OpenReview Should be Protected and Leveraged as a Community Asset for Research in the Era of LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the era of large language models (LLMs), high-quality, domain-rich, and continuously evolving datasets capturing expert-level knowledge, core human values, and reasoning are increasingly valuable. **This position paper argues that OpenReview — the continually evolving repository of peer reviews, author rebuttals, meta-reviews, and decision outcomes — should be leveraged more broadly as a core *community asset* for advancing research in the era of LLMs.** We highlight three promising areas in which OpenReview can uniquely contribute: enhancing the quality, scalability, and accountability of peer review processes; enabling meaningful, open-ended benchmarks rooted in genuine expert deliberation; and supporting alignment research through real-world interactions reflecting expert assessment, intentions, and scientific values. To better realize these opportunities, we suggest the community collaboratively explore standardized benchmarks and usage guidelines around OpenReview, inviting broader dialogue on responsible data use, ethical considerations, and collective stewardship.

1 Introduction

The past years have witnessed an extraordinary shift in the role of data within machine learning [1, 2], especially with the recent advances of large language models (LLMs) [3–5], which have progressed from task-specific tools to general-purpose reasoning engines [6–8]. As their capabilities expand across domains, the role of data for training, evaluation, and alignment becomes even more important [9–12]. The current wave of LLM development increasingly depends on high-quality, human-centered feedback [13–17], not only for fine-tuning and instruction adherence, but also for assessing model behavior, identifying failure modes, and aligning outputs with human expectations [18–21]. Yet many of the datasets used for these purposes remain limited in coverage [22], synthetic in composition [23, 24], or static in structure [15]. As a result, they often fail to capture the complexity, disagreement, and subtle reasoning that characterize authentic human judgment [25–28].

At the same time, the powerful capabilities of LLMs are beginning to reshape scientific workflows themselves [29–33]. Tools based on LLMs such as ChatGPT are making research communication, including literature reviews and even paper writing, more accessible [34–37], hence accelerating scientific output and contributing to a significant rise in the volume of submissions to top conferences. Such a transformation has intensified pressure on the peer review system [38, 39]. Conferences now receive more than 10 thousands of submissions per cycle, and the human effort required to maintain high-quality, fair, and constructive reviewing has become difficult to sustain. Given such high pressure, the need for scalable assistance tools, better evaluation data, and models that can understand or generate scholarly critique has grown [39–41]. However, large-scale, systematic exploration regarding both the datasets and methodologies that enable LLMs to capture the richness of peer review interactions is still missing [42–44].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

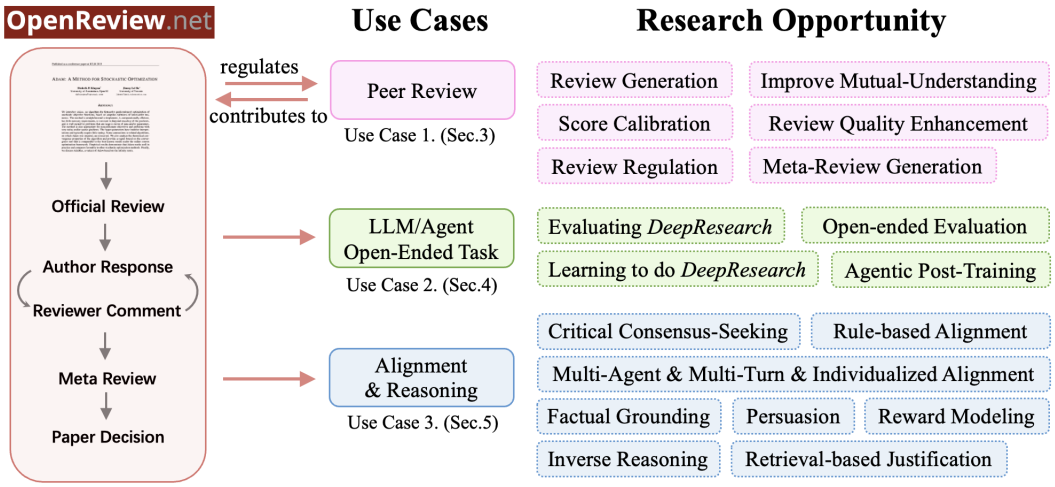

Figure 1: **Left**: an overview of the OpenReview data generation process; **mid**: this position paper argues OpenReview supports three main valuable applications — regulating peer review, empowering LLM and Agentic open-ended task research, and post-training for alignment and reasoning; **right**: highlighted research opportunities around those use cases.

OpenReview[1][45], the public review platform widely used by conferences such as ICLR, NeurIPS, and others, offers a unique opportunity to meet the needs of both sides. Contributed by the community and continually expanding over time, OpenReview hosts large-scale, structured records of scientific discussion, typically including paper submissions, reviewer assessments, author rebuttals, meta-reviews, and final decisions. These interactions span multiple rounds and involve diverse expert perspectives, making OpenReview an invaluable living dataset grounded in real-world scientific research deliberation. And has the potential to enrich both data-centric LLM research and assist the peer review system.

This position paper argues that OpenReview should be leveraged more broadly as a core community asset for advancing research in the era of LLMs. We elaborate on three areas where this dataset can provide immediate value:

1. **A data-driven approach to improve the quality and scalability of peer review.** OpenReview provides a unique source of structured, expert-generated assessments that can be used to train machine learning models to analyze and support the peer review process. Machine learning models, including the state-of-the-art general purpose language models [46–49], may learn to assist reviewers in drafting constructive feedback, calibrating scores, and identifying argumentative gaps, as well as summarizing responses, checking code, or detecting unhelpful language. In the face of rising submission volumes and reviewer fatigue, such tools could support more consistent, fair, and informative evaluations. Equally important, improving and regularizing the review process is a **prerequisite** for sustaining the long-term development of LLM-based systems that depend on high-quality expert feedback [50].
2. **Providing expert-generated benchmarks for LLM open-ended task evaluation and post-training.** Open-ended tasks such as academic writing, research evaluation, persuasion, or summarization are increasingly recognized as central to the development of general-purpose AI systems and the path toward superintelligence [51]. However, both training and evaluating models on such tasks remain challenging due to the open-ended nature and the lack of scalable, high-quality human feedback [8]. To this end, OpenReview offers a unique, high-quality resource: it contains expert-curated, multi-dimensional evaluations of research contributions grounded in real-world scientific progress. Its diverse content enables the design of benchmarks for open-ended tasks such as writing [52], research evaluation [40], persuasion [53–55], and summarization [16, 56], providing valuable data for both open-ended LLM and agentic post-training and evaluation [57].
3. **Supporting multi-dimensional alignment and reasoning research through scientific writing and discussion.** Existing benchmarks for alignment and reasoning often rely on static, synthetic, or crowd-sourced datasets that lack the depth and nuance of real expert deliberation [15, 58–62].

[1] `https://openreview.net/`

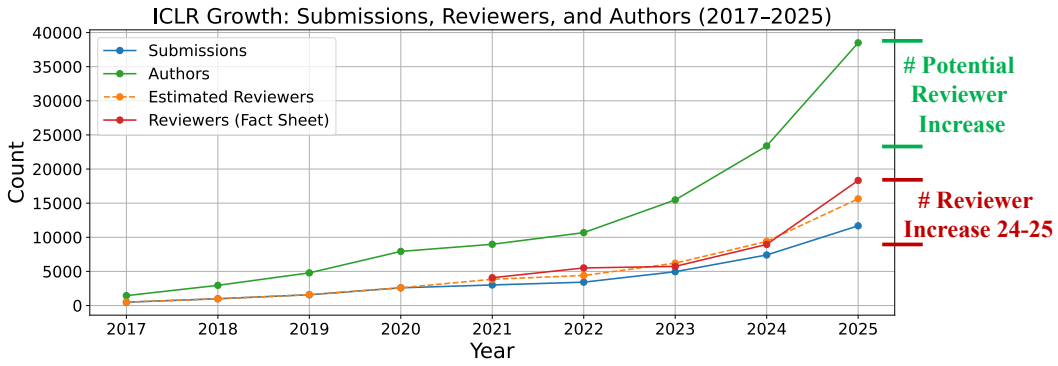

Figure 2: Growth trends at ICLR (2017–2025) in submissions, authors, and reviewers. While the number of reviewers has increased over time, it has not kept pace with the growth in submissions and authors, indicating a growing strain on the peer review process. The reviewer number estimation is calculated according to the number of submissions, the total number of reviews received, and the average reviewer workload of 3 per reviewer.

In contrast, OpenReview offers a setting that inherently involves alignment and reasoning through evidence-based debate, disagreement, revision, and consensus building. This setting enables rich evaluation tasks such as score justification via retrieval-based reasoning [63–66] and decision prediction grounded in free-form critique [67]. These tasks can serve as realistic testbeds for assessing how well LLMs can interpret, reason about, and align with expert judgments in the scientific research domain. Moreover, the dialogic nature of OpenReview — spanning rebuttals, conflicting views, and negotiated outcomes — offers a unique opportunity to study value pluralism, debate-style alignment in the wild [53, 68–71].

To help realize the potential in and beyond those outlined use cases, we propose initial directions for community-driven benchmark development and responsible data stewardship. Finally, we reflect on alternative perspectives, aiming to spark productive dialogue on the challenges and risks of leveraging the OpenReview as a core community asset.

2 The State of OpenReview Now: Scale, Opportunity, and Emerging Risks

This section examines the OpenReview platform through three perspectives. We begin with a statistical overview of its scale and evolution, using ICLR as a case study. We then highlight its value as a unique community-curated dataset for machine learning research, before turning to the structural risks that threaten its long-term quality and integrity.

2.1 The Scale and Structure of Conference Data on OpenReview

OpenReview provides a centralized platform for peer review and community discussion at major machine learning conferences, including ICLR, NeurIPS, and others. It preserves structured records of submissions, reviews, rebuttals, and decisions, creating a longitudinal archive of real-world expert deliberation under consistent guidelines.

To illustrate the scale of this platform, we focus on ICLR as a representative case. From 2017 to 2025, the number of submissions grew from fewer than 500 to over 11,600 annually. The corresponding number of authors increased from about 1,500 to 38,500, and the estimated number of reviewers rose from under 1,000 to more than 18,300. Each submission typically receives three or more expert reviews, resulting in tens of thousands of reviewer–author interactions each year. Figure 2 shows this growth trajectory in authorship, reviewing, and participation.[2]

2.2 A Rapidly Growing Community Asset for Learning

Beyond its scale, OpenReview is distinguished by its unique data quality. Unlike synthetic or crowd-sourced datasets, it captures expert-authored evaluations tied to real submissions and decisions,

[2]Data Source: ICLR 2021-2025 Fact Sheet [72–76], PaperCopilot [77].

grounded in shared scientific norms. Each paper serves as a self-contained research scenario, typically accompanied by multiple reviews, optional rebuttals, meta-reviews, and final outcomes.

Between 2017 and 2025, ICLR alone contributed over 36,000 such interaction threads, spanning both accepted and rejected submissions. These interactions provide rich examples of open-ended scientific exploration efforts. They illustrate how researchers conduct and evaluate solutions to open questions, respond to disagreement, clarify claims, and finally construct consensus, making them highly suitable for training and evaluating LLMs on scientific reasoning, argumentation, and alignment.

Moreover, OpenReview is a continuously evolving dataset. Each year brings new topics, new papers, and new debates, reflecting both the state of research and the shifting consensus of the community. This ongoing refresh ensures its competence as a benchmark for real-world LLM deployment. In Table 1, we compare relevant tasks in the LLM post-training community to demonstrate the general potential of the OpenReview dataset.

Table 1: Comparing datasets related to OpenReview. We will elaborate how to leverage OpenReview beyond those tasks in Sec.3-5.

Dataset	Task	Size	Expert	Updates	OpenEnded
See et al. [78]	Summarization	310K	✓	×	✓
Narayan et al. [79]	Summarization	226K	✓	×	✓
Yang et al. [80]	Multi-hop QA	113K QA pairs	×	×	✓
Rajpurkar et al. [81]	Comprehension	107K QA pairs	×	×	×
Fan et al. [82]	Long-form QA	270K threads	×	×	✓
Ziegler et al. [83]	Preference Modeling	60K comparisons	∼	×	✓
Bai et al. [15]	Alignment / Dialogue	170K comparisons	∼	×	✓
Köpf et al. [84]	Dialogue / Alignment	10K trees, 161K msg	∼	×	✓
Wang et al. [85]	Argumentation	1K dialogues	×	×	✓
Kang et al. [56]	Review, Decision	14.7K subs, 10.7K revs	✓	×	×
Bu et al. [86]	Aspect Rating (zh)	46.7K reviews	∼	×	×
Purkayastha et al. [87]	Argumentation	2.3K	∼	×	✓
Kennard et al. [88]	Argumentation	506 threads	✓	×	×
Ruggeri et al. [89]	Argumentation	41 dialogues	✓	×	✓
OpenReview [45]	**All above**	36K subs, 100K+ revs	✓	✓	✓

2.3 Quality Under Pressure — The Compounding Risk of Rapid Growth

While the growth of OpenReview presents significant opportunities, it also introduces structural risks. The rapid increase in submission volume has not been matched by a proportional increase in highly experienced reviewers. As conferences scale, an increasing fraction of reviews are written by newer or less engaged participants. This trend raises concerns about the consistency, reliability, and long-term stability of individual review signals, as well as the dataset quality and diversity [50].

More precisely, the concern is not only that current reviewers may deviate from academic standards, but that a growing number of untrained reviewers may internalize and reproduce biased practices, gradually compounding the problem across generations. If evaluations are learned by imitation, biased or inconsistent norms can propagate, leading to long-term degradation of review quality.

To formalize this concern, we present a Wright-Fisher model in Appendix A, which illustrates how misaligned reviewing behavior may propagate across generations.

Take Action Now. Our analysis suggests that early intervention is critical: corrective action taken before problematic patterns become institutionalized is significantly more effective than attempting to reverse them later. Proactive steps are thus essential to preserve long-term alignment between reviewing practice and community values. **Taking action now in the early stage of the field's expansion is more effective than taking action later on when substandard review practices become the norm.**

For OpenReview to remain a robust and trustworthy resource, its quality must be actively protected. This includes better reviewer recruitment and training, as well as developing scalable, practical machine learning methods for auditing and mitigating quality drift. The data itself, while valuable, is only as good as the process that generates it.

In the following sections, we will discuss three use cases of the OpenReview dataset, starting from how to leverage the dataset to improve and regulate the peer review system, such that the long-term quality of such a community asset can be guaranteed. We then highlight the potential of leveraging such an asset in LLM post-training research, ranging from open-ended to alignment tasks.

3 Assisting and Protecting the Peer Review with OpenReview

3.1 Existing LLM-Assisted Peer Review in Conferences

In the previous section, we highlighted structural risks to the quality and stability of peer review. These concerns have not gone unnoticed. In recent years, several major machine learning conferences and publishers have begun integrating LLMs into their review workflows in response.

NeurIPS 2024 introduced a checklist assistant powered by LLMs to help authors ensure ethical and methodological compliance [90]. At ICLR 2025, a Review Feedback Agent was deployed to identify vague or unconstructive reviews and suggest targeted improvements [91]. AAAI 2026 will experiment with LLM-generated reviews and discussion summaries in the first stage of review [92]. Meanwhile, several academic publishers have begun piloting AI-assisted tools for content checking and review drafting [93–95].

While most current systems operate with limited, hand-curated inputs, OpenReview provides an ideal foundation for data-driven peer review research. In this section, we focus on concrete use cases where such data can support the review system.

3.2 Practices and Opportunities for Data-Driven Support with OpenReview

We organize existing literature and potential opportunities with OpenReview according to functional categories. In the following, we will use **text boxes** to highlight **future work opportunities**. The high-level motivation of those approaches is rooted in the previous success of human-centered LLM alignment research [17, 15, 16], and data-driven decision modeling and explanation [96–100].

Principled Review Generation. Recent work has explored OpenReview for generating *realistic* peer reviews. Yuan et al. [101] and Wu et al. [102], for example, demonstrate that fine-tuning LLMs on large-scale review corpora can lead to critiques that are more calibrated and grounded than those produced by general-purpose models. These systems can be conditioned on paper content or specific review dimensions, enabling targeted and context-aware feedback. However, most current systems are designed to *mimic* human-written reviews without deeper integration with formal reviewing guidelines or accountability structures. The challenge remains to ensure that generated reviews uphold conference standards and provide actionable feedback in line with reviewer expectations.

> **Opportunity for Future Work.** LLMs should be task-specifically aligned, calibrated when leveraged in the review process. Commercial LLMs are generally optimized for user-friendliness and helpfulness, often deviating from rigorous academic review guidelines. Future work should explore structured prompting, rubric conditioning, or alignment objectives tailored for review generation [44]. In addition, LLM-generated reviews may support pre-submission preparation [42], providing anticipatory critique to authors and supporting self-assessment before formal peer review [92].

Review Quality Enhancement. Another line of research focuses on the quality of peer reviews themselves. Early work, such as Kang et al. [56], proposed metrics for review helpfulness and score prediction. More recently, classifiers trained on human preferences or meta-review feedback have been developed to detect vague, biased, or uninformative reviews [103, 104]. Studies have also examined hallucination and style inconsistencies in LLM-generated reviews [105–108]. Despite these advances, challenges remain in automatically evaluating review fairness, argument soundness, or reviewer calibration.

> **Opportunity for Future Work.** Inverse analysis techniques can help detect systematic deviation from expected standards, including overconfidence, inconsistency, or subjective bias [97]. Future efforts could explore calibration, value drift detection, and provide warning signals when the value of reviews deviate significantly from guidelines [44].

Enhancing Mutual Understanding between Reviewers and Authors. While much of the focus has been on generating or evaluating individual reviews, peer review is ultimately a dialogue. The rebuttal phase plays a crucial role in bridging perspectives between authors and reviewers. Recent datasets such as DISAPERE [109], Jiu-Jitsu [110], and ContraSciView [111] support tasks such as rebuttal generation, stance classification, and discourse structure prediction, highlighting the interactional nature of review.

Opportunity for Future Work. LLMs can serve as mediators to enhance communication in the rebuttal process. For authors, they may clarify reviewer concerns, highlight overlooked critiques, and assist in crafting respectful and persuasive responses for effective communication. For reviewers, they may help interpret rebuttals and assess whether key feedback has been adequately addressed, and effectively stimulate the discussions.

Consistency and Calibration. Efforts to correct score inconsistency across reviewers have drawn on reviewer calibration and normalization techniques. For instance, Xu et al. [112] models reviewer-specific scoring functions and applies monotonic transformations to improve comparability. These methods aim to recover more faithful rankings than simple score averaging. Nonetheless, current approaches often lack interpretability or real-time applicability. There is limited support for helping reviewers understand their own biases or dynamically recalibrate scores based on peer context.

Opportunity for Future Work. More importantly and effectively, efforts could be made to use LLM-based systems to assist reviewers in providing consistent and calibrated feedback, including providing comparative context and relevant arguments drawn from reviewer cohorts [42]. Technically, this may involve retrieval-based justification of scores and decision explanation [63, 64, 113, 114], or in-context learning reference sample selection [115, 116].

Meta-Review Generation. Finally, meta-review generation has become a growing area of interest, with benchmarks such as PeerSum [104], ORSUM [117], and MOPRD [118] targeting summarizing and concluding from multiple reviews and rebuttals. These systems must integrate conflicting reviewer perspectives, identify dominant themes, and represent area chair judgment with fidelity. Still, current general-purpose LLMs may fail to capture the nuanced reasoning behind disagreements or the weight assigned to various critiques. There is also growing concern about the potential mismatch between generated meta-reviews and actual reviewer consensus [39].

Opportunity for Future Work. Improved modeling of review disagreement and viewpoint clustering [68, 70] could enable more reliable meta-review generation. Future systems may incorporate hybrid workflows where LLMs co-author drafts with area chairs, flag unresolved conflicts, or highlight potential biases (e.g., delayed or biased feedback, ungrounded critiques) throughout the discussion period to support better decision making.

4 OpenReview for Open-Ended Task Evaluation and Post-Training

4.1 Challenges for Open-Ended LLM and Agentic Tasks

Recent progress in LLMs has enabled systems that attempt to perform complex, multi-step, and high-level tasks, often referred to as *open-ended* or *agentic* tasks [119, 120]. These tasks are characterized by the absence of a single correct answer, dependence on context, and the need for judgment, reasoning, and creativity [52]. Examples include research paper writing, paper reviewing, persuasive argumentation, hypothesis refinement, and code-based experimentation [121, 122]. Open-ended tasks are defined not by accuracy or success alone, but by depth, coherence, exploration, and alignment with human values and intentions.

This task category has received increasing attention with the rise of agent-based systems such as DeepResearch, DeepSearch, and AutoDev, which aim to position LLMs as autonomous research assistants capable of conducting literature reviews, designing experiments, debugging code, and evaluating progress [8, 123, 124]. However, a major bottleneck in building and benchmarking such systems lies in the lack of scalable, high-quality supervision. It remains difficult to evaluate whether a model has conducted a "good" literature review or proposed a "promising" research idea, particularly when using crowd-sourcing judgment [52, 121].

Scientific research, especially in machine learning, is itself an open-ended task. The process involves formulating problems, iterating on designs, running experiments, interpreting results, engaging with criticism, and ultimately persuading a community of experts. Despite this, most benchmarks for evaluating LLMs remain synthetic, short-form, or not scalable, offering little insight into how models would perform under the standards and expectations of actual research communities [121, 122].

This gap motivates our focus on OpenReview as a valuable, underutilized resource. The rich interactions on OpenReview suggest two distinct forms of supervision that are particularly suited for open-ended task development.

4.2 Two Potential Supervision Streams from OpenReview

Scientific Demonstrations: Training LLMs to Do Research. Each submitted paper on OpenReview can be viewed as a real-world demonstration of open-ended problem-solving. Papers span a wide range of topics and contain full narratives of how authors design and communicate their contributions. This includes technical framing, literature positioning, experimental results, and claim justification. In aggregate, these documents offer structured demonstrations of how research is conceived, executed, and defended [56].

Such examples can be used to train LLMs to follow the cognitive workflow of scientific research. In particular, they can support training for complex capabilities such as multi-stage planning, tool use, fact retrieval, and hypothesis revision. These capabilities align closely with the demands of emerging agentic LLM frameworks [125]. While systems like ChatDev simulate these workflows [126], few are grounded in real, high-quality demonstrations of how experts actually perform these tasks — OpenReview offers a scalable source of such supervision.

Opportunity for Future Work. OpenReview's corpus of research demonstrations can support training of LLM agents to perform multi-step scientific reasoning under real-world constraints. Future work may consider enhancing the agentic research capabilities of LLMs [8] using expert scientific research demonstrations.

Structured Evaluations: Training LLMs to Evaluate Research. In addition to research demonstrations, OpenReview also contains detailed records of how experts evaluate open-ended research work. Reviews provide constructive feedback, numerical scores, and qualitative assessments, while meta-reviews offer consensus summaries and rationales for decisions. Author responses further enrich the discourse, revealing how researchers engage with critiques and attempt to clarify or defend their contributions. These dual supervision signals are particularly valuable for developing and evaluating general-purpose models intended to reason about, participate in, and evaluate complex open-ended tasks given scientific standards. By learning from those debates, LLMs have the potential to gain the capability to comprehensively evaluate open-ended research.

Opportunity for Future Work. OpenReview's review traces can serve as supervision for LLM-based evaluators trained to judge open-ended research quality. These include automated meta-reviews, rebuttal critiques, and scoring models aligned with human preferences. With those feedback-rich reward models for open-ended tasks, future work can better anchor and be optimized for open-ended research.

5 OpenReview as High-Quality Dataset for Alignment and Reasoning

5.1 Challenges for Alignment and Reasoning Supervision

Alignment through Consensus-Seeking. Alignment research seeks to ensure that AI systems act according to human values, preferences, satisfy human intentions, and guarantee safety [22, 127]. Recent advances in reinforcement learning from human feedback (RLHF) [14–16, 128–134] have contributed to the success of LLMs in conversational systems [14]. Yet many of these advances rely on limited forms of supervision: crowd-sourcing annotations [26], synthetic preferences [59], or binary votes [135]. These sources often fail to capture the complexity, depth, and disagreement inherent in the multi-perspective and deliberative consensus-seeking processes of experts [129, 131, 136].

Reasoning beyond Binary Tasks. On the other hand, reasoning ability has become the core in enhancing the models' performance on more general assistant tasks [137]. Contributing to such progress, datasets such as GSM8K [138], MATH [139], and HotpotQA [80] have driven rapid progress in mathematical and multi-hop reasoning; techniques like (long-)chain-of-thought [66, 65, 140, 7] and retrieval-augmented methods [63, 64, 113] have significantly improved model performance on these structured tasks. However, many of these benchmarks are now nearly saturated by frontier models [141], focus on binary and verifiable tasks, and they predominantly focus on final answer correctness rather than the quality or interpretability of reasoning processes [142, 143].

More fundamentally, current reasoning tasks are often limited by narrow scope, synthetic formulation, or rigid answer structures [144–146]. Most define a single ground-truth answer, which precludes exploration of ambiguity, disagreement, or multi-agent deliberation, which are central to human reasoning, but effective in eliciting deep thinking behaviors [147, 137]. Although emerging datasets in argumentative reasoning, such as DebateSum [148] and OpenDebateEvidence [149], have expanded the scope of evaluation to include summarization and contested claims, these resources remain rare and are typically not grounded in scientific domain expert-generated contexts.

5.2 Opportunities with the OpenReview Dataset

In contrast, OpenReview offers a fundamentally different alignment and reasoning testbed. The peer review process inherently involves dialogue in which multiple parties express values, critique reasoning, and negotiate consensus. More importantly, those dialogues, in principle, should be ***objective***, centered around guidelines, and grounded in verifiable facts. Unlike existing alignment datasets, which are largely ***subjective***, static, and one-shot, OpenReview captures multi-round, multi-agent interactions grounded in real, highly verifiable, and reproducible consequences. This makes it a uniquely rich environment for alignment and reasoning research.

Learning to Reason from Expert Disagreement and Justification With OpenReview, models can be trained to infer about the logic behind review scores by learning from rationales, a form of inverse reasoning that links decisions to supporting arguments and context. The reviews themselves often present well-defined reasoning chains that connect experimental design, observed outcomes, and stated conclusions. These examples allow LLMs to practice multi-step reasoning, assess methodological soundness, and trace causal explanations. Moreover, OpenReview enables modeling how reasoning develops through multiple rounds of interaction: authors respond to critiques, reviewers clarify concerns, and final evaluations synthesize evolving viewpoints, offering a natural setting for studying the rationale behind reasoning over time.

Opportunity for Future Work. Using OpenReview in future works, it's possible to improve models' reasoning abilities by justifying numerical assessments, verifying scientific claims through factual evidence, and adapting reasoning across multi-stage interactions.

Learning to Critically Align with Individual Preferences OpenReview provides a valuable foundation for developing alignment strategies that move beyond superficial agreement. Unlike many alignment datasets that prioritize helpfulness or user-pleasing responses, peer review process demands that feedback remain grounded in correctness, rationality, and align with review guidelines, when given diverse research contexts.

Each reviewer expresses their judgment through both numerical scores and detailed commentary, guided by criteria such as novelty, technical soundness, and significance. These preferences are dynamic and can shift in response to rebuttals and clarifications, offering supervision signals for modeling alignment as a contextual and adaptive process.

Opportunity for Future Work. OpenReview enables the alignment of LLMs to offer diverse, constructive, evidence-based critique. Rather than merely affirming user input, models can learn to respectfully challenge flawed claims, explain counterarguments, and justify disagreement. This supports the development of alignment systems that emphasize factual grounding, logical reasoning, and responsible communication.

6 A Call to Create Standardized Benchmarks Based on OpenReview

In this section, we turn to the foundational infrastructure required to realize their full potential: standardized benchmarks and responsible community stewardship.

Despite its scale and richness, OpenReview remains underutilized as a research asset, primarily due to the lack of well-defined tasks and shared evaluation pipelines. To address this gap, we propose that the community collaboratively develop benchmarks in critical areas such as review quality assessment, rebuttal generation, argument grounding, and meta-review summarization. And deploy developed methods to intervene in the peer review system and improve its quality as soon as possible. These tasks are directly tied to the health of the peer review process and, by extension, the integrity of the dataset itself.

In parallel, more general tasks—including reviewer score prediction, open-ended task evaluation, post-training, alignment, and reasoning enhancement—can also be standardized to support long-term research. While these areas are essential to the development of LLMs, their delayed investigation is less likely to compromise the quality or sustainability of OpenReview as a resource.

We call upon researchers, conference organizers, and practitioners—particularly those working at the intersection of machine learning and language models—to jointly define, refine, and adopt such benchmarks. This collaborative process must also engage with broader ethical considerations, including the protection of author and reviewer privacy, responsible anonymization of sensitive content, and the mitigation of representational biases. For example, research areas with more abundant data may inadvertently dominate the training signal, potentially skewing the learned priorities of evaluation models.

Ultimately, the continued value of OpenReview as a shared academic asset depends on proactive, collective stewardship by the community it serves.

7 Alternative Views

LLM-based Review and Research. Some may argue that LLMs are becoming more and more capable of finishing scientific research and evaluation, and should eventually replace human reviewers. If models can predict review scores or generate critiques that approximate expert judgment, then preserving human oversight might appear unnecessary or even inefficient. **Our view**: We argue that peer review is not just a filtering mechanism, but a deliberative process that helps shape scientific values and standards [150, 151]. Over-reliance on automation risks eroding its collaborative and interpretive nature [152, 153]. LLMs, while powerful, are not reliable in reasoning with the same contextual grounding or responsibility as human experts [154, 155]. Human reviewers must remain responsible for interpreting and controlling LLM tools [154]. Interactions between authors and reviewers should stay dialogic and grounded in fairness, not reduced to rigid or opaque evaluations [152]. Moreover, systems must guard against hallucination, adversarial misuse, and bias propagation [156–158]. Evaluation frameworks built on OpenReview should align with scientific values rather than model evaluation metrics [22, 106].

Inconsistency of Peer Review Limits Its Usefulness for Alignment. One concern is that peer review data may be too noisy or inconsistent to serve as a reliable supervision signal [159]. Reviewers often disagree on paper quality, assign divergent scores, or emphasize different aspects of a submission. Given this subjectivity, it may be argued that using such data for alignment could reinforce inconsistent or unstable behaviors in LLMs. **Our view**: Rather than aiming for deterministic consensus, alignment in this context involves modeling disagreement, grounding claims, and reasoning for underlying conflicts. This perspective is increasingly emphasized in recent alignment literature [68, 70]

Scientific Review Tasks May be Too Narrow to Generalize. Another possible objection is that scientific reviewing and paper writing are narrow, domain-specific tasks that do not generalize to broader LLM capabilities. Models trained on OpenReview may excel at research-related tasks but fail to transfer to everyday use cases, limiting their value as general-purpose assistants. **Our view**: Research tasks serve as high-complexity instances of structured human reasoning, with grounded stakes and verifiable outcomes. Learning from these tasks offers not only domain expertise but also training in core cognitive patterns that generalize across domains. Recent success of DeepResearch-type of products [8] explicitly aim to generalize research workflows into agentic LLM behaviors.

References

[1] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42, 2025.

[2] Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Navigating data-centric artificial intelligence with dc-check: Advances, challenges, and opportunities. *IEEE Transactions on Artificial Intelligence*, 5(6):2589–2603, 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways, 2022.

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[7] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[8] OpenAI. Introducing deep research, 2025. Accessed: 2025-04-16. Deep Research is a new agentic AI capability integrated within ChatGPT that autonomously conducts multi-step web research and synthesizes comprehensive reports.

[9] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

[10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

[11] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[12] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[13] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[15] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[16] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[17] David Silver and Richard S Sutton. Welcome to the era of experience. *Google AI*, 2025.

[18] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[19] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[20] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`.

[21] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024.

[22] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[23] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[25] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. *Noise: A flaw in human judgment*. Hachette UK, 2021.

[26] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023.

[27] Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*, 2024.

[28] Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. Aligning with logic: Measuring, evaluating and improving logical consistency in large language models. *arXiv preprint arXiv:2410.02205*, 2024.

[29] Mert Karabacak and Konstantinos Margetis. Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5), 2023.

[30] Conner Ganjavi, Michael B Eppler, Asli Pekcan, Brett Biedermann, Andre Abreu, Gary S Collins, Inderbir S Gill, and Giovanni E Cacciamani. Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *bmj*, 384, 2024.

[31] Elsevier. Elsevier takes Scopus to the next level with generative AI. `https://www.elsevier.com/about/press-releases/elsevier-takes-scopus-to-the-next-level-with-generative-ai`, Aug 2023.

[32] Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert Van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.

[33] Benedikt Fecher, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. Friend or foe? exploring the implications of large language models on the science system. *Ai & Society*, 40(2):447–459, 2025.

[34] Ross Taylor, Bingqing Wen Kardas, Sid McLean, Gabriel Brown, David Agarwal, Mo outlandish Bogin, Eric Michaels, Eric Hillestad, Hongyu Jiang, Danica Dang, et al. Galactica: A large language model for science, 2022.

[35] Holly Else. By chatgpt fool scientists. *Nature*, 613:423, 2023.

[36] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

[37] Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. Autosurvey: Large language models can automatically write surveys. *Advances in Neural Information Processing Systems*, 37:115119–115145, 2024.

[38] Serge P. Horbach. The peer-review crisis: a commentary. *Scientometrics*, 118(2):699–704, 2019. doi: 10.1007/s11192-018-03031-6.

[39] Zeyuan Allen-Zhu and Xiaoli Xu. DOGE: Reforming AI Conferences and Towards a Future Civilization of Fairness and Justice. *SSRN Electronic Journal*, February 2025. doi: 10.2139/ssrn.5127931. `https://ssrn.com/abstract=5127931`.

[40] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

[41] Asheesh Kumar, Tirthankar Ghosal, Saprativa Bhattacharjee, and Asif Ekbal. Towards automated meta-review generation via an nlp/ml pipeline in different stages of the scholarly peer review process. *International Journal on Digital Libraries*, 25(3):493–504, 2024.

[42] Ryan Liu and Nihar B Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.

[43] Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, 2024.

[44] Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*, 2025.

[45] Dagstuhl Publishing Soergel. Openreview: A venue for open peer review, open publishing, open access. In *ICLR platform whitepaper*, 2013.

[46] Google DeepMind. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, 2025. Accessed: 2025-05-19.

[47] OpenAI. Gpt-4.5 system card. https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf, 2025. Technical Report.

[48] Anthropic. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family, 2024. Claude 3 Release Announcement.

[49] xAI. Grok 3 beta — the age of reasoning agents. https://x.ai/news/grok-3, 2025. Grok 3 Release Announcement.

[50] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Author correction: Ai models collapse when trained on recursively generated data. *Nature*, 640(8058):E6, 2025.

[51] Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-endedness is essential for artificial superhuman intelligence. *arXiv preprint arXiv:2406.04268*, 2024.

[52] Sian Gooding, Lucia Lopez-Rivilla, and Edward Grefenstette. Writing as a testbed for open ended agents. *arXiv preprint arXiv:2503.19711*, 2025.

[53] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

[54] Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*, 2024.

[55] Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 152–163, 2024.

[56] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*, 2018.

[57] Hao Sun, Thomas Pouplin, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. Improving llm generation with inverse and forward alignment: Reward modeling, prompting, fine-tuning, and inference-time optimization. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

[58] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[59] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

[60] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

[61] Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 1.4 million open-source distilled reasoning dataset to empower large language model training. *arXiv preprint arXiv:2503.19633*, 2025.

[62] Justus Mattern, Sami Jaghouar, Manveer Basra, Jannik Straube, Matthew Di Ferrante, Felix Gabriel, Jack Min Ong, Vincent Weisser, and Johannes Hagemann. Synthetic-1: Two million collaboratively generated reasoning traces from deepseek-r1, 2025. URL https://www.primeintellect.ai/blog/synthetic-1-release.

[63] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[64] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[65] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[67] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40, 2021.

[68] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[69] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

[70] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

[71] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951*, 2024.

[72] Iclr 2021 fact sheet. `https://iclr.cc/media/Press/ICLR_2021_Fact_Sheet.pdf`, 2021.

[73] Iclr 2022 fact sheet. `https://iclr.cc/media/Press/ICLR_2022_Fact_Sheet.pdf`, 2022.

[74] Iclr 2023 fact sheet. `https://media.iclr.cc/Conferences/ICLR2023/ICLR2023-Fact_Sheet.pdf`, 2023.

[75] Iclr 2024 fact sheet. `https://media.iclr.cc/Conferences/ICLR2024/ICLR2024-Fact_Sheet.pdf`, 2024.

[76] Iclr 2025 fact sheet. `https://media.iclr.cc/Conferences/ICLR2025/ICLR2025_Fact_Sheet.pdf`, 2025.

[77] Jing Yang. Paper copilot: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. *arXiv preprint arXiv:2502.00874*, 2025.

[78] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *ACL*, 2017.

[79] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, 2018.

[80] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

[81] Pranav Rajpurkar, Jian Zhang, Igor Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.

[82] Angela Fan, Yacine Jernite, Jason Weston, and David Grangier. Eli5: Long form question answering. In *ACL*, 2019.

[83] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, and Paul F. Christiano. Fine-tuning language models from human preferences. In *NeurIPS Workshop on Human in the Loop Learning*, 2019.

[84] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681, 2023.

[85] Xinyao Wang, Han Zhang, Noah A. Smith, and Yejin Choi. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *ACL*, 2019.

[86] Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. Asap: A chinese review dataset towards aspect category sentiment analysis and rating prediction. *arXiv preprint arXiv:2103.06605*, 2021.

[87] Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. Exploring jiu-jitsu argumentation for writing peer review rebuttals. *arXiv preprint arXiv:2311.03998*, 2023.

[88] Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. Disapere: A dataset for discourse structure in peer review discussions. *arXiv preprint arXiv:2110.08520*, 2021.

[89] Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. Argscichat: A dataset for argumentative dialogues on scientific papers. *arXiv preprint arXiv:2202.06690*, 2022.

[90] Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B Shah. Usefulness of llms as an author checklist assistant for scientific papers: Neurips'24 experiment. *arXiv preprint arXiv:2411.03417*, 2024.

[91] James Zou and Omkar Thakkar. Leveraging llm feedback to enhance review quality at iclr 2025. `https://iclr.cc/Conferences/2025/Blog/LLMReviewSupport`, 2025.

[92] Aaai launches ai-powered peer review assessment system. `https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/`, 2025.

[93] Miryam Naddaf. Ai is transforming peer review — and many scientists are worried. *Nature*, 618:456–458, 2025.

[94] Springer nature unveils geppetto ai to improve peer review integrity. `https://group.springernature.com/geppetto-ai`, 2024.

[95] James Zou. Chatgpt is transforming peer review—how can we use it responsibly? *Nature*, 635 (8037):10–10, 2024.

[96] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[97] Daniel Jarrett, Alihan Hüyük, and Mihaela Van Der Schaar. Inverse decision modeling: Learning interpretable representations of behavior. In *International Conference on Machine Learning*, pages 4755–4771. PMLR, 2021.

[98] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[99] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[100] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

[101] Shuyang Yuan, Lu Wang, and Noah A. Smith. Asap-review: Towards automating scientific paper review with argument-aware summary generation. In *Proceedings of EMNLP*, 2022.

[102] Zeyu Wu, Marcus Idahl, and Nikhil Shah. Openreviewer: A specialized llm for generating critical scientific paper reviews. *arXiv preprint arXiv:2310.12345*, 2023.

[103] Sudarshan Rao, Lexing Xie, and Sameer Singh. Detecting llm-written peer reviews. *arXiv preprint arXiv:2503.12345*, 2025.

[104] Xingchen Li, Yining Qian, Xinyu Li, and Yue Zhang. Peersum: Opinion-aware summarization of scientific peer reviews. In *Findings of EMNLP*, 2023.

[105] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[106] W Liang, Y Zhang, H Cao, B Wang, D Ding, X Yang, K Vodrahalli, S He, D Smith, Y Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. arxiv. *Preprint*, 2023.

[107] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[108] Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. Llms assist nlp researchers: Critique paper (meta-) reviewing. *arXiv preprint arXiv:2406.16253*, 2024.

[109] Alice Wang, Heeyoung Kim, and Graham Neubig. Disapere: Discourse-aware sentence-level argument pair extraction for peer review. *arXiv preprint arXiv:2402.12345*, 2024.

[110] Debanjan Purkayastha, Rahul Sarkar, et al. Jiu-jitsu: Rebuttal generation by using reviewers' comments against them. In *Findings of ACL*, 2023.

[111] Rahul Kumar, Sujay Ravi, et al. Contrasciview: Detecting contradictions in scientific peer reviews. In *Proceedings of ACL*, 2023.

[112] Chen Xu, Yao Li, Ivan Stelmakh, and Nihar B Shah. Least-squares calibration for peer reviews. In *NeurIPS*, 2021.

[113] Thomas Pouplin, Hao Sun, Samuel Holt, and Mihaela Van der Schaar. Retrieval-augmented thought process as sequential decision making. *arXiv preprint arXiv:2402.07812*, 2024.

[114] Hao Sun, Alihan Hüyük, Daniel Jarrett, and Mihaela van der Schaar. Accountable batched control with decision corpus. *Advances in Neural Information Processing Systems*, 36, 2023.

[115] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*, 2022.

[116] Tai Nguyen and Eric Wong. In-context example selection with influences. *arXiv preprint arXiv:2302.11042*, 2023.

[117] Yutong Zeng et al. Orsum: A dataset for paper meta-review generation via scientific opinion summarization. *arXiv preprint arXiv:2403.11234*, 2024.

[118] Yibo Lin, Wenhao Wang, et al. Moprd: Multidisciplinary open peer review dataset. In *Proceedings of LREC-COLING*, 2024.

[119] Maya Murad. Large language models revolutionized ai. llm agents are what's next. IBM Research Blog, 18 Jul 2024, 2024. URL: https://research.ibm.com/blog/what-are-ai-agents-llm.

[120] Yefei Liu, Prithviraj Ammanabrolu, David Demeter, et al. Agentic large language models: A survey. arXiv preprint arXiv:2503.23037, 2025.

[121] Angel Yanguas-Gil, Matthew T. Dearing, Jeffrey W. Elam, Jessica C. Jones, Sungjoon Kim, Adnan Mohammad, Chi Thang Nguyen, and Bratin Sengupta. Benchmarking large language models for materials synthesis: the case of atomic layer deposition. *Journal of Vacuum Science & Technology A*, 43(4):041401, 2025. doi: 10.1116/6.0004319.

[122] Open Philanthropy. Request for proposals: Benchmarking llm agents on consequential real-world tasks. Open Philanthropy (Feb 2024), 2024. URL: https://www.openphilanthropy.org/rfp-llm-benchmarks/.

[123] xAI. Grok 3 beta – the age of reasoning agents. xAI News, Nov 3, 2024, 2024. URL: https://x.ai/news/grok-3.

[124] Michele Tufano, Anisha Agarwal, Jinu Jang, Roshanak Zilouchian Moghaddam, and Neel Sundaresan. Autodev: Automated ai-driven development. *arXiv preprint arXiv:2403.08299*, 2024.

[125] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[126] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. *Proceedings of ACL 2024 (to appear), arXiv:2307.07924*, 2024.

[127] Rohin Shah, Alex Irpan, Alexander Matt Turner, Anna Wang, Arthur Conmy, David Lindner, Jonah Brown-Cohen, Lewis Ho, Neel Nanda, Raluca Ada Popa, et al. An approach to technical agi safety and security. *arXiv preprint arXiv:2504.01849*, 2025.

[128] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[129] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

[130] Hao Sun and Mihaela van der Schaar. Inverse-rlignment: Inverse reinforcement learning from demonstrations for llm alignment. *arXiv preprint arXiv:2405.15624*, 2024.

[131] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

[132] Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*, 2024.

[133] Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*, 2023.

[134] Yunyi Shen, Hao Sun, and Jean-François Ton. Reviving the classics: Active reward modeling in large language model alignment. *arXiv preprint arXiv:2502.04354*, 2025.

[135] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[136] Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*, 2024.

[137] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[138] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[139] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[140] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

[141] Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[142] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

[143] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[144] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

[145] Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey. *arXiv preprint arXiv: 2407.05000*, 2024.

[146] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025.

[147] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025.

[148] Allen Roush and Arvind Balaji. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining (ArgMining@COLING)*, pages 1–7, 2020.

[149] Allen Roush, Yusuf Shabazz, Arvind Balaji, Peter Zhang, Stefano Mezza, Markus Zhang, Sanjay Basu, Sriram Vishwanath, Mehdi Fatemi, and Ravid Shwartz-Ziv. OpenDebateEvidence: A massive-scale argument mining and summarization dataset. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

[150] Mohammad Hosseini and Serge P. Horbach. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res. Integr. Peer Rev.*, 8, 2023. doi: 10.1186/s41073-023-00133-5.

[151] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum.-Comput. Interact.*, 36(6):495–504, 2020. doi: 10.1080/10447318.2020.1741118.

[152] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. Towards human-AI deliberation: Design and evaluation of LLM-empowered deliberative AI for ai-assisted decision-making. arXiv preprint arXiv:2403.16812, 2024. URL https://arxiv.org/abs/2403.16812.

[153] Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-agent review generation for scientific papers. arXiv preprint arXiv:2401.04259, 2024.

[154] Iddo Drori and Dov Te'eni. Human-in-the-loop AI reviewing: Feasibility, opportunities, and risks. *J. Assoc. Inf. Syst.*, 25(1):98–109, 2024. doi: 10.17705/1jais.00867. URL https://aisel.aisnet.org/jais/vol25/iss1/7.

[155] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

[156] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, 2021. doi: 10.1145/3442188.3445922.

[157] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. doi: 10.48550/arXiv.1908.07125. arXiv:1908.07125.

[158] Laura Weidinger, John Mellor, Maribeth Rauh, and et al. Ethical and social risks of harm from language models. *Adv. Neural Inf. Process. Syst.*, 34:12872–12886, 2021. URL https://arxiv.org/abs/2112.04359.

[159] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. The neurips 2021 consistency experiment. *Neural Information Processing Systems blog post, https://blog.neurips. cc/2021/12/08/the-neurips-2021-consistency-experiment*, 2021.

A Long-lasting effect of low-quality reviews

Here, we provide a simplistic population genetic model to capture our intuition that a fast-growing reviewer body's lack of training can have a long-lasting effect even after the field matures by setting up precedent. Note that this is an extremely simplified model, and we acknowledge that reviewer quality is not binary and can depend on many factors.

We follow the standard Wright-Fisher model in population genetics. For each review round t, there are G_t "good" reviews and B_t bad reviews (in total $N_t = G_t + B_t$ reviews). In generation $t+1$, for N_{t+1} new reviews, we model them as generated randomly, with some level of preference. Formally

$$B_{t+1} \sim \text{Binomial}\left(N_{t+1}, \frac{B_t}{(1+s(t))G_t + B_t}\right) \tag{1}$$

where $s(t)$ is a factor for preference that could change over time, ideally $s(t) > 0$ so that one has a preference towards writing less bad reviews than simply replicating what was seen in the past cycle. We define $X_t = \frac{B_t}{N_t}$.

We can take a diffusion limit of the model, and the proportion of bad reviews can be approximated as a Wright-Fisher SDE

$$dX_t = s(t)X_t(1-X_t)dt + \sqrt{\frac{X_t(1-X_t)}{N(t)}}dW_t \tag{2}$$

where W_t is a one-dimensional Brownian motion.

We numerically solve the corresponding Fokker-Planck equation for different $N(t)$ and intervention $s(t)$. We assume that $N(t)$ follow a logistic growth representing the usual maturing of the field. The results are given in Fig.3. **The takeaway message is that we need to act early in stopping the trend of preferring low-quality reviews to prevent the downgrade of overall quality and setup the precedent for the next generation to follow.** The trend can still be reversed in a mid to late stage, but requires more efforts (cf. first and last row in Fig.3, we need a longer period of selection if we started late). It is useful even just to stop, instead of reverting, the current trend of preferring bad reviews (cf. second row of Fig.3). The intuition behind these results is that if a once rare bad review was fixed into the norm during the expansion of the field, it will be part of the norm and hard to be filtered out in the future when the field grows even larger.

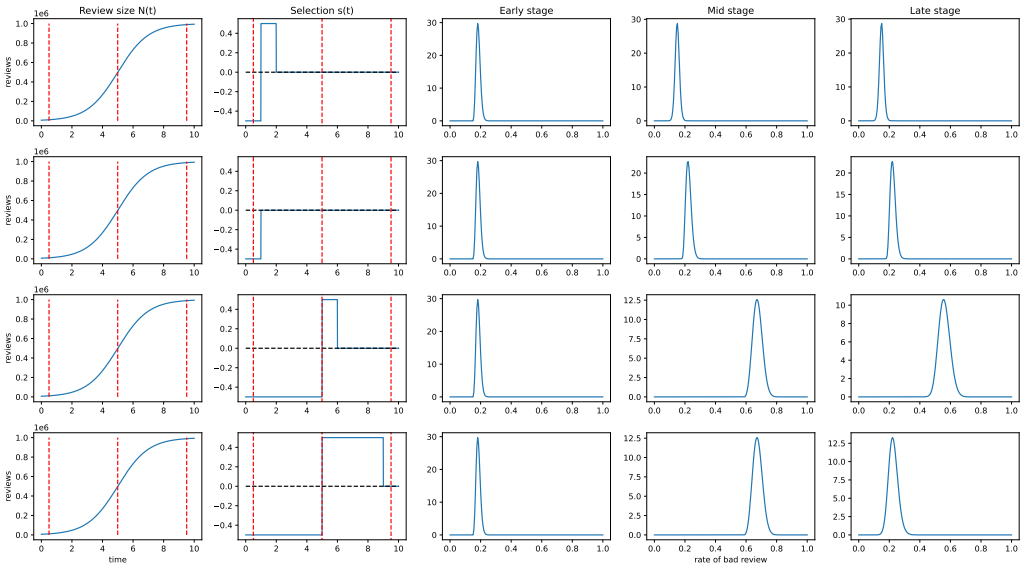

Figure 3: Distribution of frequency of bad reviews under Wright-Fisher type of selection model. The three stages of time are marked in red vertical lines in the first two panels. First column: model number of reviews, Second: what selection we put at which time, Third-last: distribution of proportion of bad reviews.