LLMs for Research Review

Presented by Michael Xu

Overview

- Introduction and Motivation
- Paper 1
- Paper 2
- Paper 3
- Conclusion
- Questions

The Peer Review Challenge

1. **Inconsistency**: Two reviewers can have completely different opinions.
2. **Reviewer Fatigue**: Too many submissions, too few experts.
3. **Quality Issues**: Reviews can be shallow, generic, or even toxic.
4. **Delays**: Turnaround time is long, which can slow down progress.

Studies show that **30%** of reviews are considered unhelpful or low quality by authors. (ACL,NeurIPS)

Enter Large Language Models

There is growing acceptance that human peer review is flawed and irreplaceable… so far.

What if LLM's could:

- Write first-pass reviews for submissions.
- Critique low-quality reviews.
- Model entire peer-review systems to study bias.

Benefit of LLM's:

- Pretrained on academic text.
- Can reason, summarize, and critique.

This Talk – Three Directions

Paper	Year	LLM Role	Task
Liang et al.	2023	Reviewer	Can GPT-4 generate useful feedback?
Du et al.	2024	Meta-Reviewer	Can LLM's critique human reviews?
Jin et al.	2024	Simulator	Can LLM's help us simulate and study peer review systems?

Paper 1

Can large language models provide useful feedback on research papers? A large-scale empirical analysis.

Weixin Liang, Yuhui Zhang, Hancheng Cao, et al.

Paper 1 – Motivation

- Scientific progress depends on **feedback** and critique.
- Effective feedback leads to the emergence of new theories by connecting individual ideas
- Feedback is laborious, resource-intensive, and complex.
- Also limited by the **exponential growth** in publications and **specialization**.
- Feedback traditionally through conferences and **peer review**.
- 100M researcher hours and $2.5B US dollars in a single year.
- Meanwhile, ICLR 960 (2018) → 4966 (2023)

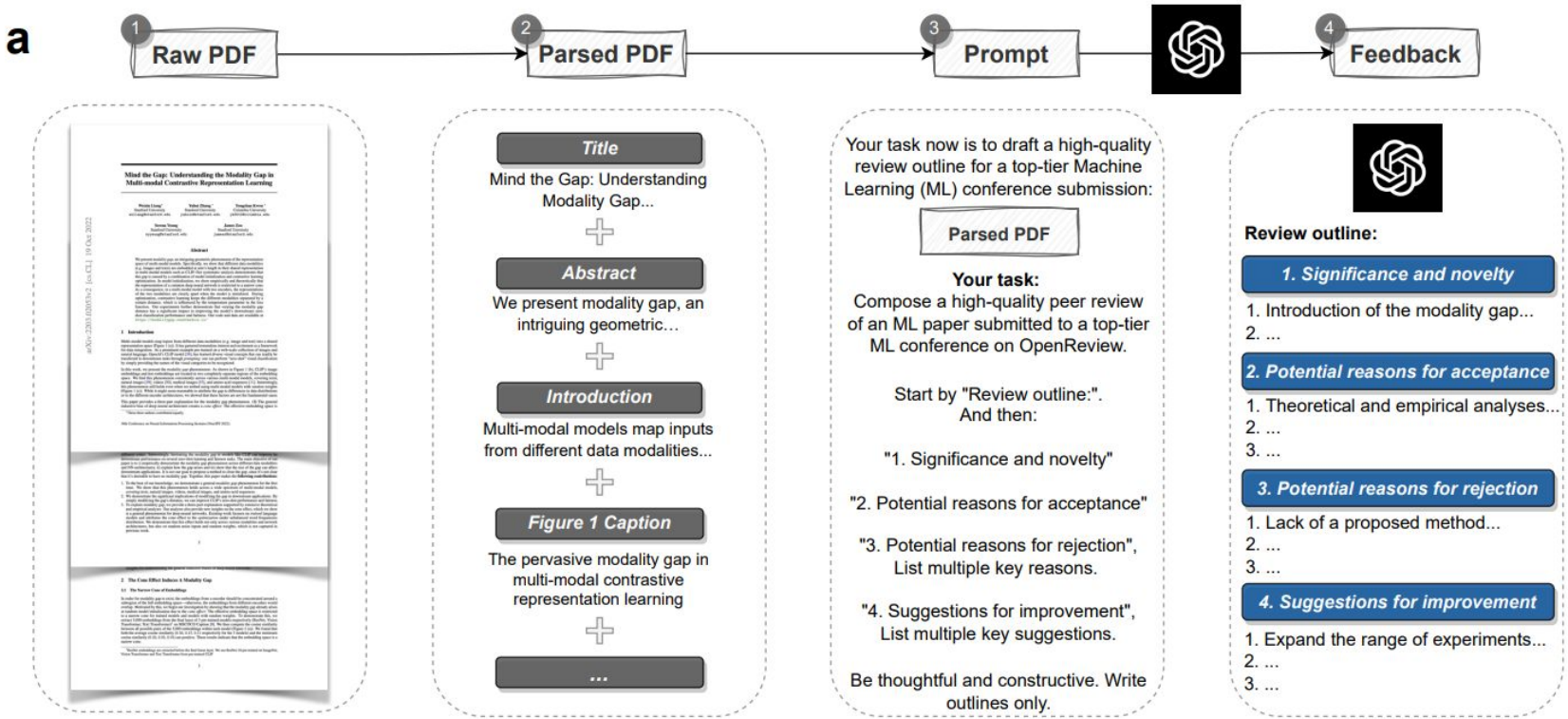
Paper 1 – Introduction

Investigate the effectiveness of GPT-4 in generating feedback for scientific papers.

- Review quality suffers from **inconsistency** and **subjectivity**.
- Reviewer quality suffers from **fatigue** and **delay**.
- LLMs can address all of these issues.
- LLMs used for paper screening, error identification, checklist validation

First large-scale systematic analysis characterizing the potential reliability and credibility of leveraging LLM for generating scientific feedback

GPT-4 based scientific feedback generation pipeline that takes the raw PDF of a paper and produces structured feedback



Input PDF → extract title, abstract, figure/table captions, and main text to construct prompt → Prompt GPT-4 to provide structured comments with four sections following leading feedback structure

Paper 1 - Introduction

Investigate the effectiveness of GPT-4 in generating feedback for scientific papers.

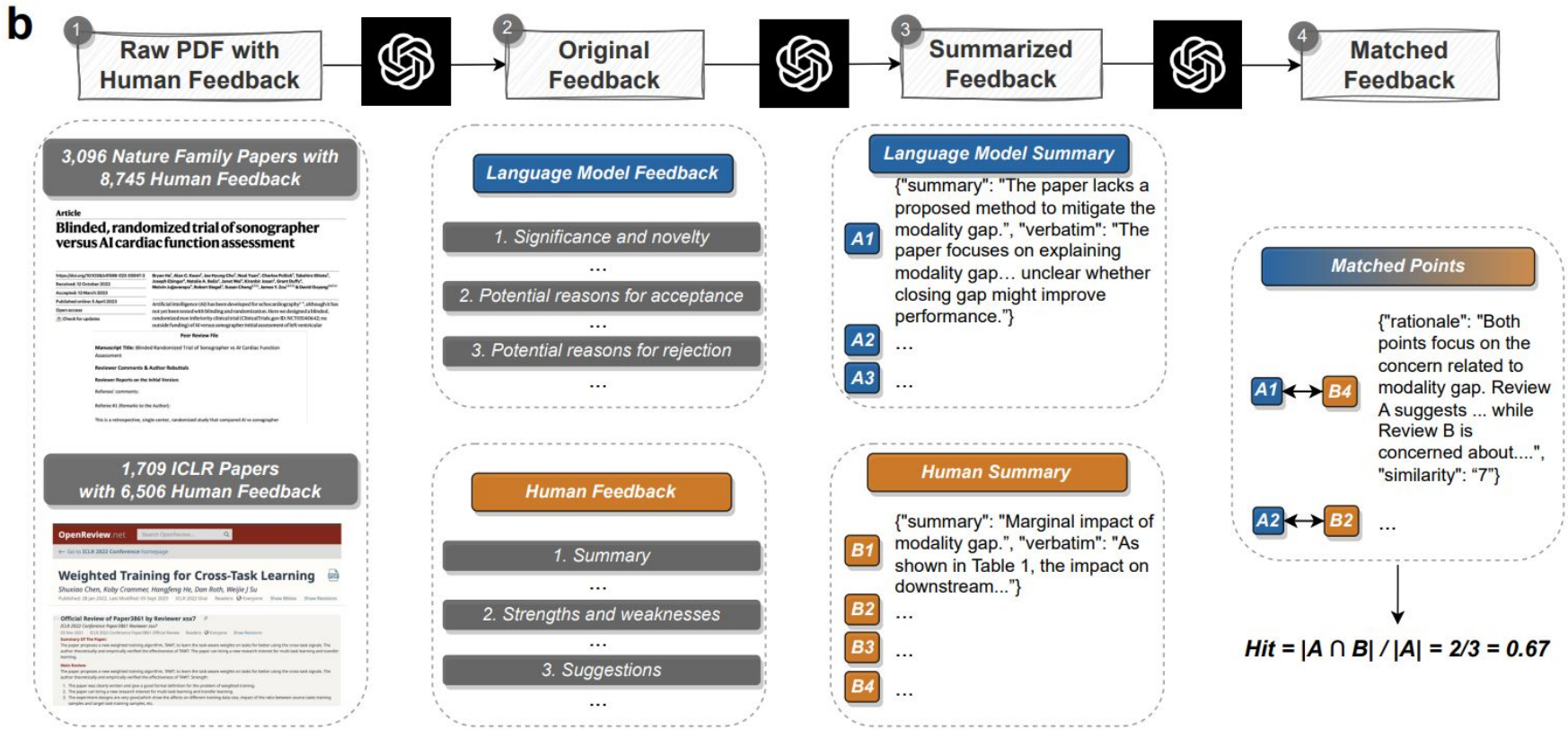
To evaluate informativeness of GPT-4 feedback: **retrospective analysis** & **prospective user study.**

Retrospective Analysis

- Apply pipeline to papers assessed by human reviewers, then compare LLM and human feedback.
- Assess **degree of overlap** between key points and compare **topic distribution**
- Two datasets containing full-text papers, meta information, and peer reviews (2022+, Nature/ICLR)
- 8,745 human comments for 3,096 accepted papers across 15 Nature family journals (breadth)
- 6,505 human comments for 1,709 papers from the ICLR (depth)

Prospective User Study

- Researchers invited to evaluate quality of GPT-4 feedback on their authored papers.
- 308 researchers from 110 US institutions in AI to computational biology.



Extractive text summarization → semantic text matching

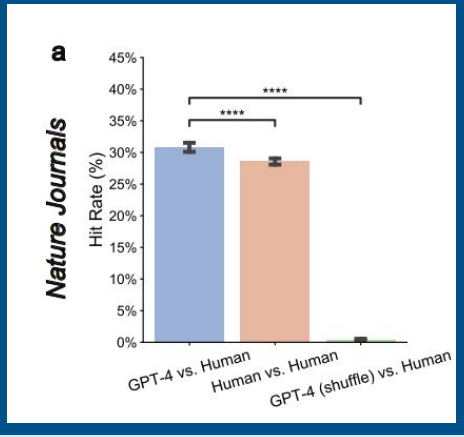
Paper 1 - Results

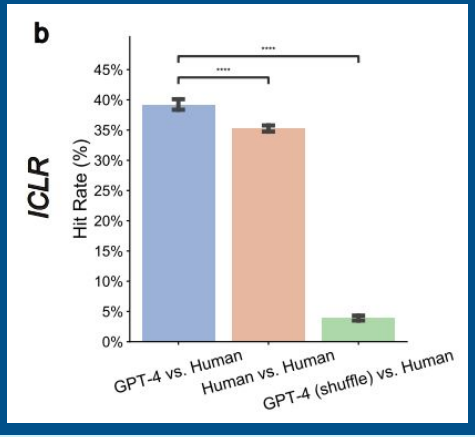
LLM feedback significantly overlaps with human-generated feedback

- **57.55%** of comments raised by GPT-4 were raised by at least one human reviewer.
- When comparing GPT-4 raised comments to each individual reviewer:
  - **30.85%** overlap. Degree of overlap between two humans was **28.58%**
  - Indicates the overlap between LLM and human is comparable to human and human.

- GPT-4 vs human - hit rate
- GPT-4 (shuffle) indicates feedback from randomly chosen paper. See if GPT-4 produces paper-specific reviews.

LLM can generate non-generic feedback





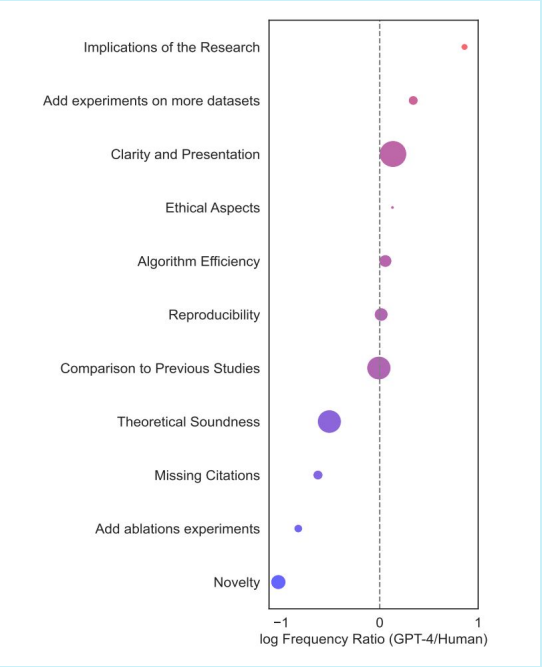
Paper 1 - Results

LLM is consistent with humans on major comments

- Comments identified by multiple human reviewers are more likely to be echoed by LLMs.
- Single reviewer - **11.39%** chance of being identified by LLMs, increased to **20.67%** for comments raised by two reviewers, and further to **31.67%** for three or more reviewers.
- LLMs are more likely to identify common issues or flaws that are consistently recognized by reviewers

LLM feedback emphasizes certain aspects more than humans

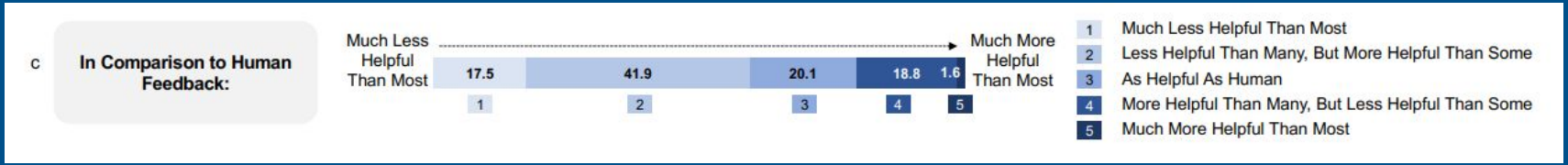
- LLM comments on implications of research 7.27x more frequently than humans
- LLM is 10.69x less likely to comment on novelty compared to humans
- Circle size indicates the prevalence of each aspect in human feedback



Paper 1 - Results

Researchers find LLM feedback helpful

- Survey on how helpful LLM feedback is in improving their work or understanding of a subject
- 50.3% helpful, 7.1% very helpful.
- 17.5% considered it inferior, 20.1% same, and 20.4% more to human feedback.



- 50.5% of researchers would reuse the system, and expressed optimism about potential improvements to the traditional human feedback process.
- 65.3% of participants think LLM feedback offers perspectives that have been underemphasized by humans

Paper 1 - Limitations

The most important limitation is its ability to generate specific and actionable feedback

- “Potential Reasons are too vague and not domain specific.”
- “GPT cannot provide specific technical areas for improvement, making it potentially difficult to improve the paper.”
- Future direction to improve the LLM based scientific feedback system is to push the system towards generating more concrete and **actionable feedback**, e.g. through pointing to specific missing work and experiments to add

Paper 2

LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing

Jiangshu Du, Yibo Wang, Wenting Zhao, et al.

Paper 2 - Introduction

Analyze the capabilities of LLMs in assisting with paper reviewing and meta-reviewing tasks

- **Claim**: This work is not advocating the use of LLMs for paper (meta-)reviewing.
- Present a comparative analysis to identify and distinguish LLM activities from human activities.
- Two research goals:
  - Enable better recognition of instances when someone implicitly uses LLMs for reviewing activities;
  - Increase community awareness that LLMs, and AI in general, are currently inadequate for performing tasks that require a high level of expertise and nuanced judgment.

How can LLMs potentially assist researchers in alleviating their heavy workload?

Paper 2 - Introduction

LLMs Assist NLP Researchers

- Examine the **effectiveness** of LLM in assisting paper (meta-)reviewing and its **recognizability**.
- Constructed the *ReviewCritique* dataset:
  - NLP papers with both human-written and LLM-generated reviews (initial submissions)
  - Each review has *deficiency labels* and corresponding explanations annotated by experts

- Explores two threads of research questions:
  - LLMs as Reviewers - LLM vs human reviews, quality and distinguishability
  - LLMs as Meta-Reviewers - Can LLMs identify deficiencies and unprofessional segments

Paper 2 - Introduction

- Still takes years train a qualified, domain-specific expert researcher.
- Researchers face increasing challenges with more papers to read, to beat, to write, and to review.

What is the potential for LLMs to work as researchers to alleviate their heavy and unhealthy workload?

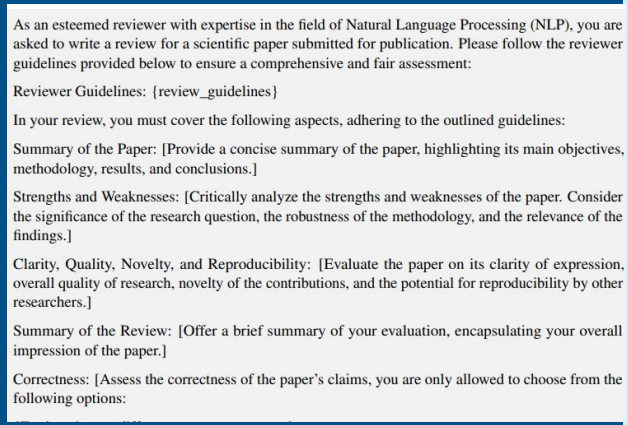
- Threefold Contribution:
  - ReviewCritique dataset serves as a valuable resource for future research.
  - Quantitative comparison of human and LLM paper reviews at the sentence level.
  - Analysis of LLMs' potential as both reviewers and meta-reviewers.

Paper 2 – Curated Dataset

- Papers are selected on the following criteria:
  - Only consider NLP papers, requires recruitment of domain specific annotators.
  - Publicly accessible human-written reviews.
  - Equal distribution of accepted and rejected papers
    - to investigate review pattern discrepancies based on the final acceptance/rejection.
- 100 papers from OpenReview (ICLR, NeurIPS, 2020-2023)
- 3-5 complete individual reviews, meta-reviews, and author rebuttals.
- Annotators flag papers that may have AI-written reviews.

Paper 2 - Curated Dataset

- To compare human and LLM reviews, select subset of 20 papers from 100. Equal accept/reject.
- Annotations are time-consuming, 20 allows for sufficient statistical comparison.
- Utilize three sota closed-source LLMs - GPT-4, Gemini-1.5, Claude Opus.
- Each LLM generates three reviews using prompts that include:
  - Standard ICLR review guidelines
  - Randomly chosen human-written reviews
  - A generation template in ICLR 2024 format
-



As an esteemed reviewer with expertise in the field of Natural Language Processing (NLP), you are asked to write a review for a scientific paper submitted for publication. Please follow the reviewer guidelines provided below to ensure a comprehensive and fair assessment:

Reviewer Guidelines: {review_guidelines}

In your review, you must cover the following aspects, adhering to the outlined guidelines:

Summary of the Paper: [Provide a concise summary of the paper, highlighting its main objectives, methodology, results, and conclusions.]

Strengths and Weaknesses: [Critically analyze the strengths and weaknesses of the paper. Consider the significance of the research question, the robustness of the methodology, and the relevance of the findings.]

Clarity, Quality, Novelty, and Reproducibility: [Evaluate the paper on its clarity of expression, overall quality of research, novelty of the contributions, and the potential for reproducibility by other researchers.]

Summary of the Review: [Offer a brief summary of your evaluation, encapsulating your overall impression of the paper.]

Correctness: [Assess the correctness of the paper's claims, you are only allowed to choose from the following options:

Paper 2 – Curated Dataset

Data Annotation

- Group of senior NLP researchers with rich Area Chairing experience, define **Deficient** review segments as follows:
  - Sentences that contain factual errors or misinterpretations of the submission.
  - Sentences lacking constructive feedback.
  - Sentences that express overly subjective, emotional, or offensive judgments, such as "I don't like this work because it is written like by a middle school student."
  - Sentences that describe the downsides of the submission without supporting evidence, for example, "This work misses some related work."
- Annotators consist of 40 NLP researchers all with multiple first-authored publications in top tier NLP venues. 16 PHD, 11 faculty members, 15 served as area chair

Paper 2 – Curated Dataset

Data Annotation continued.

- Annotation conducted on both human and LLM reviews following these steps:
  - **Paper Selection:** Annotators were allowed to choose papers that aligned with their expertise and interests
  - **Awareness of Review Scope:** Assessment focuses on reviews before rebuttal phase (first submission)
  - **Segment-Level Annotation:** Reviews were segmented by sentence. Label each segment whether it is Deficient, and provide an explanation if it is.
- Disagreements in annotations resolved by senior expert with AC experience.
- Six-month data collection process.

Paper 2 – Curated Dataset

LLM-generated reviews contain more Deficient instances compared to human reviews

	Human-written Review			LLM-generated Review		
	All	**Accepted**	**Rejected**	**All**	**Accepted**	**Rejected**
#Papers	100	50	50	20	10	10
#Reviews	380	195	185	60	30	30
w/ Deficient seg.	272	132	140	60	30	30
w/ Deficient pct. (%)	71.57	67.69	75.67	100	100	100
#Segments	11,376	6,027	5,349	1,611	812	799
Deficient	713	317	396	225	144	81
Deficient pct. (%)	6.27	5.26	7.40	13.97	17.73	10.14
#ExplainationTokens	14,773	6,957	7,816	3,877	2,584	1,293

Table 1: Statistics of ReviewCritique.

Paper 2 - Curated Dataset

Novelty of ReviewCritique

ReviewCritique differs from previous works

- Sentence level
- Annotators read first-submission, meta reviews, all reviews, and rebuttals

Benchmarking LLMs as responsible meta-reviewers

Dataset	PeerRead	PRAnalyze	Subs.PR	DISAPERE	ReviewCrit.
Sentence-level		✓	✓	✓	✓
Initial submission	✓				✓
Highly Expert-demanding					✓
Deficiency Labeling					✓
Human Review	✓	✓	✓	✓	✓
LLM review					✓
Accepted+Rejected	✓	✓			✓

Table 2: Comparison of ReviewCritique with PeerRead (Kang et al., 2018), Peer Review Analyze (Ghosal et al., 2022a), Substantiation PeerReview (Guo et al., 2023) and DISAPERE (Kennard et al., 2022).

Paper 2 – Experiments

LLMs as Reviewers

Compare LLM and human reviews by:

- Fine-grained error types if Deficient
- Fine-grained analysis for each component (summary, strength, weakness, writing)
- Considering review diversity.

Deficient → 23 fine-grained error types:

Error Type	Explanation
Misunderstanding	The reviewer misinterprets claims or ideas presented in the paper, leading to inaccurate or irrelevant comments.
Neglect	The reviewer overlooks important details explicitly stated in the paper, resulting in unwarranted questions or critiques.
Vague Critique	The review lacks specificity, claiming missing components without clearly identifying what is missing.

Error Type	Human (%)	LLM (%)
Human top-3		
Misunderstanding	22.86	9.87
Neglect	19.64	5.83
Inexpert Statement	18.23	6.73
LLM top-3		
Out-of-scope	4.35	30.49
Misunderstanding	22.86	9.87
Superficial Review	2.66	9.42

Table 3: Comparing top-3 error types between human-written and LLM-generated reviews.

Paper 2 – Experiments

LLMs as Reviewers cont. Fine-grained review analysis.

1. **Summary**: Relatively better than humans.
   Inaccurate summary segments: 0.19% vs 0.36% for all LLM vs human segments.
   LLMs don't suffer from error types like Summary Too Short or Copy-Pasted Summary.
2. **Strengths**: LLMs often accept authors' claims without critical evaluation.
   53.2% of segments in LLM reviews Strengths section are simply rephrased.
3. **Weaknesses**: Most dominant type of error in LLM reviews.
   "Need more experiments, generalizability, additional tasks, etc."
   Highlights importance of human expertise in identifying weaknesses.
4. **Writing**: LLMs may lack the ability to accurately judge writing quality
   Consistently praise writing - 15% of papers flagged for writing by humans.
5. **Recommendation Score**: 1-10 rating similar to ICLR/NeurIPS system
   LLMs average 7.43 (accepted) and 7.47 (rejected)
   Humans average 6.41 (accepted) and 4.81 (rejected)

Paper 2 – Experiments

Model	Precision / Recall / F1			
	`Labeling-All`	`Select-Deficient`	Both "No"	Either "No"
GPT-4	14.91 / 34.49 / 18.38	17.18 / 34.59 / 20.30	18.71 / 21.40 / 16.85	14.72 / 47.68 / <u>20.66</u>
Claude Opus	16.86 / 34.26 / 20.35	17.69 / 26.61 / 18.71	17.14 / 18.70 / 15.78	16.94 / 42.12 / **<u>21.99</u>**
Gemini 1.5	16.58 / 34.13 / 19.76	14.71 / 43.60 / 19.72	17.01 / 27.05 / 18.28	14.46 / 50.37 / <u>20.34</u>
Llama3-8B	7.73 / 45.95 / 12.22	11.47 / 30.29 / <u>14.88</u>	11.37 / 21.27 / 12.46	8.19 / 53.61 / 13.35
Llama3-70B	13.63 / 42.49 / 18.19	13.95 / 31.16 / 17.46	16.16 / 23.51 / 16.67	12.46 / 50.02 / <u>18.43</u>
Qwen2-72B	9.97 / 26.60 / 12.96	11.35 / 34.61 / 14.64	9.07 / 15.13 / 9.62	10.49 / 43.00 / <u>15.16</u>

Table 4: Performance of LLMs as meta-reviewers on our `ReviewCritique` dataset. The best F1 score among different prompt methods for a single model is <u>underlined</u>. The best F1 score across all models is also **bold**.

Labeling-All: (id, deficient or not, explanation)
Select Deficient: (id, explanation)

Ensemble: Both "No"
Either "No"

Paper 2 - Weaknesses

- LLMs struggle to accurately judge the paper writing quality submission and tend to provide superficial reviews.
- LLMs are prone to generate out of scope reviews indicating a tendency to hallucination
- ReviewCritique focuses on the textual information from the submissions and does not include figures, tables, or other visual elements.
- Dataset is restricted to NLP domain. Could expand to other domains to test generalizability.
- Focuses on pre-rebuttal/first-submission phase of peer review process.

Paper 3

AGENTREVIEW: Exploring Peer Review Dynamics with LLM Agents

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, et al.

Paper 3 - Introduction

LLM based peer review simulation framework

- **AGENTREVIEW** - the first large language model (LLM) based peer review simulation framework.
- LLMs are capable of:
  - Realistic simulations of societal environments.
  - Providing high quality feedback on academic literature.
- AGENTREVIEW is open and flexible, captures the **multivariate** nature of peer review
- Features customizability:
  - Characteristics of reviewers, authors, and ACs
  - Reviewing mechanisms

This adaptability allows for the exploration and **disentanglement** of the distinct roles in peer review.

Paper 3 - Introduction

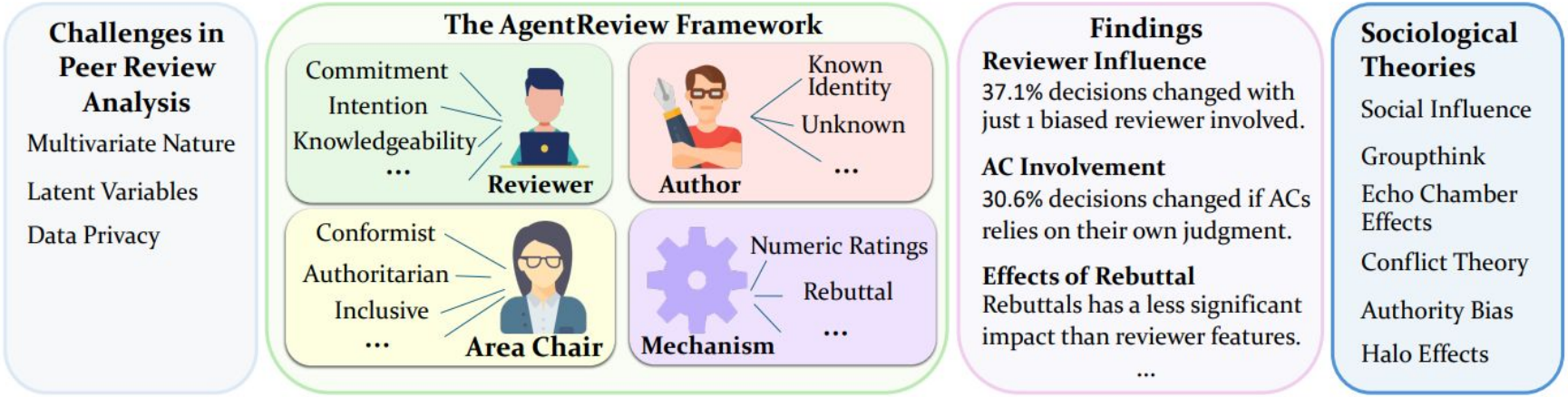


Figure 1: AGENTREVIEW is an open and flexible framework designed to realistically simulate the peer review process. It enables controlled experiments to *disentangle* multiple variables in peer review, allowing for an in-depth examination of their effects on review outcomes. Our findings align with established sociological theories.

Paper 3 - Framework

AGENTREVIEW integrates three roles – reviewers, authors, and ACs. All powered by LLM agents.

Reviewers:

- Commitment - reviewer's dedication and responsibility. Carefully constructed feedback.
- Intention - motivation behind the reviews. Reviewer may have biases/conflict of interests.
- Knowledgeability - reviewer's expertise in domain.

Responsible/Irresponsible, Benign/Malicious, Knowledgeable/Unknowledgeable

These categorizations are set by prompts and fed into our system as fixed characteristics

Paper 3 - Framework

Authors: Submit papers and provide rebuttals to initial reviews during Reviewer-AC period.

- Reviewers can be aware of authors' identities due to public release, or can remain unknown.
- Allows exploration of anonymity on review process.

Area Chairs (ACs): Ensure integrity of the review outcomes.

- Multiple roles including facilitating reviewer discussions, synthesizing feedback into meta-reviews, making final decisions.
- Three styles of ACs based on their involvement strategies:
  - *Authoritarian* - dominate decision making, prioritizing their own evaluations over reviewers.
  - *Conformist* - rely heavily on reviewers' evaluations, minimizing their own influence.
  - *Inclusive* - consider all available feedback to make well-rounded decisions.

Paper 3 - Framework

Review Process Design

5 phase pipeline that simulates peer review process:

1. **Reviewer Assessment**:
   Three reviewers evaluate manuscript. No cross-influence between reviewers.
   Significance and novelty, potential reasons for acceptance/rejection, suggestions for improvement, along with 1-10 rating.
2. **Author-Reviewer Discussion**:
   Respond to each review with a rebuttal document.
3. **Reviewer-AC Discussion:**
   AC initiates discussion between reviewers asking them to reconsider and update reviews after rebuttals.
4. **Meta-Review Compilation**:
   AC integrates insights from Phases I-III, their own evaluations, and ratings into a meta-review.
   Provides synthesized assessment of manuscript's strengths and weaknesses.
5. **Paper Decision:**
   AC reviews all meta-reviews and makes informed decision on acceptance/rejection. Adopt fixed rate of 32%.
   AC makes decisions for batch of 10 papers and accepts ~ 3 to 4.

Paper 3 - Framework

Five-phase pipeline

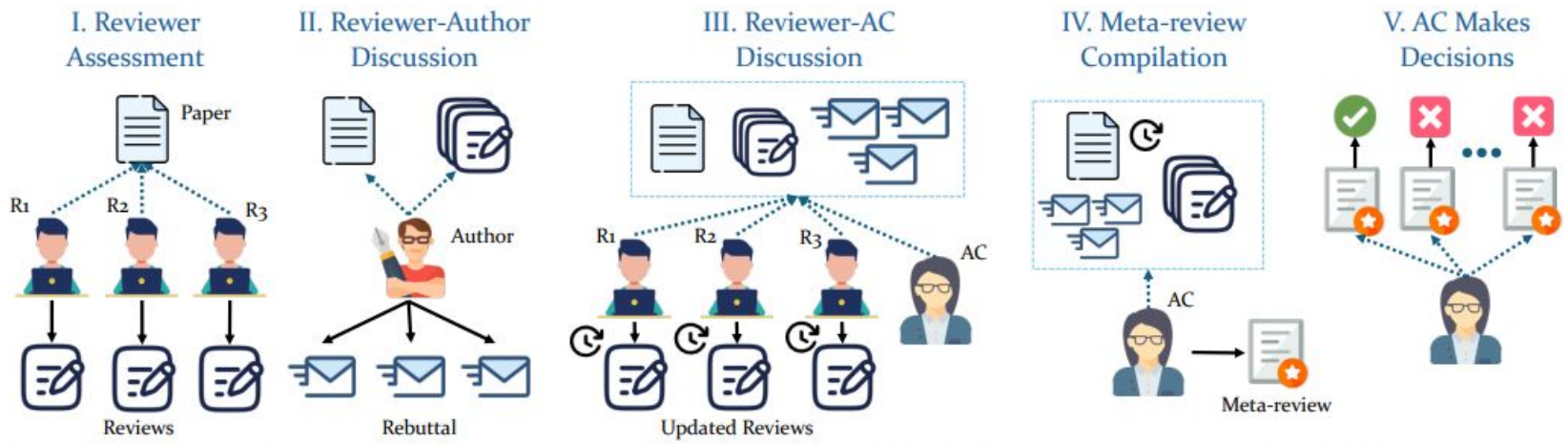


Figure 2: Our paper review pipeline consists of 5 phases. Solid **black** arrows → represent authorship connections, while **blue** dashed arrow → indicate visibility relations.

Paper 3 - Data Selection

The paper data for AGENTREVIEW is sourced from real conference submissions with four criteria:

1. Conference must have international impact. Papers have real world impact.
2. Papers must be publicly available
3. Quality of papers reflects real-world distribution. Accepted/Rejected.
4. Papers span a broad time range and cover variety of topics.

Select ICLR due to its leading status and retrieve papers using OpenReview API (2020-2023)

350 rejected papers, 125 posters, 29 spotlights, 19 orals.
Finally, extract title, abstract, figure and table captions, and main text → LLM agents.

Paper 3 - Results

- Baseline with no specific characteristics
- Start with replacing a normal reviewer with responsible/irresponsible, then increase the number of reviews.
- Agent-based reviewers demonstrate classic phenomena in sociology such as:
  Social influence, echo chamber, halo effects.

Setting	Initial (Phase I) Avg.	Initial (Phase I) Std.	Final (Phase III) Avg.	Final (Phase III) Std.
baseline	5.053	0.224	5.110	0.163
responsible	4.991	0.276	5.032	0.150
irresponsible	4.750	0.645	4.815	0.434
benign	4.990	0.281	5.098	0.211
malicious	4.421	1.181	4.368	1.014
knowledgeable	5.004	0.260	5.052	0.152
unknowledgeable	4.849	0.479	4.987	0.220

Table 1: Summary of results. We report the reviewer scores before & after Reviewer-Author Discussion (Phase III in Figure 2). 'Initial' & 'Final' indicate the reviewer ratings in Phase I & III, respectively.

Paper 3 - Results

Social Influence: individuals in a group tend to revise their beliefs towards a common viewpoint.

- Across all settings std. declines after Reviewer-AC discussion indicating **conformity.**
- Emphasized when a highly knowledgeable reviewer is in discussion.

Reviewer Fatigue/Peer Effect: paper review is time consuming and unpaid. Reviewers often feel their voluntary efforts are unrecognized → reduced commitment, superficial assessments.

- Presence of just one irresponsible reviewer leads to decline in overall commitment.
- Average word count drops 18.7% between baseline and irresponsible
- One subpar reviewer can lower performance of others.

Paper 3 - Results

🙂 normal reviewers				😪 irresponsible reviewers			
#	Initial	Final	+/−	#	Initial	Final	+/−
3	5.053 ± 0.623	5.110 ± 0.555	+0.06	0	/	/	/
2	5.056 ± 0.633	5.015 ± 0.546	−0.04	1	4.139 ± 1.121	4.416 ± 0.925	+0.27
1	5.256 ± 0.896	5.005 ± 0.630	−0.25	2	4.548 ± 0.925	4.543 ± 0.872	−0.01
0	/	/	/	3	4.591 ± 0.912	4.677 ± 0.745	+0.09

Table 3: Average reviewer ratings when varying numbers of 🙂 *normal* reviewers are replaced by 😪 irresponsible reviewers. '#' represents the number of reviewers of each type. 'Initial' & 'Final' refer to the average ratings in Phase I & III. The left and right side of the table shows average ratings from 🙂 normal reviewers and 😪 irresponsible reviewers, respectively. +/− indicates the change in average ratings after rebuttals.

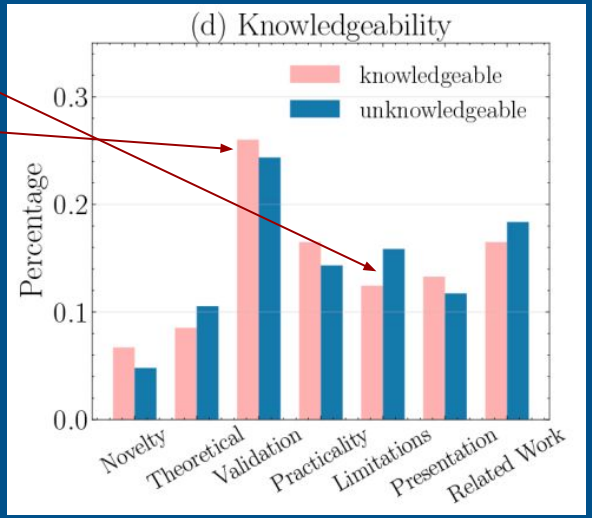
Highlights a noticeable decline in review ratings under influence of irresponsible reviewers.
2 irresponsible reviewers leads to drop of 0.25, 5.256 → 5.005
Baseline shows improvement in final ratings by 0.06, 5.053 → 5.110

Paper 3 - Results

Reviewer Knowledgeability

- Despite efforts at matching expertise, review assignments are often imperfect or random.
- Recent surge in submissions leads to expansion of reviewer pool.
- Less knowledgeable reviewers are 24% more likely to mention insufficient discussion or limitations.
- Expert reviewers address these basic aspects and also provide 6.8 % more critiques on experimental validation

- Distribution of reasons for rejection →



Paper 3 - Results

Involvement of ACs

- Alignment between Reviews and Meta-reviews are quantified by BERTScore and Embedding Sim.
- *Inclusive* ACs most aligned with Baseline
  - Demonstrates their effectiveness in maintaining integrity of the review process by balancing different viewpoints.
- *Authoritarian* ACs have significantly lower correlation with Baseline.
  - Indicates decisions may be skewed by individual biases.
- *Conformist* ACs have high semantic overlap with Baseline but might lack independent judgement

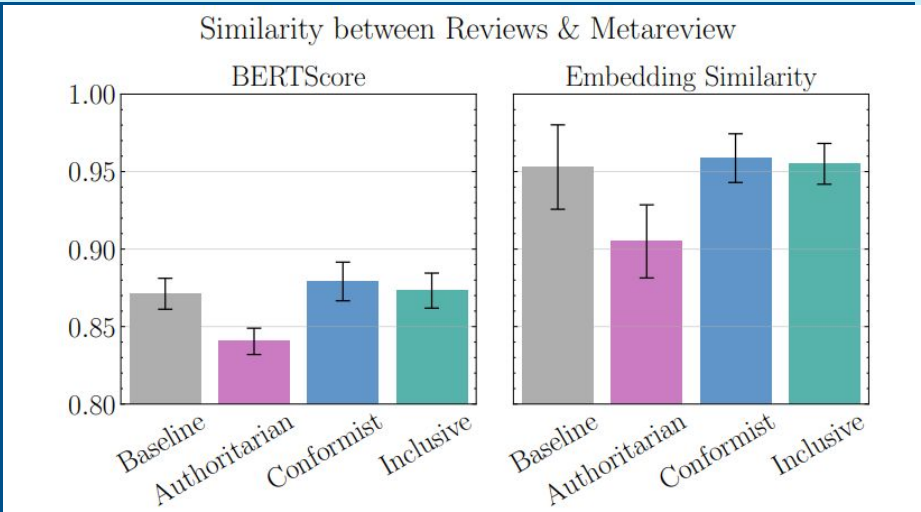


Figure 5: Similarities between reviews and meta-reviews w/ various intervention strategies from AC. Left: BERTScore, right: sentence embedding similarity.

Paper 3 - Limitations

- AGENTREVIEW is unable to dynamically incorporate or adjust experimental results in response to reviewer comments during Reviewer-Author Discussion (Phase II) as LLMs lack capability to generate new empirical data.
- Analysis isolates and examines individual variables of the peer review process, such as reviewer commitment or knowledgeability.
- Real world peer reviews involve multiple interacting dimensions.
- Simulation outcomes were not compared with actual peer review results.
- Use of baseline reviewer is challenging due to wide variability in human reviewer characteristics.

Thank you for listening.
Any Questions?