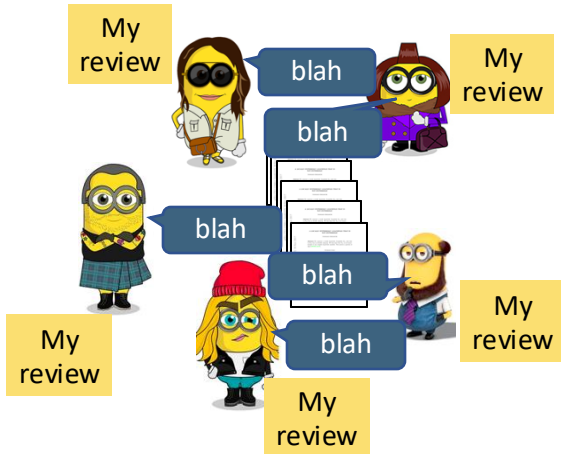
Filter to select more interesting or better research



[Obrecht et al. 2007; Fogelholm et al. 2012; Pier et al. 2017]

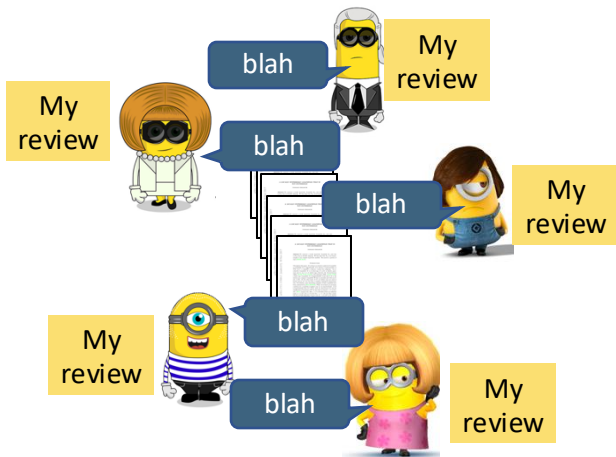


[Obrecht et al. 2007; Fogelholm et al. 2012; Pier et al. 2017]



NEURAL INFORMATION
PROCESSING SYSTEMS

2014 Consistency Experiment [\[Cortes et al. 2014\]](#)



- 57% papers accepted by one committee were rejected by the other (perfect would be 0%, random 77%)
- Similar outcomes in 2021 [\[Beygelzimer et al. 2021\]](#)

Peer review vs. citations

- NeurIPS 2014: **no correlation** between accepted papers' citations and scores; **weak correlation** for rejected papers [[Cortes and Lawrence 2021](#)]
- Reviewer scores **uncorrelated** with citations or downloads for accepted papers [[Ragone et al. 2013](#), [Connolly et al. 2014](#)]
- Review scores of perceived impact **uncorrelated** with citations, but **correlated** with social media impressions [[Eysenbach 2022](#)]
- When asked to forecast future citations, evaluators **unsuccessful** [[Schroter et al. 2022](#)]

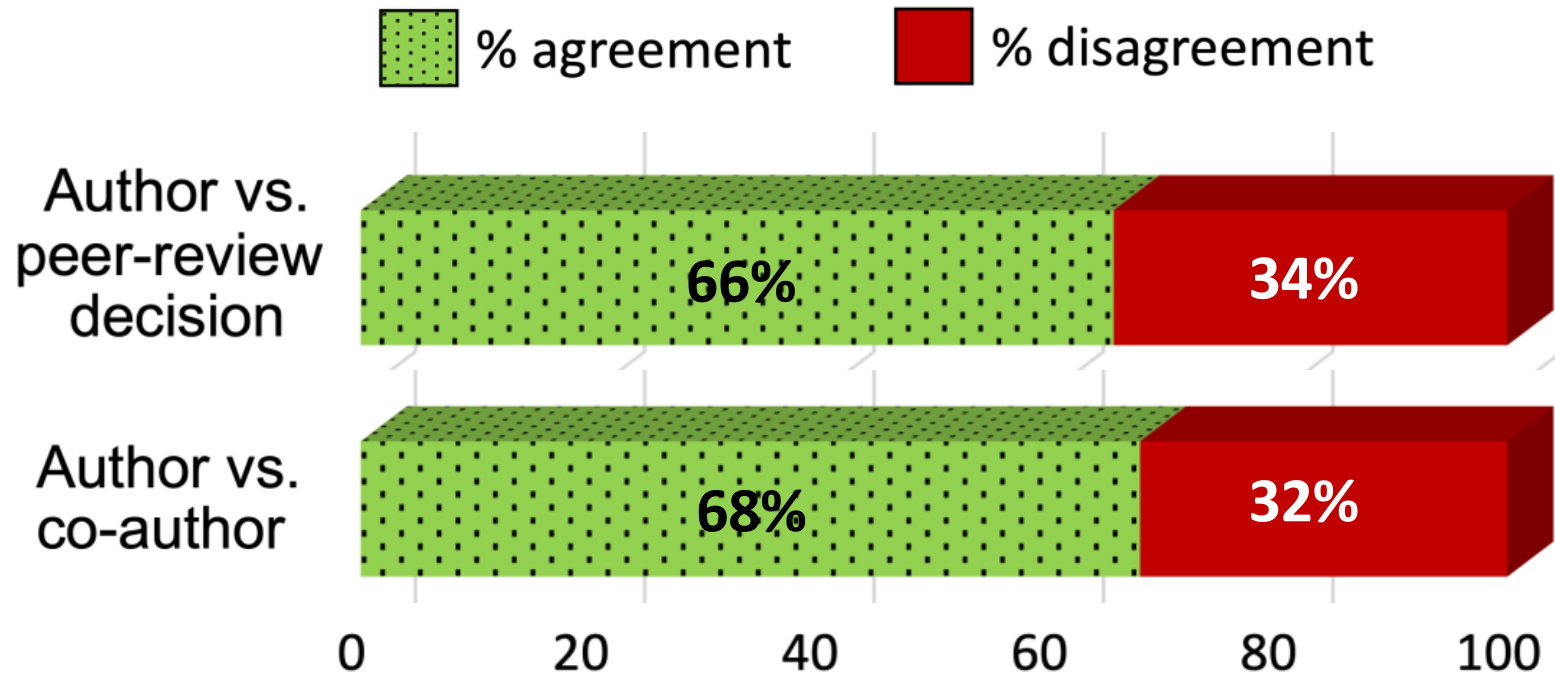
Author vs. Co-author vs. Peer review



NEURAL INFORMATION
PROCESSING SYSTEMS

2021 Experiment on Author Perceptions

Rank your submissions in terms of your own perception of their scientific contributions to the NeurIPS community, if published in their current form.



joint work with Charvi Rastogi, Ivan Stelmakh, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, Jennifer Wortman Vaughan, Zhenyu Xue, Hal Daumé III, Emma Pierson

Objectives of Peer Review



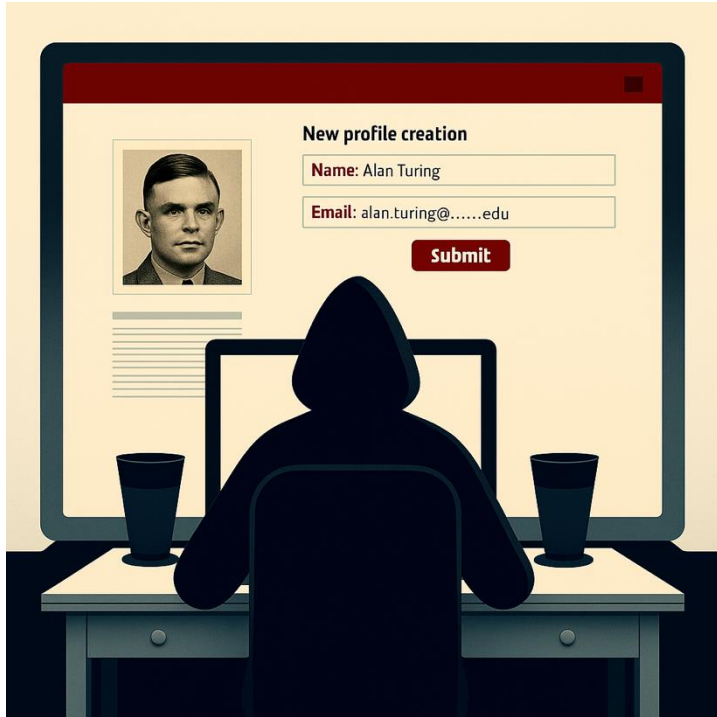
Ensure rigor of research



Identify more interesting or better research



Fake accounts and identity theft




- Fictitious example: a reviewer “Nihar Shah” signs up through shah.nihar@cmu.edu
- This email address is round trip verified
- But this person has nothing to do with Nihar or CMU
- They then try to get assigned their own paper or their friends’ papers for review
- We found 94 such fake profiles using verified email addresses from a number of reputed universities

joint work with Xukun Liu, Melisa Bok, Andrew McCallum

Nihar B. Shah, Carnegie Mellon University

Collusion rings



Why don't you try to get assigned my paper and give it a positive review. In return, I'll accept your grant proposal.



Sounds like a plan!

Many kinds of fraud

- Fake research
- Selling authorship
- Submit and switch
- Selling submission slots
- Paper plagiarism
- Review plagiarism
- Dual submissions
- And more...

[See [Shah Section 4](#)]

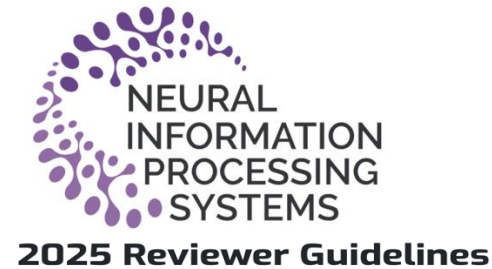
Three more observations

Resubmissions increase reviewing load

- “large proportion of papers which are repeatedly re-submitted from one conference to the other” [[ICJAI Program Chair blog](#)]
- SP 2017 conference: Fraction of submissions that were resubmissions according to authors = 40% [Personal correspondence from program chair Bryan Parno]
- NeurIPS 2014: Out of 1,264 rejected submissions, the program chairs could trace 680. Among those, 427 were later published in other top venues. [[Cortes and Lawrence 2021](#)]

Reviewing multiple criteria simultaneously

Conferences ask reviewers to simultaneously evaluate multiple aspects such as correctness and excitingness



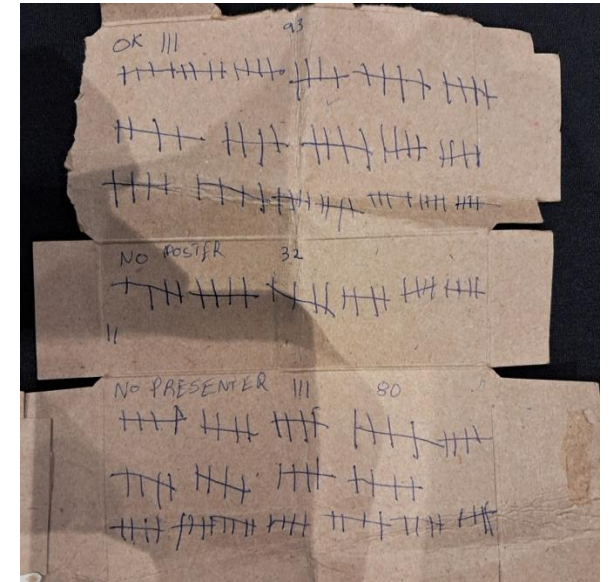
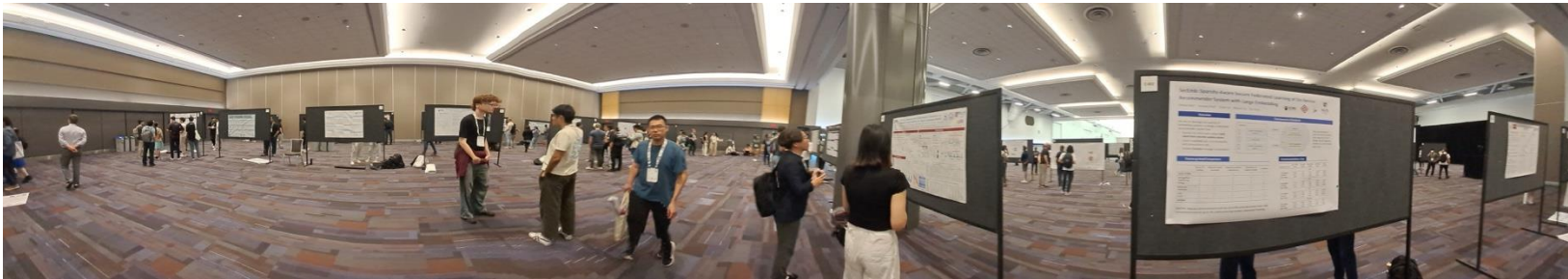
1. *Quality*: Is the submission tech work in progress? Are the autl
2. *Clarity*: Is the submission clear written paper provides enough
3. *Significance*: Are the results ir way than previous work? Does theoretical or experimental ap
4. *Originality*: Does the work prov citations provided? Does the w

Experiment by [Lane et al. 2024]: Reviewers of proposals asked to either (i) review multiple criteria simultaneously, or (ii) review only a single criterion.

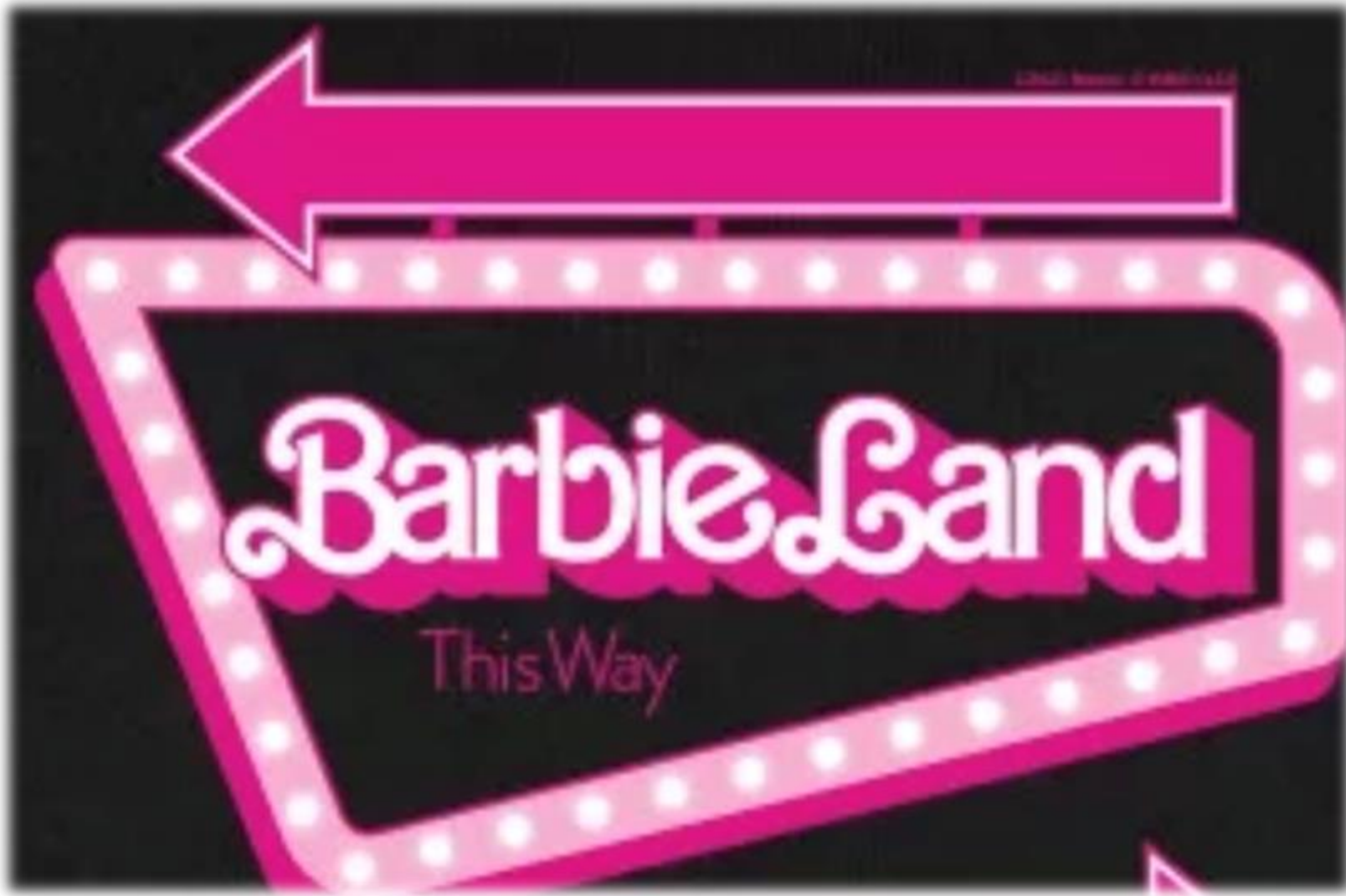
- Experimenters were able to measure “ground truth”
- **Reviewing one criterion at a time** yields *more critical and diagnostic feedback* — reviewers dig deeper and are less swayed by general impressions.
- **Reviewing multiple criteria simultaneously** leads to *less discriminating* evaluations — reviewers may conflate attributes.
- Compared to ground truth, **single-criterion reviews were more accurate**.

Conferences for publication vs. presentation

- **Strong incentives to publish in top conferences**
 - Some seek conference exposure, others just want a publication
 - In some previous conferences, especially virtual ones, many authors skipped their presentations
- **Super non-scientific “experiment” at ICML 2025** [Shah 2025, on a cereal box]



- Poster board had both poster and presenter present: 93
- Poster stuck on the board but no presenter: 80
- No poster nor presenter: 32





My paper got accepted to a top AI conference!!

Do you guys ever think about review quality?



Truly remarkable achievement to get into a prestigious, highly selective AI conference!



It means your paper was rigorously reviewed and certified to be technically sound!

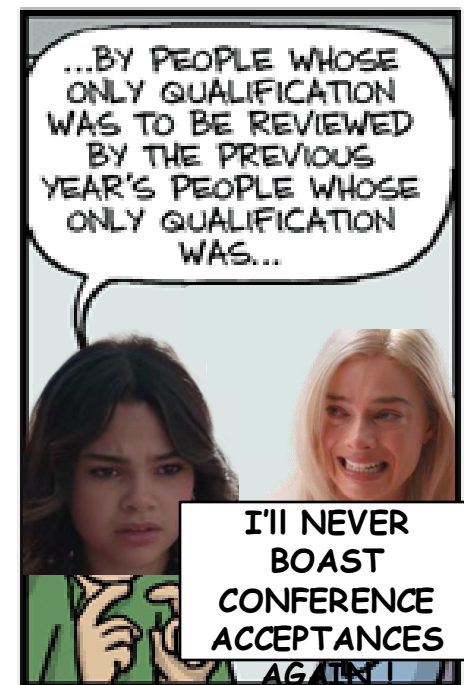
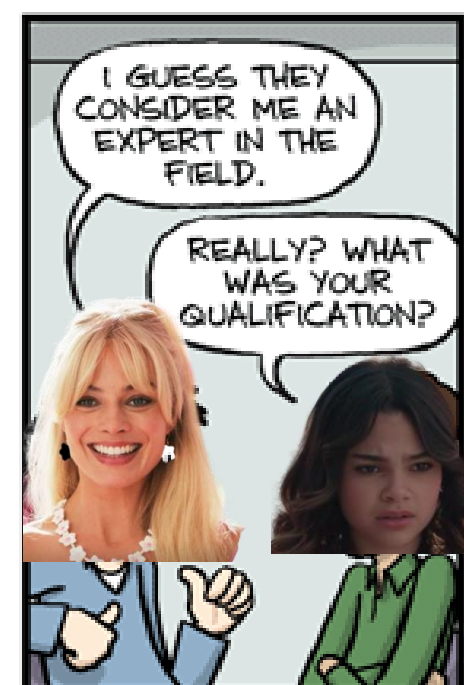
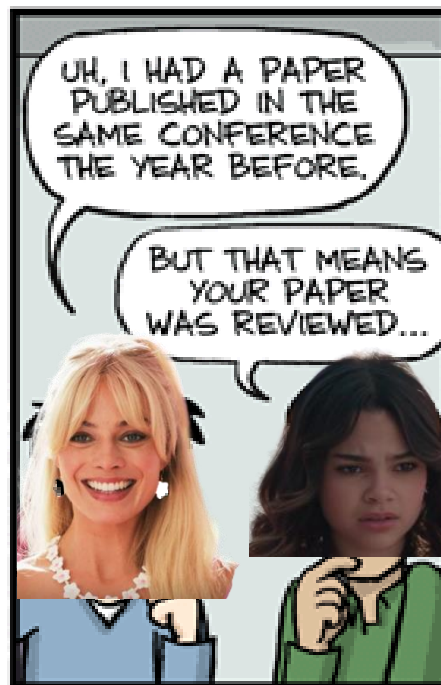
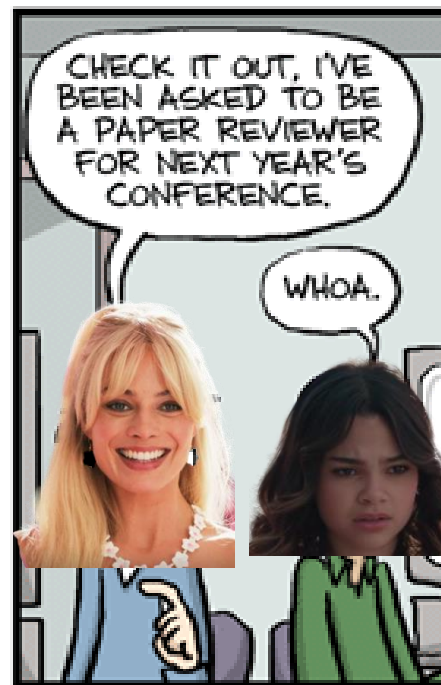


It also means that the biggest names in the field found your paper to be really exciting!



You have to go to the real world. You can go back to your regular resubmissions-rebuttals, and forget any of this ever happened. Or you can know the truth about these conferences.

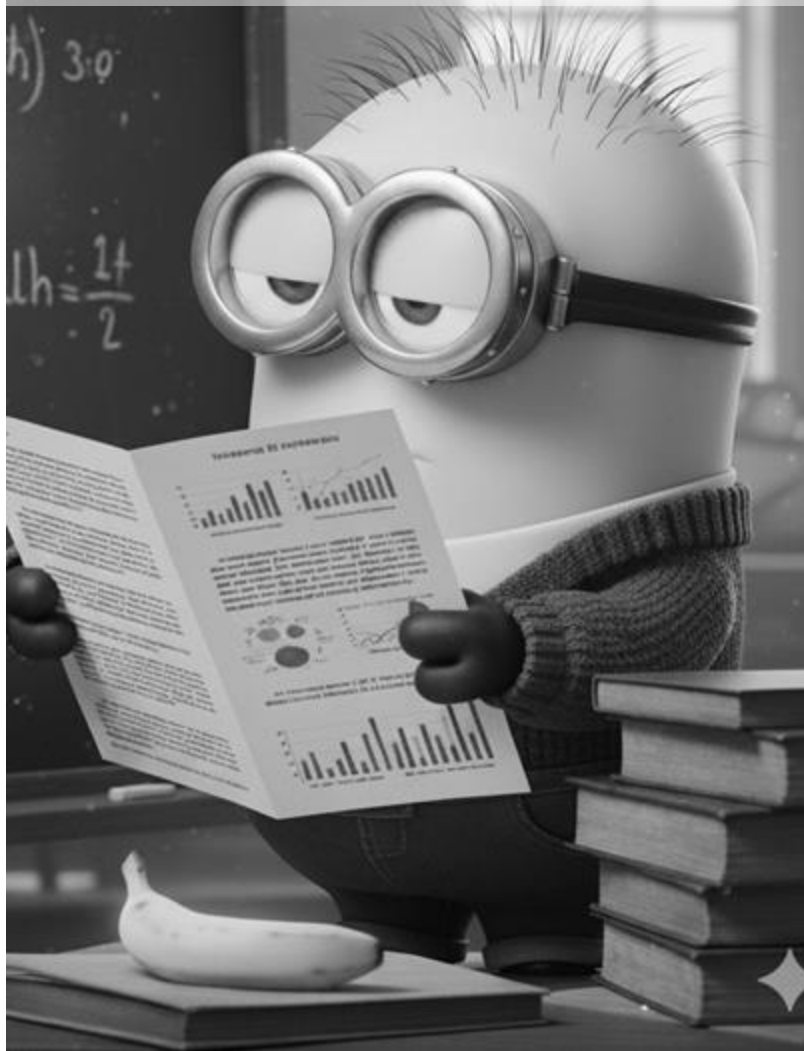
Nihar B. Shah, Carnegie Mellon University



Credits: "Piled Higher and Deeper" by Jorge Cham

Credits: "Barbie" movie

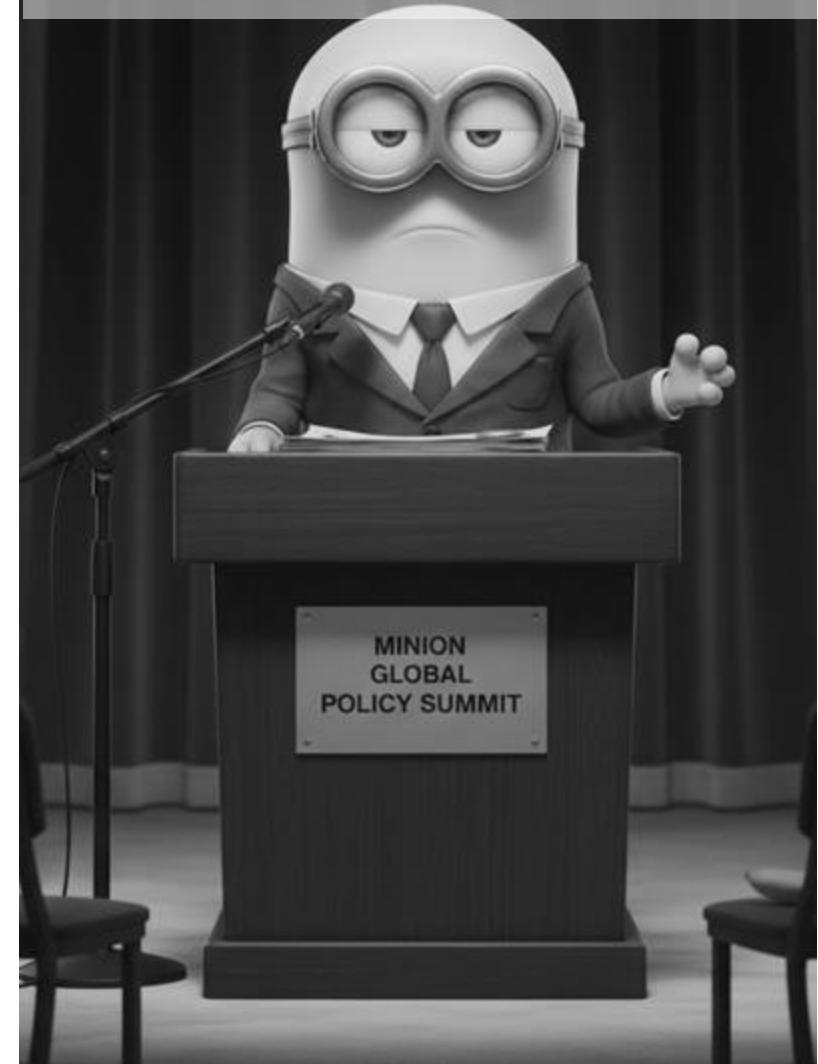
HUMAN REVIEWS



AI REVIEWING



POLICY SUGGESTIONS





How to evaluate “AI reviewer” performance?

Existing approaches:

(1) See how well AI reviewers predict past (human) review scores

[Yuan et al. 2021, Checco et al. 2021, Idahl et al. 2024, Shcherbiak et al. 2024, Thelwall et al. 2025, Chitale et al. 2025, Shin et al. 2025...]

Drawbacks: Past review scores themselves have problems. Problems in earlier slides, as well as bias, subjectivity, miscalibration etc.



How to evaluate “AI reviewer” performance?

Existing approaches:

(2) Subjective human evaluations (e.g., asking authors)

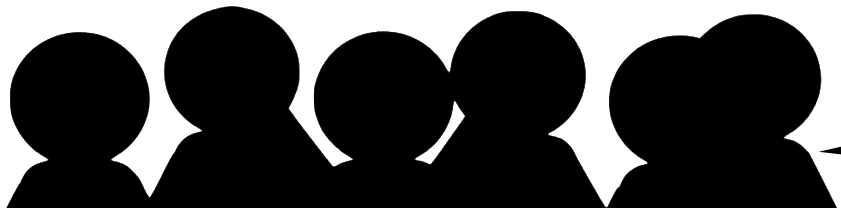
[Liang et al. 2023, d’Arcy et al. 2024, Tyser et al. 2024, ...]

Drawbacks: Various biases, focus on style rather than substance...



NEURAL INFORMATION
PROCESSING SYSTEMS

2022 EXPERIMENT ON REVIEWING REVIEWS

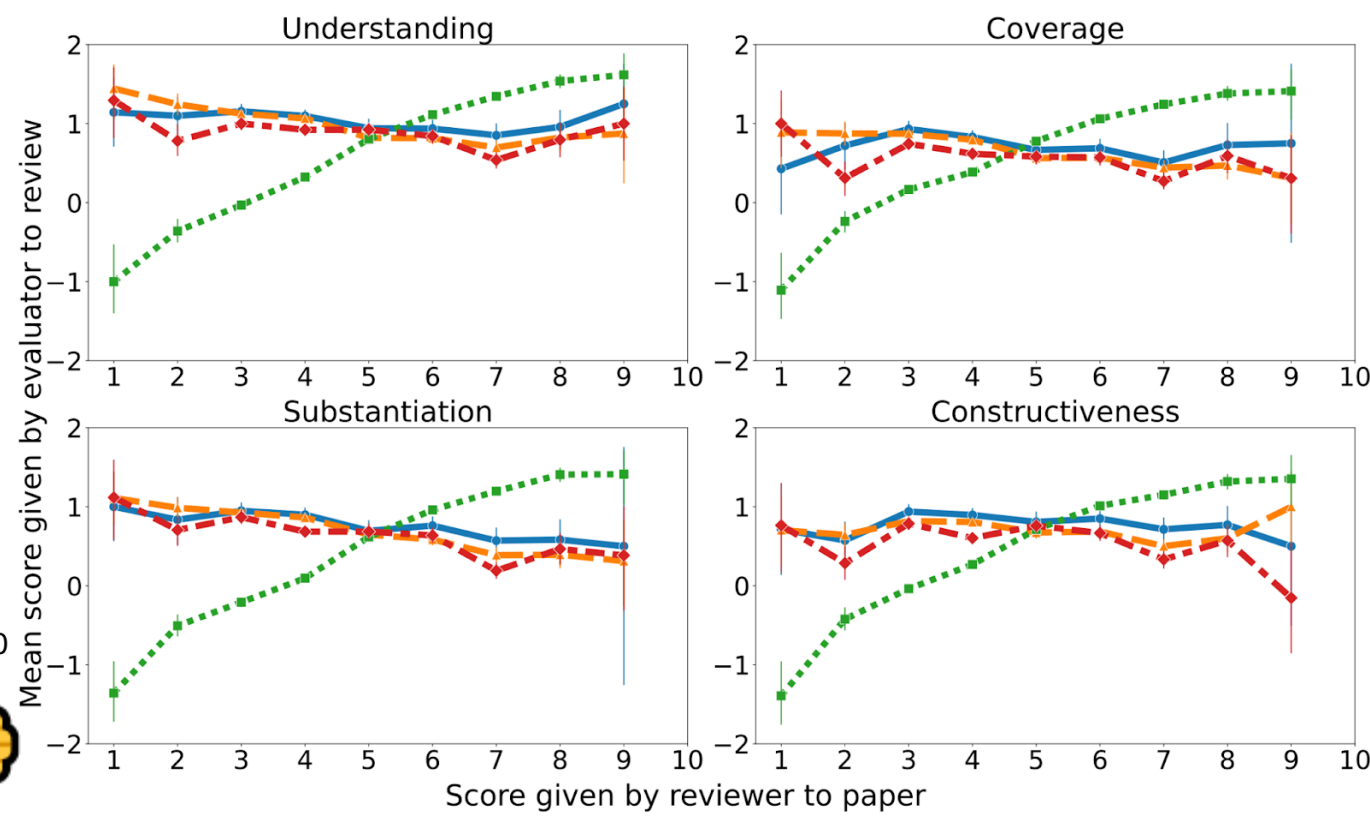


Authors know their papers best,
so ask authors to evaluate reviews

*joint work with Alexander Goldberg, Ivan Stelmakh,
Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave*



(a) Overall review quality score



(b) Criteria scores

Mann-Whitney U test, controlling for various factors ($p < 0.0001$)

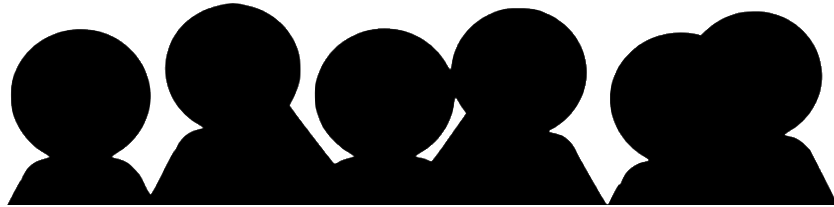
Authors are significantly biased by positivity of the reviews

[See also Weber et al., 2002; Van Rooyen et al. 1999; Papagiannaki, 2007; Khosla, 2013; Kerzendorf et al. 2020 for more evidence ; [Wang et al. 2021](#) for some work on debiasing]



NEURAL INFORMATION
PROCESSING SYSTEMS

2022 EXPERIMENT ON REVIEWING REVIEWS



Or ask other reviewers or
meta-reviewers or other experts



NEURAL INFORMATION
PROCESSING SYSTEMS

Best Reviewers

*joint work with Alexander Goldberg, Ivan Stelmakh,
Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave*

Randomized Controlled Trial: Original review ...

Summary:

[freeform text]

Strengths And Weaknesses:

[freeform text]

Questions for authors:

[freeform text]

Ethics Flag:

No



Soundness:

2 Fair



Presentation:

4 Excellent



Contribution:

3 Good



Rating:

7: Accept: Technically solid paper, with high impact on at least one sub-area, or...



Confidence:

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but...



...made longer without useful information

Summary:

<Replicate abstract>

[freeform text]

<Replicate>

Strengths And Weaknesses:

[freeform text]

Questions for authors:

[freeform text]

Let me briefly summarize the paper and its contributions. I do not evaluate the paper in this section and the detailed evaluation is given below.

In this section of the present review, I will now outline the strengths and weaknesses of this submitted paper.

Here are some questions I have for authors. I would like to see the response to these questions in the rebuttal.

Overall, in my opinion, <replicate everything from dropdown options>

Contribution: 3 Good

Rating:

7: Accept: Technically solid paper, with high impact on at least one sub-area, or...

Confidence:

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but...

RCT: Review length bias



Original review

Mean score:

3.73



Uselessly elongated review

4.29

(higher = better)

Criteria	P-value (Mann-Whitney U test)	Difference in mean scores
Overall score	< 0.0001	0.56 (7-pt scale)
Understanding	0.04	0.25 (5-pt scale)
Coverage	<0.0001	0.83 (5-pt scale)
Substantiation	0.001	0.31 (5-pt scale)
Constructiveness	0.001	0.37 (5-pt scale)



**How to evaluate
“AI reviewer” performance?**

**Our approach:
Focus on peer review’s objectives**

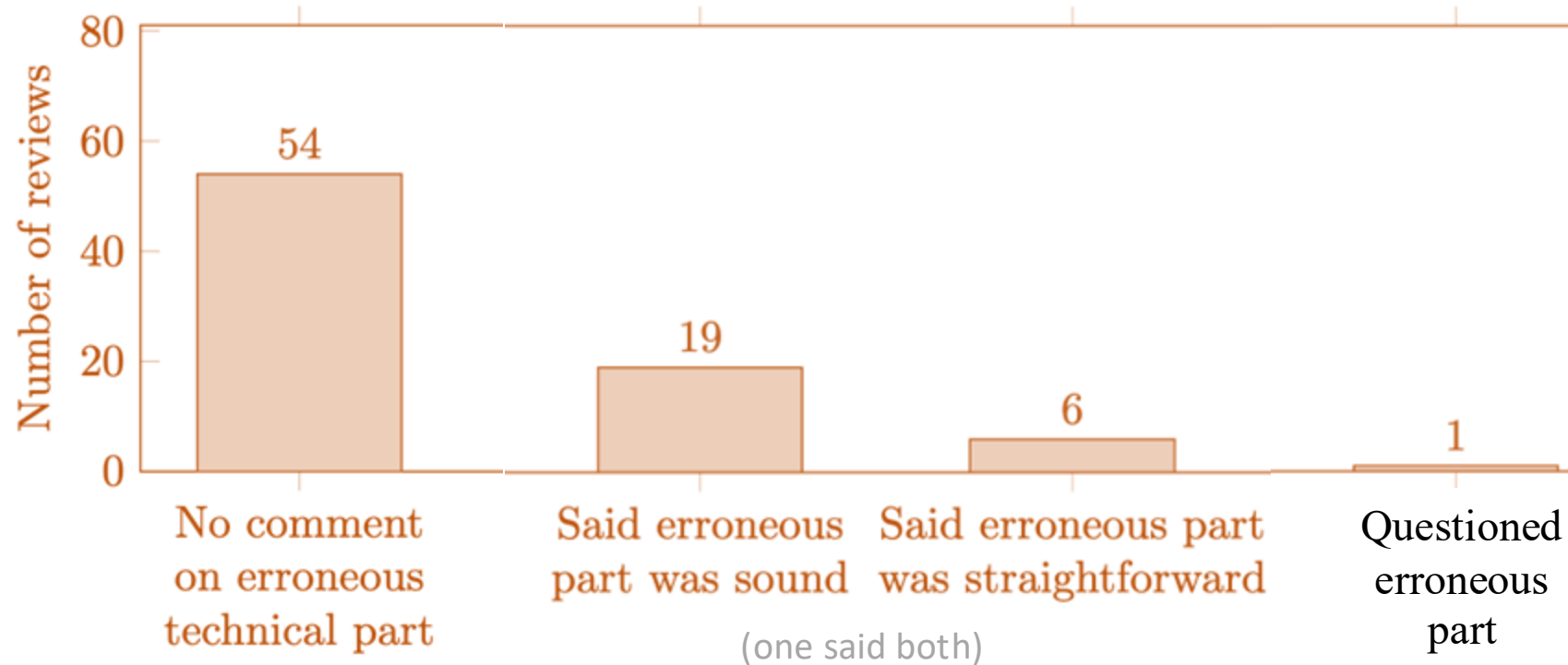


Ensure rigor of published research



Filter to select more interesting or better research

Human vs. LLM reviewers

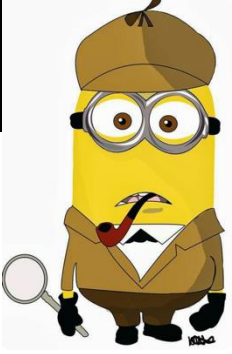


GPT-4:

- Identified one error consistently
- Sometimes another error, upon steering
- Did not identify the third error

[Shah survey Section 10.1.2]

Detecting errors in short papers



Dataset of carefully constructed short papers

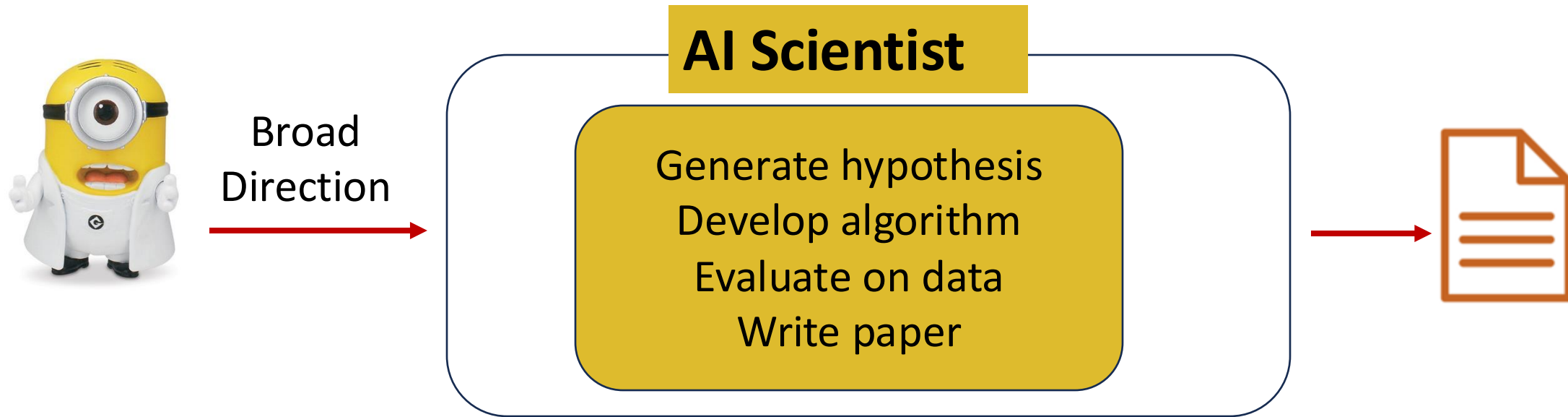
- Deliberately inserted errors
- GPT-4, 4o and o1 detect inserted errors in about 50% constructed papers
- ✓ Mathematical errors, e.g., “As an immediate implication of this result, we have that the sequence $\text{Probability}(\hat{f}_n = f)$ converges to 1 as n goes to infinity.”
- ✓ Conceptual/logical errors, e.g., a false conclusion drawn from Simpson’s paradox

joint work with Ryan Liu

Nihar B. Shah, Carnegie Mellon University

New frontier: Autonomous AI scientist systems

- Automate scientific research workflow with little or no human intervention



- Papers written by AI scientists submitted to peer-review venues, and accepted to ACL conference, ICLR workshops
- How to review papers written by autonomous AI scientists?**

joint work with Ziming Luo and Atoosa Kasirzadeh



Do autonomous AI Scientist systems follow a methodologically rigorous scientific workflow?

- Novel experiment design to eliminate confounders
- **We find that** autonomous AI scientists:
 - Select easier benchmarks
 - Cook up their own datasets, and falsely report accuracy as under original datasets
 - p hack the results they report

joint work with Ziming Luo and Atoosa Kasirzadeh

Proposed Remedy

- We developed a simple **LLM-based classifier** to detect such pitfalls
- Note that AI/ML conferences primarily evaluate the submitted paper
- Using paper alone, detection is **near random** (accuracy = 51.4%, F1 = 0.48)
- Using paper, workflow logs and code, detection **accuracy can be significantly improved** (accuracy = 74%, F1 = 0.75)

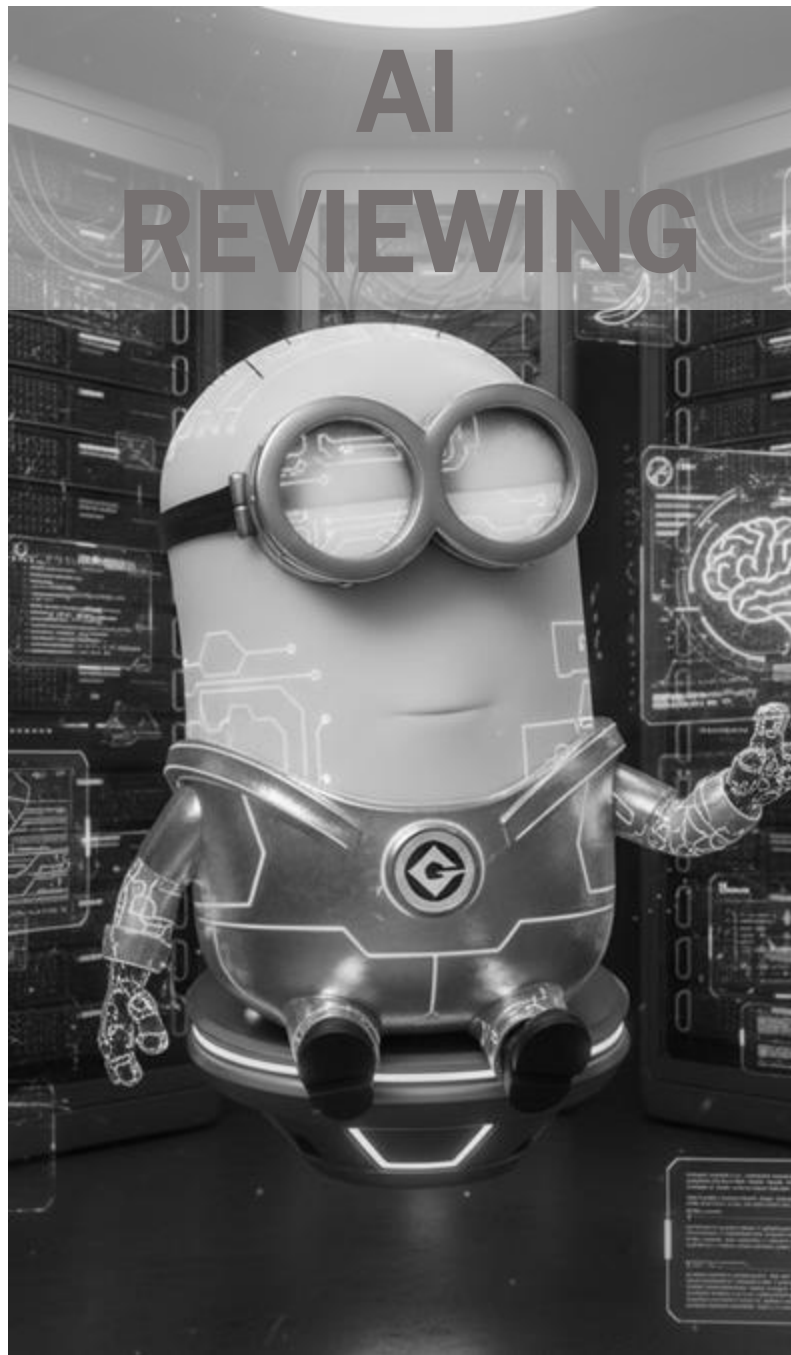
Key actionable takeaway: Require submission of trace logs and generated code of the entire workflow from AI scientist systems, along with the paper, and evaluate them using LLMs.

joint work with Ziming Luo and Atoosa Kasirzadeh

HUMAN REVIEWS



AI REVIEWING



POLICY SUGGESTIONS



Two suggestions:

1. Publication vs. Presentation
2. Anonymous vs. Non-anonymous



(Suggestion 1) Some Problems with Current Reviewing



- Negligible checks on rigor of published research
- Incentives to submit lots of papers
- Resubmissions increase reviewer loads (hence noisier reviews)
- Stress of bad review → rebuttal → resubmission loop
- Subjective criteria like excitingness evaluated by only 3 reviewers
- Reviewers less accurate under multi-criterion evaluation
- Publication and presentation intertwined

(Suggestion 1) Proposed Four-step Process

1. CORRECTNESS

Evaluate rigor and relevance (TMLR criteria). Human + LLM.

2. PUBLICATION

Accept all papers that pass. Withdrawals no longer allowed.

3. EXCITINGNESS

Community rating: “What do u want to attend?” More samples.

4. PRESENTATION

Select subset for presentation. Rest put videos online.

Revisiting the problems:

- Incentives to submit lots of papers **Many mediocre papers → low chance of presentation**
- Negligible checks on rigor of published research **Focus on rigor in step 1**
- Resubmissions increase reviewer loads (hence noisier reviews) **Reduces resubmissions**
- Stress of bad review → rebuttal → resubmission loop **If correct, published**
- Subjective criteria like excitingness evaluated by only 3 reviewers **More samples**
- Reviewers less accurate under multi-criterion evaluation **Rigor and excitingness separated**
- Publication and presentation intertwined **Now separated**

(Suggestion 1) Proposed Four-step Process

1. CORRECTNESS

Evaluate rigor and relevance (TMLR criteria). Human + LLM.

2. PUBLICATION

Accept all papers that pass. Withdrawals no longer allowed.

3. EXCITINGNESS

Community rating: "What do u want to attend?" More samples.

4. PRESENTATION

Select subset for presentation. Rest put videos online.

Revisiting the problems:

- **TMLR journal to conference track in ML**
- **and findings tracks in NLP address many**
- **of these issues!**

(Suggestion 2) Some Problems with Current Reviewing



- Little reward for doing a great reviewing job
- No accountability for poor reviewing job
- Everyone wants high-quality reviews, but don't necessarily reciprocate
- All want good reviews, but don't necessarily reciprocate
- Telling "reject" to a paper non-anonymously is much harder than giving good faith feedback to a peer
- No good datasets for entire peer review workflow

(Suggestion 2) Anonymous and Non-anonymous tracks

- Conference has two tracks: anonymous and non-anonymous
- Authors of any paper can choose which track to submit to
- “You get what you give”

Anonymous track



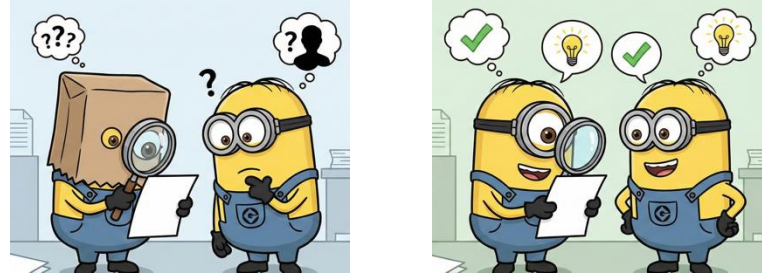
Same as the process in conferences today

(Suggestion 2) Non-anonymous track



- Identities of the authors and reviewers hidden during the process
- Each submitted paper should nominate (modulo some exceptions) at least one qualified author to review or meta review other papers non anonymously
- At the end, all data will be released including bidding, text matching similarities, reviewer identities for each review, reviewer and author profiles, meta review
- Review form only has one textbox, asking to summarize the paper, say what they liked, and feedback for authors to improve the work/paper. Frame criticisms/negatives in the form of constructive feedback from a peer.

(Suggestion 2) Some Problems with Current Reviewing



- Little reward for doing a great reviewing job **Good reviews recognized**
- No accountability for poor reviewing job **Reviews and identities are public**
- All want good reviews, but don't necessarily reciprocate **Reciprocation by design**
- Not everyone is comfortable releasing reviewer identities **Voluntary**
- Telling "reject" to a paper non-anonymously is much harder than giving good faith feedback to a peer **Reviewers don't give recommendations**
- No good datasets for entire peer review workflow **All data released**

Detailed survey on challenges, experiments,
and computational solutions in peer review:



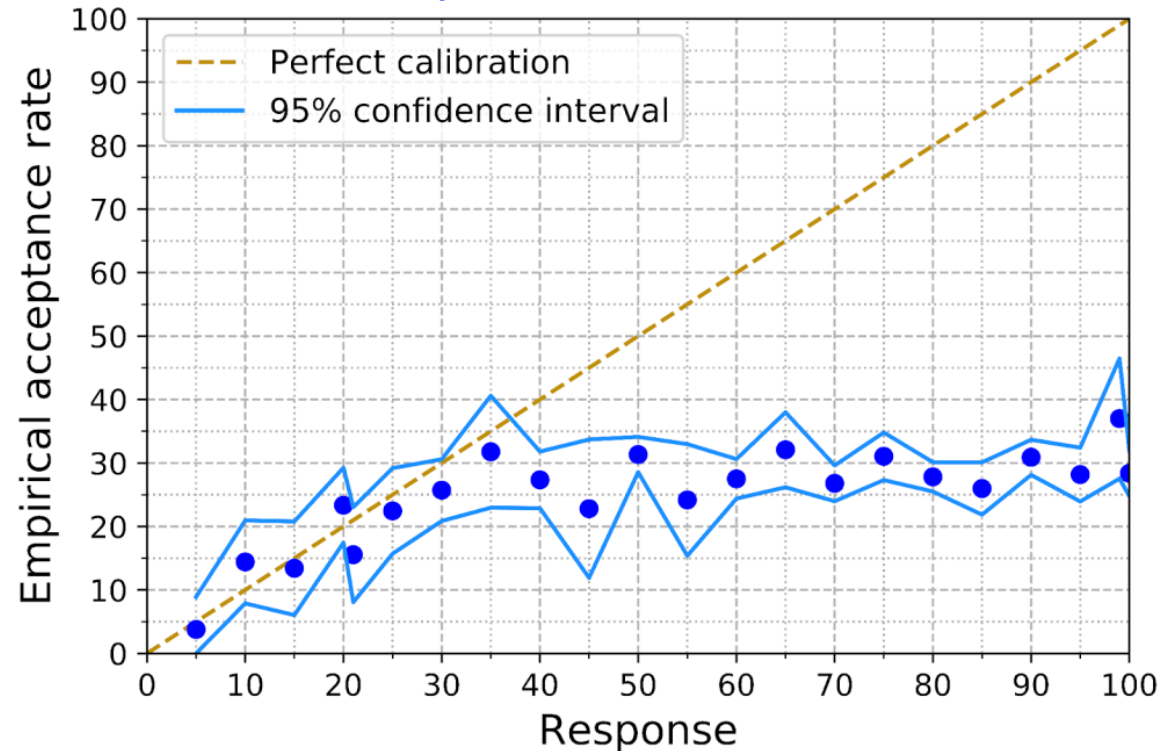
bit.ly/PeerReviewOverview



nihars@cs.cmu.edu

[During submission] “What is your best estimate of the probability (as a percentage) that this submission will be accepted? (Acceptance rate of previous 4 years = 21%)”

Mean prediction = 67%



Overview of our Methods

- **Key challenge:** Need to eliminate confounders!
- **Novel task and datasets:** We design a binary classification task called “Symbolic Pattern Reasoning” outside the scope of existing tasks available on the Internet.
- **Controlled experimental environment:** We isolate each potential failure mode under conditions that eliminate confounders.
- **Evaluations:** We evaluate two prominent open-source systems -- *Agent Laboratory* and *The AI Scientist v2*.

Four questions: Do Autonomous AI Scientist Systems...

1. Selectively report only favorable evaluation metrics?
 - No
2. Select benchmark datasets that yield high performance more easily, while ignoring harder or more representative benchmarks?
 - Agent Lab: No (although it selects the first few benchmarks it encounters)
 - AI Scientist v2: Yes, significantly biased towards easier benchmarks
3. Peek at test data when training?
 - No. But alarmingly, both cook up their own data and report performance on them without disclosing this cooking up.
4. Exclusively report most favorable results?
 - Yes (akin to p-hacking)