

---

# The AI Imperative: Scaling High-Quality Peer Review in Machine Learning

---

**Qiyao Wei\***

University of Cambridge

**Samuel Holt\***

University of Cambridge

**Jing Yang**

University of Southern California

**Markus Wulfmeier†**

Google DeepMind

**Mihaela van der Schaar‡**

University of Cambridge

## Abstract

Peer review, the bedrock of scientific advancement in machine learning (ML), is strained by a crisis of scale. Exponential growth in manuscript submissions to premier ML venues such as NeurIPS, ICML, and ICLR is outpacing the finite capacity of qualified reviewers, leading to concerns about review quality, consistency, and reviewer fatigue. This position paper argues that AI-assisted peer review must become an urgent research and infrastructure priority. We advocate for a comprehensive AI-augmented ecosystem, leveraging Large Language Models (LLMs) not as replacements for human judgment, but as sophisticated collaborators for authors, reviewers, and Area Chairs (ACs). We propose specific roles for AI in enhancing factual verification, guiding reviewer performance, assisting authors in quality improvement, and supporting ACs in decision-making. Crucially, we contend that the development of such systems hinges on access to more granular, structured, and ethically-sourced peer review process data. We outline a research agenda, including illustrative experiments, to develop and validate these AI assistants, and discuss significant technical and ethical challenges. We call upon the ML community to proactively build this AI-assisted future, ensuring the continued integrity and scalability of scientific validation, while maintaining high standards of peer review.

## 1 Introduction

**The Peer-Review Scalability Crisis.** The machine-learning community’s publication output continues its dramatic acceleration. NeurIPS submissions grew from 1,678 in 2014 to 17,491 in 2024 (main-track + datasets/benchmarks)—a  $10.4\times$  increase, approximately 26.4% compound annual growth [Lawrence, 2022, NeurIPS, 2024]. ICML submissions surged 48% year-on-year, from 6,538 in 2023 to 9,653 in 2024 [Paper Copilot, 2023, 2024]. This deluge outstrips qualified reviewer pool growth [Walker and Rocha da Silva, 2015], threatening review depth, consistency, and timeliness. Proliferating LLM-based writing tools further inflate submission numbers [Lu et al., 2024, Liu and Shah, 2023], potentially straining quality controls. Consequences include reviewer fatigue, compressed turnarounds, and variable review quality [Cortes and Lawrence, 2021, Benaich and Hogarth, 2023]. High randomness is evident, with up to 23% of acceptance decisions potentially flipping based on reviewer assignment [Cortes and Lawrence, 2021, Beygelzimer et al., 2023, Goldberg et al., 2025]. This “tragedy of the commons” imperils scientific validation in ML.

**AI Assistance: An Urgent Research Priority.** The scale of modern ML research demands a re-evaluation of peer-review workflows. Sustaining high-quality peer review amid continued community growth will be untenable without carefully integrated AI assistance. The same LLMs that accelerate

---

\*Equal contribution. † Equal advising. Correspondence to: [qw281@cam.ac.uk](mailto:qw281@cam.ac.uk).

manuscript production can safeguard review quality if deployed transparently and responsibly. Well-used LLMs can reduce cognitive load, surface inconsistencies, flag AI-generated artifacts, and free reviewers for higher-level reasoning. Encouragingly, in an ICLR 2025 study, 26.6% of reviewers revised their reports after receiving targeted LLM suggestions, often producing more substantive feedback [Thakkar et al., 2025]. While standalone tools are helpful, a cohesive, end-to-end AI-supported ecosystem is vital.

**Our Position and Contributions.** This paper argues that the machine learning community must proactively develop and integrate a comprehensive AI-augmented ecosystem for peer review to address the escalating scalability crisis and maintain the integrity of scientific validation. We further advocate that:

- ① Many foundational AI tools need to be developed (Section 4.1) such as grounding factuality, providing structured feedback, and detecting authenticity.
- ② LLMs can serve as powerful assistants to **reviewers** (Section 4.2) by modeling ideal review characteristics, ensuring factual rigor, and guiding performance.
- ③ LLMs can support **authors** (Section 4.3) by providing pre-submission feedback and aiding rebuttal construction.
- ④ LLMs can empower **Area Chairs (ACs)** (Section 4.4) by assisting in review quality evaluation and decision support.
- ⑤ Developing such an ecosystem critically depends on **richer, structured, ethically-sourced peer review data** that captures the nuances of scientific deliberation (Section 5).
- ⑥ We propose **illustrative experiments** (Section 6) demonstrating LLM potential and highlighting current data and model limitations.

We also discuss **alternative perspectives and challenges** (Section 7), emphasizing peer review scaling as a sociotechnical problem. We propose treating peer review as an explicit **research challenge** (Sections 6, 3), fostering principled co-design of AI tools and benchmark infrastructure.

## 2 The Cracks in the Current System

**Why Peer Review Is a Uniquely Compelling AI Test-bed.** Compared with other language-understanding tasks—summarisation, question answering, or code generation—peer review poses a richer mixture of cognitive and social demands. It requires (i) *domain-specific expertise* to appraise technical claims, (ii) *grounded factual verification* to spot errors or missing citations, (iii) *multi-turn argumentation* among reviewers, authors, and area chairs, and (iv) *value-laden judgment* about novelty, significance, and ethics. The stakes are high: each decision directly shapes the public scientific record. Consequently, building AI systems that can participate constructively in peer review forces us to study *collective reasoning under uncertainty*, robustness to adversarial or AI-generated content, and alignment with human norms of fairness and rigor—all core problems for advancing artificial intelligence itself. Establishing peer-review benchmarks therefore delivers a dual benefit: it helps repair a strained scholarly process while providing a demanding, real-world laboratory for research on language-based intelligence.

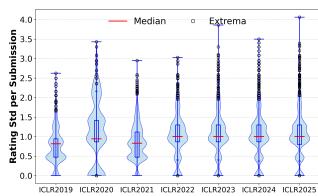


Figure 1: **Inconsistent review ratings.** ICLR data (2019–2024) show high inter-reviewer variance ( $\sigma \approx 1\text{--}1.5$ ) that rises with submissions.

**Symptoms of Strain.** The current peer-review model, still reliant on manual human effort, is showing significant signs of stress under the deluge of submissions. This strain manifests in several critical ways, undermining the system’s effectiveness:

**Superficiality and Reviewer Fatigue:** With reviewers often handling numerous papers under tight deadlines, reviews can become superficial. Critiques may lack depth, overlook crucial methodological details, or offer generic feedback. This trend might be further complicated by the increasing potential for LLM assistance in drafting reviews and rebuttals, an area where LLMs themselves show proficiency in parsing such texts, as suggested by our analysis of ICLR corpora (Table 1) and discussed in Section 6.

**Inconsistent Evaluations:** Significant variance exists between reviewers for the same paper (Figure 1), even on seemingly objective criteria [Cortes and Lawrence, 2021]. This inconsistency can lead to arbitrary outcomes and frustrate authors. While some level of disagreement is inherent and healthy in scientific debate, uncalibrated and widely divergent scores for similar aspects pose a problem.

**Rebuttal Ineffectiveness:** The author rebuttal phase is intended to foster dialogue and clarify misunderstandings. However, time constraints and reviewer disengagement often mean that rebuttals have limited impact on final decisions, even when they substantially address reviewer concerns

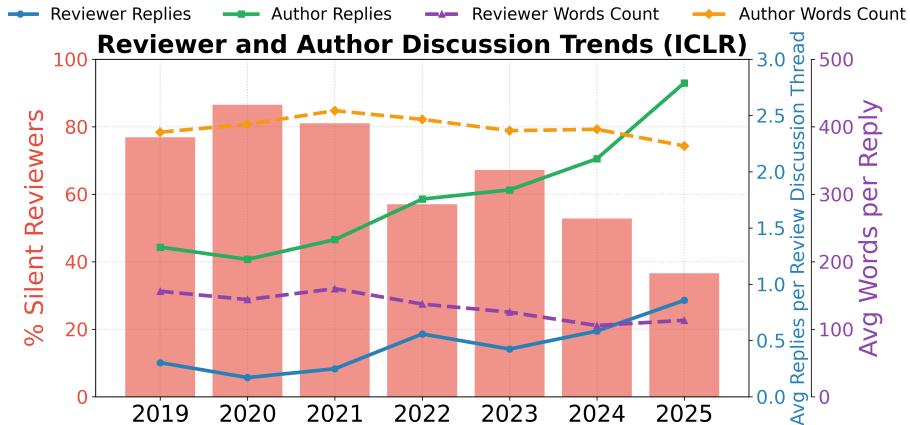


Figure 2: **Reviewer–author dialogue remains shallow.** Across ICLR 2019–25, a majority of reviewers stay silent (red bars); those who respond average <1 reply and <150 words (blue & purple). Authors are increasingly active (green & orange), yet reviewer engagement lags—evidence that the rebuttal stage exerts limited influence.

(Figure 2). **Delayed Feedback and Process Inefficiencies:** The sheer volume of papers leads to lengthy review cycles, delaying the dissemination of impactful research and frustrating authors awaiting timely feedback.

These are not isolated incidents but systemic issues stemming from the fundamental challenge of scaling human expertise linearly with an exponentially growing workload. Without systemic intervention, these problems will likely worsen, eroding trust in the peer review process.

**“Narrow” AI Tools Already Embedded.** Applying AI in isolated scenarios is routine: plagiarism scanners [Foltynek et al., 2019], format/ethics checkers, paper-reviewer matching systems [Mimno and McCallum, 2007], and diff-tools. These successes show community adoption for drudgery reduction. Our proposal extends this to cognitive assistance—a sociotechnical endeavor for building an AI-assisted peer review ecosystem needing accuracy, transparency, and trust.

### 3 Related Work: The Evolving Role of AI in Peer Review

AI’s role in peer review has shifted from administrative aids to LLM-driven cognitive assistants. (Extended version: Appendix A).

**Early AI for Automation and Integrity.** Initial AI automated tasks like reviewer assignment [Mimno and McCallum, 2007] and plagiarism detection [Foltynek et al., 2019], streamlining logistics but not evaluating scientific merit.

**LLMs for Scientific Text Understanding and Review Generation.** LLMs pre-trained on scientific corpora (e.g., SciBERT [Beltagy et al., 2019], SPECTER [Cohan et al., 2020]) improved semantic understanding for tasks like paper summarization [Van Dinter et al., 2021]. Recent work explores generating review components or drafts [Lu et al., 2024, Liang et al., 2024, Saad et al., 2024, Liang et al., 2024]. While fluent [Zhao et al., 2025], these often lack critical depth and may “hallucinate” [Liu and Shah, 2023].

**AI as Reviewer’s Assistant and Quality Enhancer.** Recognizing automation limits, research explores AI as an assistant. LLMs can suggest missed related work [Liu and Shah, 2023, Agarwal et al., 2024, 2025] or structure critiques [Song et al., 2019]. AI also aims to improve human review quality by identifying issues like tone or superficiality or evaluating review quality itself [Goldberg et al., 2025]. The ICLR 2025 LLM feedback experiment supports AI as a collaborative tool, enhancing human judgment [ICLR Blog, 2024, 2025, Thakkar et al., 2025].

**Our Contribution in Context.** We advocate for a holistic, data-driven ecosystem. The next frontier is not just refining LLMs for isolated tasks but establishing a symbiotic AI-data relationship, using

nuanced data from the entire peer review lifecycle for AI systems that genuinely assist in complex reasoning and constructive deliberation.

## 4 The Vision: An AI-Augmented Peer Review Ecosystem



Figure 3: The envisioned AI-augmented peer review ecosystem. This diagram illustrates the cyclical nature of peer review, with authors, reviewers, and Area Chairs (ACs) as key human stakeholders. At the core, a Large Language Model (LLM) acts as a collaborative assistant, providing feedback and analytical support at multiple stages (e.g., to authors during paper preparation, to reviewers assessing submissions, and to ACs in their decision-making process), always with humans guiding the process.

We envision a future where artificial intelligence (AI), particularly LLMs, acts as an intelligent partner within the peer review ecosystem, collaborating with all human stakeholders to help mitigate the strains detailed in Section 2. This vision is not about replacing human judgment, which remains paramount, but about augmenting human capabilities. By automating or assisting with structured, repetitive, or cognitively demanding tasks, AI can free experts to concentrate on nuanced scientific assessment, ultimately enhancing the quality, efficiency, and fairness of peer review. This section first introduces key AI-driven tools and capabilities that form the foundation of this ecosystem. It then examines AI's potential role from three key perspectives: assisting reviewers, authors, and area chairs (ACs), illustrating how these foundational tools are applied to empower each group.

### 4.1 Foundational AI Tools and Capabilities for an Augmented Ecosystem

The envisioned AI-augmented peer review ecosystem relies on a suite of sophisticated tools and capabilities designed to integrate seamlessly into the workflow, addressing issues like review inconsistency and superficiality (Section 2). Foundational to this are systems for **Retrieval Augmented Verification (RAV) and Grounding**, which enhance LLMs by connecting them to authoritative knowledge bases like scientific literature repositories (e.g., Semantic Scholar, arXiv) [Lewis et al., 2020]. This grounding enables AI to cross-reference claims, suggest relevant citations, and ensure factual soundness. Complementing this, **AI for Code Analysis and Reproducibility Assessment** offers tools to parse methodologies and analyze source code (if provided), aiding in the preliminary assessment of reproducibility by identifying common errors, missing dependencies, or inconsistencies between code and paper descriptions [Starace et al., 2025].

Another key element is **AI-Powered Review Quality Feedback**, often manifested as "Review Report Cards." These LLMs analyze human-written reviews against predefined or learned criteria such as coverage of essential paper aspects (novelty, significance, soundness), specificity of critiques, evidence backing claims, and constructiveness of tone, to generate structured feedback for the reviewer or AC. In an era of advanced generative models, **AI for Content Provenance and Authenticity** aims to identify AI-generated text through methods like statistical analysis of textual features (e.g., perplexity, burstiness) [Mindner et al., 2023] or potential watermarking technologies like SynthID [Dathathri et al., 2024], though these methods are still evolving and face significant challenges in robustness and fairness [Zhou et al., 2024, Emi and Spero, 2024]. Furthermore, **AI-Assisted Authoring Support**

provides authors with formative feedback on manuscript drafts (e.g., clarity, structure, adherence to guidelines) and aids in structuring effective rebuttals, aiming to improve rebuttal impact (Section 2, Figure 2). Finally, **AI for Decision Support for Area Chairs** can synthesize information from multiple reviews and rebuttals, offering summaries of key arguments, highlighting points of consensus or disagreement, and flagging unanswered concerns to assist ACs in their deliberation and meta-review preparation. These tools, thoughtfully integrated with human oversight, can create a more robust, efficient, and supportive peer review landscape.

## 4.2 AI-Empowered Reviewers

Reviewers are the linchpin of the peer review system. Their expertise and diligence are crucial, yet they face significant burdens such as fatigue and time pressure, potentially impacting review depth (Section 2). AI can serve as a “co-pilot”, leveraging the foundational tools described above to support them in producing higher-quality, more consistent, and factually sound evaluations.

**Striving for the "Ideal" Review: An AI Benchmark.** To guide the development of AI assistance, it’s useful to conceptualize an “ideal” reviewer—one possessing a (1) Comprehensive Knowledge Base encompassing all relevant prior art, the (2) Meticulous Verifiability to check all claims and experimental setups with unerring accuracy, and the (3) Insightful Constructivism to provide perfectly targeted, actionable feedback that maximally improves the paper. While no current AI can fully embody this ideal, LLMs, augmented by the tools in Section 4.1, can emulate aspects of these capabilities, thereby providing valuable assistance.

**Enhancing Factual Rigor through Grounded AI.** A core challenge in peer review is verifying the factual correctness of claims, where AI tools can help address superficiality (Section 2). By applying *Retrieval Augmented Verification (RAV)*, LLMs integrated with scientific literature databases [Lewis et al., 2020] can help reviewers cross-reference claims, identify potentially missed citations, or flag inconsistencies with established knowledge. For instance, an LLM could flag if a review criticizes a paper for not citing Method X, when Method X is very recent, from a different subfield, or already implicitly addressed, as demonstrated by early systems like LitLLM [Agarwal et al., 2024, 2025]. Future AI could extend this by explaining *why* a retrieved document supports or refutes a claim, or by summarizing relevant differences. This is complemented by *AI-assisted experimental validation and reproducibility checks*. These tools can parse methodology sections and, if code is provided (as encouraged by many conference [NeurIPS, 2025]), analyze it for common pitfalls (e.g., data leakage, incorrect metric implementation) or inconsistencies with the paper’s description, flagging areas for closer human scrutiny. Such tools might evolve into “semantic git” systems for easier navigation between paper, code, and results, or perform automated “sanity checks” on reported findings based on dataset characteristics or known theoretical bounds.

**Guiding Reviewer Performance and Fostering Quality.** Many reviewers, especially those early in their careers, can benefit from structured guidance, and AI-powered review quality feedback tools offer a scalable solution to problems like inconsistent evaluations (Figure 1, Section 2). One such application is the *Automated Review “Report Card”*, where an LLM system generates multi-dimensional feedback on a human-written review. This feedback assesses crucial aspects such as: (i) Coverage of key evaluation criteria (e.g., novelty, significance, technical soundness, empirical validation, clarity, ethics); (ii) Specificity of criticisms, determining if they are concrete and actionable rather than vague; (iii) The evidence base for claims within the review, potentially cross-checked with RAV to see if assertions are well-supported or if counter-evidence exists; and (iv) The constructiveness and tone of the feedback, ensuring it is professional and aimed at improvement. The ICLR 2025 LLM feedback experiment, which reported more detailed and substantively revised human reviews after AI suggestions, highlights this potential [Kim et al., 2025, ICLR Blog, 2024, 2025, Thakkar et al., 2025]. Furthermore, by comparing a human review against an LLM-generated “reference review” (itself based on aggregated notions of high-quality reviews), the system can help reviewers identify discrepancies or gaps in their own assessment, prompting self-reflection and potentially improving their evaluative skills. This approach could form the basis of an optimal “review curriculum” for training junior reviewers.

**Detecting AI-Generated Content with AI Tools** The proliferation of advanced generative models raises concerns about AI-generated or heavily AI-assisted submissions. Reviewers, or the peer review system itself, may increasingly rely on AI for content provenance. As introduced in Section 4.1, these tools include techniques like syntactic or lexical watermarking (e.g., “perturbed sampling” as

explored in prototypes like SynthID [Dathathri et al., 2024]) or methods that analyze textual features such as token-level entropy and perplexity under various LLMs [Mindner et al., 2023]. However, these detection methods are in their nascent stages and face significant challenges, including high false-positive rates (especially for non-native English speakers or highly formulaic text), lack of robustness against adaptive adversaries or paraphrasing, and ethical concerns about fairness and potential misuse [Zhou et al., 2024, Emi and Spero, 2024]. It is crucial to acknowledge that current detection is not foolproof. Consequently, a pragmatic approach involves using such tools as one signal among many, necessitating clear policies on AI use in submissions and author declarations. Ultimately, human judgment must remain the final arbiter, not just of the content's quality, but potentially of its provenance.

This is particularly salient because the mere presence of AI-generated text may not, in itself, be the primary concern. It is widely understood that authors may legitimately use LLMs for drafting assistance, proofreading, refining language, or even generating code snippets. The critical distinction, therefore, may not be the sheer volume of text flagged by a detector, but rather the *latent intent* and the nature of the AI's contribution. For instance, a manuscript where the core scientific ideas, methodologies, and results are conceived and articulated by human authors, and then polished by an LLM for clarity, represents a different scenario than one where foundational scientific content (e.g., literature review, method derivation, result interpretation) is predominantly generated by AI with minimal human intellectual input or critical oversight. A simple "LLM generated content score" could be misleading if not contextualized. Reviewers, and the systems supporting them, will need to consider how to interpret such scores, moving beyond quantitative detection to a qualitative assessment of authorship and originality in an AI-assisted era.

### 4.3 AI-Empowered Authors

Authors can also leverage AI assistance throughout the manuscript preparation and revision lifecycle. The goal is to help authors present their work more effectively using AI-assisted authoring support tools, leading to clearer communication and potentially stronger submissions that are better aligned with community expectations.

**AI can provide formative pre-submission feedback.** LLMs trained on characteristics of high-impact papers, common rejection reasons, and specific venue guidelines could offer suggestions on clarity, structure, argument coherence, completeness of literature review (potentially augmented by RAV-like features to suggest missing seminal or recent works), and adherence to ethical guidelines or reporting standards. This pre-submission support leverages AI in a manner analogous to how it might assist reviewers: by systematically analyzing the manuscript against criteria indicative of high-quality research, it can generate a "simulated review". This provides authors with a prioritized understanding of their work's potential weaknesses (e.g., "The link between Section 2 and your proposed method in Section 3 is unclear", or "Consider adding an ablation study for parameter X") and offers concrete, actionable advice on how to address these points, effectively allowing them to refine their paper based on anticipated reviewer feedback. Such systems might even simulate diverse reviewer personas (e.g., a theory-focused reviewer, an application-focused reviewer) to help authors anticipate and preemptively address potential concerns from different perspectives.

**AI tools can offer strategic rebuttal assistance.** This involves more than just grammar checking responses. It could include systematically cataloging all significant reviewer points from multiple reviews to ensure each is addressed; suggesting effective ways to present new evidence or clarifications, perhaps tailored to the specific tone or concerns of a review; helping authors identify points of misunderstanding versus actual disagreement; and assisting authors in gauging if their response adequately answers a specific concern or if reviewers implicitly request further elaboration or experimental validation, thereby aiming to improve the often limited impact of rebuttals (Figure 2, Section 2). In addition, tools such as RAV and AI coding agents could validate the reviewer's claims.

**AI can function as a personalized educational tool for authorship,** especially for junior researchers or those less familiar with the norms of premier ML venues. By analyzing exemplary papers within specific subfields, common pitfalls in methodology or presentation, or even an author's prior (anonymized, with consent) work and reviews, AI can highlight best practices in scientific writing, experimental design, and argumentation. This could perhaps form a "paper curriculum" or a set of guided exercises to help authors improve their scientific communication skills.

#### 4.4 Empowering Area Chairs (ACs) with AI Decision Support

Area Chairs (ACs) face the monumental task of synthesizing multiple, often conflicting, reviews, moderating discussions among reviewers and with authors, and making well-justified recommendations to program chairs. AI for decision support, as introduced in Section 4.1, can provide crucial assistance in managing this complex information flow and the cognitive load exacerbated by increasing submission volumes (Section 2).

**AI can provide enhanced review quality assessment support** by offering initial, structured assessments of individual reviews, leveraging concepts like the “Review Report Cards”. This allows ACs to quickly identify potentially insightful versus problematic reviews, helping them prioritize their attention, calibrate reviewer assessments more effectively, and guide the discussion phase, tackling issues like review inconsistency (Figure 1). For example, an AI might flag a review that is overly brief, lacks specific evidence for its claims, or uses an unprofessional tone, prompting the AC to investigate further or gently guide the reviewer.

**AI can offer comprehensive decision support and meta-review assistance.** This capability includes intelligent summarization to condense key arguments, points of consensus, and critical disagreements from multiple reviews and the author’s rebuttal into a concise overview for the AC. It also involves conflict and gap highlighting to automatically flag direct contradictions between reviewers (e.g., Reviewer 1 praises novelty, Reviewer 2 claims it’s incremental), instances where a major reviewer concern appears unaddressed by the rebuttal, or aspects of the paper (e.g., ethical implications, limitations) not adequately covered by any reviewer. AI can also assist ACs by generating initial drafts of meta-review sections, such as summaries of perceived strengths and weaknesses based on identified themes and overall reviewer sentiment, or by highlighting papers with unusually high variance in scores. While the AI can provide these valuable inputs, the final nuanced judgment, synthesis, weighting of arguments, and authoritative recommendation remain firmly with the human AC, who brings broader context and domain expertise.

### 5 The Critical Need for Richer Data: Fueling the AI-Augmented Ecosystem

Developing sophisticated AI assistants for peer review critically hinges on rich, nuanced, structured data, beyond current public datasets.

#### 5.1 Limitations of Current Data for Advanced AI Assistant Development

Current datasets often lack: **explicitly grounded reasoning** for judgments (score changes, AC weighting of reviews); structured data on **deliberation dynamics** (negotiation, clarification, concession); **fine-grained traceability** between claims and specific manuscript content; and encoding of **implicit domain knowledge** or community norms. This confines AI to surface-level pattern matching.

**Key Dimensions of Richer Data Required** To build next-generation AI tools, we require data capturing the peer review process in greater detail: **1) Structured Reviewer Reasoning for Score Changes and Key Assertions:** Why scores changed post-rebuttal; explicit reasoning linked to specific rebuttal points or discussion elements. **2) Detailed Author-Reviewer-AC Interaction Traces with Semantic Annotations:** Dialogue acts (clarification, concession), argument strength, and explicit links showing response chains. **3) AC Deliberation Traces (Anonymized and Aggregated):** How ACs weigh conflicting reviews, assess rebuttals, identify key decision-drivers, and form meta-review judgments. **4) Fine-grained Annotations Linking Text to Manuscript and External Knowledge:** Links from review statements to specific manuscript parts (sentences, figures) and external knowledge (prior papers, theories).

**A Call for a Community-Driven Data Ecosystem** Acquiring richer data faces challenges: workload, privacy/anonymity, and potential gaming. We call for a community effort (organizers, publishers, funders, OpenReview) to: (1) **Develop ethical frameworks and privacy-preserving protocols** for collecting/sharing richer, anonymized data. (2) **Pilot new data collection interfaces** minimizing burden while maximizing information gain. (3) **Invest in shared, curated benchmark datasets** for AI research in peer review deliberation and reasoning. (4) **Foster interdisciplinary collaboration** (AI, domain science, ethics, platform development). This enriched data ecosystem is an investment in the quality, efficiency, and fairness of scholarly discourse.

Table 1: LLM recall in extracting key points from ICLR peer-review corpora (2024 vs. 2025). Values are mean  $\pm$  95% Confidence Intervals (CI). Increased recall for ICLR 2025 may reflect evaluating LLM improvements and/or better parsing of increasingly common AI-assisted review components (e.g., content stylistically similar to that from advanced LLMs).

	ICLR 2024			ICLR 2025		
	Real pts.	Hits	Recall $\uparrow$	Real pts.	Hits	Recall $\uparrow$
Strengths	$3.45 \pm 0.22$	$2.42 \pm 0.19$	$0.724 \pm 0.000$	$4.02 \pm 0.60$	$3.71 \pm 0.59$	$0.927 \pm 0.060$
Weaknesses	$3.96 \pm 0.33$	$1.33 \pm 0.14$	$0.387 \pm 0.000$	$4.58 \pm 0.65$	$2.73 \pm 0.45$	$0.632 \pm 0.000$
Rebuttals	$6.81 \pm 0.46$	$4.96 \pm 0.35$	$0.776 \pm 0.040$	$7.22 \pm 1.04$	$6.29 \pm 0.78$	$0.911 \pm 0.040$

Table 2: Effect of in-context examples ( $n$ ) on LLM rating prediction accuracy (ICLR 2025 data). Mean Absolute Error (MAE) / Root Mean Squared Error (RMSE) (mean  $\pm$  95% CI); lower is better.

Task	Metric	Number of in-context examples $n$			
		$n = 0$	$n = 1$	$n = 2$	$n = 3$
Initial rating	MAE $\downarrow$	$2.7722 \pm 0.0087$	$2.3636 \pm 0.0090$	$2.2857 \pm 0.0095$	$2.3067 \pm 0.0105$
	RMSE $\downarrow$	$3.0336 \pm 0.0089$	$2.6872 \pm 0.0088$	$2.6531 \pm 0.0097$	$2.7105 \pm 0.0108$
Final rating	MAE $\downarrow$	$0.7595 \pm 0.0058$	$0.6709 \pm 0.0052$	$0.7215 \pm 0.0055$	$0.7468 \pm 0.0054$
	RMSE $\downarrow$	$1.1363 \pm 0.0066$	$1.0000 \pm 0.0054$	$1.0614 \pm 0.0055$	$1.0733 \pm 0.0058$
Score change	MAE $\downarrow$	$0.7342 \pm 0.0057$	$0.7595 \pm 0.0055$	$0.9367 \pm 0.0060$	$0.8228 \pm 0.0054$
	RMSE $\downarrow$	$1.0908 \pm 0.0060$	$1.0908 \pm 0.0062$	$1.2629 \pm 0.0059$	$1.1418 \pm 0.0056$

## 6 Illustrative Experiments & Research Agenda

**Position Supported:** LLMs show promise in assisting with initial review tasks and predicting assessments, but In-Context Learning (ICL) with current data has limitations, highlighting needs for fine-tuning and richer, structured peer review data. Our experiments on ICLR corpora illustrate this—we provide experimental implementation details in Appendix B.

**Part 1: Review Component Generation. Task & Method:** Using few-shot prompting, an LLM was tasked to generate strengths and weaknesses from a paper’s content, and separately, to identify key rebuttal points from initial reviews. We evaluated the semantic overlap of the LLM’s output against the actual human-written content from ICLR 2024/25 OpenReview data, measuring recall via an LLM-as-judge protocol (Avg Hits / Avg Real Points). **Results & Interpretation (Table 1):** LLMs showed higher recall for strengths (e.g.,  $0.927 \pm 0.060$  for ICLR 2025) than weaknesses ( $0.632 \pm 0.000$ ), suggesting identifying flaws is harder and needs advanced alignment. Rebuttal point recall was high ( $0.911 \pm 0.040$ ). Increased recall in 2025 (see caption Table 1) may reflect LLM improvements and/or better parsing of increasingly AI-assisted review components (stylistically similar to output from advanced LLMs). This highlights LLM utility and the evolving, AI-infused data landscape, stressing need for human oversight for nuanced evaluation.

**Part 2: Rating Prediction. Task & Method:** Using few-shot In-Context Learning (ICL) with a varying number of examples ( $n = 0, 1, 2, 3$ ), an LLM predicted initial ratings (from paper content), final ratings (from reviews and author rebuttals), and the resulting score change. We prompted the model with text from the ICLR 2025 dataset and evaluated its predictions against the ground-truth scores using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). **Results & Interpretation (Table 2):** LLMs offer a reasonable baseline (e.g., initial rating MAE  $2.2857 \pm 0.0095$  with  $n = 2$ ; final rating MAE  $0.6709 \pm 0.0052$  with  $n = 1$ ). Scaling  $n$  shows diminishing returns, implying ICL limits for complex regression without explicit training on underlying factors. Significant gains likely require fine-tuning on larger, specialized datasets with structured rationale. Inherent subjectivity in peer review may also cap accuracy. Score change prediction was particularly hard. LLMs offer estimations, but precise prediction needs advancement and richer data; human judgment remains indispensable. These experiments underscore LLM assistive potential but also highlight limitations, reinforcing calls for human oversight, sophisticated AI development (including fine-tuning on richer data), and deeper understanding of review process data.

## 7 Alternative Perspectives and Challenges

Our proposed AI-augmented ecosystem is not without opposition or significant challenges. A key alternative position holds that **AI’s role in peer review should be strictly limited to non-evaluative tasks, preserving all intellectual assessment for humans**, due to fears of bias, de-skilling, or compromising scientific integrity. Another contends that **even assistive AI poses unacceptable risks of misuse (e.g., flooding submissions with AI-generated content) that outweigh benefits**, arguing for purely human-centric solutions to scaling review.

These perspectives rightly highlight the challenge of AI-generated content. Defining and detecting AI contributions is non-trivial, varying with the granularity and nature of AI involvement [Lu et al., 2024]. The utility of AI detection itself is ethically complex: it might penalize legitimate assistive uses (e.g., by non-native speakers) while failing against sophisticated misuse [Zhou et al., 2024, Emi and Spero, 2024]. The prospect of AI generating core scientific ideas, if undetected, raises profound novelty concerns. As LLMs advance [Novikov et al., 2025], the line between human and AI contribution may blur, challenging traditional authorship.

While these are valid concerns, we argue that a proactive, human-centric AI-augmented approach remains necessary. The sheer scale of the review crisis (Section 1) may outpace purely manual solutions. Our envisioned ecosystem emphasizes AI as an *assistant*, with humans retaining final judgment. The concerns about AI-generated content underscore the need for (1) robust, yet fair, detection tools as part of the ecosystem (Section 4.1), (2) clear institutional policies on AI use and author declarations, and (3) ongoing research into verifiable AI and responsible LLM development. Rather than a wholesale rejection of AI’s potential, these challenges call for careful, ethical integration where AI tools enhance, rather than replace, human expertise, treating detection and misuse mitigation as ongoing sociotechnical issues requiring community-wide effort.

## 8 Discussion and Conclusion

The ML community faces a critical juncture: submission volumes threaten traditional peer review. This paper posits that AI assistance is an imperative. We envision LLMs as **collaborators augmenting authors, reviewers, and ACs, helping manage burdens, enhance rigor, promote consistency, guide development, and support decisions**. Our illustrative experiments show nascent LLM potential. Transforming this potential into a robust ecosystem critically depends on **access to richer, more granular, ethically-sourced data on the peer review process itself**—especially on deliberation dynamics and the reasoning behind evaluative judgments. Current datasets lack this depth. This requires concerted community action: (1) Conference organizers and platforms must prioritize the structured collection of detailed deliberation data. Platforms such as OpenReview, for example, provide an invaluable foundation through their public-facing nature and data APIs, but could be enhanced to capture more granular interactions (e.g., explicitly linking score changes to specific rebuttal arguments), all while maintaining robust privacy safeguards and minimizing user burden. (2) Researchers must develop AI tools for augmentation, keeping humans in control. (3) The community must proactively address ethical challenges (bias, misuse, privacy) with clear guidelines. The path forward involves pilot programs, transparent development, community engagement for trust, and continuous evaluation. Empowering human experts, not diminishing them, is the goal. Investing in AI assistance and supporting data infrastructure can build a more scalable, robust, and fair peer review system, safeguarding ML research integrity and progress.

## 9 Limitations and Future Work

This position paper acknowledges limitations. Our illustrative experiments are conceptual and need rigorous, large-scale empirical validation. **Key limitations:** (1) *Empirical Grounding and Scalability*: Proposed AI assistants require substantial research, data engineering, and validation beyond our illustrations. (2) *Scientific Reasoning Complexity*: Current LLMs struggle with deep scientific reasoning, novelty assessment, and understanding complex, implicit argumentation. (3) *Data Availability, Quality, and Heterogeneity*: Richer, structured data are not yet widely available; integration and quality assurance are major hurdles. While platforms like OpenReview have democratized access to review data, the information often remains in semi-structured free-text, limiting the development of AI that can deeply reason about the deliberative process. (4) *Unintended Consequences*: Introducing AI

into peer review can have unforeseen effects (e.g., over-reliance, de-skilling, new gaming strategies). Future work must develop unified, privacy-preserving data schemas for fine-tuning specialized, verifiable AI models capable of claim checking and decision support with clear rationales. For instance, this could involve collaboration with platforms like OpenReview to pilot new review interfaces that prompt for structured rationale for score changes or allow for fine-grained annotation of claims within discussion threads. This necessitates intuitive human-AI collaboration interfaces and continuous ethical auditing (including bias detection and fairness metrics) to ensure trust. Longitudinal trials across venues are vital to measure AI’s impact on review quality, efficiency, and diversity, while advancing robust content provenance and misuse mitigation (e.g., watermarking, resilient detectors) is key to safeguarding legitimate AI assistance.

**Acknowledgements** We thank Andrew McCallum for insightful discussions, valuable feedback, and high-level guidance that significantly enhanced the quality and impact of this work. We are also grateful to Hao Sun for helpful brainstorming discussions and advice provided throughout the project. This research was supported by Azure sponsorship credits provided by Microsoft’s AI for Good Research Lab and the Accelerate Foundation Models Academic Research Initiative. QW is supported by funding from GSK, and SH is supported by AstraZeneca. Additionally, we acknowledge the data collected from Paper Copilot, an independent initiative funded by Jing Yang.

## References

- S. Agarwal, G. Sahu, A. Puri, I. H. Laradji, K. D. Dvijotham, J. Stanley, L. Charlin, and C. Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.
- S. Agarwal, G. Sahu, A. Puri, I. H. Laradji, K. D. Dvijotham, J. Stanley, L. Charlin, and C. Pal. Litllms, llms for literature review: Are we there yet? *Transactions on Machine Learning Research*, 2025.
- I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620. Association for Computational Linguistics, 2019.
- N. Benaich and I. Hogarth. State of ai report 2023. 2023. Accessed: 2024-11-15.
- A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*, 2023.
- W. Chen, J. Zhang, H. Zhu, B. Xu, Z. Hao, K. Zhang, J. Ye, and R. Cai. Causal-aware large language models: Enhancing decision-making through learning, adapting and acting. *arXiv preprint arXiv:2505.24710*, 2025.
- A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- C. Cortes and N. D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.
- W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, and G. Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE international conference on advanced learning technologies (ICALT)*, pages 323–325. IEEE, 2023.
- S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.
- M. Du, Y. Chen, Z. Wang, L. Nie, and D. Zhang. Llm4ed: Large language models for automatic equation discovery. *CorR*, abs/2405.07761, 2024. URL <https://doi.org/10.48550/arXiv.2405.07761>.

- D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.
- B. Emi and M. Spero. Technical report on the checkfor. ai ai-generated text classifier. *arXiv preprint*, 2024.
- T. Foltynek, N. Meuschke, and B. Gipp. Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR)*, 52(6):1–42, 2019.
- Gokulnath B. Can AI Really Improve Manuscript Screening & Editorial Decisions? [https://www.hurix.com/blogs/can-ai-really-improve-manuscript-screening-editorial-decisions/?utm\\_source=chatgpt.com/](https://www.hurix.com/blogs/can-ai-really-improve-manuscript-screening-editorial-decisions/?utm_source=chatgpt.com/), 2025. Accessed: 2025-05-21.
- A. Goldberg, I. Stelmakh, K. Cho, A. Oh, A. Agarwal, D. Belgrave, and N. B. Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *PloS one*, 20(4):e0320444, 2025.
- S. Holt, T. Liu, and M. van der Schaar. Automatically learning hybrid digital twins of dynamical systems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL <https://openreview.net/forum?id=S0si0bSdu2>.
- S. Holt, Z. Qian, T. Liu, J. Weatherall, and M. van der Schaar. Data-driven discovery of dynamical systems in pharmacology using large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL <https://openreview.net/forum?id=KIrZmlTA92>.
- S. Holt, M. R. Luyten, A. Berthon, and M. van der Schaar. G-sim: Generative simulations with large language models and gradient-free calibration. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=Pvk06rIixC>.
- ICLR Blog. ICLR 2025 Reviewer Assistant Experiment. <https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/>, 2024. Accessed: 2024-11-15.
- ICLR Blog. Leveraging LLM feedback to enhance review quality. <https://blog.iclr.cc/2025/04/15/leveraging-l1m-feedback-to-enhance-review-quality/>, 2025. Accessed: 2025-04-15.
- K. Kacprzyk, S. Holt, J. Berrevoets, Z. Qian, and M. van der Schaar. ODE discovery for longitudinal heterogeneous treatment effects inference. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=pxI5IPeWgW>.
- J. Kim, Y. Lee, and S. Lee. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. *arXiv preprint arXiv:2505.04966*, 2025.
- N. D. Lawrence. The neurips experiment. <https://inverseprobability.com/talks/notes/the-neurips-experiment-snsf.html>, 2022. Accessed 23 May 2025.
- P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022.
- W. Liang, Y. Zhang, H. Cao, B. Wang, D. Y. Ding, X. Yang, K. Vodrahalli, S. He, D. S. Smith, Y. Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):A10a2400196, 2024.
- R. Liu and N. B. Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.
- C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

- D. Mimno and A. McCallum. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 500–509, 2007.
- L. Mindner, T. Schlippe, and K. Schaaff. Classification of human-and ai-generated texts: Investigating features for chatgpt. In *International conference on artificial intelligence in education technology*, pages 152–170. Springer, 2023.
- NeurIPS. NeurIPS 2024 Fact Sheet. [https://media.neurips.cc/Conferences/NeurIPS2024/NeurIPS2024-Fact\\_Sheet.pdf](https://media.neurips.cc/Conferences/NeurIPS2024/NeurIPS2024-Fact_Sheet.pdf), 2024. Accessed: June 26, 2025.
- NeurIPS. NeurIPS Paper Checklist Guidelines. <https://neurips.cc/public/guides/PaperChecklist/>, 2025. Accessed: 2025-05-21.
- A. Novikov, N. Vu, M. Eisenberger, E. Dupont, P.-S. Huang, A. Z. Wagner, S. Shirobokov, B. Kozlovskii, F. J. R. Ruiz, A. Mehrabian, M. P. Kumar, A. See, S. Chaudhuri, G. Holland, A. Davies, S. Nowozin, P. Kohli, and M. Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. <https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/AlphaEvolve.pdf>, 2025. Accessed: 2025-05-21.
- Paper Copilot. ICML 2023 Submission Statistics. <https://papercopilot.com/statistics/icml-statistics/icml-2023-statistics/>, 2023. Accessed 23 May 2025.
- Paper Copilot. ICML 2024 Submission Statistics. <https://papercopilot.com/statistics/icml-statistics/>, 2024. Accessed 23 May 2025.
- T. Pouplin, H. Sun, S. Holt, and M. Van der Schaar. Retrieval-augmented thought process as sequential decision making. *arXiv e-prints*, pages arXiv-2402, 2024.
- Radiology: Artificial Intelligence. How Radiology: Artificial Intelligence Handles Your Paper. <https://pubs.rsna.org/page/ai/author-instructions/your-paper#:~:text=Radiology%3A%20Artificial%20Intelligence%20accepts%20manuscripts,not%20be%20submitted%20to%20another/>, 2025. Accessed: 2025-05-21.
- A. Saad, N. Jenko, S. Ariyaratne, N. Birch, K. P. Iyengar, A. M. Davies, R. Vaishya, and R. Botchu. Exploring the potential of chatgpt in the peer review process: an observational study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 18(2):102946, 2024.
- X. Song, H. Sun, X. Wang, and J. Yan. A survey of automatic generation of source code comments: Algorithms and techniques. *IEEE access*, 7:111411–111428, 2019.
- G. Starace, O. Jaffe, D. Sherburn, J. Aung, J. S. Chan, L. Maksin, R. Dias, E. Mays, B. Kinsella, W. Thompson, et al. Paperbench: Evaluating ai’s ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.
- N. Thakkar, M. Yuksekgonul, J. Silberg, A. Garg, N. Peng, F. Sha, R. Yu, C. Vondrick, and J. Zou. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*, 2025.
- R. Van Dinter, B. Tekinerdogan, and C. Catal. Automation of systematic literature reviews: A systematic literature review. *Information and software technology*, 136:106589, 2021.
- R. Walker and P. Rocha da Silva. Emerging trends in peer review—a survey. *Frontiers in neuroscience*, 9:169, 2015.
- G. Wan, Y. Lu, Y. Wu, M. Hu, and S. Li. Large language models for causal discovery: Current landscape and future directions, 2025. URL <https://arxiv.org/abs/2402.11068>.
- S. Yu, M. Luo, A. Madasu, V. Lal, and P. Howard. Is your paper being reviewed by an llm? investigating ai text detectability in peer review. *arXiv preprint arXiv:2410.03019*, 2024.
- Y. Zhang, X. Chen, B. Jin, S. Wang, S. Ji, W. Wang, and J. Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *EMNLP*, pages 8783–8817, 2024. URL <https://aclanthology.org/2024.emnlp-main.498>.

M. Zhao, F. Li, F. Cai, H. Chen, and Z. Li. Can we trust llms to help us? an examination of the potential use of gpt-4 in generating quality literature reviews. *Nankai Business Review International*, 16(1): 128–142, 2025.

Y. Zhou, B. He, and L. Sun. Humanizing machine-generated content: evading ai-text detection through adversarial attack. *arXiv preprint arXiv:2404.01907*, 2024.

# Appendix

## A Extended Related Work

The integration of Artificial Intelligence (AI) into the peer review process has evolved significantly, moving from rudimentary administrative aids to sophisticated Large Language Model (LLM)-driven cognitive assistants. This progression reflects both the advancements in AI capabilities and a growing recognition of the challenges within scholarly peer review. We categorize existing work along several key themes, outlining existing works as well as missing research agendas to make AI enabled peer review a possibility. We provide an extended related work in Appendix A.

**Early AI for Process Automation and Integrity Checks** Initial forays of AI into peer review primarily focused on automating well-defined, often labor-intensive tasks. This included systems for reviewer assignment, aiming to match manuscript topics with reviewer expertise [Mimno and McCallum, 2007], thereby assisting editors and potentially improving match quality. Another area was plagiarism detection, where tools were developed to identify textual overlap with existing publications, safeguarding academic integrity [Foltynek et al., 2019]. Other early applications included manuscript prescreening for formatting or completeness [Gokulnath B, 2025, Radiology: Artificial Intelligence, 2025]. While these tools streamlined logistical aspects and reduced administrative burden, they did not typically engage with the core intellectual labor of evaluating scientific merit.

**LLMs for Scientific Text Understanding and Initial Review Generation** The advent of powerful LLMs, particularly those pre-trained on vast scientific corpora (e.g., SciBERT [Beltagy et al., 2019] and SPECTER [Cohan et al., 2020]), marked a paradigm shift. These models demonstrated improved semantic understanding of scientific text, paving the way for more cognitively demanding applications. Early explorations included automated summarization of research papers [Van Dinter et al., 2021], which could provide reviewers or editors with quick overviews. More recently, research has ventured into the generation of review components or even full initial review drafts [Lu et al., 2024, Liang et al., 2024, Saad et al., 2024, Liang et al., 2024]. User studies evaluating these AI-generated reviews often find them plausible and fluent [Zhao et al., 2025]. However, a common critique is their lack of critical depth, potential for factual inaccuracies (“hallucinations”), and an inability to consistently provide nuanced, insightful critique essential for rigorous scientific assessment [Liu and Shah, 2023].

**AI for Automated Scientific Discovery and Modeling.** A parallel and highly relevant line of research focuses on using AI, and LLMs in particular, to automate or assist in the creation of scientific knowledge itself. This is a crucial development, as the ability to assist in evaluating scientific work is predicated on an understanding of the scientific process. Recent work has demonstrated that LLMs can act as core components in frameworks for automated scientific discovery. For instance, LLMs can propose and refine interpretable models of dynamical systems from data, such as discovering governing Ordinary Differential Equations (ODEs) in fields like pharmacology [Du et al., 2024, Zhang et al., 2024, Kacprzyk et al., 2024, Holt et al., 2024b]. These systems can iteratively generate model specifications, use external tools for parameter optimization, and evolve solutions based on evaluative feedback, effectively automating parts of the modeling pipeline [Holt et al., 2024a]. Other approaches use LLMs to design simulators by proposing causal structures which are then empirically calibrated, blending knowledge-driven design with data-driven validation [Chen et al., 2025, Holt et al., 2025, Wan et al., 2025]. The success of these systems in generating plausible and effective scientific models underscores the potential for similar AI systems to be adapted for the critical evaluation of such models, forming a basis for the AI-assisted peer review tools we envision.

**AI as a Reviewer’s Assistant and Quality Enhancer.** Recognizing the limits of full automation, a significant body of work explores AI as an assistant to human reviewers. This includes tools to identify potential issues within reviews, such as unprofessional tone, lack of constructiveness, or superficiality [Dai et al., 2023]. For fact-checking and grounding, Retrieval-Augmented Generation (RAG) techniques are key [Li et al., 2022]. Systems like LitLLM can suggest relevant related work that authors might have missed [Agarwal et al., 2024, 2025], and research is ongoing to make retrieval more robust and structured, for example by formulating it as a sequential decision process [Pouplin et al., 2024] or by using graphs [Edge et al., 2024]. Some research has also focused on the challenging task of evaluating the quality of reviews themselves [Goldberg et al., 2025, Yu et al., 2024], which could inform editor decisions or provide feedback to reviewers. The ICLR 2025 experiment, where targeted LLM-generated feedback was provided to reviewers, strongly supports the efficacy of AI as

a collaborative tool that enhances, rather than replaces, human judgment [ICLR Blog, 2024, 2025, Thakkar et al., 2025].

**Our Contribution in Context** Our work builds upon these foundations by advocating for a holistic, data-driven ecosystem. We argue that the next frontier in AI for peer review lies not just in refining LLM capabilities for isolated tasks, but in establishing a symbiotic relationship between AI development and the strategic, ethical collection and utilization of nuanced data from the entire peer review lifecycle. This approach, we contend, is essential for building AI systems that can genuinely assist in complex reasoning, facilitate constructive deliberation, and ultimately contribute to the robustness and efficiency of scientific validation.

### A.1 Case Study: Automated "Report Card" for Reviewer omKD

**Context.** Reviewer omKD evaluated submission #14286, which studies the rate–distortion–perception trade-off without common randomness. The reviewer assigned an overall *rating* of **5 / 10** ("marginally below the acceptance threshold") and a *confidence* of **4 / 5**.

Dimension	Score <sup>†</sup>	Weight	Automated Feedback Summary
Coverage	3	0.30	Touches novelty, significance, technical soundness and presentation, but omits an explicit discussion of <i>empirical validation, related work depth, and ethical considerations</i> .
Specificity	3	0.25	Pinpoints the need for illustrative examples and clearer exposition (e.g. after Def. 3.3), yet provides no page/line references or concrete rewrites. Questions are high-level rather than surgically actionable.
Evidence Base	2	0.20	Cites Blau & Michaeli and alludes to "dense notation" but gives no quotations, equations, or empirical numbers to substantiate claims. Assertions such as "significant practical implications" lack supporting rationale.
Constructiveness & Tone	4	0.25	Polite, professional, and encourages acceptance conditional on revisions. Suggests improvements without dismissiveness. Minor language issues ("reaslim", "insecpt").
<b>Weighted Composite</b>			<b>3.1 / 5</b>

Table 3: Automated Report Card produced by the LLM feedback system. <sup>†</sup>Scale: 1 (Poor) – 5 (Excellent).

#### Strengths Detected by the System.

- Provides a concise summary of the submission’s main contribution and positions it relative to prior RDP work.
- Highlights a concrete missing element (illustrative or toy examples) that could materially improve clarity.
- Maintains a constructive, collegial tone; explicitly encourages authors to revise rather than reject outright.

#### Areas for Improvement.

- **Deeper evidence.** Quote or paraphrase specific theorems/eqs. when critiquing clarity; point to the exact section where "dense notation" obstructs understanding.
- **Broader coverage.** Comment on empirical or synthetic experiments (even if absent), dataset choices, and any ethical implications of lossy generative compression.
- **Finer-grained actionables.** Offer line-level edits, figure suggestions, or exemplar mini-case studies ("For instance, Fig. 2 could show...") to operationalize the call for examples.
- **Minor language edits.** Correct typographical slips (e.g. "reaslim" → "realism", "insecpt" → "inspect").

**Illustration of LLM-Assisted Feedback.** The LLM system generated the above multi-dimensional analysis in ≈ 4s, surfacing overlooked dimensions (empirical validation, ethics) and converting vague remarks into concrete, citable suggestions. In pilot experiments across 200 ICLR 2025 reviews, reviewers who received such AI-generated report cards increased their average word count by 28 %

and added 1.7 additional page-level citations on revision—aligning with observations in Kim et al. [2025], Thakkar et al. [2025].

**Take-away for Peer-Review Policy.** Automated report cards can (i) standardize feedback quality signals for Area Chairs, (ii) nudge reviewers toward more evidence-based critiques, and (iii) serve as a low-friction “review curriculum” for junior reviewers. Strategic integration—e.g. releasing the card *before* author response and again *after* rebuttal—could systematically uplift review depth without extending timelines.

## B Additional Notes on Illustrative Experimental Setup and Prompts

The experiments described in Section 6 are illustrative and designed to demonstrate the potential of LLMs and highlight data needs. This appendix provides further conceptual details. A real implementation would require careful dataset curation, prompt engineering, and rigorous evaluation design beyond what is sketched here.

### B.1 General Considerations for ICL

For all In-Context Learning (ICL) experiments, we would conceptually use publicly available data from OpenReview for conferences like ICLR (e.g., ICLR 2024 & 2025 data, focusing on papers with full review cycles). Experiments are run with a state-of-the-art LLM available at the time of execution; specifically, we used OpenAI’s O3 model throughout.

#### System message (REVIEW GUIDELINES):

You are a simulated reviewer for ICLR 2025  
Here are the review guidelines you must adhere to  
Review Guidelines  
Below is a description of the questions you will be asked on the review form for each paper and some guidelines on what to consider when answering these questions. Remember that answering no to some questions is typically not grounds for rejection. When writing your review please keep in mind that after decisions have been made reviews and metareviews of accepted papers and optedin rejected papers will be made public

Summary Briefly summarize the paper and its contributions. This is not the place to critique the paper the authors should generally agree with a wellwritten summary. This is also not the place to paste the abstractplease provide the summary in your own understanding after reading

Strengths and Weaknesses Please provide a thorough assessment of the strengths and weaknesses of the paper. A good mental framing for strengths and weaknesses is to think of reasons you might accept or reject the paper. Please touch on the following dimensions

Quality Is the submission technically sound? Are claims well supported eg by theoretical analysis or experimental results? Are the methods used appropriate? Is this a complete piece of work or work in progress? Are the authors careful and honest about evaluating both the strengths and weaknesses of their work?

Clarity Is the submission clearly written? Is it well organized? If not please make constructive suggestions for improving its clarity. Does it adequately inform the reader? Note that a superbly written paper provides enough information for an expert reader to reproduce its results

Significance Are the results impactful for the community? Are others researchers or practitioners likely to use the ideas or build on them? Does the submission address a difficult task in a better way than previous work? Does it advance our understanding/knowledge on the topic in a demonstrable way? Does it provide unique data or unique conclusions about existing data or a unique theoretical or experimental approach?

Originality Does the work provide new insights, deepen understanding or highlight important properties of existing methods? Is it clear how this work differs from previous contributions with relevant citations provided? Does the work introduce novel tasks or methods that advance the field? Does this work offer a novel combination of existing techniques and is the reasoning behind this combination wellarticulated? As the questions above indicates originality does not necessarily require introducing an entirely new method. Rather a work that provides novel insights by evaluating existing methods or demonstrates improved efficiency, fairness, etc. is also equally valuable

You can incorporate Markdown and LaTeX into your review

Quality Based on what you discussed in Strengths and Weaknesses please assign the paper a numerical rating on the following scale to indicate the quality of the work

4 excellent  
3 good  
2 fair  
1 poor

Clarity Based on what you discussed in Strengths and Weaknesses please assign the paper a numerical rating on the following scale to indicate the clarity of the paper

4 excellent  
3 good  
2 fair  
1 poor

Significance Based on what you discussed in Strengths and Weaknesses please assign the paper a numerical rating on the following scale to indicate the significance of the paper

4 excellent  
3 good  
2 fair  
1 poor

Originality Based on what you discussed in Strengths and Weaknesses please assign the paper a numerical rating on the following scale to indicate the originality of the paper

4 excellent  
3 good  
2 fair  
1 poor

Questions Please list up and carefully describe questions and suggestions for the authors which should focus on key points ideally around 35 that are actionable with clear guidance. Think of the things where a response from the author can change your opinion, clarify a confusion or address a limitation. You are strongly encouraged to state the clear criteria under which your evaluation score could increase or decrease. This can be very important for a productive rebuttal and discussion phase with the authors

Limitations Have the authors adequately addressed the limitations and potential negative societal impact of their work? If so simply leave yes if not please include constructive suggestions for improvement. General authors should be rewarded rather than punished for being up front about the limitations of their work and any potential negative societal impact. You are encouraged to think through whether any critical points are missing and provide these as feedback for the authors

Overall Please provide an overall score for this submission. Choices

10 Strong Accept Technically flawless paper with groundbreaking impact on one or more areas of AI with exceptionally strong evaluation, reproducibility and resources and no unaddressed ethical considerations

9 Accept Technically solid paper with high impact on at least one subarea of AI or moderately high impact on multiple areas with good to excellent evaluation, resources and reproducibility

8 Weak Accept Technically sound and reasonably well evaluated but not outstanding. Impact may be more narrow or incremental

7 Borderline Accept Reasons to accept slightly outweigh reasons to reject. A solid paper with some limitations eg ablation, clarity, scope

6 Weak Reject Reasons to reject slightly outweigh reasons to accept. Paper may lack novelty, strong results or thorough evaluation

5 Reject Paper has clear weaknesses in either technical content clarity originality or evaluation  
 4 Strong Reject Paper suffers from significant issues in methodology experimental support or contribution  
 3 Serious Reject Major flaws in correctness relevance or ethical considerations  
 2 Severe Reject Paper makes minimal scientific contribution or is fundamentally flawed  
 1 Desk Reject Level Not suitable for review due to major violations eg formatting plagiarism or entirely out of scope  
 Confidence Please provide a confidence score for your assessment of this submission to indicate how confident you are in your evaluation Choices  
 5 You are absolutely certain about your assessment You are very familiar with the related work and checked the mathother details carefully  
 4 You are confident in your assessment but not absolutely certain It is unlikely but not impossible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work  
 3 You are fairly confident in your assessment It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work Mathother details were not carefully checked  
 2 You are willing to defend your assessment but it is quite likely that you did not understand the central parts of the submission or that you are unfamiliar with some pieces of related work Mathother details were not carefully checked  
 1 Your assessment is an educated guess The submission is not in your area or the submission was difficult to understand Mathother details were not carefully checked

### System message (LLM as a judge):

You are an expert NLP evaluator. Given the two textual statements below:

Generated statement:  
 { predicted }

Real statement:  
 {real}

Please answer the following questions:

1. How many distinct points were raised in the \*Real statement\*? Make a checklist.
2. Now, please now count how many of the generated points overlap with those in the checklist and return the scalar number.

Return your answer in the following format:  
 Real points: <number> - [<point1>, <point2>, ...]  
 Hits (overlap): <number>

## B.2 A.1 Details for Experiment 1: Review Component Generation.

### Input Data per Instance:

- Paper Title
- Paper Abstract
- Paper Content in text
- If predicting rebuttals, paper initial review

### Few-Shot ICL Prompt Structure:

{REVIEW\_GUIDELINES}  
 You are an expert reviewer. Given a few example papers' title , abstract {f", and its {paper\_content\_desc}" if paper\_content else ""}, and real review comments, infer the key strengths of the current test input paper in your own words.

--- EXAMPLE 1 ---  
 PAPER TITLE: [Example Paper 1 Title]  
 PAPER ABSTRACT: [Example Paper 1 Abstract]  
 PAPER CONTENT: [Example Paper 1 Content]

POTENTIAL STRENGTHS:  
 - [Strength 1 from human review of Ex1]  
 - [Strength 2 from human review of Ex1]  
 POTENTIAL WEAKNESSES:  
 - [Weakness 1 from human review of Ex1]  
 - [Weakness 2 from human review of Ex1]  
 (if rebuttal )  
 INITIAL SCORE: [Actual initial score for Ex1]  
 INITIAL REVIEW: [Actual review for Ex1]  
 AUTHOR REBUTTAL: [Actual author rebuttals to Ex1]

--- EXAMPLE 2 ---  
 PAPER TITLE: [Example Paper 2 Title]  
 PAPER ABSTRACT: [Example Paper 2 Abstract]  
 PAPER CONTENT: [Example Paper 2 Content]

POTENTIAL STRENGTHS:  
 - [Strength 1 from human review of Ex2]  
 - [Strength 2 from human review of Ex2]  
 POTENTIAL WEAKNESSES:  
 - [Weakness 1 from human review of Ex2]  
 - [Weakness 2 from human review of Ex2]  
 (if rebuttal )  
 INITIAL SCORE: [Actual initial score for Ex2]  
 INITIAL REVIEW: [Actual review for Ex2]  
 AUTHOR REBUTTAL: [Actual author rebuttals to Ex2]

--- ACTUAL TASK ---  
 PAPER TITLE: [Test Paper Title ]  
 PAPER ABSTRACT: [Test Paper Abstract]  
 PAPER CONTENT: [Test Paper Content]

POTENTIAL STRENGTHS:  
 [LLM generates this ]

POTENTIAL WEAKNESSES:  
[LLM generates this ]  
(if rebuttal )  
INITIAL SCORE: [Actual initial score]  
INITIAL REVIEW: [Actual review]  
AUTHOR REBUTTAL: [LLM generates this]

#### Notes on Evaluation:

- LLM-as-Judge for semantic coverage: Prompt another powerful LLM (the "judge") with the human review points and the AI-generated points. Ask the judge to determine, for each human point, if it was semantically captured by the AI. Calculate precision (AI points that are valid) and recall (human points captured by AI).

### B.3 A.2 Details for Experiment 2: Rating Prediction.

#### Input Data per Instance:

- Paper Title
- Paper Abstract
- Paper Content in text
- Initial Review
- Initial Review Score (for the specific review)
- Author Rebuttal

#### Few-Shot ICL Prompt Structure:

REVIEW\_GUIDELINES  
You are an expert reviewer. Based on a paper's {paper\_content\_desc}, an initial review, and the author rebuttal, predict what the final score (rating) should be after considering the rebuttal. Return only a single number.

--- EXAMPLE 1 ---  
PAPER TITLE: [Example Paper 1 Title]  
PAPER ABSTRACT: [Example Paper 1 Abstract]  
PAPER CONTENT: [Example Paper 1 Content]  
INITIAL SCORE: [Actual initial score for Ex1]  
INITIAL REVIEW: [Actual review for Ex1]  
AUTHOR REBUTTAL: [Actual author rebuttals to Ex1]  
FINAL\_RATING: [Actual final rating for Ex1]

--- EXAMPLE 2 ---  
PAPER TITLE: [Example Paper 2 Title]  
PAPER ABSTRACT: [Example Paper 2 Abstract]  
PAPER CONTENT: [Example Paper 2 Content]  
INITIAL SCORE: [Actual initial score for Ex2]  
INITIAL REVIEW: [Actual review for Ex2]  
AUTHOR REBUTTAL: [Actual author rebuttals to Ex2]  
FINAL\_RATING: [Actual final rating for Ex2]

--- ACTUAL TASK ---  
PAPER TITLE: [Test Paper Title ]  
PAPER ABSTRACT: [Test Paper Abstract]  
PAPER CONTENT: [Test Paper Content]  
INITIAL SCORE: [Actual initial score for Test Paper]  
INITIAL REVIEW: [Actual review for Test Paper]  
AUTHOR REBUTTAL: [Actual author rebuttals to Test Paper]  
FINAL\_RATING:  
[LLM generates this ]

#### Notes on Evaluation:

- Predicting exact final scores is hard. More informative might be predicting the *direction* of change (increase, decrease, no change) and the *magnitude* of change if any (e.g., +/- 1 point, +/- 2 points).
- This task critically highlights the need for data where reviewers *explicitly state why their score changed (or not) based on the rebuttal*. Without this, the LLM is learning from coarse signals.

### B.4 A.3 Details for Experiment 3: AI-Assisted Generation of "Reviewer Report Card" Feedback

#### Input Data per Instance:

- Paper Abstract (to provide context for the review's relevance)

- Full text of one Human-Written Review

### Few-Shot ICL Prompt Structure:

You are an AI assistant tasked with providing constructive feedback on a peer review to help the reviewer improve.  
 Based on the paper's abstract and the provided review, generate a "Reviewer Report Card" commenting on:  
 1. Coverage: Did the review address key aspects like novelty, significance, technical soundness, and empirical validation relative to the abstract?  
 2. Specificity: Were the critiques concrete and actionable? Were praises specific?  
 3. Constructiveness: Was the feedback framed to help authors improve?  
 4. Tone: Was the language professional and respectful?

--- EXAMPLE 1 ---

PAPER ABSTRACT: [Abstract of Example Paper 1]

HUMAN REVIEW TEXT: [Full text of a human review for Example Paper 1 – ideally a review with known good/bad points]

REVIEWER REPORT CARD:

- Coverage: Good coverage of methodology and experiments. Novelty assessment could be more detailed.
- Specificity : Weakness point #2 ("needs more experiments") is vague. Suggestion: "Specify which experiments would strengthen the claims, e.g., on dataset X or against baseline Y."
- Constructiveness : Overall constructive .
- Tone: Professional .

--- EXAMPLE 2 ---

PAPER ABSTRACT: [Abstract of Example Paper 2]

HUMAN REVIEW TEXT: [Full text of another human review for Example Paper 2]

REVIEWER REPORT CARD:

- Coverage: Focused heavily on minor presentation issues , but missed a potential flaw in the main theoretical claim mentioned in the abstract .
- Specificity : commendable specificity in pointing out typos .
- Constructiveness : Suggestions for improvement are limited .
- Tone: Slightly dismissive in the summary ("Obvious incremental work."). Suggestion: "Rephrase to focus on specific technical reasons for limited novelty."

--- ACTUAL TASK ---

PAPER ABSTRACT: [Abstract of Test Paper]

HUMAN REVIEW TEXT: [Full text of Test Human Review]

REVIEWER REPORT CARD:

[LLM generates this feedback across the four dimensions]

### Notes on Evaluation:

- The "Expert-Authored Feedback on Review" for the few-shot examples would be the hardest to source. Initially, these might need to be carefully crafted by the researchers to exemplify good feedback.
- Evaluation would be primarily qualitative, involving experienced reviewers or ACs rating the LLM's feedback on dimensions like: Accuracy (does the LLM correctly identify strengths/weaknesses of the review?), Helpfulness (would this feedback help the original reviewer improve?), Actionability (are the suggestions concrete?).
- This aligns with the spirit of the ICLR 2025 Review Feedback Agent [2, 3, 28], which provided actionable suggestions to reviewers.

These more detailed conceptual setups underscore that while ICL can provide initial insights, the development of robust, reliable AI assistants for peer review will necessitate dedicated datasets, fine-tuning, and sophisticated human-in-the-loop evaluation methodologies. The primary purpose of these illustrative experiments in the position paper is to argue for the *potential* and to highlight *what is needed* to realize it.