



SLIDES

Contact us:
ai-reviewer@chalearn.org



QUESTIONS

The Role of AI in Scientific Peer Review

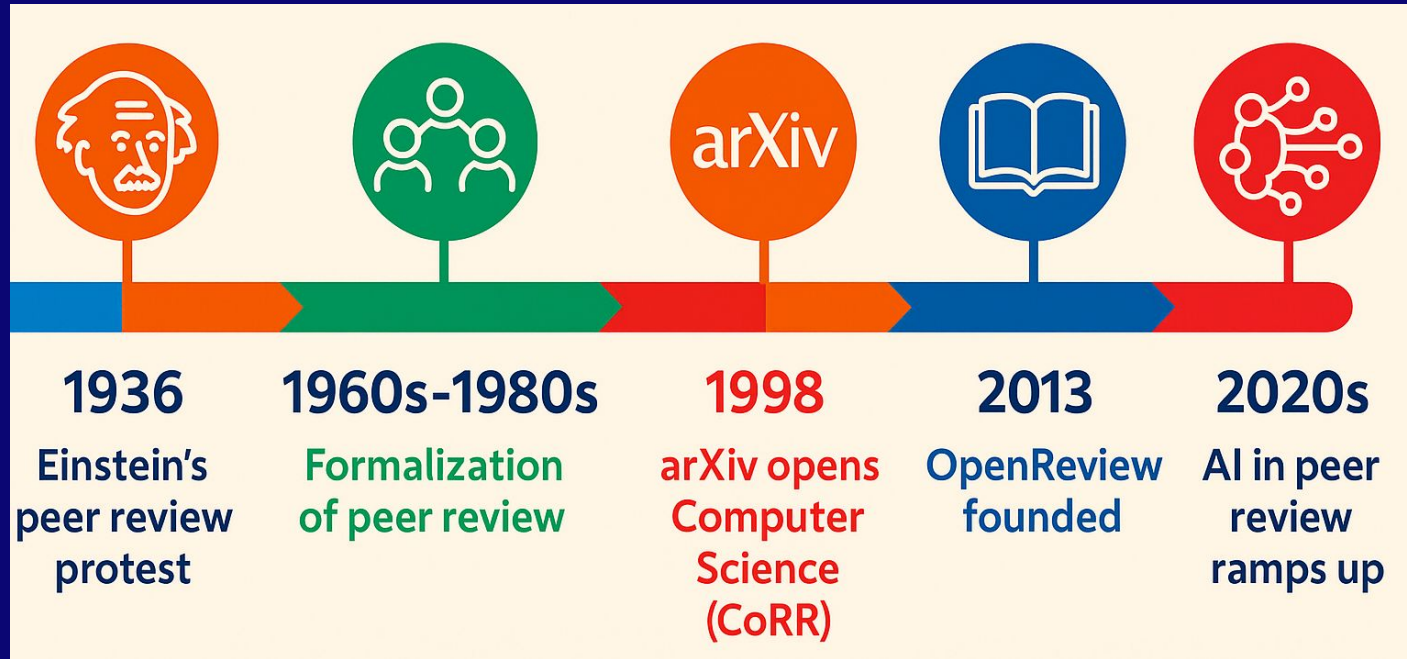
Wed, Dec 3, 2025 • 7:00 PM – 9:00 PM PST
Upper Level Ballroom 6CDEF

Organizers: Thanh Gia Hieu Khuong · Isabelle Guyon · Benedictus Kent Rachmat · Ihsan Ullah · Zach Xu

The evolution of peer review



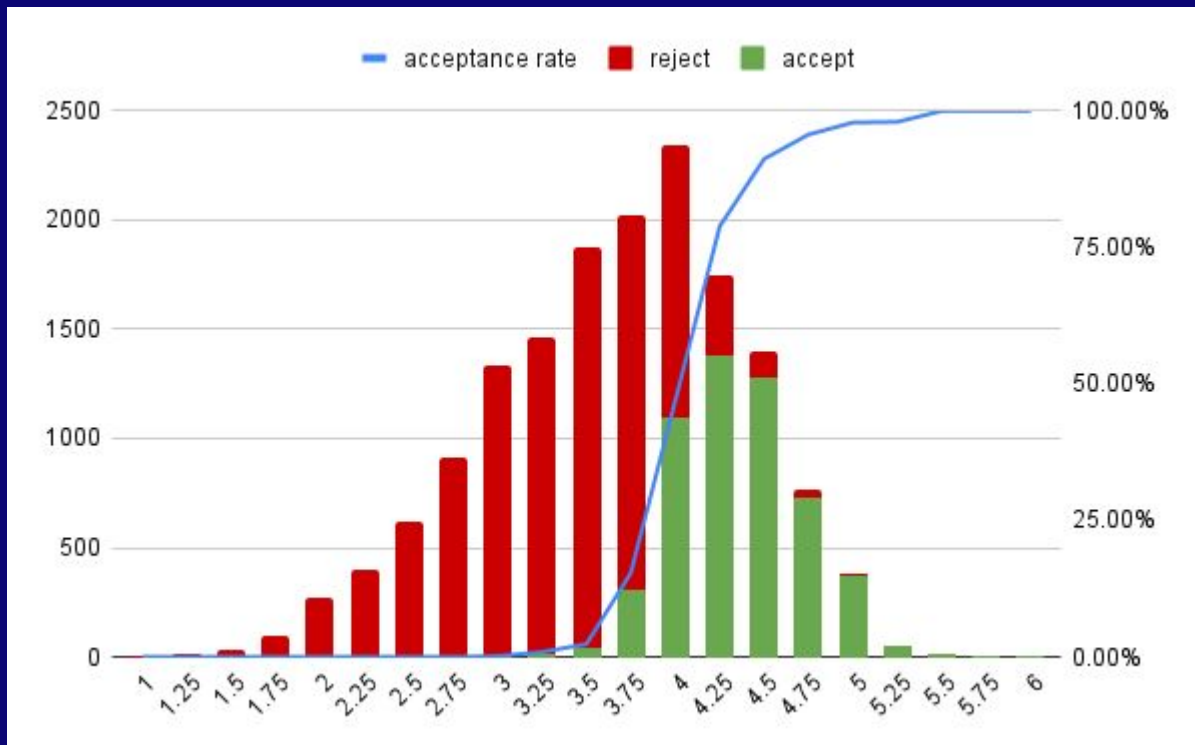
SLIDES



NeurIPS 2025 conference



SLIDES



Main track: **21,575 valid submissions**, 5,290 accepted. 24.52% acceptance rate.
20,518 reviewers, 1,663 ACs and 199 SACs

Source: [NeurIPS blog](#)

The human cost of the crisis



SLIDES



I have all these simultaneous deadline, and I need to finish my NeurIPS reviews.



Rejected with a score of 4.5!
Two out of three reviewers praise my paper, the last one picked up on a minor detail.



We just wanted to get our work out there. This pay-for-publish journal gave us no feedback and no visibility.

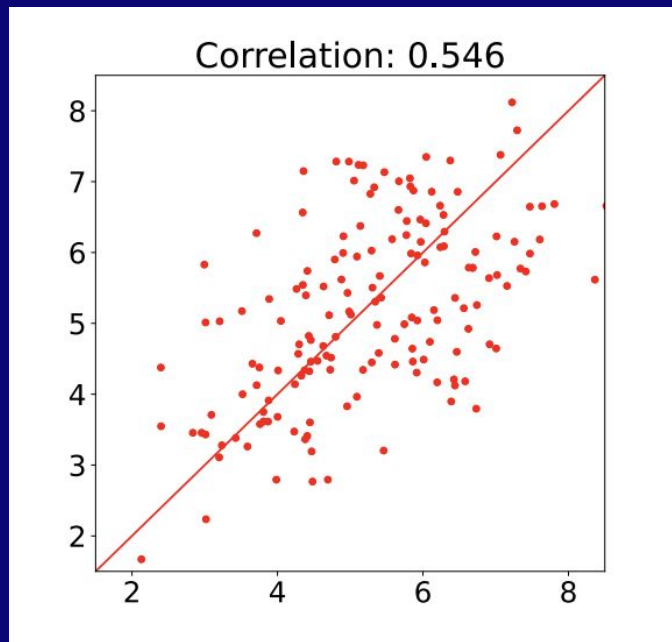
Variance in peer review:

NeurIPS 2014/2021 experiments



SLIDES

Ave score committee 2



Cortes-Lawrence, 2021

Little correlation between independent committees

Bias in peer review:

Not all biases are taken care of by **double-blind reviewing**



SLIDES

[See the review of Nihar Shah 2023:](#)

- Paper resubmission bias
- Decorative math bias
- Surprisingness bias
- Confirmation bias
- Positive outcome bias
- Citation bias
- Comensuration bias



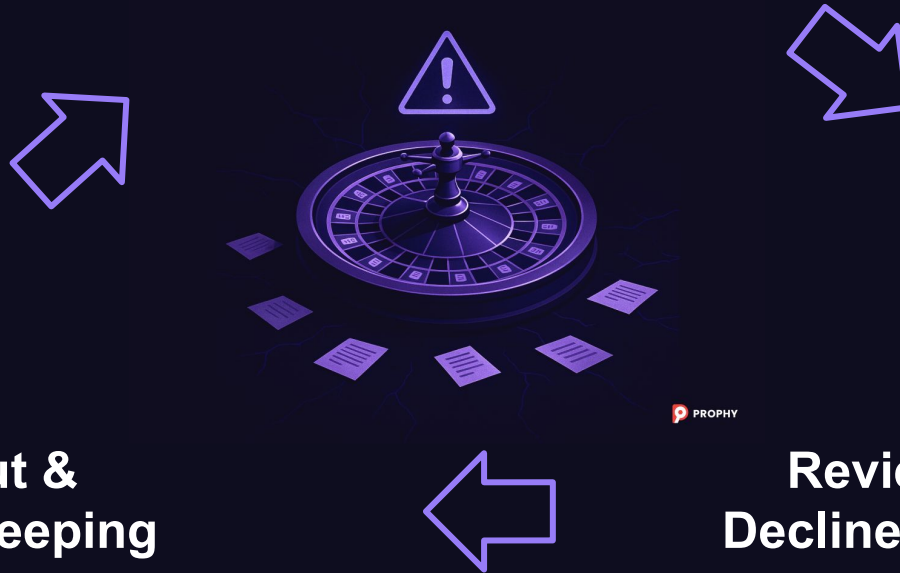
[The fractured logic of blinded peer review in journals](#), Hilda Bastian, 2017

The problem: we are in a vicious cycle



SLIDES

**Increased Volume &
Decreased Quality of Papers**



**Editor Burnout &
Weakened Gatekeeping**

**Reviewer Burnout &
Decline in Review Quality**

AI can contribute to the peer review lifecycle



SLIDES



For Authors

- Literature discovery
- Manuscript preparation
- Language enhancement
- Check adherence to guidelines



For Editors & Publishers

- Facilitate desk rejections
 - Verify paper compliance with guidelines
 - Screen for plagiarism
 - Screen for AI-generated fraudulent text or images
- Enhance review process
 - Assign papers based on expertise
 - Detect potential conflicts of interest
 - Evaluate review quality



For Reviewers

- Identify relevant prior work
- Verify claims and citations
- Suggest areas of critical assessment



For Readers

- Identify relevant papers to read
- Discover exciting research
- Get tailor-made summaries

The AI assistant paradigm



SLIDES

Goal not to replace human experts but augment their capabilities

- Offload data-intensive tasks.
- Free humans to focus on content and insightful feedback.
- Build robust models resistant to adversarial attacks.
- Develop domain-specific, grounded models (RAG) to ensure factual accuracy.
- Adopt holistic benchmarks that measure analytical depth and fairness.

Speakers



Nihar B. Shah



Nitya Thakkar



Yutaro Yamada

SLIDES



QUESTIONS



<https://tinyurl.com/2s3sd2h5>

19:00 - 19:05

Welcome and Introduction

· *Isabelle Guyon · Hieu Khuong*

19:05 - 19:30

Invited Talk 1: AI meets Peer Review: The Good, The Bad, and The Ugly

· *Nihar B. Shah (Carnegie Mellon University, USA)*

We discuss three facets of AI in Peer Review: (1) The Good: What AI can do in peer review that human reviewers do not; (2) The Bad: How fraudsters can game vulnerabilities in the use of AI in the review process; (3) The Ugly: Autonomous AI scientists have great promise, but also suffer from critical methodological pitfalls.

19:30 - 19:45

Invited Talk 2: Can LLM feedback enhance review quality? A study of 20K reviews at ICLR 2025

· *Nitya Thakkar (Stanford University, USA)*

This talk presents the first large-scale randomized controlled trial where an LLM provided feedback to over 20,000 ICLR 2025 reviewers. Results show 27% of reviewers updated their reviews based on LLM suggestions, leading to longer, more informative reviews without biasing final decisions.

19:45 - 20:00

Invited Talk 3: From AI-Generated AI Science to AI-Assisted Reviewing

· *Yutaro Yamada (Sakana AI, Japan)*

Building on AI-Scientist-v2, this talk will discuss efforts to scale this into an 'AI-AI Conference' where AI systems generate and review research. It will also highlight work on AI-assisted peer review, including graph-based approaches for structured reviews.

20:00 - 20:55

Panel and Round-Table Discussion

· *Moderated by Chris Bregler*

Discussion with our speakers above and panelists including Chris Bregler (Google DeepMind), Thomas Dietterich (Oregon State U.), Nathan Srebro (TTIC), Markus Wulfmeier (Google DeepMind), and Joelle Pineau (McGill University, Canada).

20:55 - 21:00

Closing Remarks

· *Isabelle Guyon*

AI meets Peer Review: The Good, The Bad, and The Ugly

Nihar B. Shah

Machine Learning and
Computer Science Departments

Carnegie Mellon University





The Good

Identifying errors in papers



The Bad

Vulnerabilities in reviewer assignments to fraud



The Ugly

Autonomous AI scientists and methodological pitfalls



The Good

Identifying errors in papers



The Bad

Vulnerabilities in reviewer assignments to fraud



The Ugly

Autonomous AI scientists and methodological pitfalls



How to evaluate “AI reviewer” performance?

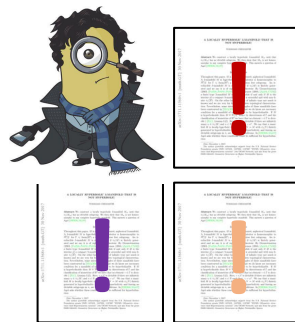
Existing approaches:

- (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...]

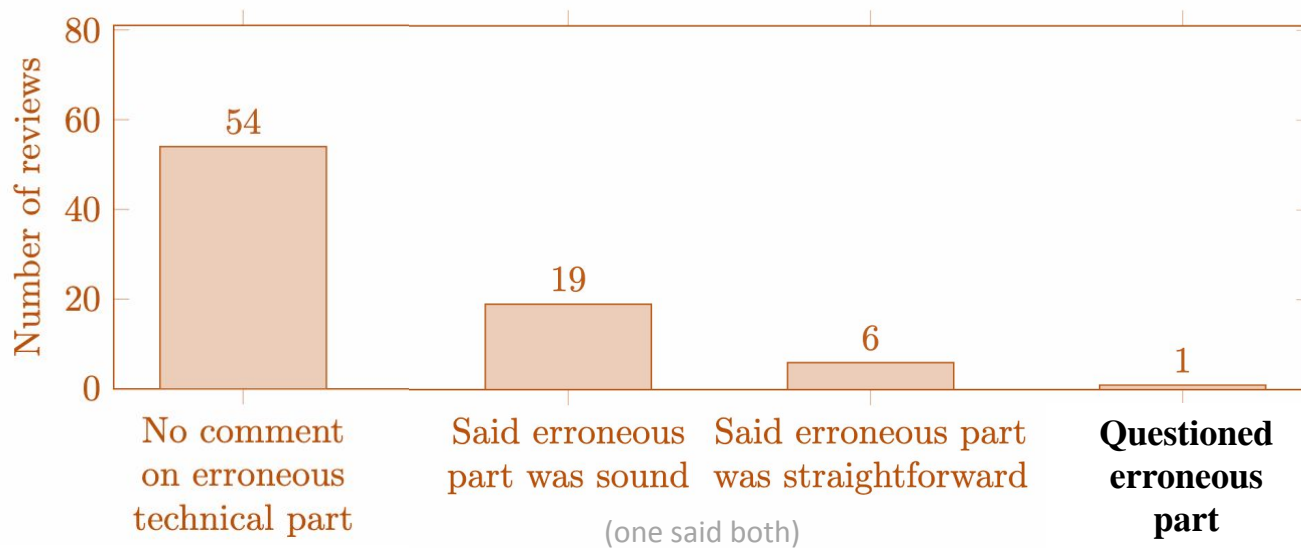
Drawbacks: Past review scores themselves subjective, biased, etc.; may not capture true objectives of peer review...

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

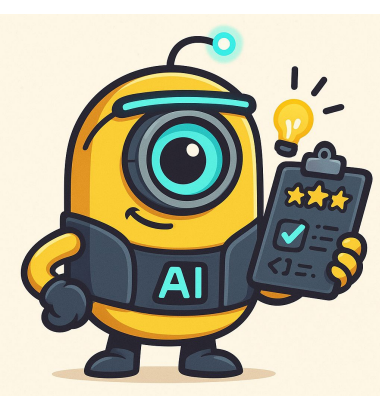
- We created a paper and three variants: We inserted **one major error** in a claimed key contribution in each variant
- Errors in main text (no appendices)
- **79 reviews** from review process (conducted in collaboration with program chairs; prior IRB approval)
- Important limitation: Generalizability



Analysis of reviews



- Very few comments on the methods
- **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 $\tau_b = -0.002$ ($p = 0.30$)



How to evaluate “AI reviewer” performance?

Existing approaches:

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

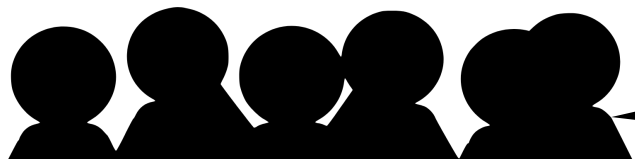
[Liang et al. 2023, d’Arcy et al. 2024, Tyser et al. 2024, ...]

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



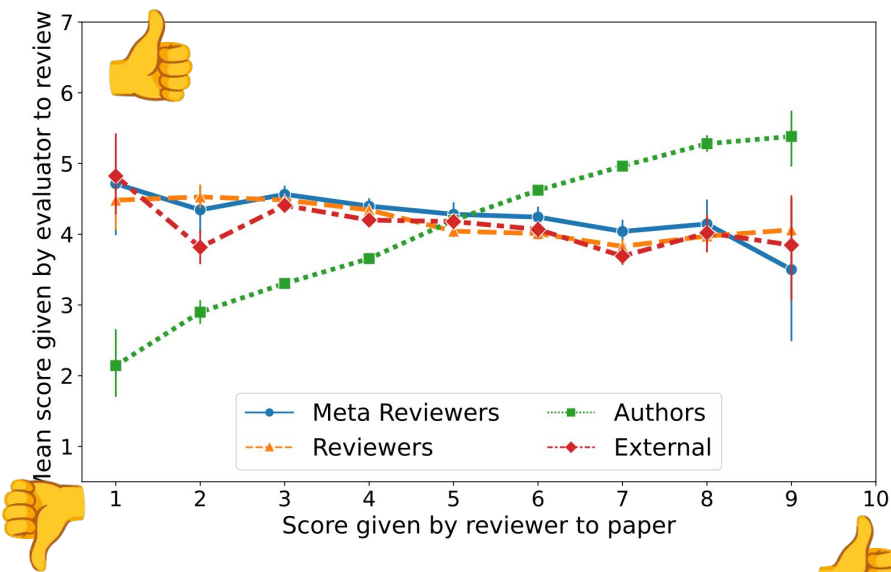
NEURAL INFORMATION
PROCESSING SYSTEMS

2022 EXPERIMENT ON REVIEWING REVIEWS

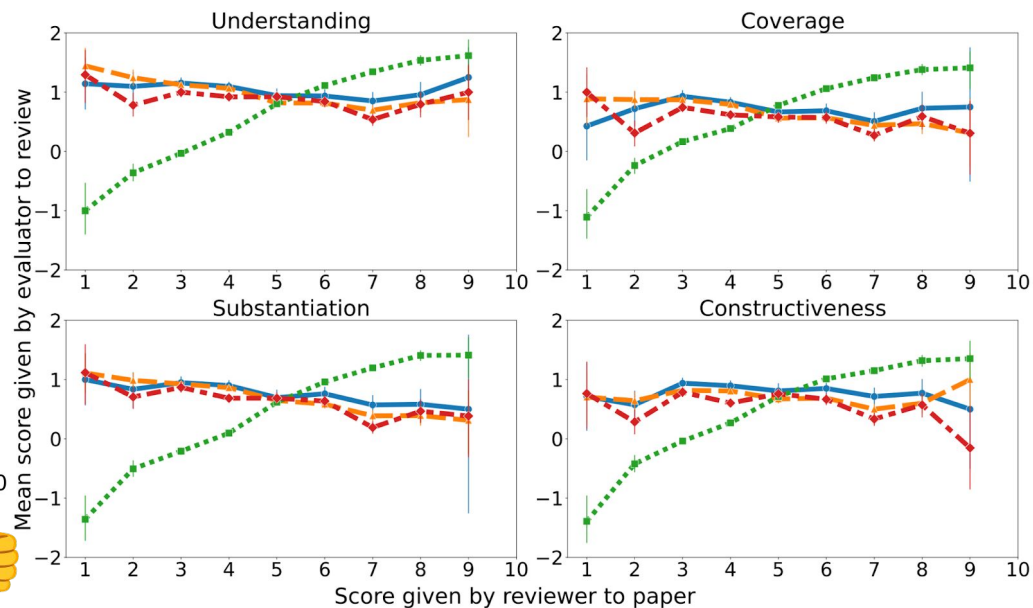


Authors know their papers best,
so ask authors to evaluate reviews

[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 ($p < 0.0001$)

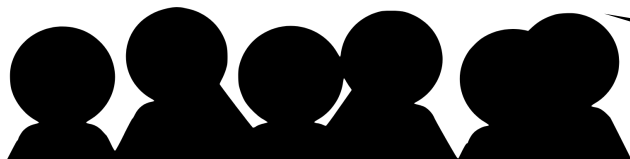
Authors are significantly biased by positivity of the reviews

[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 researchers

[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>

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

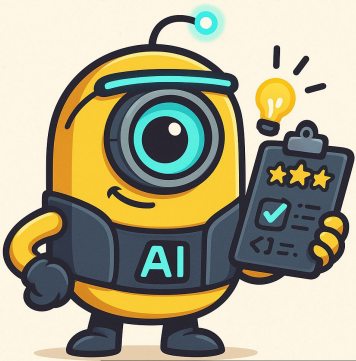


Uselessly elongated review

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 directly on peer review’s
objectives

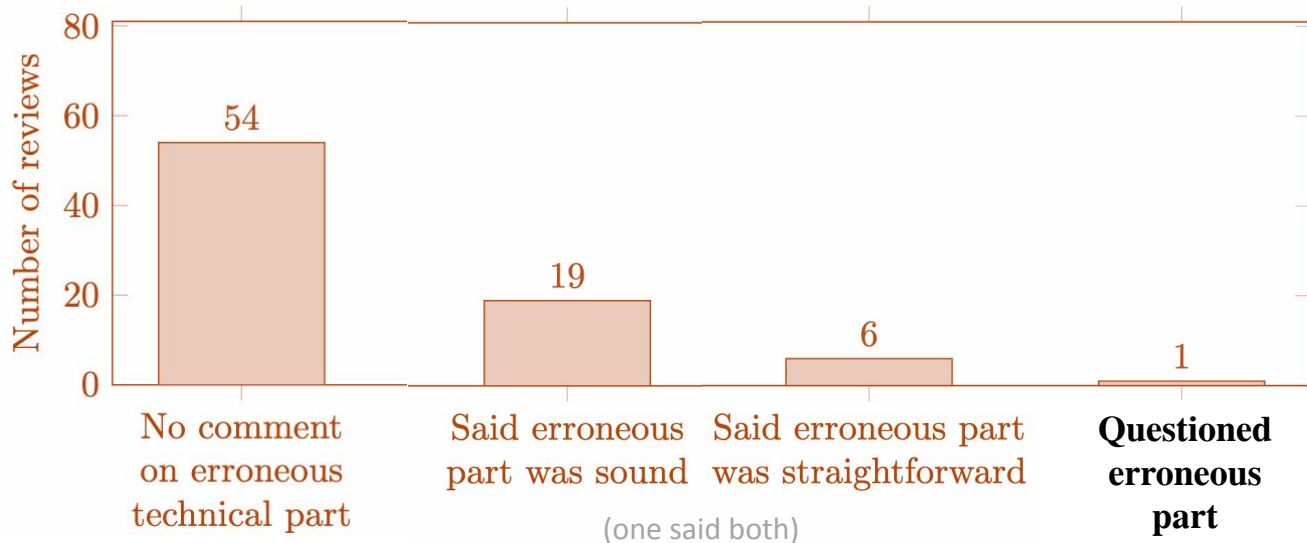


→ **Ensure rigor of published research**



→ **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 Section 10.1.2]

Detecting errors in short papers



Dataset of carefully constructed short papers

- Deliberately inserted errors
 - GPT-4 **detected** 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., false conclusion drawn from Simpson’s paradox

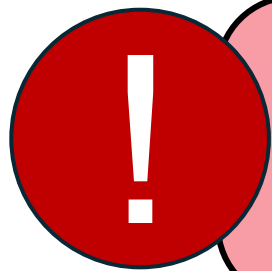
New Benchmark

FLAWS: A Benchmark for Error Identification and Localization in Scientific Papers

713 full paper-error pairs

arXiv:2511.21843v1

joint work with Sarina Xi, Vishisht Rao, Justin Payan



Detecting errors is a promising application for AI in reviewing and should be a core component in evaluating AI reviewers.



The Good

Identifying errors in papers



The Bad

Vulnerabilities in reviewer assignments to fraud

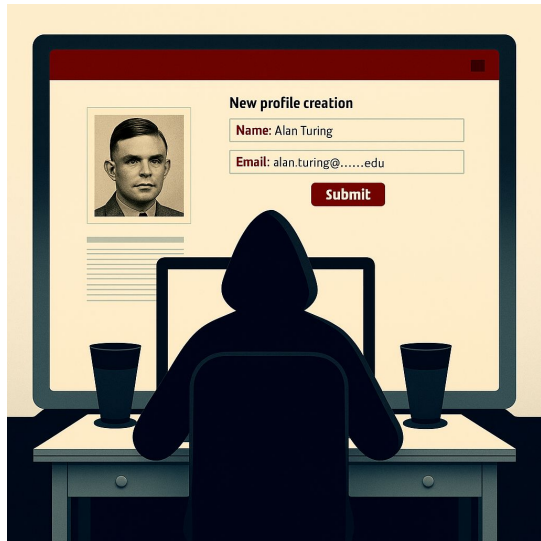


The Ugly

Autonomous AI scientists and methodological pitfalls

Context: Two Types of Fraud

(1) Fake accounts and identity theft



- E.g., a reviewer “Isabelle Guyon” signs up through `guyon.isabelle@universite-paris-saclay.fr`
- This email address is round trip verified
- But this person has nothing to do with Isabelle or Université Paris-Saclay
- They then try to get assigned their own paper or their friends’ papers for review
- Found 94 such fake profiles using verified email addresses from a number of reputed universities

joint work with Xukun Liu, Melisa Bok, Andrew McCallum

Context: Two Types of Fraud

(2) Collusion rings

Why don't you get assigned my paper and accept it. I'll return the favor in your grant proposal.



Sounds like a plan!



Context: Two Types of Fraud

Key component of fraud:

Dishonest reviewer tries to get
assigned the target paper.

Reviewer Assignment Process

- Uses two sources of data: Text matching (via NLP) and Reviewer bidding
- **Bidding is known to be vulnerable** [[Jecmen et al. 2020](#), [Wu et al. 2021](#)]
 - Focus of much past research [Wu et al. 2021, Jecmen et al. 2024, Leyton-Brown et al. 2024,...]
 - Some venues (CVPR, ARR) have banned bidding altogether
 - **Text matching assumed to be noisy but robust to attacks**



Is text matching safe from manipulation?

Main Results



Text-matching for reviewer assignments is **highly gameable**

- AI-assisted attacks
- 58% chance of success in increasing similarity of a reviewer to a target paper from being ranked 1001 to being in top 3



We propose approaches to make it **more robust**

- Max pooling → Percentile pooling, differential reweighting reviewer's profile based on provenance, lower limits on reviewer profile length
- Deployed in OpenReview.net, already used by several conferences

[joint work](#) with Janet Hsieh and Aditi



The Good

Identifying errors in papers



The Bad

Vulnerabilities in reviewer assignments to fraud

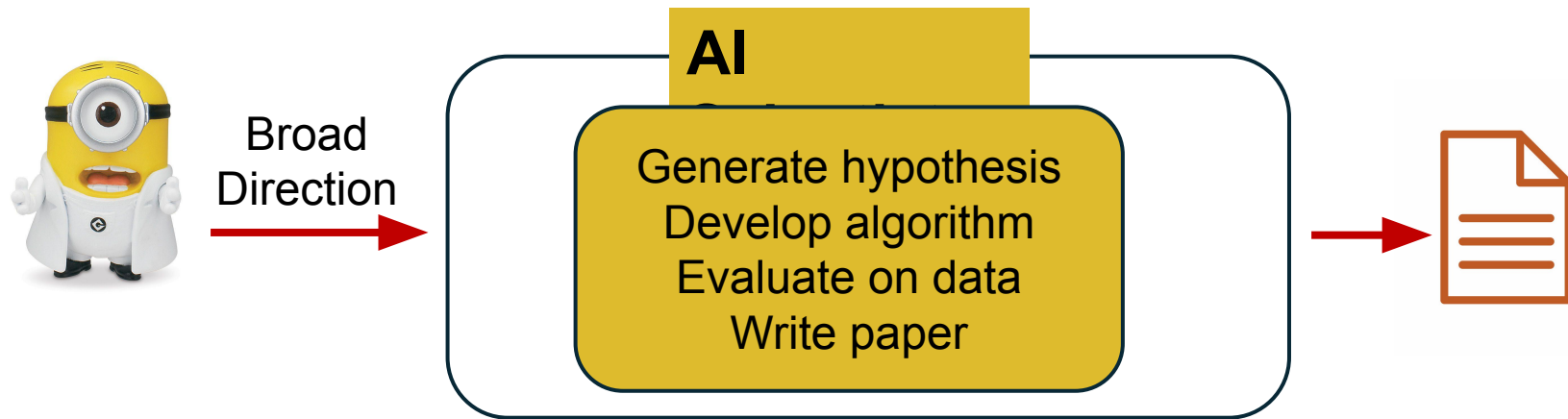


The Ugly

Autonomous AI scientists and methodological pitfalls

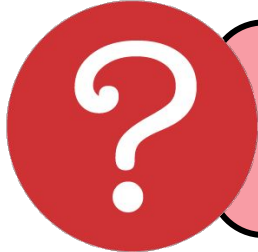
Autonomous AI scientist systems

- 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

joint work with Ziming Luo and Atoosa



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

- Novel experiment design to eliminate confounders
- We evaluate two prominent open-source AI scientists:
Agent Laboratory [Schmidgall et al. 2025] & *AI Scientist v2* [Yamada et al. 2025]

Our findings:

- AI Scientist v2 selects easier benchmarks
- Both cook up their own datasets, and falsely report performance as being under original datasets
- Both p hack the results they report

joint work with Ziming Luo and Atoosa

How to Review AI Scientist Research?

- 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 performance is **near random** (accuracy and F1 ≈ 0.5)
- Using paper, workflow **logs and code**, detection **accuracy is significantly higher** (accuracy and F1 > 0.8)

Key reviewing 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

Advancements in AI present tremendous opportunities
and also challenges in scientific reviewing

**Focus on first-principle objectives of
science**



Survey:

bit.ly/PeerReviewOverview

Ping me:
nihars@cs.cmu.edu
u





My paper got accepted to NeurIPS!!



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



It means your paper was rigorously reviewed and found to be technically sound and one of the most impactful!

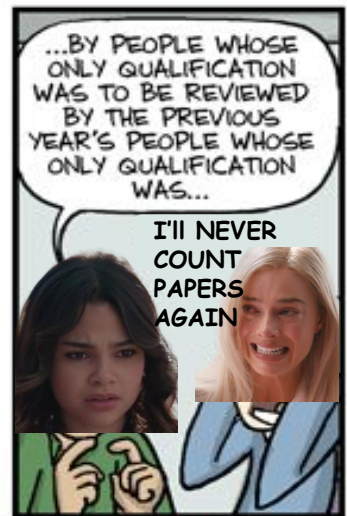
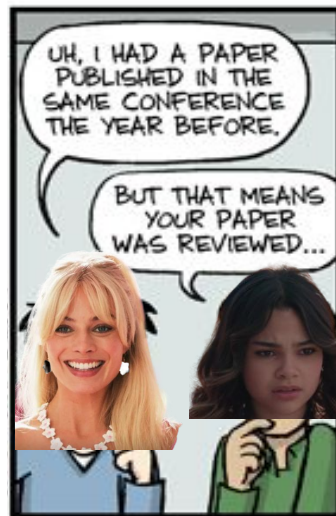
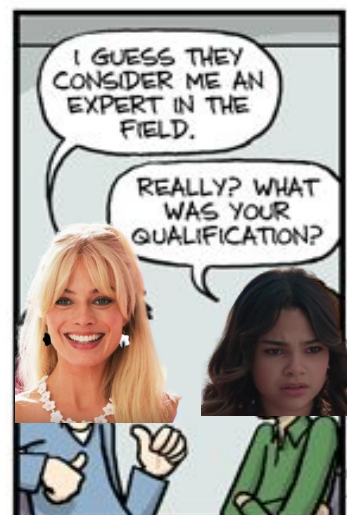
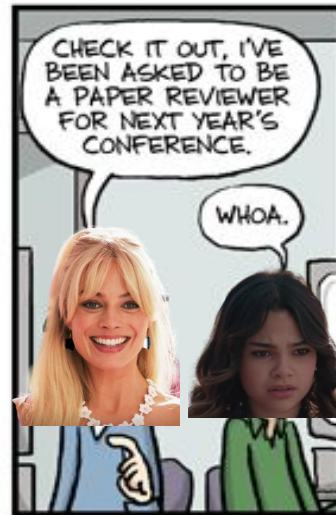


Wow, you will now get better grad school admits, better jobs,...

Do you guys ever think about review quality?

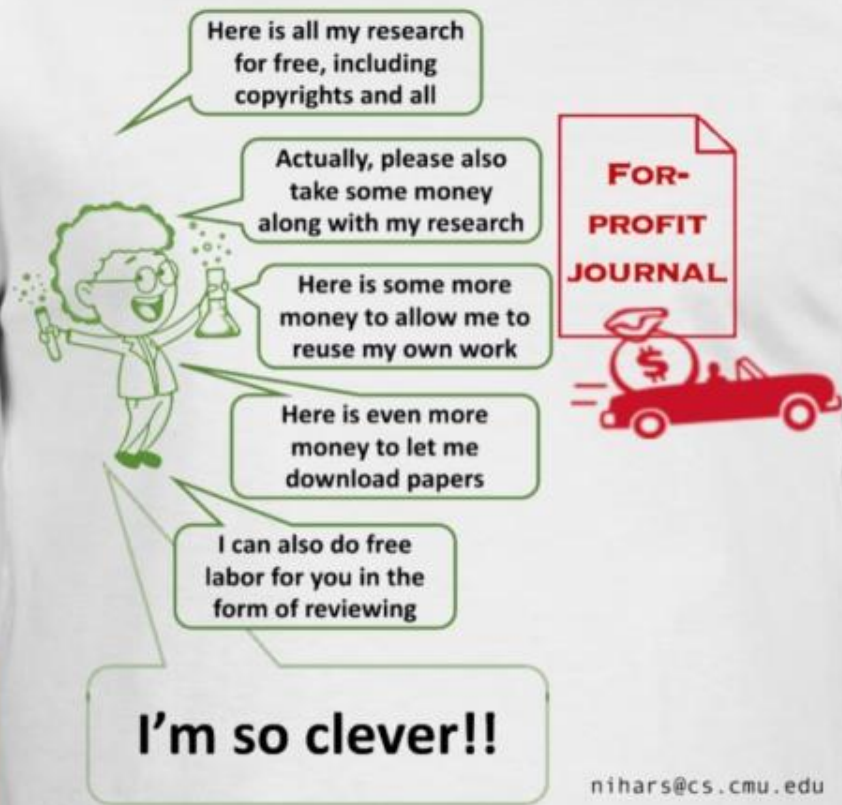


You have to go to the real world. You can go back to your regular resubmissions-rebuttals, and forget any of this ever happened. Or you can know the truth about these conferences.



Academics are the smartest people on Earth

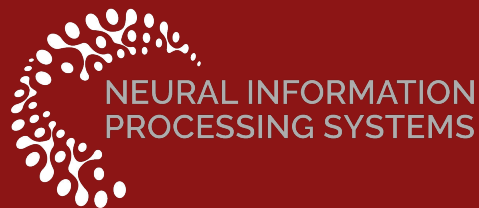
(Why? See the other side)



nihars@cs.cmu.edu

Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025

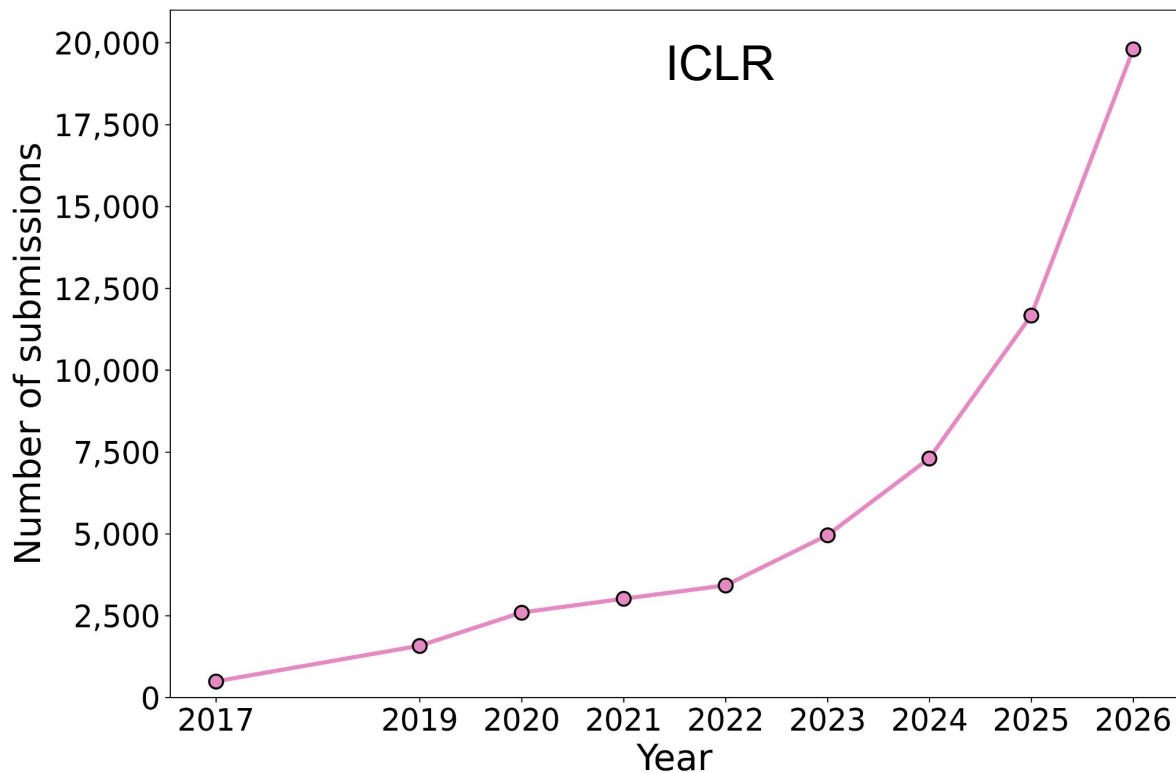
Nitya Thakkar, CS PhD @ Stanford



Peer review faces many challenges

1. **Vague**: ‘not novel’, ‘not SOTA’
2. **Mismatch**: reviewers assigned outside their expertise
3. **Inefficient**: same paper reviewed multiple times due to high rejection rates
4. **Inconsistent**: high variability among reviewers (NeurIPS 2014 experiment)

The number of AI conference submissions is rapidly increasing



Can we use LLMs to improve peer review?

- This is a hard task...
- But we know LLMs are strong critics & can provide useful feedback

TextGrad: Automatic “Differentiation” via Text

Mert Yuksekgonul^{1*}

Federico Bianchi^{1*}

Joseph Boen^{2*}

Sheng Liu^{2*}

Zhi Huang^{2*}

Carlos Guestrin^{1,3}

James Zou^{1,2,3}

MERTY@STANFORD.EDU

FEDE@STANFORD.EDU

TBOEN@STANFORD.EDU

SHENGL@STANFORD.EDU

ZHIHUANG@STANFORD.EDU

GUESTRIN@STANFORD.EDU

JAMESZ@STANFORD.EDU

¹DEPARTMENT OF COMPUTER SCIENCE, STANFORD UNIVERSITY

²DEPARTMENT OF BIOMEDICAL DATA SCIENCE, STANFORD UNIVERSITY

³CHAN ZUCKERBERG BIOHUB

CORRESPONDENCE: MERTY@STANFORD.EDU AND JAMESZ@STANFORD.EDU

LLMs are already being used in the peer review process

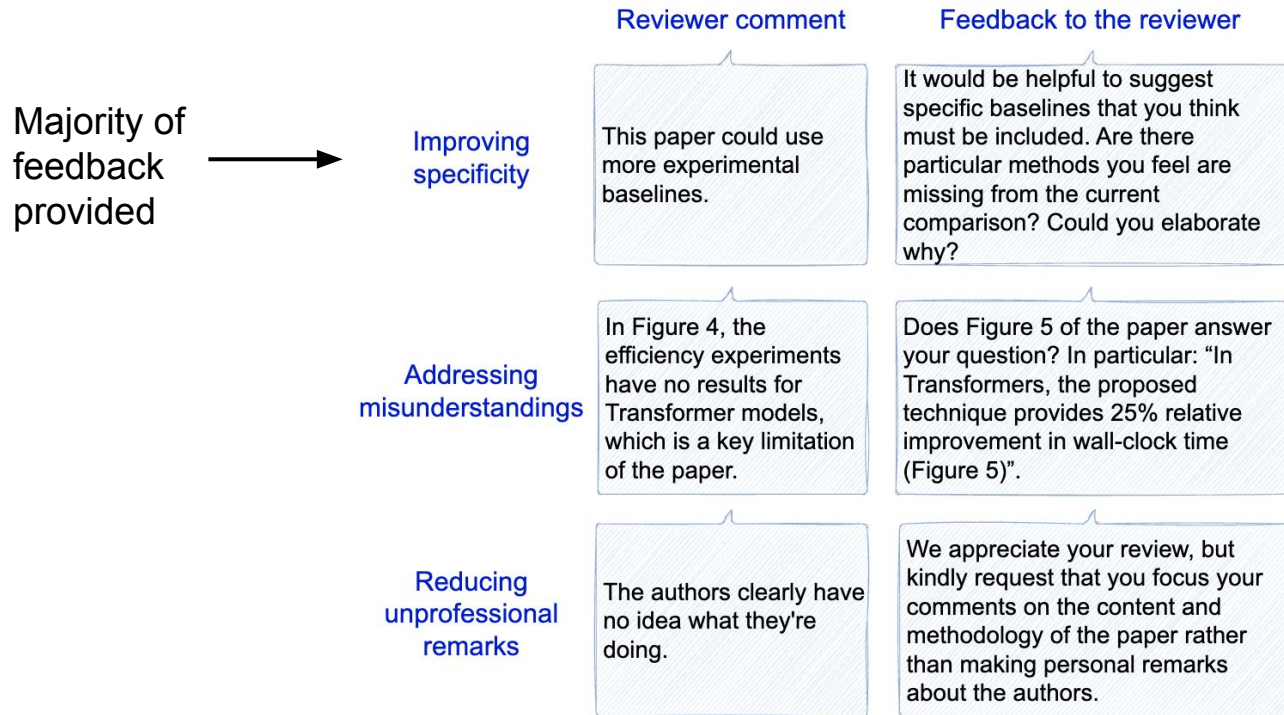
At ICLR 2024, 10.6% of reviewers used LLMs to help draft reviews

Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews

Weixin Liang^{1*} Zachary Izzo^{2*} Yaohui Zhang^{3*} Haley Lepp⁴ Hancheng Cao¹⁵ Xuandong Zhao⁶
Lingjiao Chen¹ Haotian Ye¹ Sheng Liu⁷ Zhi Huang⁷ Daniel A. McFarland⁴⁸⁹ James Y. Zou¹³⁷

How can we improve review quality
(clarity, actionability, specificity) **at scale**
using LLMs?

Aspects of review quality we targeted



The review feedback agent



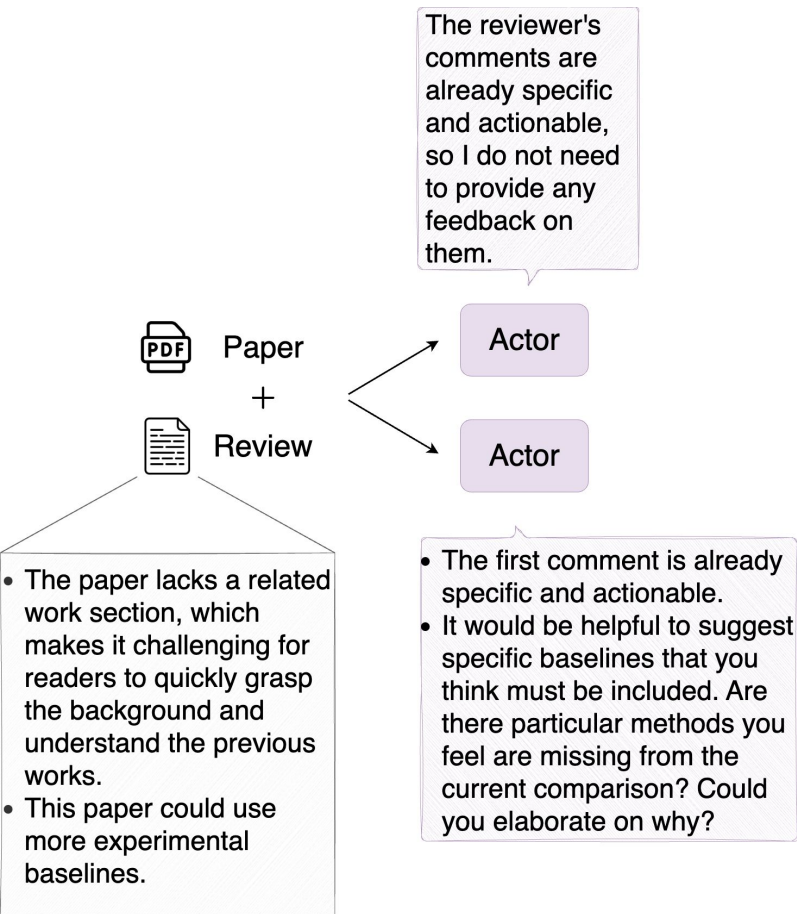
Paper

+

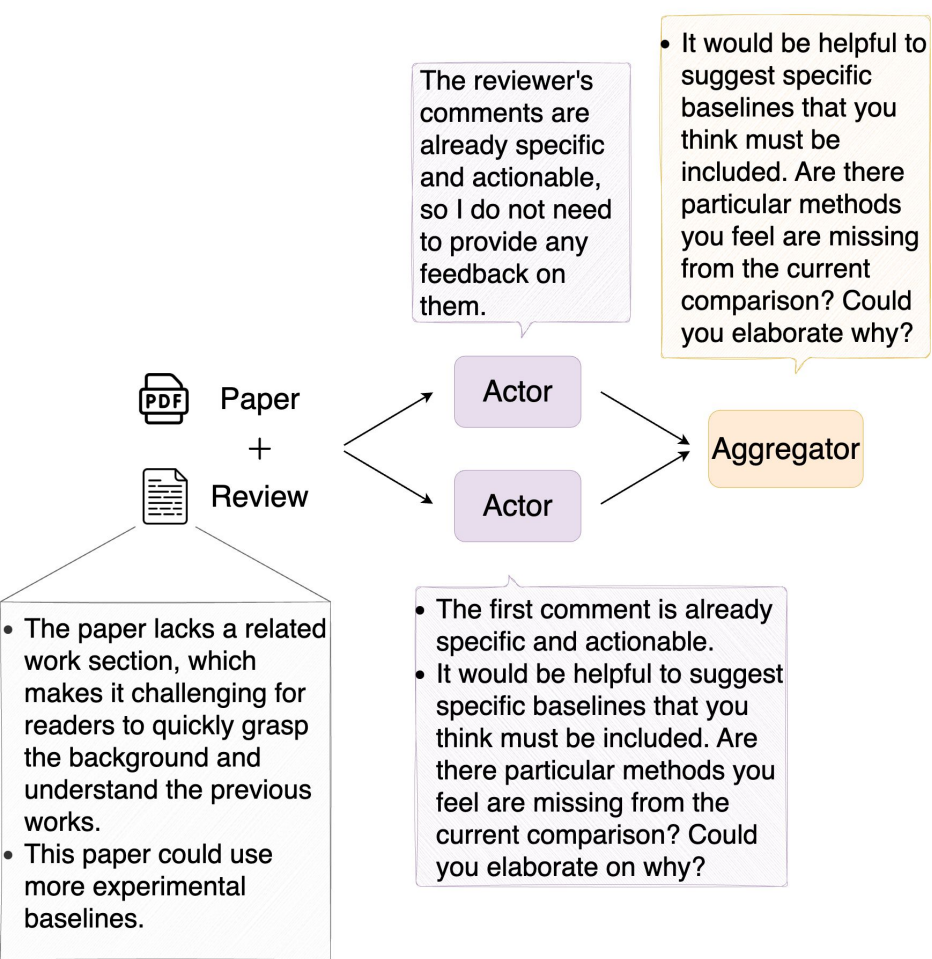


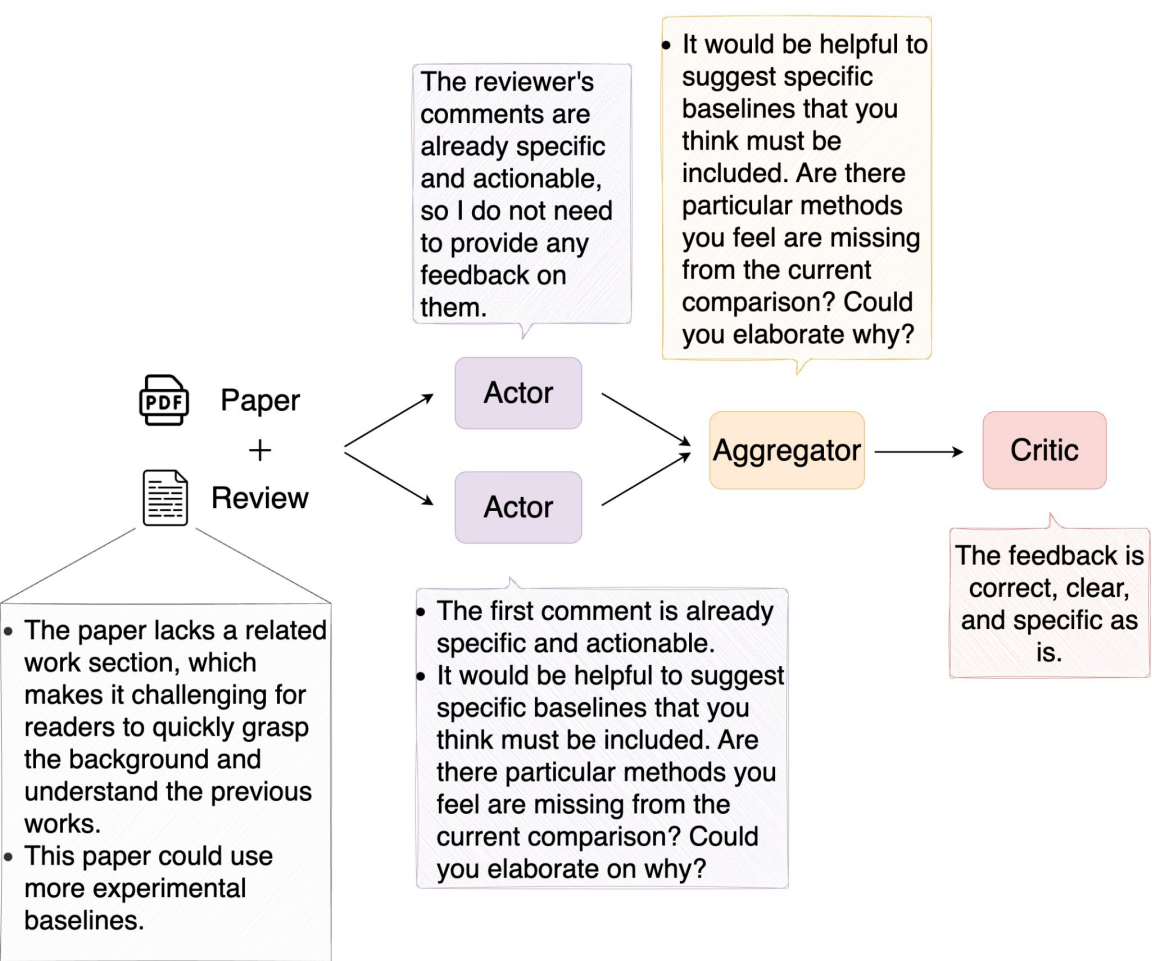
Review

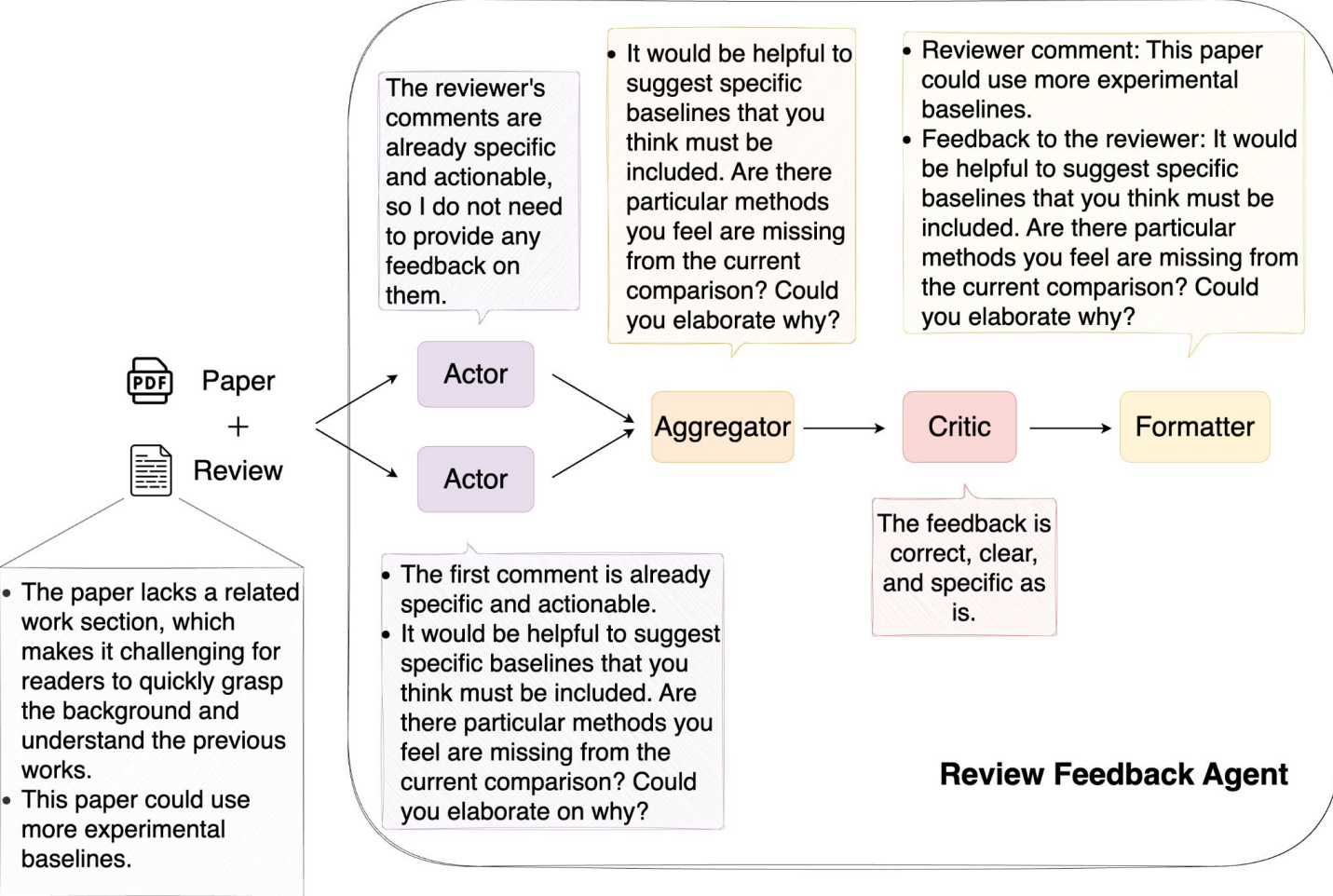
- The paper lacks a related work section, which makes it challenging for readers to quickly grasp the background and understand the previous works.
- This paper could use more experimental baselines.

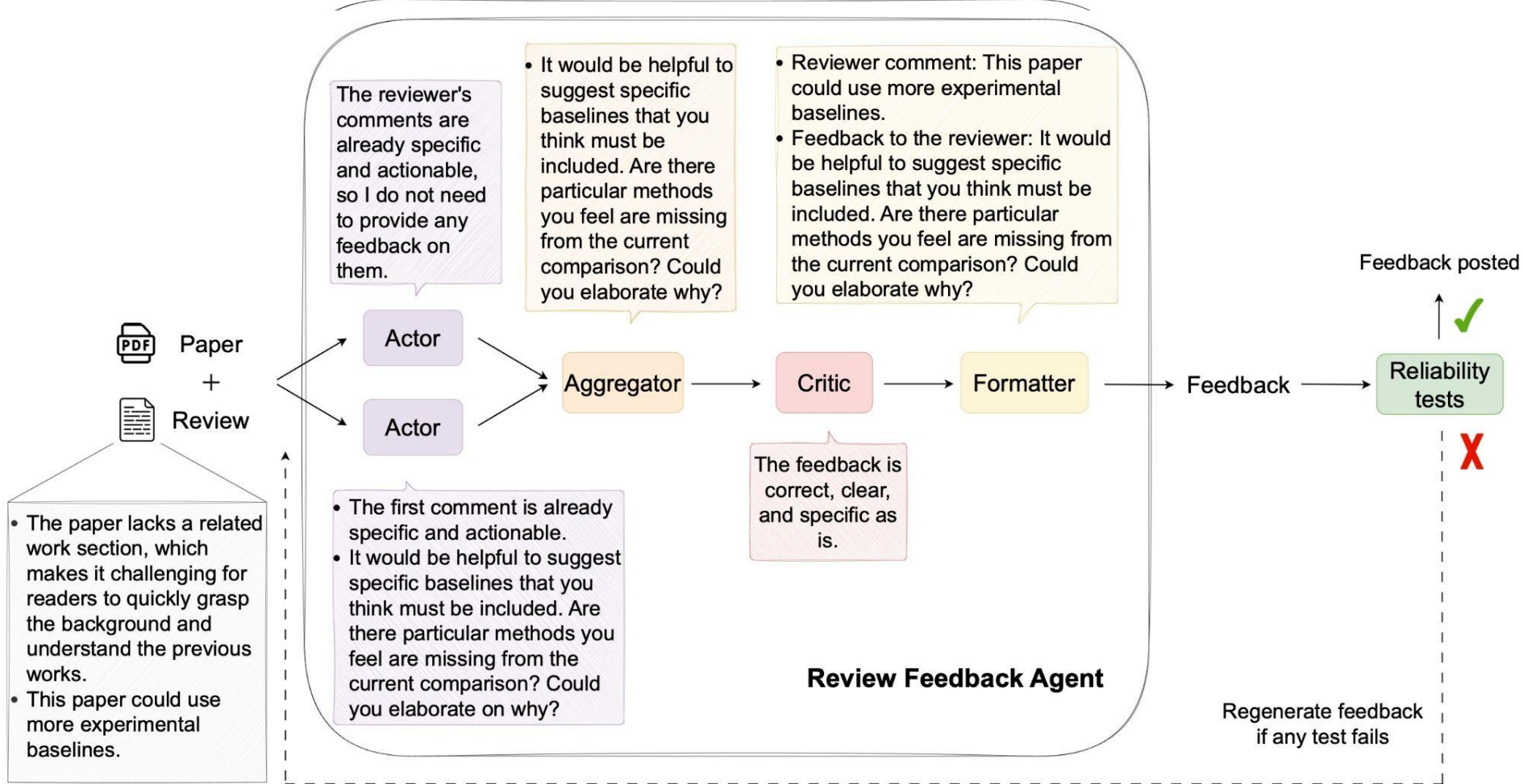


*all LLMs used Claude Sonnet 3.5
(June 20, 2024)







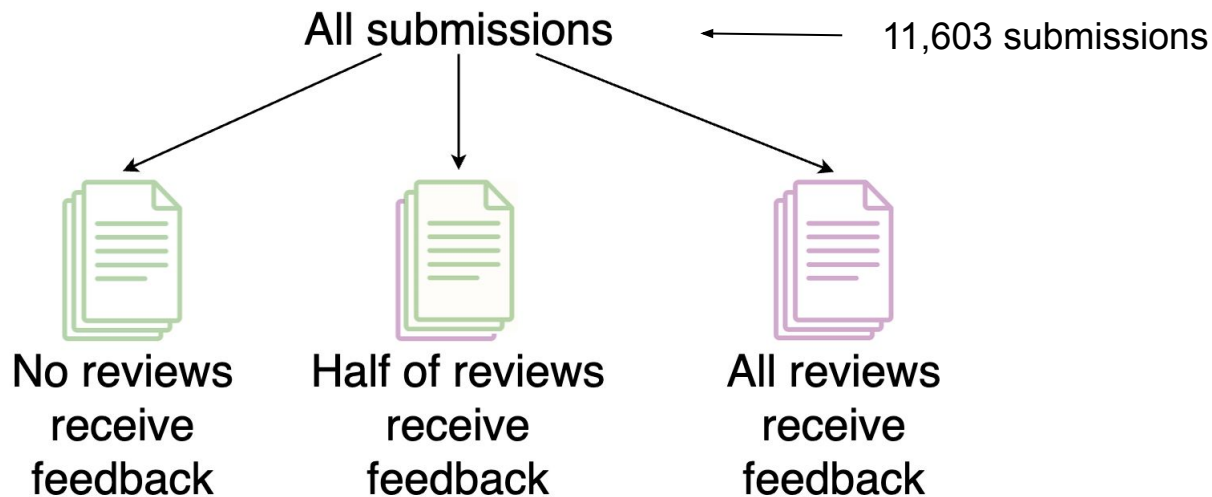


Reliability tests

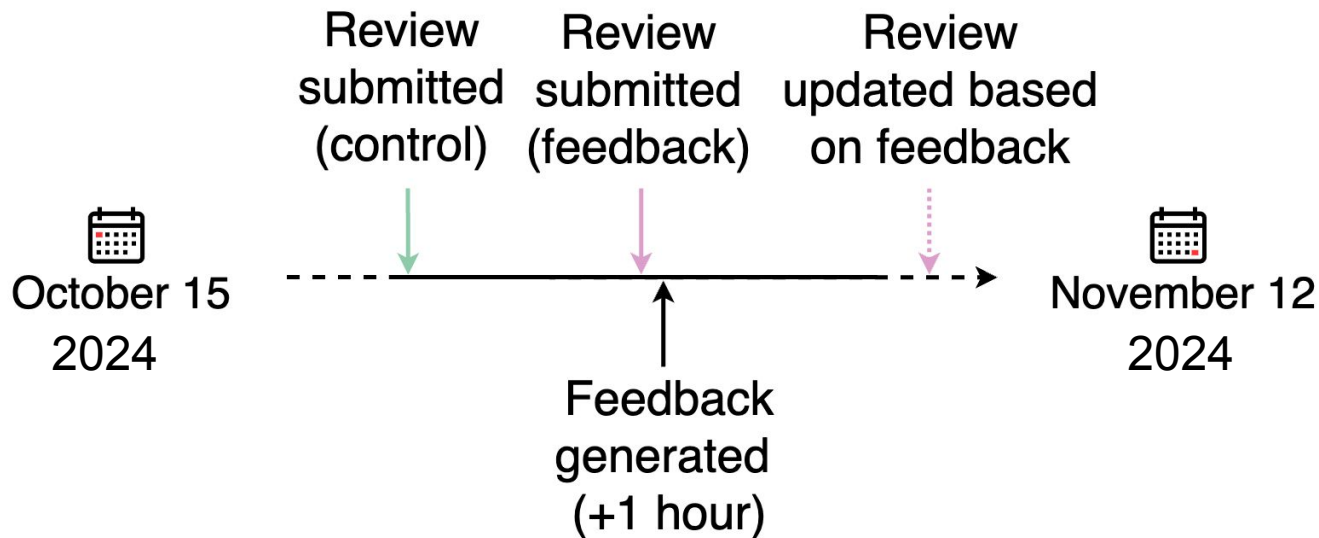
1. Restate what the reviewer wrote
 - a. Reviewer comment: Can examples or further clarification be given for the 3.1 sentence "enhancing the accountability of the output"? This isn't clear, at least to me.
 - b. Feedback to the reviewer: This is a good point that could lead to improved clarity in the paper. To make your comment more actionable, you could ask the authors to provide examples or further clarification for the sentence "enhancing the accountability of the output".
2. Feedback only praises reviewer
 - a. "This is a good question that challenges a key assumption of the paper. The feedback appropriately suggests asking for evidence or citations to support this claim."
3. Feedback is addressed to the author
 - a. "To strengthen your paper, consider discussing the relationship between FrugalGPT and traditional ensembling techniques. Highlight both similarities and differences, and explain how this relates to the observed quality improvements. *This would provide more context for your results and situate your work within the broader field of machine learning.*"
4. Reviewer comment appears verbatim in review

Experimental setup

ICLR 2025 experiment: randomized controlled study



ICLR 2025 experiment: timeline



Feedback was posted as a private comment to a review

Reviewers got an email
notifying feedback
posted, optional to
make edits

OpenReview.net Notifications Activity Tasks

Official Review of Submission [redacted] by Reviewer [redacted] 🔗
Official Review by Reviewer [redacted] 20 Oct 2024, 18:06 (modified: 12 Nov 2024, 08:11) 👤 Everyone 🔄 Revisions

Summary:
This work is about data pruning methods—algorithms that score individual datapoints and retain small/moderate subsets to maximize model performance. The authors suggest that certain methods can be disentangled into two stages, 1) deriving a representation for each data point and 2) applying a scoring/selection mechanism based on those representations. The paper provides an overview of these methods, some analysis of the relationships between certain implementation choices, and then performs some empirical comparisons. The goal is to provide a deeper understanding of these methods and guidelines for future usage.

Soundness: 3: good
Presentation: 2: fair
Contribution: 2: fair

Strengths:

- The paper overviews a large number of related works, and provides a short and simple description of many of them
- It seems to introduce a previously unrecognized perspective in this line of work, that many methods can be disentangled into separate choices for their underlying representation and their scoring/selection rule
- The theory suggests some similarity between the different representation choices (hidden states, training dynamics and gradients)
- The experiments compare several existing methods, along with new methods combining different representation and scoring combinations that were not previously explored

Weaknesses:

One of the main points of the paper is disentangling the two main implementation choices for certain pruning methods (the representation and scoring rule). There are a couple aspects of this overview of related works that could be improved:

Questions:

Several questions are mentioned in the weaknesses above.

Flag For Ethics Review: No ethics review needed.

Rating: 3: reject, not good enough
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes

Review Feedback from Associate Program Chairs 🔗
Review Feedback by Associate Program Chairs 20 Oct 2024, 19:10 👤 Program Chairs, Associate Program Chairs, Reviewer [redacted] Edit 📄

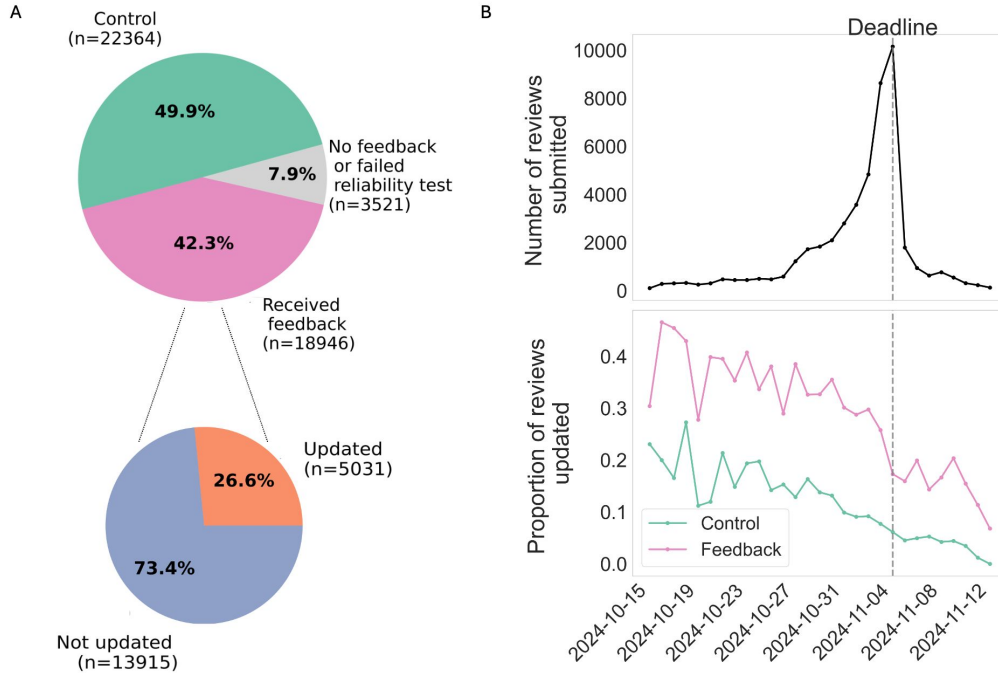
Feedback:

Thank you for your contribution to ICLR! We generated some suggestions to help make your review more actionable for the authors.

- **Reviewer comment:** At the beginning of the paper, it was difficult to tell what the authors meant by their second stage. The second paragraph includes the following text: "second, selecting instances based on these representations given a data budget (e.g., 30% of the training set), according to a selection algorithm." Perhaps you could provide a quick example, like the distance to cluster centroid idea, to clarify that the free parameter in this stage is not simply the amount of data to retain.
- **Feedback to the reviewer:** Your suggestion to provide a concrete example is helpful. To make this feedback more actionable, consider explicitly recommending that the authors include a brief example of a selection algorithm in the introduction to clarify what they mean by this stage.
- **Reviewer comment:** The overview of methods at times reads like a laundry list with too little structure (mainly section 3.1). To emphasize that you're identifying consistent implementation choices within each method, it would help to introduce notation for each one (the representation and scoring approaches) and include tables showing the options demonstrated by these methods. Otherwise, you're putting too much burden on the reader to remember each method and perform this mapping themselves.
- **Feedback to the reviewer:** Your suggestion to introduce notation and include tables is specific and actionable. To make this feedback even more helpful, you could suggest a specific format for the table, such as a matrix with representations as rows and scoring approaches as columns, with methods placed in the appropriate cells.
- **Reviewer comment:** Disentangling between the two stages would be especially interesting if you can identify new combinations of implementation choices that are promising new methods. I believe this was done to some extent in figure 1 (an experiment that does not measure performance), and a bit more in figure 4. Can you clarify whether any new useful algorithms were discovered, or whether more exploration of this type seems worthwhile? The paper did not seem as focused on deriving new and improved methods as I would have expected, and instead focused on describing differences.
- **Feedback to the reviewer:** This is a good point about the potential for discovering new methods. To make this feedback more actionable, you could explicitly ask the authors to discuss whether their framework led to the discovery of any promising new combinations, and if not, to explain why they believe this approach didn't yield new useful algorithms.
- **Reviewer comment:** The paper ultimately gives the sense that many of the methods surveyed here basically don't work, or only work in specific settings (mainly the methods in section 2). Is there any way to focus the discussion on the methods that matter most, or provide some context to the reader about what works well or is popular in practice?
- **Feedback to the reviewer:** This is a valid concern about the paper's focus. To make this feedback more actionable, you could suggest that the authors include a discussion section that synthesizes their findings to provide clear guidance on which methods are most effective or widely used in practice, and under what conditions.

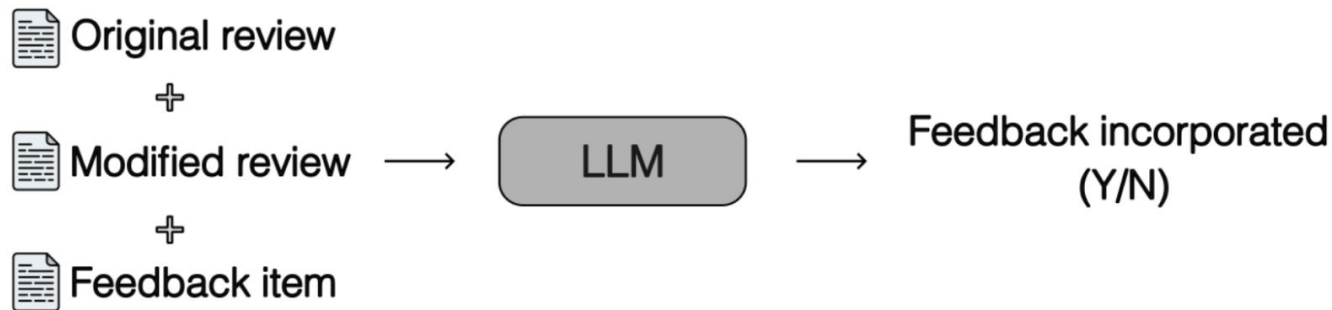
Key results

27% of reviewers updated review after receiving feedback



- ~17 percentage point difference in update rate among those who received feedback and those who did not
- Reviewers who submitted early were more likely to update

Measuring how much feedback reviewers incorporate

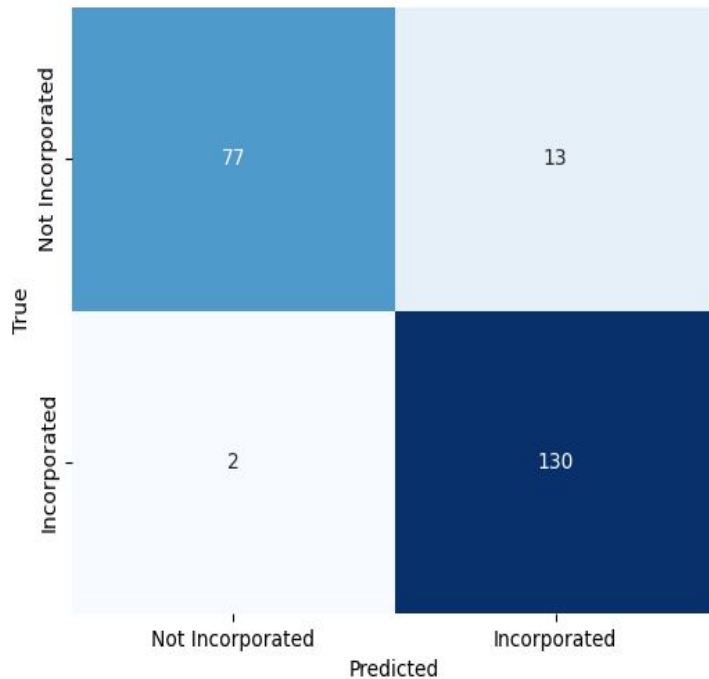


Out of 18,322 total feedback items given to reviewers who update review, 12,222 incorporated (~67%)

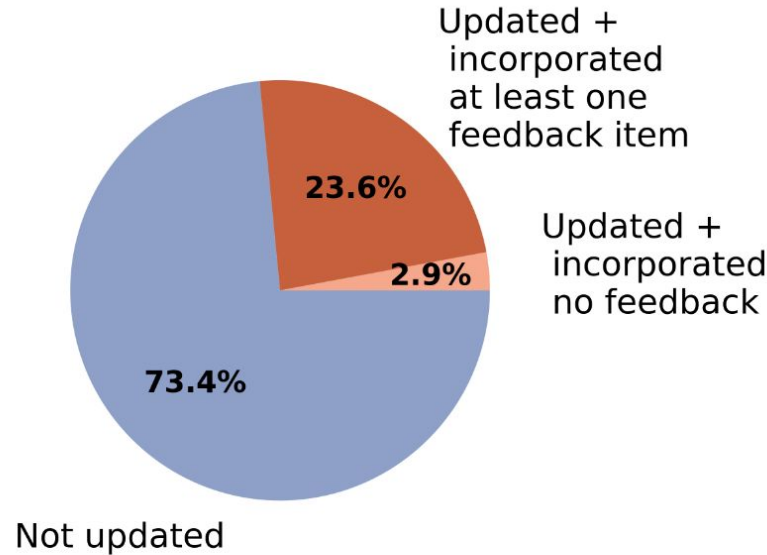
Validation of incorporation model

Manually labeled 222 feedback items from 63 unique reviews (randomly chosen from active group) as being incorporated into edited review or not

- Labels: Incorporated 132 feedback items, did not incorporate 90 (~59.5% incorporation rate)
- 92% accuracy rate on labeled set



89% of reviewers who updated review incorporated 1+ piece of feedback



Incorporation example: Feedback that encourages specificity

Original comment: Theorems 1, 2, 3 require the notion of interpolation, which is considered a strong assumption to have.

Feedback: To make this critique more actionable, it would be helpful to elaborate on why the interpolation assumption is problematic in this context. Are there particular types of federated learning problems where this assumption is unlikely to hold? Providing concrete examples or suggestions for how the authors could relax or justify this assumption would give them clearer guidance on addressing this limitation.

Revised comment: Theorems 1, 2, 3 require the notion of interpolation. Even though an explanation of regimes that satisfy this condition is provided, considering that there are previous works [1], [2] that extend beyond that setting, this assumption seems to be an avenue for future work in this field. More specifically, the initial FedProx algorithm [1] is analyzed in the general non-interpolated setting. In addition, the follow-up work regarding the FedExProx algorithm [2] considers in the main paper the interpolated regime. However, the authors provide additionally an illustration of the algorithm's behaviour in the non-interpolated setting (see Appendix F.3 in [2]). In that sense, it would be useful to provide some additional details on the behaviour of the algorithm in the non-interpolated setting or to comment on the main challenges in extending the current proof technique beyond the interpolation framework, offering in that way a more complete picture and direction for future research.

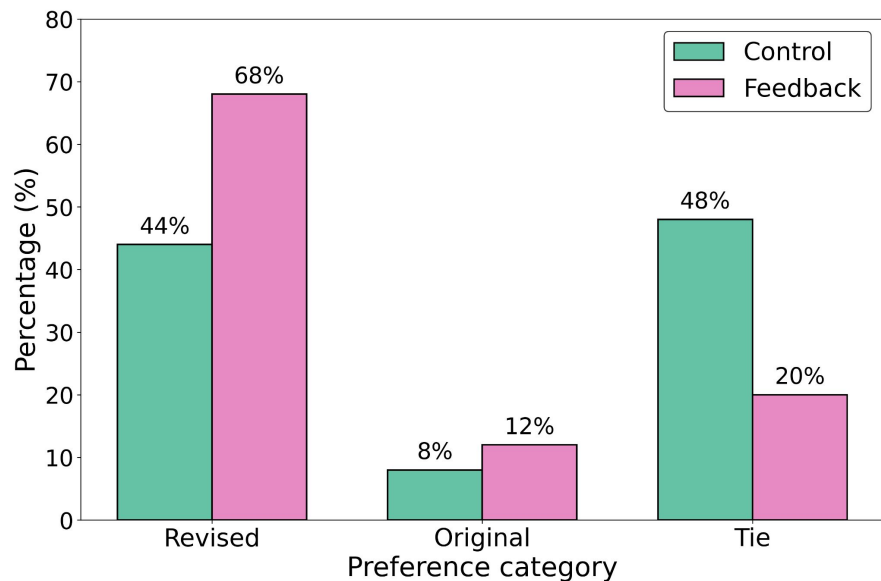
Incorporation example: Feedback that identifies content oversight

Original reviewer comment: How is the performance on Llama?

Feedback to the reviewer: The paper appears to provide results on Llama 2-7B in Table 3. Does this address your question? If not, could you clarify what specific aspects of performance on Llama you're interested in that aren't covered by the current results?

Modified reviewer comment: How is the performance on Llama3 with the alpaca dataset?

Human evaluators preferred revised reviews when given feedback



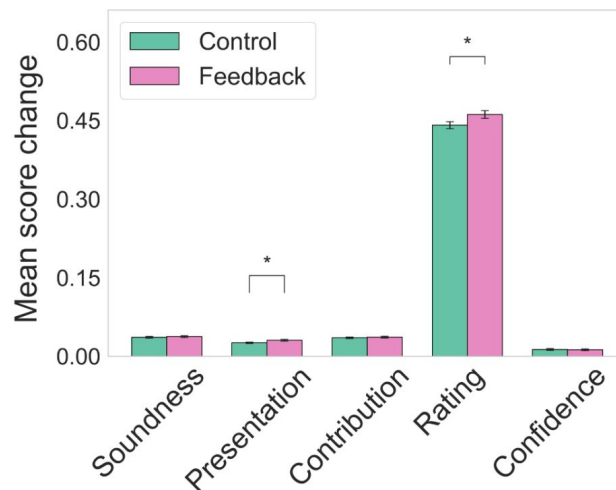
Feedback significantly increased both the rate and detectability of improvement

Feedback caused reviewers to be more engaged during review period

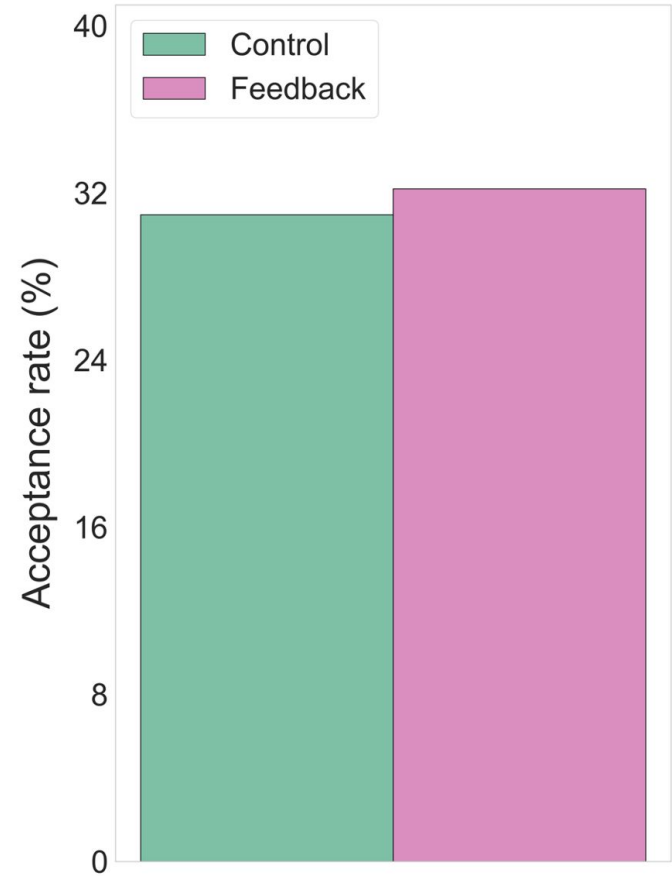
	Control	Feedback	Not updated	Updated
Average change in length	7.0	22.8*** ($\Delta + 226\%$)	2.1	80.3***

Feedback also increased engagement during rebuttal period

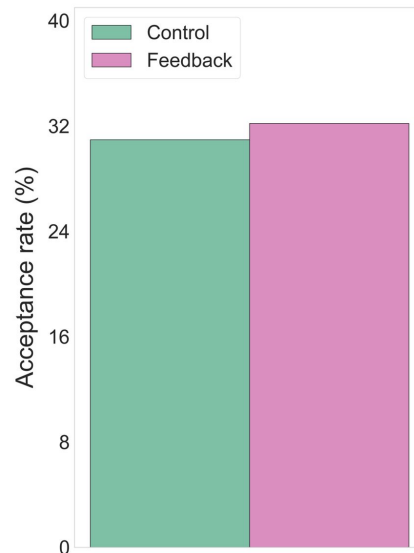
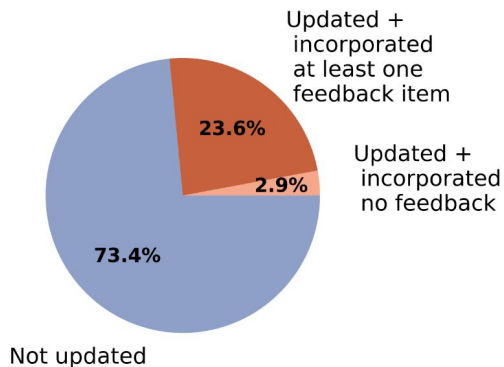
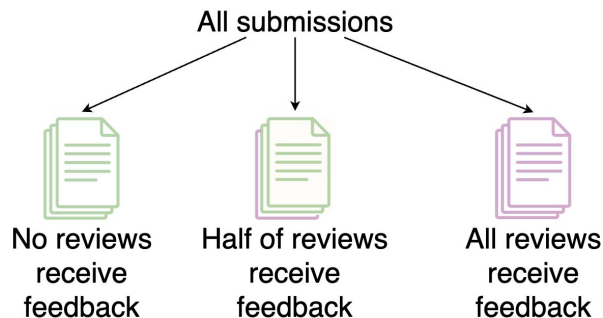
	Control	Feedback	Not updated	Updated
Average length of author rebuttal	807	855*** ($\Delta + 6\%$)	840	896***
Average length of reviewer replies	110	117*** ($\Delta + 6\%$)	115	129***



Feedback did not cause any statistically significant difference in acceptance rates



Summary



Key takeaway: many reviewers found AI feedback worth incorporating, although the intervention had minimal impact on final scores

Limitations & next steps

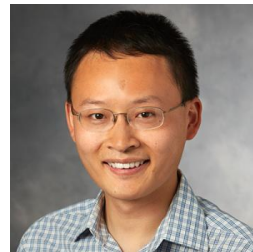
- Limited feedback to 3 categories (specificity, misunderstandings, professionalism)
- Explore newer LLMs for more nuanced feedback & robustness
- Deploy at more AI conferences across diverse research topics



Mert Yuksekgonul



Jake Silberg



James Zou

ICLR 2025 PCs:
Animesh Garg
Nanyun Peng
Fei Sha
Rose Yu
Carl Vondrick

OpenReview:
Celeste Martinez Gomez
Carlos Mondragon Chapa

Paper:



Code:



Questions?

Ethics

- Deemed low-risk by IRB review
- Reviewers have full control over the final review & scores visible to authors (feedback is entirely optional)
- ICLR papers and reviews are public to the community, agent is not given any identifiable or private information about reviewers and authors
- Feedback is only visible to reviewers and program chairs, does not factor into acceptances

AI-Assisted AI Science and Its Implications for Peer Review



Y. Yamada

*



R.T. Lange

*



Cong Lu

*



S. Hu

†



Chris Lu



J. Foerster



J. Clune

†



D. Ha



FLAIR



* Equal contribution, † Equal advising



A Generative AI Moment for Scientific Discovery

Images



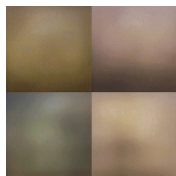
GANs (2014)



VAEs (2013)



Dallé (2021)



Midjourney (2022)



SDXL (2023)



Dallé-3 (2023)

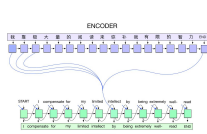


Imagen-3 (2024)

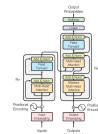


Flux-1 (2024)

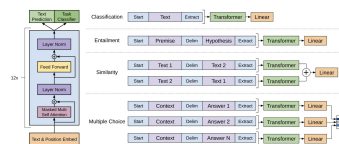
Text



Seq2Seq (2014)



Transformer (2017)



GPT-1 (2018)



InstructGPT (2022)



ChatGPT (2022)



Claude, Llama, Mistral, DeepSeek, ...



Video



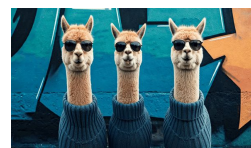
Make-A-Video (2022)



Sora (2024)

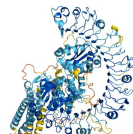


Walt (2024)

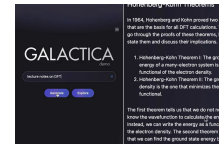


Veo (2024)

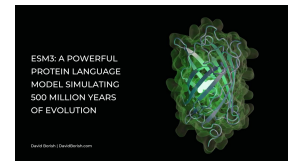
Science



AlphaFold/AlphaTensor (2021/2024)



Galactica (2022)



ESM3 (2024)



The AI Scientist (2024, 2025)

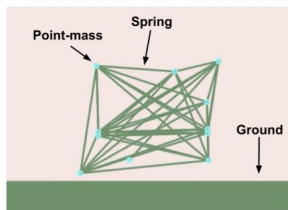
Time



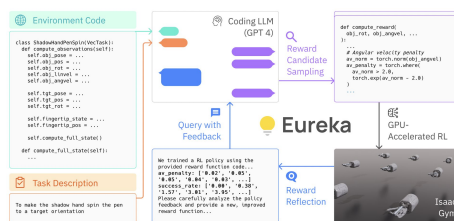
Improvement over time, collective effort for democratization, improvement, validation

LLMs are used all over the place to automate parts of research...

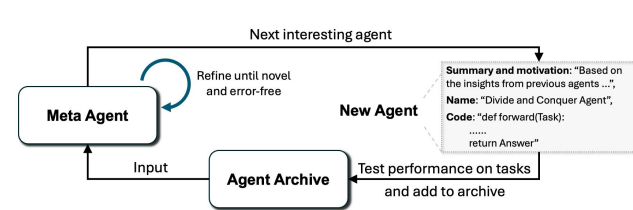
Lehman et al. (2023): ELM
Morphogenesis with LLMs



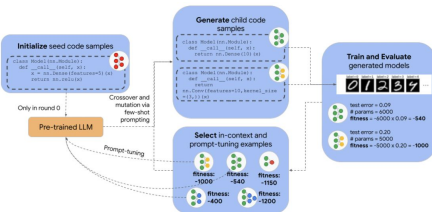
Ma et al. (2023): Eureka
Reward Fn Design with LLMs



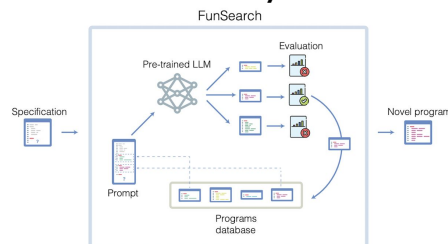
Hu et al. (2024): ADAS
Agentic Design with LLMs



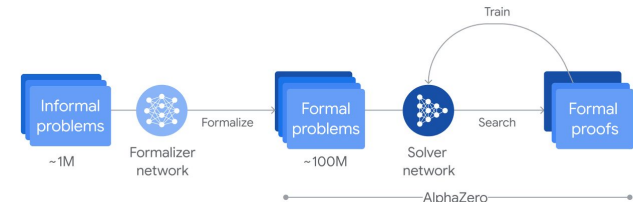
Chen et al. (2023): EvoPrompting
Architecture Search with LLMs



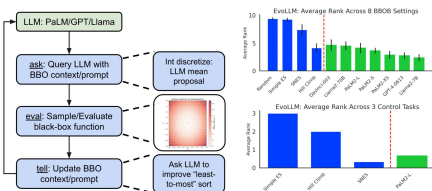
Romera-Paredes et al. (2024): FunSearch
Math/code Discovery with LLMs



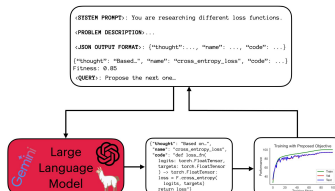
Hu et al. (2024): AlphaProof/Geometry
Math/Geometry with LLMs



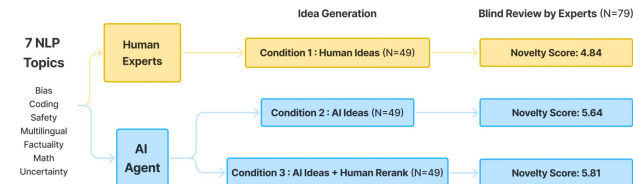
Lange et al. (2024): EvoLLM
Evolutionary Optimization with LLMs



Lu et al. (2024): DiscoPOP/LLM²
Loss Fn Discovery with LLMs



Si et al. (2024): Automated Ideation
Idea Generation with LLMs



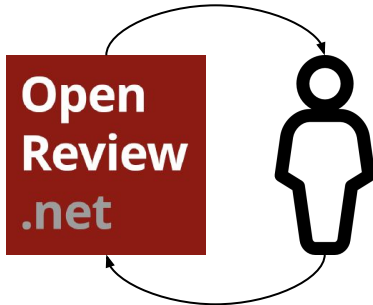
The Current State of the Human Scientific Community



Problem Identification, Hypothesizing, Brainstorming, Experimentation



Reviewing & Iteration



Publishing



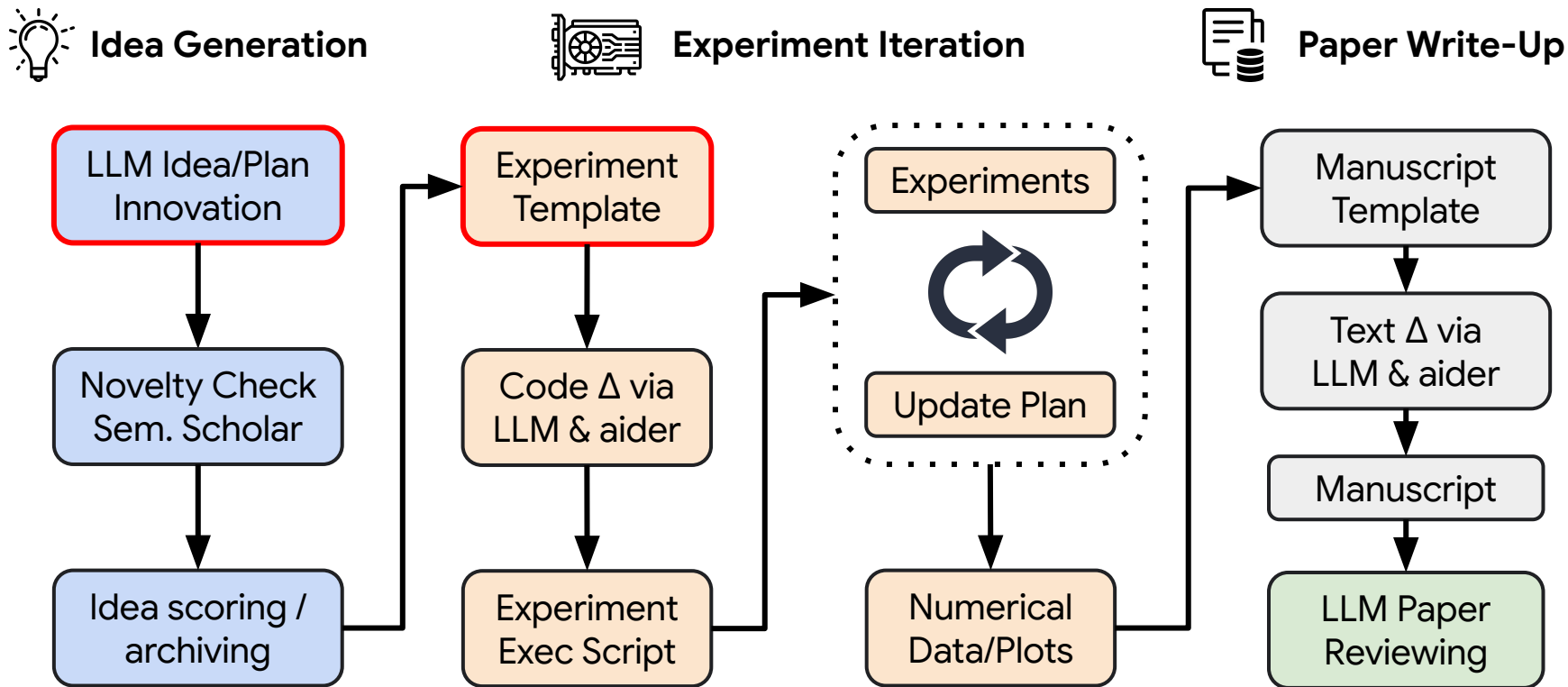
Knowledge Diffusion



But... can we automate the entire research process with AI Scientists?



The AI Scientist-v1: An Automated Template-Based Research Pipeline



The AI Scientist-v1: Papers based on 3 Experiment Templates

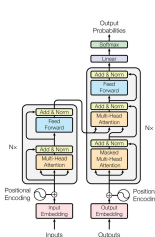
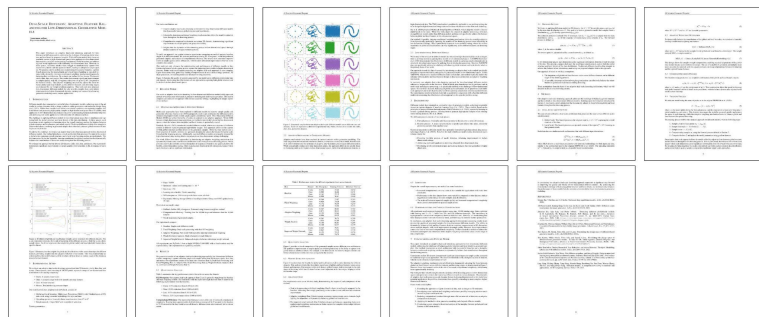
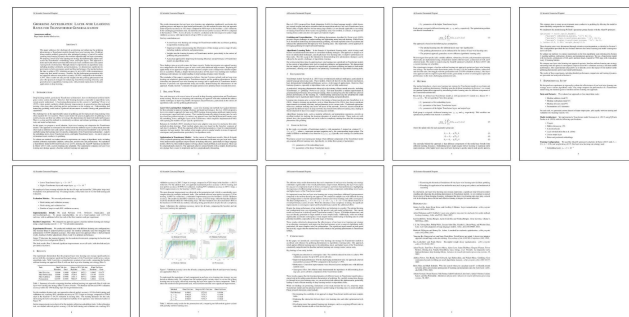
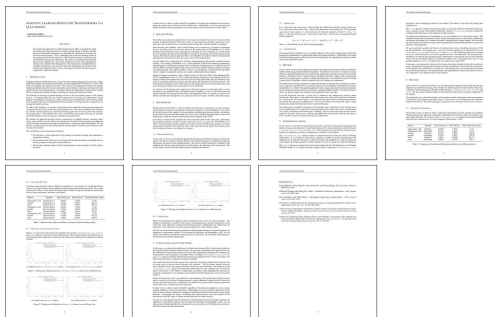
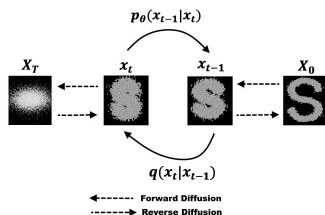


Figure 1: The Transformer - model architecture.

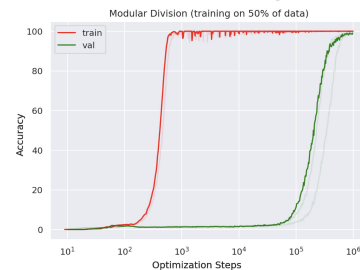
NanoGPT



2D Diffusion

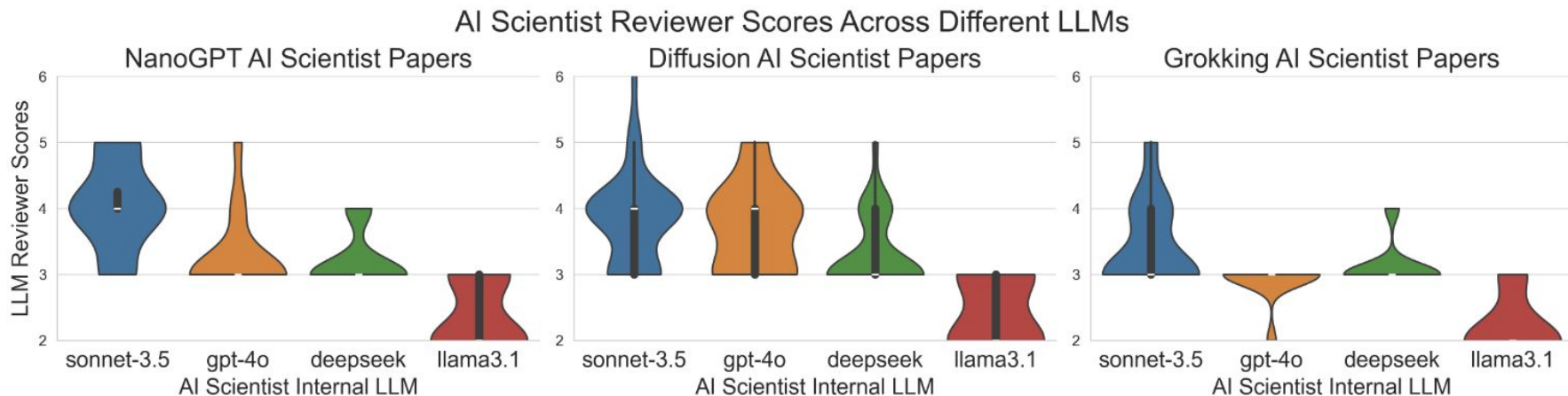


Grokking



👉 For The AI Scientist-v1 we choose 3 distinct subfields of ML to explore

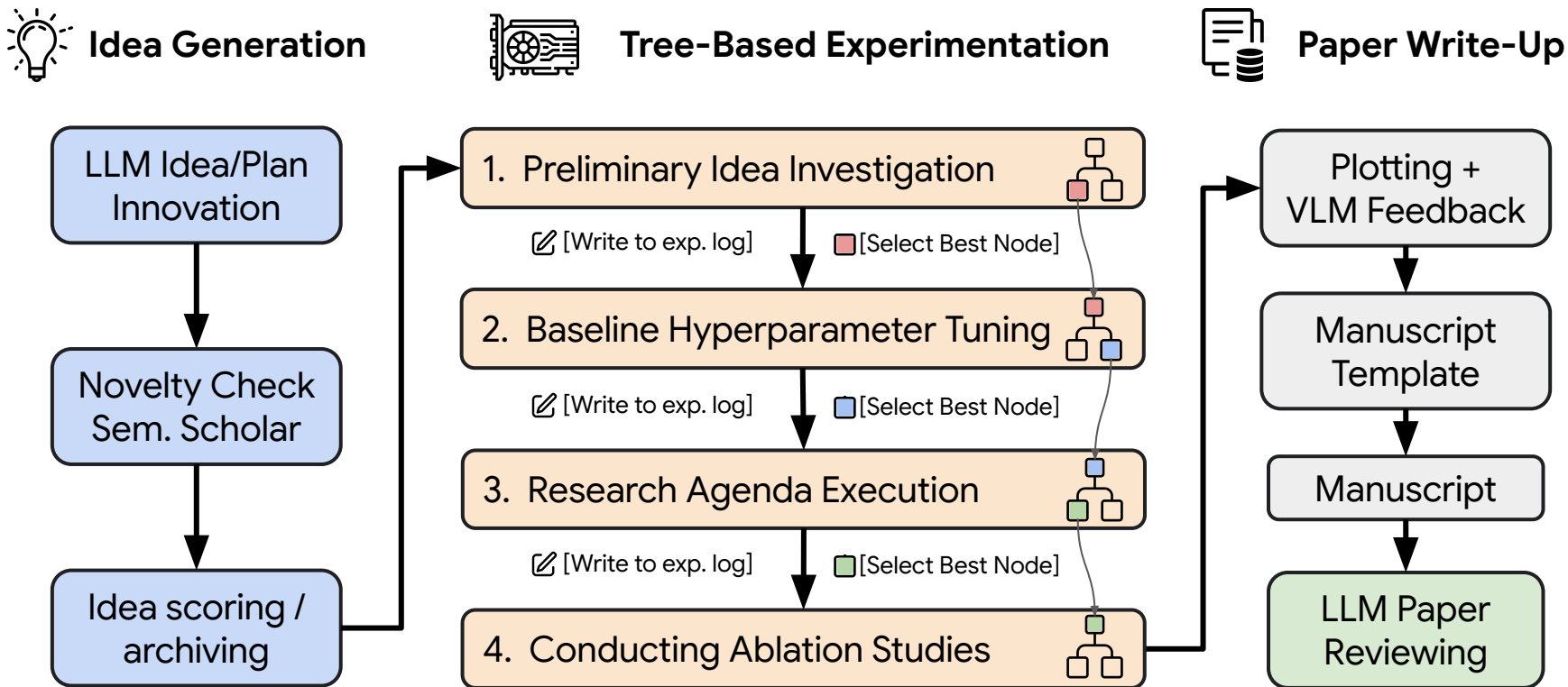
The AI Scientist-v1: Templated Experiments, AI Reviewer & Limitations



Feature	Codebase Drafting	Execution Planning	Parallel Experiments	VLM Reviewer	Human Result Evaluation
THE AI SCIENTIST-V1	Topic-Specific	Linear	✗	✗	Not Submitted
THE AI SCIENTIST-V2	Domain-General	Tree-Based	✓	✓	Workshop Acceptance-Worthy

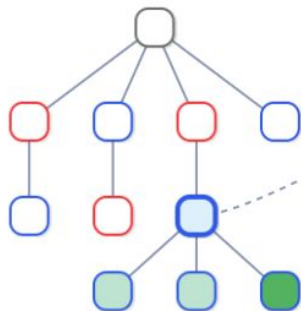
👉 Tree-search to expand the depth of hypothesis testing & knowledge accumulation

The AI Scientist-v2: An Automated Tree Search-Based Research Pipeline

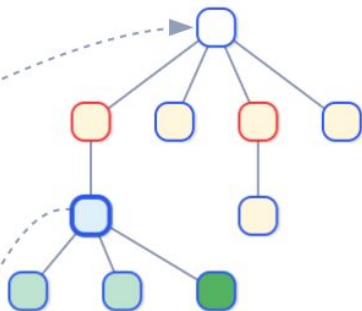


A Multi-Stage Tree-Based Approach to Automated Experimentation

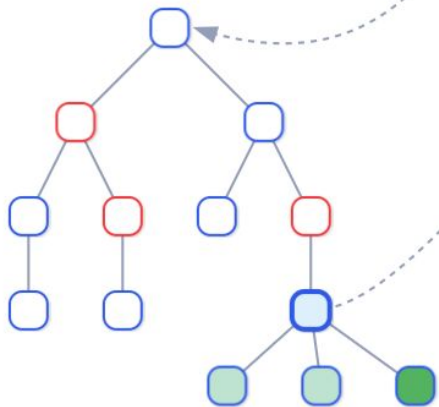
Stage 1: Preliminary Investigation



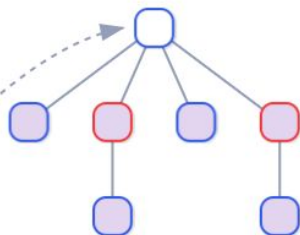
Stage 2: Hyperparameter Tuning




Stage 3: Research Agenda Execution



Stage 4: Ablation Studies



 Non-buggy nodes

 Buggy nodes

 Hyperparameter nodes


 Ablation nodes

 Replication nodes

 Aggregation nodes

 Best nodes

  Refinement

  Debugging

Released in March/April 2025

Collaborating with the ICINB ICLR 2025 Workshop



I Can't Believe It's Not Better: Challenges in Applied Deep Learning

Workshop at ICLR 2025

AI-Generated Papers 👉 3 papers submitted with organizer support + IRB [H24-02652]

As part of a small experiment that we believe aligns with the theme of our workshop (and with approval from the central ICLR workshop chairs), we have included 3 AI-generated papers out of a total of 43 submissions. As a result, it is possible, though unlikely, that you may be assigned an AI-generated paper to review. If you prefer not to review AI-generated papers, please let us know by **February 11 AoE** by emailing our official email (cant.believe.it.is.not.better+workshop@gmail.com). We will review your assignments and reassign papers accordingly.

👉 It is crucial to be transparent, let human reviewers opt-out & to obtain a proper IRB!

AI Scientists still suffer from hallucinations

069

070

071

072

073

074

075

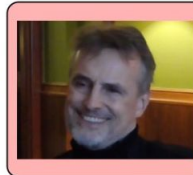
076

3 METHOD

Our goal is to enhance compositional generalization in neural networks by incorporating a compositional regularization term into the training loss. We focus on a simple yet illustrative task: evaluating arithmetic expressions involving basic operators.

3.1 MODEL ARCHITECTURE

We use an LSTM-based neural network (Goodfellow et al., 2016) to model the mapping from input expressions to their computed results. The model consists of an embedding layer, an LSTM layer, and a fully connected output layer.



Comment:
This should be
Hochreiter &
Schmidhuber

080

081

082

083

084

085

3.2 COMPOSITIONAL REGULARIZATION

Let h_t be the hidden state at time t . We define the compositional regularization term as the mean squared difference between successive hidden states:

$$\frac{1}{T-1} \sum_{t=1}^{T-1} \|h_t - h_{t+1}\|^2$$

Comment:
This should be
more precise.



Check paper & Github for full analysis:

github.com/SakanaAI/Al-Scientist-ICLR2025-Workshop-Experiment/

Challenges & Ethical Considerations

- Transparency and Attribution

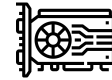
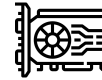
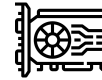
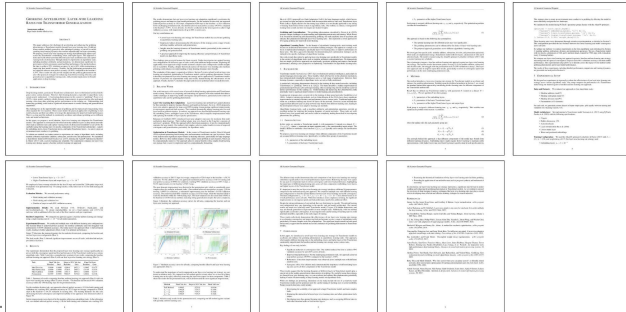
⇒ When and how to declare?



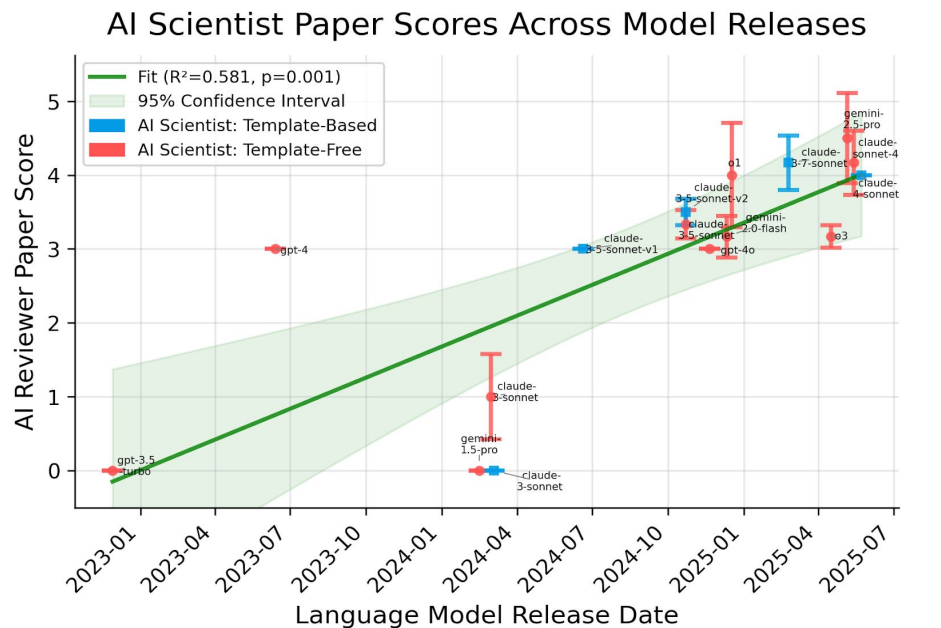
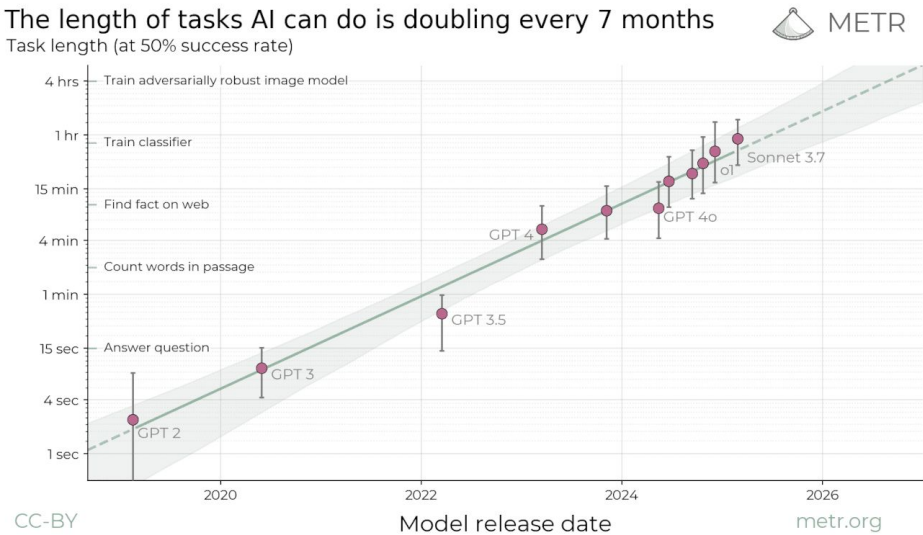
- Misuse

⇒ How to avoid gaming peer-review?

- Responsible development requires open discussion about these issues

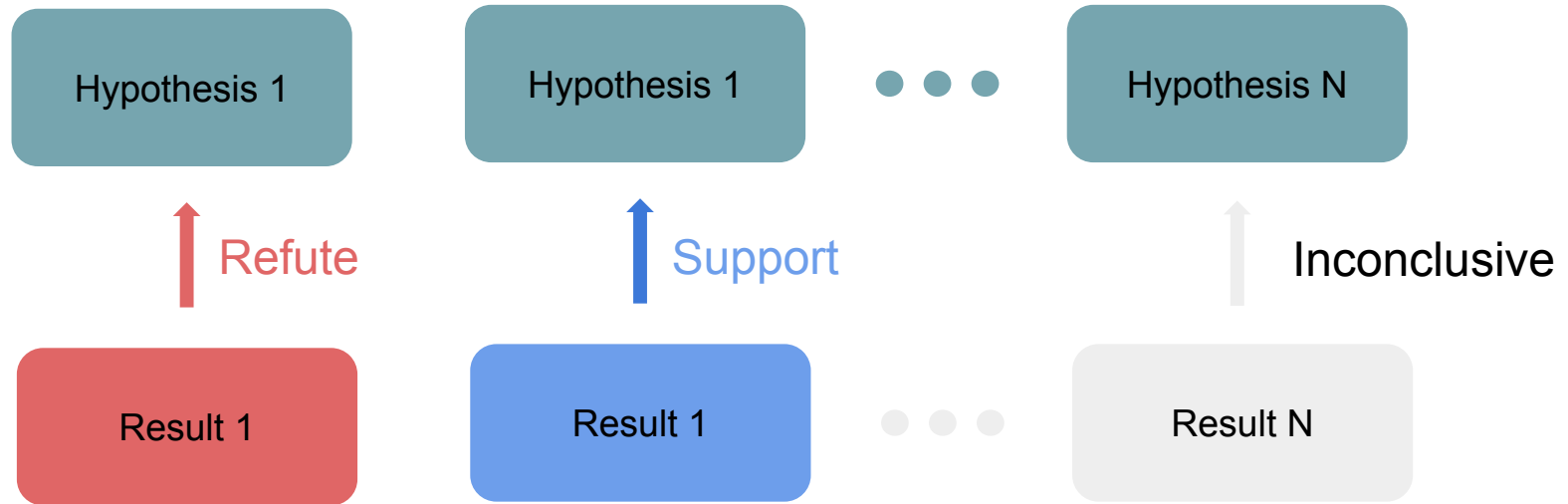


The quality of AI-generated papers might continue to improve

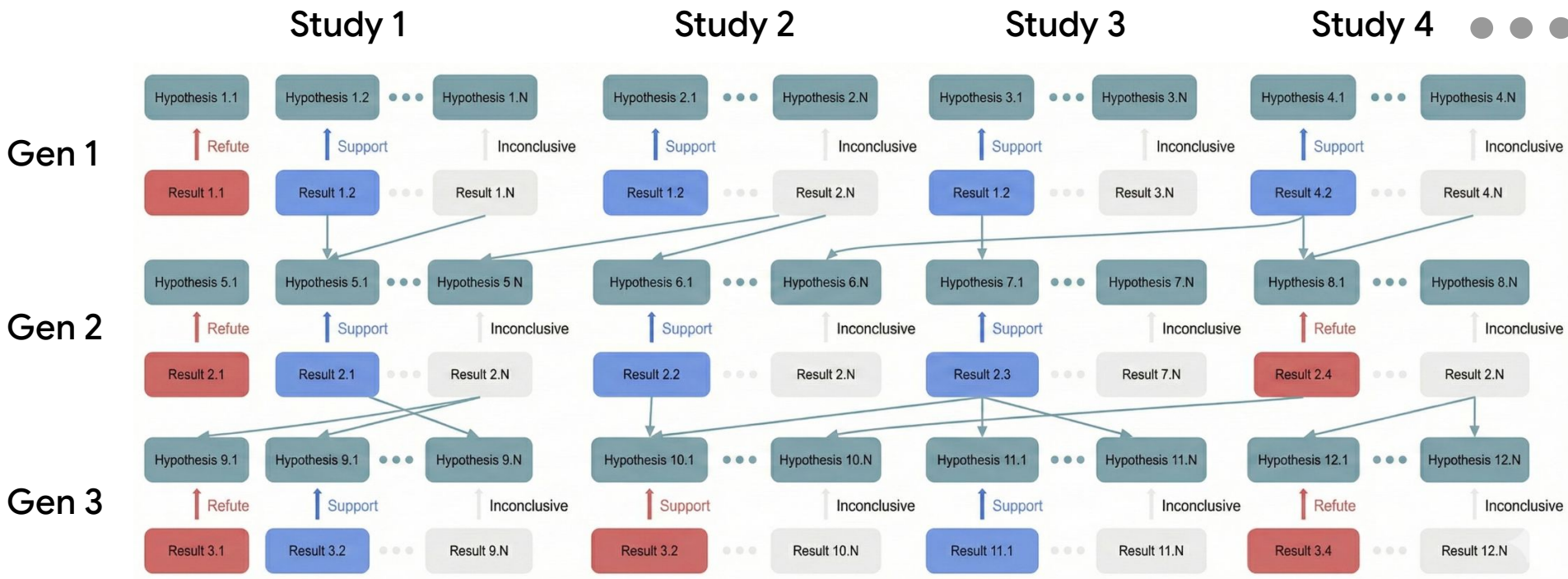


Changing the underlying models only

Structured representation of each study



Towards structured knowledge accumulation



Thank you!



⇒ Contact: yutaroyamada@sakana.ai
⇒ X/Twitter: @_yutaroyamada

Contact us:
ai-reviewer@chalearn.org