Let Papers Flow: AI Conferences Should Embrace Submission Explosion via Autonomous Review Pipelines

Chaoyue He *, Xin Zhou, Di Wang, Hong Xu, Wei Liu, Chunyan Miao

Posted Date: 13 April 2026

doi: 10.20944/preprints202604.0797.v1

Keywords: peer review; AI conferences; autonomous review pipelines; submission explosion; AI-assisted review; large language models; continuous certification; research acceleration; science of science; sustainable AI research

 doi:10.20944/preprints202604.0797.v1

Article

Let Papers Flow: AI Conferences Should Embrace Submission Explosion via Autonomous Review Pipelines

Chaoyue He [1], Xin Zhou [1], Di Wang [1], Hong Xu [1], Wei Liu [2] and Chunyan Miao [1]

[1] Alibaba–NTU Global e-Sustainability CorpLab (ANGEL), Singapore
[2] Alibaba Group, Hangzhou, China
* Correspondence: cyhe@ntu.edu.sg

Abstract

As AI tools accelerate literature search, experimentation, and drafting, AI conferences face a growing, structural submission explosion. This position paper argues that to survive this era of accelerated research, venues must formally adopt **autonomous review pipelines** as their core control plane. By deploying machine-first, human-governed systems to run the scalable first pass of review—handling routine evidence construction, integrity checks, and routing—conferences can stop treating abundance as a pathology. Making this automated control plane the spine of the review process preserves scarce human attention for high-stakes escalation, calibration, and edge cases. Furthermore, we show that establishing this infrastructure carries a profound downstream implication: it enables conferences to decouple baseline technical certification from seasonal curation, ultimately absorbing more submissions, shortening time-to-feedback, and tightening quality standards without breaking the human review commons or requiring perpetual committee inflation.

Keywords: peer review; AI conferences; autonomous review pipelines; submission explosion; AI-assisted review; large language models; continuous certification; research acceleration; science of science; sustainable ai research

1. Introduction

AI research has entered a regime of accelerated supply. Tools that assist with literature search, coding, experimentation, drafting, and even parts of scientific discovery are compressing research cycles and increasing the rate at which plausible submissions can be produced [1–4]. The institutional bottleneck is therefore shifting. The main scarcity is no longer how many PDFs the community can produce; it is how much trusted human attention the field can still devote to careful evaluation, calibration, appeals, and reading.

In Herbert Simon's classic formulation, an information-rich world consumes attention [5]. At flagship AI-conference scale, that scarcity is now visible across the entire stack. ICLR 2025 reported 11,603 submissions and 18,325 reviewers, with 99.98% of papers receiving at least three reviews [6]. NeurIPS 2025 reported 21,575 valid main-track submissions handled by 20,518 reviewers, 1,663 ACs, and 199 SACs, and its program chairs explicitly described a process that is getting noisier and harder to calibrate at scale [7,8]. The scarce resource is not just reviewer time. It is also AC and SAC time for synthesis, PC time for governance, and reader time for deciding what the field will actually absorb. As Figure 1 later shows, the 2025 reviewer-and-chair stacks already run into the tens of thousands of named committee roles once these layers are summed, so the question is no longer only whether one more cycle can be staffed, but whether perpetual committee enlargement is itself a sustainable review architecture.

Venues are already experimenting with machine assistance. ICLR 2025 deployed a reviewer-feedback agent [9–11] and piloted a TMLR partnership [12]; ICML 2026 piloted Google's Paper

 doi:10.20944/preprints202604.0797.v1

Assistant Tool (PAT) on the author side [13,14]; and AAAI-26 launched an official AI-assisted peer-review pilot in which the AI review gives no score or recommendation and all decisions remain human [15]. At the same time, broader norms are unstable. Nature has documented both growing unease around AI-mediated review and substantial undisclosed or policy-violating use, while prompt-injection attacks and illicit AI review have already become concrete governance problems [16–19].

We therefore use **autonomous review pipeline** to mean a machine-run first-pass control plane that produces structured evidence and routing signals, not a sovereign AI judge. The point is not to add one more reviewer bot. It is to let conferences officially adopt and govern machine-first review infrastructure so that routine evidentiary work scales with compute, human depth is reserved for escalation and certification, and the research loop speeds up rather than stalling at review. Table 1 defines the paper's core vocabulary; Figure 1 sketches the architecture.

Position. To survive accelerated research, AI conferences must adopt autonomous review pipelines as their first-pass control plane. By scaling routine evidentiary work with compute, this machine-first spine reserves scarce human attention for high-stakes escalations. Ultimately, it decouples continuous technical certification from selective curation, accelerating scientific iteration while preserving a sustainable review commons.

Table 1. Core vocabulary for the architecture advocated in this paper. Figure 1 then shows why the attention-allocation problem is already visible at current venue scale.

Term	Working definition	Why it matters here
Submission explosion	Sustained growth in submission volume and variety as paper production becomes cheaper.	Volume should be designed for rather than feared.
Autonomous review pipeline	A machine-run first-pass review stack that generates structured evidence and routing signals without acting as the final judge.	Review becomes scalable without removing humans from legitimacy.
Attention budget	The limited human time available for reviewers, chairs, and readers.	The scarce resource is depth, not PDFs.
Escalation	Explicit transfer of a paper or cluster from machine-first review to deeper human scrutiny.	Human effort can be concentrated where uncertainty is highest.
Rolling certification	Continuous judgment that work is sound enough for the technical record, potentially with typed uncertainty notes.	Certification need not wait for a single seasonal batch.
Conference curation	Selective allocation of scarce visibility such as talks, posters, awards, and synthesis slots.	Visibility can be separated from baseline certification.

2. Motivation: Suppressing Volume Is the Wrong Response

The submission explosion is the fundamental premise that forces a redesign of conference architecture. The usual instinct in a review crisis is to press down on this supply: stricter desk rejection, per-author caps, anti-overlap rules, or a thicker social norm that authors should "submit less." Some abuse control is necessary, but volume suppression misdiagnoses the bottleneck. The real problem is that a human-first seasonal process still tries to do too many jobs at once: admissibility checks, novelty reconstruction, literature positioning, artifact inspection, conflict resolution, and visibility allocation.

We unpack this dynamic across three dimensions: Section 2.1 details how major venues already show this curve, Section 2.2 explains how abundance improves the scientific record, and Section 2.3 argues fairness must attach to due process rather than equal labor.

 doi:10.20944/preprints202604.0797.v1

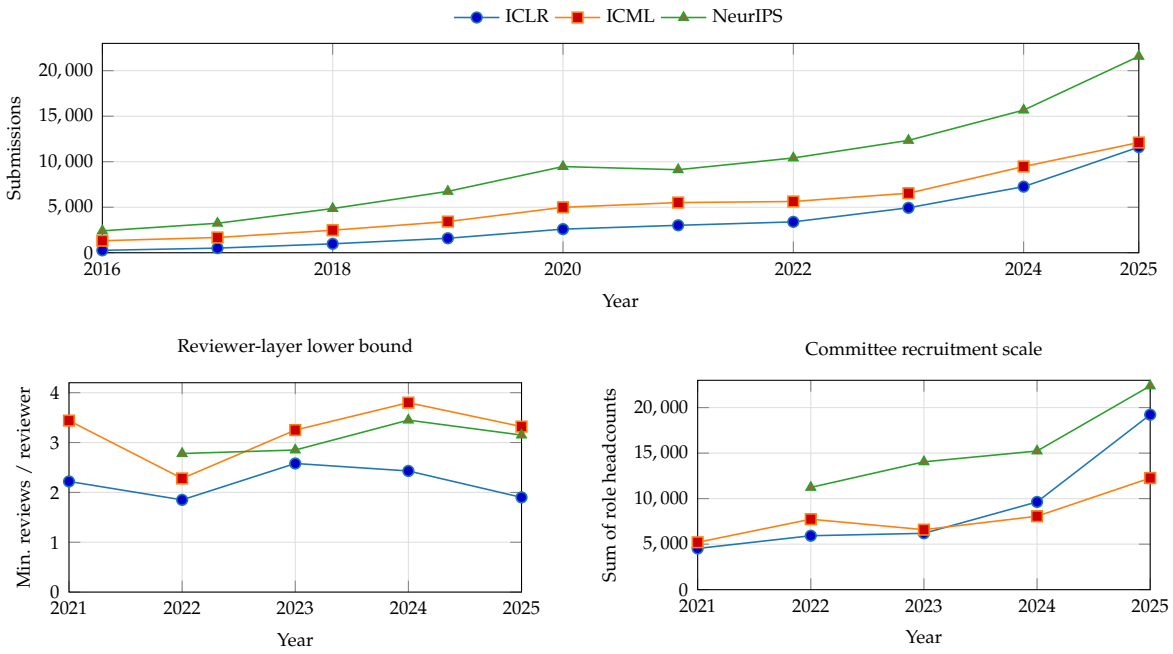

Figure 1. Submission growth and the sustainability of the human review stack. **Top:** 2016–2025 main-track submission trajectories for ICLR, ICML, and NeurIPS. **Bottom-left:** A conservative lower bound on first-round reviewer demand, computed as $3 \times$ submissions/reviewers. **Bottom-right:** The sum of reported reviewer, AC/MR, SAC/SMR, and PC headcounts over time. Exact headcounts, along with derived ratios and absolute layer-specific breakdowns, appear in Appendix B.

2.1. The Big-Three Venues Already Show the Curve

Figure 1 separates three trends: rising submission volume, the conservative floor for reviewer demand, and the escalating scale of annual committee recruitment. The top panel shows sustained, structural growth across venues, with ICLR growing fastest relatively and NeurIPS remaining largest absolutely.

The lower-left panel establishes a conservative lower bound for reviewer workload ($3N/R$), assuming a standard minimum of three reviews per paper before accounting for emergency backfills or second-round checks [6,20–22]. Maintaining this baseline requires continuous, massive reviewer expansion, which directly incurs quality-control costs. For instance, ICLR 2022 instituted a mentorship pipeline to promote inexperienced reviewers [23], and NeurIPS 2025 explicitly linked scale-driven recruitment to increased calibration noise [7]. Expanding the pool is therefore a quality-management challenge, not just a headcount problem.

The lower-right panel exposes the core sustainability threat: the current human-first system stabilizes only through perpetual committee enlargement. By 2025, the combined sum of named committee roles (reviewers, handling chairs, senior synthesis chairs, and PCs) reached roughly 19.2k at ICLR, 12.3k at ICML, and 22.4k at NeurIPS (Table A6). The institutional warning is clear: as AI lowers the cost of submissions, relying on annual, massive recruitment drives is an unsustainable control plane. While reviewer pools scale elastically, the synthesis and governance layers remain critically thin and hard to replenish (detailed historically in Appendix B, Tables A4–A10 and Figures A1–A2).

2.2. Abundance Can Improve the Scientific Record

Restriction-first responses also miss something scientifically valuable: abundance is not only a burden on review; it is evidence that the field has become more experimentally expressive. As AI tools accelerate research and lower the cost of drafting, more of the real search process can be externalized: replications, narrow but useful observations, artifact fixes, negative results, and domain-specific variants that previously lost out to a lab's scarce "best shot." Publication bias against negative or

 doi:10.20944/preprints202604.0797.v1

null results is already well documented [24,25]. A review architecture that can absorb more work can therefore make the record both broader and faster to correct.

Reader attention is the second bottleneck. Even if a venue could somehow review every paper carefully, the community would still need to decide what to read, cite, reproduce, and build on. Accepted-paper lists merely move overload downstream. This is why curation and interfaces belong inside the review architecture. Machine-first review can emit claim cards, cluster pages, artifact status, and contrastive summaries that help readers navigate abundance rather than drown in it.

2.3. Fairness Should Attach to Due Process, Not Equal Labor per PDF

At scale, fairness cannot mean giving every PDF the same quantity of human labor. That norm becomes gamable once submission cost falls and papers become highly correlated. A fairer principle is that every submission receives the same transparent rules, the same structured first pass, and the same accessible escalation path, while scarce human depth goes where uncertainty, disagreement, risk, or potential impact are greatest.

The current system already shows the strain. NeurIPS 2025 emphasized rising noise and calibration difficulty, explicitly linking larger-scale recruitment to a greater chance of less experienced reviewers and ACs [7]. ICLR 2025 launched a reviewer-feedback agent because review quality was harder to sustain under rapidly growing volume [9,10]. Review policies have become more operationally interventionist as well: ICLR warns that low-quality placeholder reviews can trigger desk rejection of the reviewers' own papers, while ICML and ICLR still discuss emergency-review backfills when the three-review floor is not met [20,21,26]. The NeurIPS 2026 Position Track blog post also reported that roughly 35% of reviewers in the previous year's position track were unresponsive, forcing emergency intervention and delays [27–30]. These are not marginal frictions. They are signs that the system is already rationing trusted attention and spending growing effort just to keep the human-first pipeline functional.

3. The Core Thesis: Autonomous Review Pipelines Should Be the Control Plane

To survive submission abundance, AI conferences must formally adopt *autonomous review pipelines* as their default control plane. By *control plane*, we mean a machine-run, venue-governed layer that ingests submissions, constructs structured evidence, triggers interventions, and routes scarce human attention [31]—not a sovereign AI judge.

This scalable first pass should systematically answer bounded questions before humans spend depth: Is the submission policy-compliant? What are its central claims and missing baselines? Are artifacts executable? Does the work belong to a larger portfolio of highly correlated submissions? Figure 1 sketches six core functions handling these checks. By universally applying these baseline diagnostics, the minimum floor of evidentiary depth rises, allowing human reviewers to enter the process with a richer dossier and concentrate exclusively on high-stakes escalations (e.g., difficult claims, disputed novelty, policy violations).

This architecture is multi-sided. Authors receive pre-submission diagnostics (missing controls, clarity checks); venues gain robust routing and cluster-detection tools; and readers inherit durable discovery assets like claim cards and contrastive summaries [10,32,33]. Crucially, this pipeline aligns with the accelerated tempo of AI research [2]. Treating incoming papers not as isolated PDFs, but as potentially correlated portfolios or rapid experimental iterations, allows the venue to group related submissions, eliminate duplicated manual reconstruction, and dramatically shorten the latency between idea, evidence, and critical feedback.

To replace opaque scoring with actionable evidence, the concrete output of this first pass must be a structured pipeline report exposing the same underlying data to authors, reviewers, and readers. At minimum, it should contain: (1) **Policy summary** (anonymity, formatting, disclosures); (2) **Claim cards** (extracted core contributions); (3) **Literature positioning** (retrieved baselines); (4) **Artifact status** (code execution results); (5) **Cluster context** (grouping metadata for parallel work); (6) **Uncertainty flags** (machine-generated risk notes); and (7) **Escalation recommendation** (explicit routing triggers). These

 doi:10.20944/preprints202604.0797.v1

structured objects make feedback legible enough for authors to repair issues while the work is still live [34].

To ground this architectural claim, Table 2 traces a hypothetical submission through this exact pipeline. Rather than a static "accept" or "reject" verdict, the machine handles routine validation and evidence structuring, producing an auditable report that intentionally routes the paper to a human expert only when novelty bounds or evaluation robustness become uncertain. An extended operational sketch of this pipeline is provided in Appendix C.

Table 2. A worked example of a submission flowing through the autonomous review pipeline. The machine handles routine validation and evidence structuring, but intentionally routes the paper to a human expert when baselines and evaluation robustness become uncertain.

Pipeline Stage	Output for Hypothetical Submission: *Semantic Per-Pair DPO (SP2DPO)*
Policy Screen	**Pass.** Anonymity preserved; format compliant; anonymous GitHub link provided; required compute disclosures present.
Extracted Claims	**C1:** SP2DPO improves win-rate on AlpacaEval by 12% over standard DPO. **C2:** SP2DPO mitigates response-length bias by balancing semantic density across preference pairs.
Literature Positioning	Accurately positions against standard DPO and cDPO. **Flag:** Missing comparison to Offset DPO (ODPO), which is highly relevant for length-bias mitigation claims.
Artifact Probe	**Execution Pass.** Repository cloned successfully. Dependencies installed. Training script dry-run executed without syntax errors.
Uncertainty Flags	**Moderate Risk:** Claim 1's 12% win-rate improvement is statistically large but highly sensitive to the specific prompt templates used in AlpacaEval evaluation.
Escalation Decision	**Route to Human Expert.** *Justification:* While technically sound (code runs) and policy compliant, the missing ODPO baseline and the prompt-sensitivity of the evaluation require domain-expert judgment to verify robustness.
Typed Certification	*Policy:* Compliant. *Artifact:* Probed (Pass). *Claims:* Escalated.
Reader Claim Card	*(Post-Certification output generated for readers)* **Method:** Semantic Per-Pair DPO (SP2DPO). **Core Claim:** Reduces length bias in preference tuning. **Status:** Certified (with attached reviewer notes on template sensitivity).

Ultimately, fairness in this regime attaches to uniform due process rather than equal manual labor per PDF. Every submission receives the same transparent rules and structured first pass, while human authority becomes more meaningful precisely because it is no longer diluted across routine screening.

4. Strict Standards Can Coexist with Faster Research Cycles

The primary objection to this architectural shift is that exploding submissions will inevitably degrade accepted quality. We argue no: if pipelines operate as certification stacks rather than synthetic reviewers, submission abundance widens intake without relaxing the evidentiary bar.

First, machine-first review raises both the **floor** and **ceiling** of rigor. It guarantees uniform baseline checks (anonymity, policy compliance, artifacts) that currently rely on reviewer goodwill, while redirecting expert human attention exclusively to high-impact, disputed, or uncertain edge cases (Appendix Table A2).

Second, a pipeline-first venue makes standards highly **auditable**. ML peer review is notoriously noisy and difficult to calibrate [35–39]. A governed pipeline mitigates this opacity by transparently logging module triggers, escalation flags, and human overrides.

Third, faster review cycles are essential for scientific health. Slower evaluation loops become increasingly costly when AI compresses the pace of proposing and executing research [1,2,40]. By accelerating the first pass, venues tighten the feedback loop, ensuring earlier error correction and better overall throughput [41].

Finally, structured first passes enable a crucial semantic shift toward *typed certification*. Rather than hiding judgment behind an opaque scalar verdict, venues can issue multidimensional certificates detailing policy compliance, artifact reproducibility, and claim support. This nuanced record is far more honest and actionable, successfully isolating technically reliable execution from empirically uncertain impact [33,36,38,42].

5. Human Roles, Governance, and the Economics of Attention

Autonomous review pipelines redeploy rather than eliminate human judgment: reviewers become escalation specialists, ACs/SACs focus on synthesis and calibration, PCs handle governance, and

 doi:10.20944/preprints202604.0797.v1

readers gain structured summaries (Appendix Table A3). However, explicit governance is critical to prevent correlated bias, where monocultural models might narrow collective scientific focus or over-favor standard paradigms [3]. Venues must ensure pluralistic backends, audit logs, and escalation policies designed to protect novel or out-of-paradigm work.

Simultaneously, the submission artifact must evolve beyond the transitional PDF wrapper [43]. Future submissions should integrate structured claim and executable artifact layers, enabling venues to directly verify dependencies and replication hooks, akin to a continuous-integration system for science [33,44].

Crucially, this redesign addresses the broken social contract of review. AI-accelerated research weakens or further erodes the historic symmetry between the cost of writing and reviewing a paper, deeply straining current incentive structures [45–47]. By reframing the economic tradeoff from "paid compute versus free volunteer labor" to "paid compute versus scarce human attention," a machine control plane acts as an essential shock absorber. It offloads routine verification to compute, preserving expensive human depth for edge cases where it is actually needed [34]. Just as importantly, it offers a path away from a conference ecosystem that must re-recruit an ever larger human committee every cycle merely to keep the seasonal workflow intact. We formalize this dynamic in Section 5.1 and address failure modes in Section 5.2.

5.1. A Falsifiable Model of Institutional Attention

To transition from vision to testable institutional theory, we formalize the allocation of scarce attention. Let N represent the total submissions; ω the shadow price of one hour of trusted human reviewer time; h_0 the average human hours required for a traditional manual review; and c_m the amortized computational cost per paper for the machine-first pass. In a pipeline architecture, the system escalates a fraction $p_e \in (0, 1)$ of papers to human experts, requiring h_e hours per escalated paper (where $h_e > h_0$ due to concentrated effort on difficult claims). The system incurs penalties for false clearances (rate λ_{FC}, cost K_{FC}) and false escalations (rate λ_{FE}, cost K_{FE}). Figure 1's reviewer panel provides a conservative empirical proxy for the first-round human term: once a venue guarantees at least three reviews, reviewer demand cannot fall below $3N/R$ reviews per reviewer, before any emergency backfill or discussion-phase follow-up [6,20,22]. The figure's committee-recruitment panel then adds the parallel institutional signal: even when the three-review floor is met, the human-first system does so by repeatedly enlarging the named committee stack.

Under traditional human-first review, expected attention cost is simply $C_{\text{human}} = Nh_0\omega$. A pipeline-first architecture, handling the first pass computationally and routing exceptions to humans, incurs a total expected cost of:

$$C_{\text{pipe}} = N\Big(c_m + p_e h_e \omega + \lambda_{FC} K_{FC} + \lambda_{FE} K_{FE}\Big) \tag{1}$$

The pipeline dominates ($C_{\text{pipe}} < C_{\text{human}}$) when the required escalation rate satisfies:

$$p_e < \frac{h_0}{h_e} - \frac{c_m + \lambda_{FC} K_{FC} + \lambda_{FE} K_{FE}}{h_e \omega} \tag{2}$$

This inequality clearly specifies failure modes: excessive noise (λ_{FE}), inability to clear routine work (p_e), or compute costs (c_m) eclipsing the value of human time. Crucially, as the shadow price of human attention spikes in the era of accelerated research ($\omega \to \infty$), the dominance condition approaches the asymptotic limit $p_e < h_0/h_e$. Provided the machine can safely handle enough routine work to offset the deeper human depth (h_e) spent on edge cases, conferences can sustain rigorous evaluation without exhausting the reviewer commons.

5.2. Failure Modes and Safeguards

Because autonomous review pipelines act as structural gatekeepers, they create new systemic risks that require explicit governance and safeguards:

 doi:10.20944/preprints202604.0797.v1

Automation bias. Reviewers or chairs may over-trust fluent machine outputs. The safeguard is to make uncertainty legible, require explicit human sign-off on escalated decisions, and audit whether humans are simply copying automated recommendations.

Gaming and adversarial optimization. Authors may learn to target the surface cues of the control plane. Venues should therefore rotate checks, combine independent modules, maintain adversarial evaluation sets, and reserve sanctions for deliberate manipulation.

Cluster errors. Portfolio clustering can over-merge genuinely distinct work or miss related thin slices. Cluster detections should trigger human review or author clarification, not irreversible penalties.

Reader-interface distortions. Once machine-generated summaries shape discovery, poor summaries can bias reader attention just as badly as poor review scores. Reader-facing interfaces should be evaluated directly as part of the review system, not as a downstream cosmetic feature.

Unequal access to tooling. If some authors can privately use strong paper assistants while others cannot, hidden adoption amplifies inequality rather than reducing communal workload fairly. A pipeline architecture addresses this by making bounded diagnostic tools venue-provided and broadly accessible before submission [13].

6. Downstream Implication: From Seasonal Batches to Continuous Review

A pipeline-first control plane naturally weakens the logic of seasonal batch review, shifting the optimal architecture toward continuous intake, rolling certification, and periodic curation. Deadlines remain useful for synchronized visibility, but lose their status as the sole gateway to the technical record. Existing precedents—such as TMLR [42,48], ACL Rolling Review [49], the Journal-to-Conference track [50], and NeurIPS's AC pilot [51], as well as distinctions between main tracks and dataset tracks [52,53]—already demonstrate the viability of decoupling baseline certification from seasonal curation.

A continuous regime aligns better with accelerated research by drastically reducing error-correction latency. Authors receive structured first-pass diagnostics early, enabling them to address missing controls or broken artifacts while code and experimental contexts remain fresh. This tightens the claim-criticism-repair loop without lowering standards [1,2,40]. Low-risk work seamlessly enters a searchable certified stream, while contested cases wait for deeper human review. Consequently, conference deadlines pivot to govern strictly *curation* (allocating scarce talks, posters, or awards) rather than existential admissibility.

This shift profoundly upgrades both the reader and reviewer experience. For readers navigating an accelerating literature, flat proceedings are insufficient; venues must provide claim cards, cluster pages, and typed certification states to direct attention intentionally. For reviewers, the burden shifts from synchronized seasonal bursts of thousands of volunteers to a smaller, continuously active pool of escalation experts, making their service more legible, intellectually valuable, and easier to sustain than annual committee inflation.

Ultimately, if agentic workflows generate drafts and ablations on a near-daily cadence, seasonal review spikes become structurally misaligned with the tempo of knowledge production [1–3]. Conferences face a stark choice: enforce harsh supply suppression, or build a governed shock absorber. A continuous autonomous review pipeline offers the sustainable path forward, absorbing research velocity while keeping final legitimacy, appeals, and accountability strictly human-governed [40].

7. A Staged Path to Venue Adoption

No major venue should jump to a fully automated control plane overnight. Moving toward a pipeline-first architecture requires bounded, auditable steps judged by institutional metrics—turnaround, reviewer burden, escalation quality, and reader utility—rather than mere accept/reject classification accuracy. The objective is to build infrastructure that matches the tempo of accelerated research while preserving human accountability.

Stage 1: Officialization of bounded assistance. Venues must replace hidden private usage with transparent, task-level policies specifying allowed AI use, required disclosures, and human-only

 doi:10.20944/preprints202604.0797.v1

actions. As recent pilots and violations show, formalizing access equalizes the playing field, preventing well-resourced labs from quietly leveraging superior private tools [15,26,54–57].

Stage 2: Author- and intake-side automation. The lowest-risk starting point is deploying PAT-style diagnostics, format checks, overlap screening, and basic claim extraction before human assignment [13,14,58]. This catches avoidable issues while code and experimental context are fresh, responsibly accelerating research by shifting criticism earlier in the loop [1,2].

Stage 3: Venue-side routing and portfolio escalation. Once first-pass reports are reliable, the infrastructure must integrate directly into conference workflows (e.g., OpenReview) [33,44,59–61]. Rather than asking models for free-form reviews, chairs use the control plane for cluster detection, uncertainty surfacing, and risk-based escalation, treating highly correlated submissions as joint portfolios rather than isolated PDFs.

Stage 4: Decoupling certification from curation. Building on existing rolling pathways [42,48–51], venues can finally separate continuous intake from seasonal visibility. Routine, sound work enters a searchable certified stream rapidly, while conferences reserve scarce talk slots, posters, and synthesis sessions strictly for high-attention curation.

Staging lowers institutional risk without ignoring the accelerating supply of papers. Conferences do not need to wait for a flawless synthetic judge; they only need governed infrastructure that speeds up routine evidence construction and sharper escalation, keeping final legitimacy and appeals safely in human hands [3,34,40]. Appendix G summarizes how existing pilots inform this path.

8. Alternative Views

The proposal to adopt autonomous review pipelines as a central control plane naturally invites comparisons to several alternative reform paradigms currently debated within the community.

View 1: The labor deficit model.

This view treats the crisis as a mere staffing shortage, solvable by aggressively recruiting larger reviewer pools and improving incentives. **Rebuttal:** While better incentives help and Figure 1 shows recruitment can buffer the first-pass layer, this relies on repeated large-scale expansion, emergency backfilling, and continuous influxes of less experienced reviewers. The deeper failure is sustainability. A workflow demanding perpetual committee enlargement is not a stable control plane for machine-accelerated paper supply. The true bottleneck is not raw headcount, but the highly inelastic supply of trusted attention at the synthesis and governance layers needed to evaluate evidence, calibrate noise, and defend decisions.

View 2: Supply suppression.

This perspective aims to protect review quality by artificially making submissions scarcer—enforcing strict per-author caps, aggressive desk-rejection quotas, or penalizing narrow extensions. **Rebuttal:** Artificial scarcity sacrifices the scientific benefits of accelerated research. Aggressive filtering disproportionately targets the exact artifacts an accelerated field needs to solidify its epistemic foundation: replications, negative results, specialized findings, and exploratory variants. We should build infrastructure that absorbs abundance, rather than punishing researchers for increased productivity.

View 3: The egalitarian fallacy.

On this view, fairness dictates that every submitted PDF must receive roughly the same amount of human manual labor and depth. **Rebuttal:** In a high-volume regime, this baseline is easily gamed; it allows well-resourced labs to appropriate communal attention by flooding the queue with highly correlated variants. True equity lies in uniform due process—applying the exact same automated evidentiary standards and policy checks to all submissions—while reserving scarce human depth for papers that exhibit genuine uncertainty, novelty, or impact.

 doi:10.20944/preprints202604.0797.v1

View 4: The sovereign AI assessor.

The fully automated endpoint argues for transitioning directly to LLM judges that synthesize reviews and issue final accept/reject decisions. **Rebuttal:** This collapses helpful assistance into unaccountable sovereignty. Our architecture advocates for a *control plane*, not a synthetic judge. We must bound the machine's authority to evidence construction, literature positioning, and routing, keeping legitimacy, policy calibration, and final accountability strictly in human hands.

View 5: Fused certification and curation.

This view insists that maintaining the prestige of the field requires conference acceptance to simultaneously mean technical validation and high-visibility presentation. **Rebuttal:** This conflates establishing the scientific record with allocating community attention. Rolling venues (e.g., TMLR, ARR) already demonstrate that these functions can be decoupled. Once reader attention is recognized as the ultimate scarcity, conferences should certify broadly for technical soundness, but curate narrowly for visibility.

View 6: Post-conference publication.

This argument treats the centralized conference model itself as obsolete, advocating a shift entirely to decentralized preprints (arXiv) and overlay journals. **Rebuttal:** While decentralized models remove the submission bottleneck, they drastically exacerbate the discovery bottleneck. Conferences remain vital as curators of scarce human attention and synthesizers of field-level agendas. They simply should not remain the sole, choked gateway through which all technically valid work must pass.

9. Research Agenda

The position outlined in this paper is normative, but it demands a concrete, falsifiable empirical agenda. Appendix H supports this agenda by mapping out relevant literature (Section H.1), systems (Section H.2), and implementation hooks (Section H.3). To operationalize the autonomous review pipeline, the community must address challenges across three distinct tracks:

1. Granular evaluation frameworks.

Modular benchmarks: We must move beyond monolithic "AI reviewer" datasets that optimize for matching a human's final accept/reject score. The field needs specialized benchmarks for bounded evidentiary tasks: claim extraction fidelity, literature retrieval relevance, artifact execution success, and accurate cluster detection. **Evaluating typed certification:** Researchers should test whether multidimensional, typed certificates (e.g., separating policy compliance, empirical reproducibility, and theoretical novelty) provide better downstream utility to readers and meta-researchers than opaque scalar scores.

2. Institutional policy and governance.

Calibrated escalation policies: Formal evaluations must compare deterministic risk thresholds, learned routing policies, and human-in-the-loop triage to safely route contested, anomalous, or high-impact claims. **Governance of attention:** As human roles shift from routine screeners to escalation and certification specialists, empirical work must design new incentive structures, audit automation bias, and evaluate formalized appeal channels.

3. Systems and deployment.

Epistemic impact of reader interfaces: If conferences shift toward curation, the user interface becomes an epistemic tool. We must measure how AI-generated claim cards, cluster pages, and contrastive summaries alter citation spread, code adoption, and reading behavior compared to traditional flat proceedings. **Live operational pilots:** The decisive experiments must be in production. Venues should launch bounded, continuous-review pilots. These pilots must publicly track turnaround latency,

 doi:10.20944/preprints202604.0797.v1

queue lengths by risk tier, false-escalation rates, human workload relief, and the size and composition of the human committee they still require in order to move this architecture from theory to practice.

10. Conclusions

The submission explosion driven by AI-accelerated research is not a pathology to be suppressed, but a structural shift demanding a scalable response. By embracing autonomous review pipelines as a machine-first, human-governed control plane, conferences can systematically absorb this abundance without compromising rigor. This architecture accelerates critical feedback and automates routine evidentiary checks, ultimately redirecting scarce human attention away from basic screening and toward high-stakes escalation, continuous certification, and meaningful scientific curation. Just as importantly, it offers a path toward a sustainable review commons that does not depend on perpetual committee enlargement to keep seasonal review alive.

Acknowledgments: This research is supported by the RIE2025 Industry Alignment Fund (Award I2301E0026) and the Alibaba–NTU Global e-Sustainability CorpLab.

Appendix A. Design Layers

Table A1 outlines the primary users and typical outputs across different design surfaces of the review pipeline.

Table A1. Autonomous review pipelines should serve authors, venues, and readers, not only reviewers.

Surface	Primary users	Typical outputs	Main benefit
Author-side preflight	Authors	Clarity checks, missing-comparison warnings, artifact prompts, and claim-level feedback.	Improves papers before costly human review.
Venue-side control plane	Chairs, reviewers, editors	Routing signals, cluster detection, evidence reports, and escalation triggers.	Reduces routine human load and improves consistency.
Reader-side discovery	Readers, meta-researchers, curators	Claim cards, cluster pages, topic maps, and comparative summaries.	Makes large literatures more navigable.

Table A2. Why submission explosion need not lower the quality bar. The claim is not that machines should decide legitimacy alone, but that a machine-first control plane can make minimum rigor checks more universal and auditable.

Rigor dimension	Human-first seasonal review	Pipeline-first certification
Coverage of minimum checks	Many routine checks are reviewer-dependent and uneven under heavy load.	Every submission can receive the same policy, disclosure, overlap, and artifact-availability checks.
Evidence completeness	Claim reconstruction, related-work positioning, and artifact probing vary sharply with reviewer time and expertise.	Claim cards, retrieval-based literature positioning, and bounded artifact probes can be required before human certification.
Boundary-case scrutiny	Expert time is diluted by routine first-pass screening.	Human depth is concentrated on uncertain, high-impact, disputed, or appealed cases.
Decision semantics	One scalar verdict often hides which dimension actually failed.	Typed certificates can record policy, evidence, artifact, and escalation status separately.
Auditability	Failure modes are dispersed across private reviews and hard to diagnose later.	Module outputs, overrides, and escalation thresholds can be logged, audited, and recalibrated.

 doi:10.20944/preprints202604.0797.v1

Table A3. Division of labor in a pipeline-first venue.

Layer	Dominant question	Autonomous review pipeline responsibility	Human responsibility
Intake	Is the submission admissible?	Format, anonymity, disclosure, overlap, and integrity screening.	Appeals on flags, policy exceptions, and sanctions.
Evidence build	What does the paper claim and what evidence surrounds it?	Claim extraction, literature positioning, artifact probing, and cluster formation.	Inspect high-risk or high-impact claims and request clarifications.
Escalation	Where is scarce human depth most valuable?	Risk scoring, uncertainty estimates, portfolio context, and disagreement signals.	Add review depth, adjudicate edge cases, and redirect resources.
Certification	What enters the technical record?	Structured report with typed evidence and uncertainty notes.	Accountable final judgment on escalated or borderline cases.
Curation	What gets scarce visibility?	Reader summaries, cluster pages, topic maps, and candidate highlight lists.	Talks, posters, awards, synthesis sessions, and agenda setting.

Appendix B. Figure 2 Data, Sources, and Derived Ratios

Figure 1 now combines a long-run submission trend, a conservative reviewer-demand floor, and an absolute committee-recruitment series. We use the multiplier 3 in the reviewer panel because the three venues publicly target or report at least three reviews per paper; fourth-reviewer assignments, emergency backfills, second-round reviewing, and discussion-phase follow-on work only increase the true burden [6,20–22,62]. A 3.5-review or 4-review assumption would simply scale every reviewer-load value by 1.1667 or 1.3333, so we choose 3 to avoid overstating the case. Table A4 records the 2016–2025 submission histories used in the top panel. Table A5 records the reviewer, handling-chair, senior-layer, and program-chair headcounts we collected beyond 2023. Table A6 then derives the summed committee-recruitment proxy used in the lower-right main panel. Table A7 contains the reviewer-load lower-bound series used in the revised main figure, while Tables A8–A10 report the appendix handling-layer, senior-layer, and governance-layer ratios. Figures A1 and A2 provide the corresponding role-specific absolute and ratio decompositions.

Table A4. Main-track submission histories used in the top panel of Figure 1. Sources: official ICLR fact sheets for 2021–2025, supplemented by historical OpenAccept series for earlier ICLR years and for the ICML and NeurIPS historical trajectories; when an official recent count differed slightly from a community-maintained tracker, we use the official venue count [6,7,62–68].

Venue	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2016→2025
ICLR	265	507	981	1,591	2,594	3,014	3,391	4,938	7,262	11,603	43.79x
ICML	1,320	1,676	2,473	3,424	4,990	5,513	5,630	6,538	9,473	12,107	9.17x
NeurIPS	2,406	3,240	4,856	6,743	9,467	9,122	10,411	12,343	15,671	21,575	8.97x

 doi:10.20944/preprints202604.0797.v1

Table A5. Historical headcounts collected for the layers most relevant to the revised Figure 1. ICML uses *meta-reviewers* and *senior meta-reviewers*, which we harmonize with AC-like and SAC-like layers for cross-venue comparison. For NeurIPS 2022, the committee counts are paired with the official public program-stats snapshot reporting 9,634 full submissions. When an official page listed names but not an aggregate total, we report the exact count of named role entries on that official page. '—' indicates that we did not robustly recover a directly comparable official archival count for that year. Sources: official archived reviewer pages, fact sheets, committee pages, and review-process reports [6,7,62–65,69–77,77–98].

Venue	Layer	2021	2022	2023	2024	2025
ICLR	Reviewer pool	4,072	5,507	5,734	8,950	18,325
ICLR	Handling chair (AC)	450	394	413	624	823
ICLR	Senior synthesis (SAC)	—	20	40	60	71
ICLR	Program chairs	3	3	4	4	4
ICML	Reviewer pool	4,807	7,403	6,035	7,474	10,943
ICML	Handling chair (meta-reviewer)	342	281	504	492	1,161
ICML	Senior synthesis (senior meta-reviewer)	52	53	47	97	155
ICML	Program chairs	2	3	3	4	4
NeurIPS	Reviewer pool	—	10,406	12,974	13,640	20,518
NeurIPS	Handling chair (AC)	—	742	968	1,393	1,663
NeurIPS	Senior synthesis (SAC)	—	82	98	195	199
NeurIPS	Program chairs	—	4	4	4	4

Table A6. Derived committee-recruitment proxy used in the lower-right panel of Figure 1, computed as the sum of the reported reviewer, handling-chair, senior-synthesis, and program-chair headcounts. This should be read as a sum of named role headcounts, not as a deduplicated count of unique individuals. It therefore serves as a venue-level proxy for how much human committee recruitment each cycle requires.

Venue	2021	2022	2023	2024	2025
ICLR	4,525	5,924	6,191	9,638	19,223
ICML	5,203	7,740	6,589	8,067	12,263
NeurIPS	—	11,234	14,044	15,232	22,384

Table A7. Conservative lower-bound reviewer demand used in the lower-left panel of Figure 1, computed as $3 \times$ submissions/reviewers. We use 3 because it is the common public review floor across the major venues. Where venues seek fourth reviewers, second-round reviewers, or emergency backfills, the realized human burden is higher. A 3.5-review or 4-review sensitivity would multiply every entry in this table by 1.1667 or 1.3333, respectively.

Venue	2021	2022	2023	2024	2025
ICLR	2.22	1.85	2.58	2.43	1.90
ICML	3.44	2.28	3.25	3.80	3.32
NeurIPS	—	2.78	2.85	3.45	3.15

 doi:10.20944/preprints202604.0797.v1

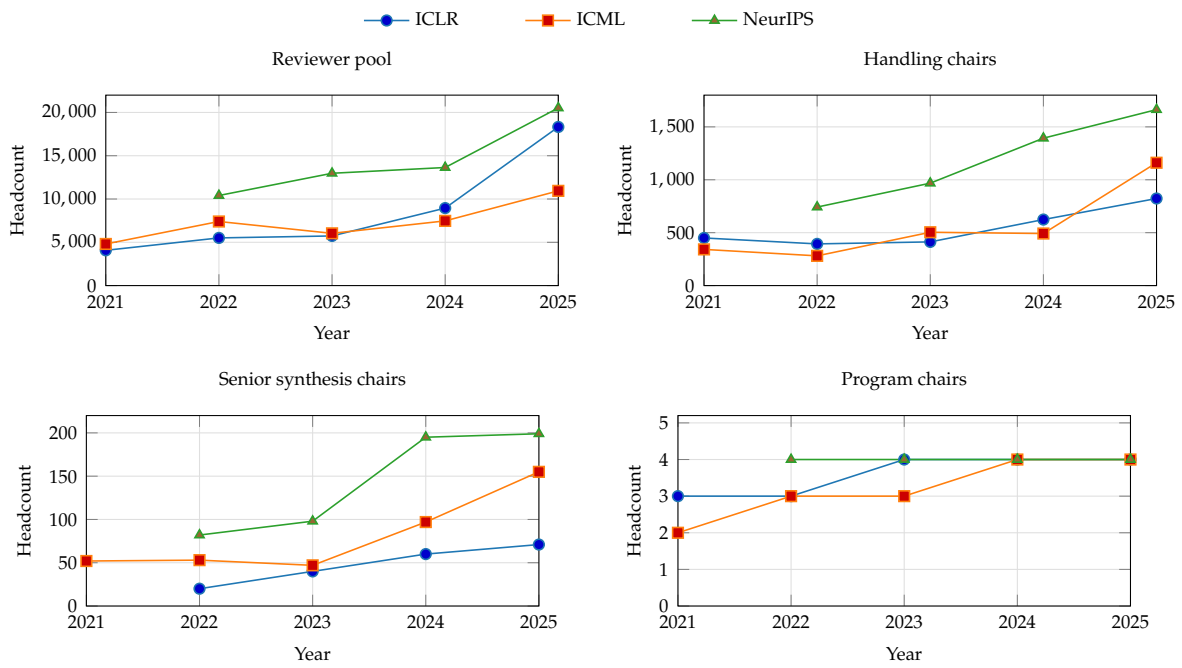

Figure A1. Role-specific absolute headcounts behind the sustainability reading of Figure 1. The same color/marker mapping is used throughout: ICLR (blue circles), ICML (orange squares), and NeurIPS (green triangles). Reviewers scale most elastically, while the handling, senior-synthesis, and governance layers remain much thinner. ICML uses the terms *meta-reviewer* and *senior meta-reviewer*; we plot them here as AC-like and SAC-like layers for cross-venue comparison. Missing values are left blank rather than inferred.

Table A8. Submissions divided by handling-chair headcount, reported for appendix reference and plotted in Figure A2. For ICML, the handling layer is the meta-reviewer layer. These ratios are informative, but committee expansions can reset them sharply from one year to the next.

Venue	2021	2022	2023	2024	2025
ICLR	6.70	8.61	11.96	11.64	14.10
ICML	16.12	20.04	12.97	19.25	10.43
NeurIPS	—	12.98	12.75	11.25	12.97

Table A9. Historical senior-synthesis-layer ratios collected for completeness: ICLR SACs, ICML senior meta-reviewers, and NeurIPS SACs. These counts are useful context and are plotted in Figure A2, but we do not use them as the main visual signal because the role was introduced, renamed, or publicly reported at different times across venues.

Venue	2021	2022	2023	2024	2025
ICLR	—	169.55	123.45	121.03	163.42
ICML	106.02	106.23	139.11	97.66	78.11
NeurIPS	—	117.49	125.95	80.36	108.42

 doi:10.20944/preprints202604.0797.v1

Table A10. Submissions divided by program-chair headcount, reported for appendix completeness. These are institutional governance-load proxies rather than direct assignment counts.

Venue	2021	2022	2023	2024	2025
ICLR	1,004.67	1,130.33	1,234.50	1,815.50	2,900.75
ICML	2,756.50	1,876.67	2,179.33	2,368.25	3,026.75
NeurIPS	—	2,408.50	3,085.75	3,917.75	5,393.75

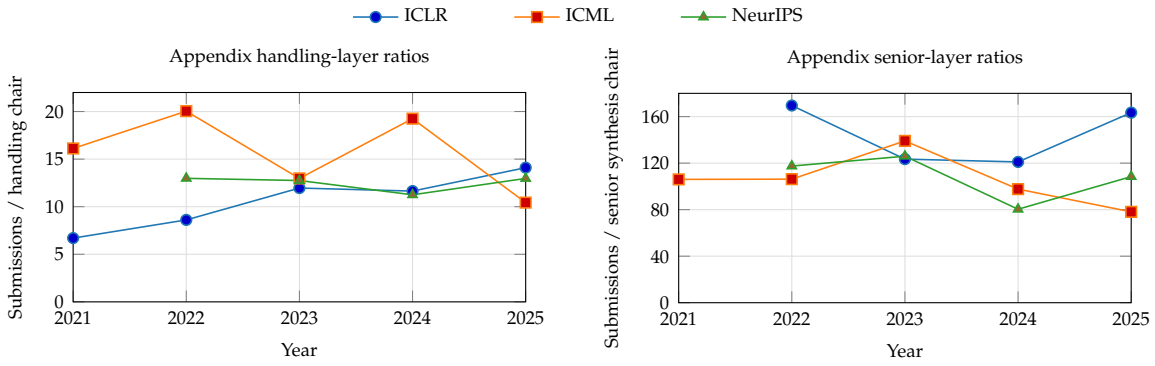

Figure A2. Supplementary layer-specific ratios beyond the main Figure 1. Left: submissions per handling chair (ICLR ACs, ICML meta-reviewers, NeurIPS ACs). Right: submissions per senior synthesis chair (ICLR SACs, ICML senior meta-reviewers, NeurIPS SACs). These layers are informative but less clean as primary cross-venue signals because role definitions changed over time and staffing surges can temporarily reset their ratios.

Appendix C. Extended Operational Sketch

A scalable pipeline-first venue can be described as a queueing and escalation system rather than as a uniform review ritual. Suppose N submissions enter the venue over some period, and let p_e be the fraction that require meaningful human escalation after the automated first pass. In a traditional human-first design, human effort scales roughly with N because nearly every paper must be independently screened, contextualized, and scored by several people. In a pipeline-first design, the goal is to make the expensive human term scale with $p_e N$, where p_e is kept small by better evidence gathering, clustering, and routing, without concealing uncertainty or denying appeal rights.

One concrete lifecycle is as follows.

1. **Continuous intake.** Authors submit at any time. The system checks anonymity, formatting, required disclosures, overlap, and other basic policy conditions.
2. **Evidence construction.** The system extracts major claims, clusters nearby papers, retrieves relevant literature, and probes released artifacts where feasible.
3. **Structured report.** Each submission receives a pipeline report with claim cards, uncertainty notes, and routing signals. Authors may be asked for clarifications or artifact fixes before any human escalation.
4. **Selective escalation.** Humans are assigned mainly to uncertain, high-impact, contested, or appeal-triggered cases.
5. **Rolling certification.** The venue records whether the paper is technically certified, provisionally certified with uncertainty notes, or unresolved pending further review.
6. **Conference curation.** Periodic conference decisions allocate scarce visibility among certified papers through talks, posters, awards, and synthesis sessions.

This lifecycle is compatible with portfolio-aware handling without making portfolios the only organizing principle. If the control plane detects a cluster of related submissions, it can request a compact portfolio clarification from the authors or route the cluster to one escalation team. The key point is that correlation is surfaced when needed, instead of being discovered manually and inconsistently.

 doi:10.20944/preprints202604.0797.v1

Appendix D. Schema for the Submission Evidence Graph

To operationalize the evidence build layer discussed in Section 3, the autonomous review pipeline must extract entities and relations into a formal structure [99,100]. Table A11 outlines a minimal working schema for a submission-level knowledge graph. This structure allows the control plane to query relationships across thousands of concurrent submissions.

Table A11. A proposed knowledge graph schema for representing submission evidence. By transforming PDFs into structured graphs, the pipeline enables automated literature positioning and precise cluster routing.

Node / Edge Type	Example Entities / Relations	Function in the Review Pipeline
Claim Node	"Method X reduces inference latency by 20%"	Central unit of evaluation; requires human or automated verification.
Method Node	"Semantic Per-Pair DPO", "LoRA"	Identifies the core algorithmic or architectural contributions.
Dataset Node	"ImageNet", "Custom Web Scrape"	Triggers artifact checks and flags potential data contamination.
Metric Node	"Accuracy", "Throughput (tokens/sec)"	Standardizes comparisons across similar papers in the same cluster.
`USES_METHOD`	Submission → Method Node	Links a paper to its underlying techniques, enabling portfolio clustering.
`CLAIMS_OVER`	Method Node → Baseline Node	Explicitly maps the authors' stated improvements against prior art.
`EVALUATES_ON`	Method Node → Dataset Node	Maps the empirical boundary of the paper's claims.

Appendix E. Big-Three Conference Scale Statistics

Table A12 consolidates the main-track counts used in Figure 1. The point of this table is not to imply perfect apples-to-apples comparability across venues; the venues differ in what they call valid submissions, how they report review roles, and whether track-level statistics are published in fact sheets or later reflections. The point is that every major venue is already staffing review at industrial scale [6,7,62,65,67,72–74,82,83]. Table A6 then makes the sustainability implication explicit by summing the reported role headcounts into a single committee-recruitment proxy.

Because the program-chair layer is tiny relative to reviewer, AC, and SAC staffing, we record recent counts separately in Table A13; the longer historical series used for Figure 1 appears in Table A5. The contrast matters substantively: submission growth is pressuring not only the review workforce but also the very small governance layer that must supervise policy, appeals, and calibration. This is why the paper's main sustainability signal is the total committee-recruitment series in Table A6, with the thinner governance layer retained as appendix support.

 doi:10.20944/preprints202604.0797.v1

Table A12. Main-track scale statistics used in Figure 1. Program-chair counts are summarized separately in Table A13 because they sit at a much smaller governance scale than the review-layer staffing counts in this table.

Venue	Year	Submitted	Accepted	Acc. rate	Reviewers	ACs	SACs	Source scope
ICLR	2023	4,938	1,574	31.88%	5,734	413	40	Official fact sheet (research track).
ICLR	2024	7,262	2,260	31.12%	8,950	624	60	Official fact sheet (research track).
ICLR	2025	11,603	3,704	31.92%	18,325	823	71	Official fact sheet (research track).
ICML	2023	6,538	1,827	27.94%	6,035	504	47	OpenAccept main-track history for submissions/accepts; official reviewers page for committee counts.
ICML	2024	9,473	2,609	27.54%	7,474	492	97	OpenAccept main-track history plus official fact sheet for committee counts.
ICML	2025	12,107	3,260	26.93%	10,943	1,161	155	OpenAccept main-track history plus official fact sheet for committee counts.
NeurIPS	2023	12,343	3,218	26.07%	12,974	968	98	Official fact sheet (main track only).
NeurIPS	2024	15,671	4,037	25.76%	13,640	1,393	195	Official fact sheet (main track only).
NeurIPS	2025	21,575	5,290	24.52%	20,518	1,663	199	Official main-track review reflection.

Table A13. Recent program-chair counts from official organizing-committee pages. The longer historical series used for Figure 1 appears in Table A5. The top governance layer stays nearly flat relative to submission growth.

Venue	2023 PCs	2024 PCs	2025 PCs
ICLR	4	4	4
ICML	3	4	4
NeurIPS	4	4	4

Appendix F. Related Work

This paper sits at the intersection of peer-review reform, meta-science on review uncertainty, and recent work on AI-assisted reviewing. First, a growing literature documents that peer review is noisy, inconsistent, difficult to calibrate, and often too slow for modern scientific communication [35–38,40]. Second, reform proposals inside the ML community have emphasized reviewer incentives, author feedback, and mechanisms that exploit author-side information such as owner-assisted scoring or self-rankings [45–47,101–103]. Third, work on partial randomization in adjacent allocation settings is a useful reminder that overloaded evaluation systems must decide what to do near noisy boundaries rather than pretending the ordering is perfectly stable [104,105]. Fourth, recent AI-reviewing work and surveys ask what parts of review can be standardized, simulated, or evidence-grounded by machine systems [32,33,44,58,61,106–109]. The call for modular review benchmarks is also consistent with a broader benchmark-design trend toward specialized text and multimodal evaluation suites rather than

 doi:10.20944/preprints202604.0797.v1

one aggregate score [110,111]. Our position differs in emphasis from all of these strands. We do not argue only for better incentives, better score aggregation, or better AI assistants inside the existing workflow. We argue that the review system's center of gravity should shift to a machine-run control plane plus human-governed escalation. Tables A16, A17, and A18 then situate that position in the existing research and systems landscape.

Appendix G. Existing Pilots and Why Broad Deployment Remains Narrow

The field already has bounded precedents for machine assistance in peer review, but they still stop short of a broadly implemented conference policy. Table A14 summarizes the most relevant conference-side examples. The common pattern is revealing: what venues are willing to deploy today are tightly scoped systems that either coach humans, assist authors, or provide one bounded AI signal under strict human authority.

Table A14. Representative precedents for machine assistance in peer review. The key gap is not the absence of building blocks, but the absence of a shared conference norm that makes a machine-first, human-governed first pass official, bounded, and auditable.

Venue / example	Workflow location	What the machine does	Human authority retained	Why this does not yet amount to broad conference adoption
ICLR 2025 reviewer-feedback agent [9,10,108]	Reviewer side	Flags vague, redundant, or unprofessional reviews and offers quality-improving feedback.	Humans still write the actual reviews and make all paper judgments.	It improves reviewer behavior, but it does not become the venue's first-pass control plane for every submission.
ICML 2026 PAT [13,14]	Author side	Provides manuscript feedback, prompts, and diagnostics before or around submission.	Humans still conduct the formal review process.	It improves drafts before review, but it is not a conference-wide first-pass review pipeline applied to all submissions at decision time.
AAAI-26 AI-assisted peer-review pilot [15]	Venue side	Adds one AI review and an AI-generated summary under explicit pilot rules.	AI gives no score or recommendation; all decisions remain human.	The pilot is deliberately narrow and cautious, reflecting unresolved governance questions about legitimacy, privacy, and robustness.
Rolling pathways such as TMLR, ARR, and J2C [42,48–50]	Certification timing	Decouple parts of certification and presentation timing; support more continuous review flows.	Human editors and reviewers remain in charge of certification.	These pathways relax deadline pressure, but they are not themselves a unified machine-first conference review architecture.
Representative research prototypes such as SEA, ReviewerGPT, quality-checkers, and FactReview [32,33,44,58,61,106]	Experimental systems	Generate review-style feedback, find critical problems, structure debate, or verify claims against evidence.	Used as experimental tools rather than institutional arbiters.	The technical direction is promising, but conferences still lack shared governance for how these systems should be integrated, disclosed, and audited at scale.

The reason broader adoption remains narrow is therefore not only that current models are imperfect. It is that conferences still face unresolved institutional gaps. Table A15 turns those gaps into a policy map. This is the heart of the position argument: if the blockers are institutional, then the response should also be institutional. Conferences need explicit governance rather than a quiet equilibrium of hidden usage, fragmented rules, and ad hoc pilots.

 doi:10.20944/preprints202604.0797.v1

Table A15. Why broad deployment remains narrow even though conference pilots and research prototypes already exist. The blockers are institutional as much as technical, which is why the paper argues for policy-level endorsement and governance rather than for a stand-alone tool.

Institutional gap	Why it blocks broad deployment	Conference-level response
Confidentiality and privacy	Review involves unpublished manuscripts, sensitive artifacts, and reputation-sensitive judgments. Venues need to know where content goes, what models retain, and what third-party processing is acceptable.	Prefer venue-operated tooling or tightly governed vendors, require explicit data-handling terms, and separate allowed bounded assistance from prohibited external processing.
Legitimacy and accountability	Once an automated signal materially affects triage or escalation, conferences must answer who is responsible when the signal is wrong, harmful, or biased.	Define bounded machine authority in writing, keep final responsibility for borderline legitimacy, sanctions, and appeals in named human roles, and publish override pathways.
Policy fragmentation	Venues currently permit, prohibit, or partially regulate AI usage in different ways, which makes community trust and cross-venue norms unstable.	Publish task-level policies that specify what is machine-assisted, what must be disclosed, what remains human-only, and what logs are retained for audit.
Adversarial manipulation	Hidden prompts, prompt injection, and optimization against system heuristics can distort triage or reviewing once authors know the machine layer matters.	Red-team the pipeline, rotate checks, use multiple signals rather than one brittle model output, and maintain author appeal channels for questionable flags or cluster assignments.
Unequal access to tooling	If some authors can privately use strong paper assistants while others cannot, hidden adoption amplifies inequality rather than reducing communal workload fairly.	Provide conference-operated assistance to all participants, especially on the author side and first-pass review side, so the official pipeline narrows asymmetry instead of widening it.
Weak evaluation and audit	Without logging, post-cycle review, and public dashboards, conference communities cannot tell whether a deployed system actually reduces effort, raises quality, or simply moves failure around.	Treat deployment as an auditable policy intervention: log module outputs, escalation triggers, overrides, appeals, and publish post-cycle dashboards on burden, error, and reader usefulness.

Appendix H. Systems, Platforms, and Repositories for Autonomous Review Pipelines

The ecosystem needed for an autonomous review pipeline is not hypothetical. It is emerging across three layers: research on automated or AI-assisted reviewing, venue infrastructure for rolling or machine-mediated workflows, and open systems for parsing, metadata retrieval, and matching. Table A16 highlights representative papers, while Table A17 and Table A18 collect concrete systems, platforms, and implementation hooks.

 doi:10.20944/preprints202604.0797.v1

Appendix H.1. Representative Research Literature

Table A16. Representative literature relevant to autonomous review pipelines. The papers differ in what they automate; together they illustrate a design space broader than "LLM writes the final review."

Work	Primary contribution	Why it matters for this paper
SEA [61]	Generates review-style feedback with standardization, evaluation, and analysis modules.	Shows early end-to-end automation of paper feedback, while stopping short of venue legitimacy.
Quality-checker framing [58]	Recasts LLMs as manuscript quality checkers rather than substitute reviewers.	Closely matches our argument that machines should gather evidence and flag problems.
Live review-feedback study [108]	Studies the ICLR 2025 reviewer-feedback deployment in a real review process.	Provides direct evidence that AI can assist review while creating governance tensions.
ReViewGraph [44]	Models automated paper reviewing through structured reviewer–author debate graphs.	Suggests richer representations for disagreement and escalation.
FactReview [33]	Combines claim extraction, literature positioning, and execution-based claim verification.	Exemplifies the evidence-grounded, claim-level style of autonomous review pipeline advocated here.
LLM-ASPR survey [109]	Synthesizes tasks, datasets, systems, and failure modes for automated scholarly paper review.	Helps position this paper's argument in the broader technical and governance landscape.

Appendix H.2. Representative Systems, Platforms, and Repositories

Table A17. Selected systems and repositories that could form part of an autonomous review pipeline stack. This list is representative rather than exhaustive.

Category	Examples	Relevance
Venue workflow infrastructure	OpenReview platform and API [59,112,113]	Provides an existing open review platform, API layer, and venue workflow primitives for machine-mediated review.
Reviewer matching infrastructure	OpenReview expertise models [60]	Supplies paper–reviewer affinity tools that could feed escalation and expert routing.
Rolling or continuous review venues	TMLR, ACL Rolling Review, Journal-to-Conference [42,48–50]	Demonstrate that review, certification, and conference presentation can already be partially decoupled.
Document parsing and structuring	GROBID [114]	Converts scholarly PDFs into structured representations useful for claim extraction and policy checks.
Scholarly metadata and retrieval	OpenAlex and Semantic Scholar APIs [115,116]	Enable literature positioning, citation retrieval, cluster building, and reader-facing comparisons.
Venue-side AI assistance	ICLR review-feedback agent, ICML PAT [9,13,14]	Show that mainstream venues are already piloting machine assistance on both the reviewer side and the author side.

Appendix H.3. Implementation Hooks and Repository Map

Table A18 makes the deployment story more concrete by mapping common autonomous review pipeline functions to reusable public infrastructure.

 doi:10.20944/preprints202604.0797.v1

Table A18. Concrete repositories, APIs, and workflow components that can support an initial autonomous review pipeline deployment.

Function	Concrete resource	Typical use in an autonomous review pipeline pilot
Submission intake and state transitions	OpenReview platform plus API v2 and `openreview-py` [59,112,113]	Ingest submissions, store machine reports, attach decision metadata, and expose venue-specific workflow actions.
Reviewer or editor matching	OpenReview expertise models [60]	Compute affinity scores, seed escalation routing, and support specialist assignment once a paper leaves the automated queue.
PDF-to-structure conversion	GROBID [114]	Convert scholarly PDFs into structured sections, references, and metadata for claim extraction, overlap checks, and policy validation.
Literature retrieval and citation context	OpenAlex and Semantic Scholar APIs [115,116]	Build neighborhood graphs, retrieve likely comparisons, and generate reader-facing cluster or claim pages.
Venue-side author assistance	ICML PAT and reviewer-feedback agents [9,13,14]	Provide pre-submission diagnostics for authors and structured coaching for reviewers without delegating final legitimacy.
Rolling certification pathways	TMLR, ACL Rolling Review, and Journal-to-Conference [42,48–50]	Supply external certification channels that can feed conference curation after the machine-first control plane has done initial triage.

The main gap is therefore not whether building blocks exist, but whether the community is willing to combine them into a review system whose center of gravity is automated evidence construction plus human-governed escalation, rather than universal human-first screening.

References

1. Gottweis, J.; Weng, W.H.; Daryin, A.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* **2025**.
2. Bubeck, S.; Coester, C.; Eldan, R.; Gowers, T.; Lee, Y.T.; Lupsasca, A.; Sawhney, M.; Scherrer, R.; Sellke, M.; Spears, B.K.; et al. Early science acceleration experiments with GPT-5. *arXiv preprint arXiv:2511.16072* **2025**.
3. Hao, Q.; Xu, F.; Li, Y.; Evans, J. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature* **2026**, pp. 1–7.
4. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. The AutoResearch Moment: From Experimenter to Research Director **2026**.
5. Simon, H.A. Designing organizations for an information-rich world. *International Library of Critical Writings in Economics* **1996**, *70*, 187–202.
6. ICLR 2025. ICLR 2025 Fact Sheet. Conference Fact Sheet, 2025. Accessed March 23, 2026.
7. NeurIPS 2025 Program Committee Chairs. Reflections on the 2025 Review Process from the Program Committee Chairs. NeurIPS Blog, 2025. Accessed March 23, 2026.
8. Koniusz, P.; Chen, N.; Ghassemi, M.; Pascanu, R.; Lin, H.T.; Aroyo, L.; Locatello, F.; Palla, K. Responsible Reviewing Initiative for NeurIPS 2025. NeurIPS Blog, 2025. Accessed March 23, 2026.
9. Zou, J.; Vondrick, C.; Yu, R.; Peng, V.; Sha, F.; Garg, A. Assisting ICLR 2025 Reviewers with Feedback. ICLR Blog, 2024. Accessed March 23, 2026.
10. Thakkar, N.; Yuksekgonul, M.; Silberg, J.; Garg, A.; Peng, N.; Sha, F.; Yu, R.; Vondrick, C.; Zou, J. Can LLM feedback enhance review quality? A randomized study of 20k reviews at ICLR 2025. *arXiv preprint arXiv:2504.09737* **2025**.
11. Zou, J.; Thakkar, N.; Vondrick, C.; Yu, R.; Peng, V.; Sha, F.; Garg, A. Leveraging LLM Feedback to Enhance Review Quality. ICLR Blog, 2025. Accessed March 23, 2026.

 doi:10.20944/preprints202604.0797.v1

12. Vondrick, C. Extended Partnership Pilot with TMLR for ICLR 2025. ICLR Blog, 2024. Accessed March 23, 2026.
13. Jayaram, R.; Cohen-Addad, V.; Agarwal, A.; Dudik, M.; Li, S.; Jaggi, M. ICML Experimental Program using Google's Paper Assistant Tool (PAT). ICML Blog, 2026. Accessed March 23, 2026.
14. Kamath, G.; Jayaram, R.; Cohen-Addad, V.; Agarwal, A.; Dudik, M.; Li, S.; Jaggi, M. Retrospective on PAT x ICML 2026 AI Paper Assistant Program. ICML Blog, 2026. Accessed April 7, 2026.
15. AAAI-26. FAQ for the AI-Assisted Peer-Review Process Pilot Program. Conference FAQ / PDF, 2025. Accessed April 8, 2026.
16. Naddaf, M. AI is transforming peer review—and many scientists are worried. *Nature* **2025**, *639*, 852–854.
17. Naddaf, M. More than half of researchers now use AI for peer review—often against guidance. *Nature* **2026**, *649*, 273–274.
18. Gibney, E. Scientists hide messages in papers to game AI peer review. *Nature* **2025**, *643*, 887–888.
19. Gibney, E. Major conference catches illicit AI use-and rejects hundreds of papers. *Nature* **2026**.
20. ICML. ICML 2025 Peer Review FAQ. Conference Website, 2025. Accessed April 9, 2026.
21. ICLR 2025. ICLR 2025 SAC Guide. Conference Website, 2025. Accessed April 9, 2026.
22. NeurIPS 2024. NeurIPS 2024 SAC Guidelines. Conference Website, 2024. Accessed April 9, 2026.
23. Brockmeyer, B. Mentorship Program for New Reviewers at ICLR 2022. ICLR Blog, 2022. Accessed April 9, 2026.
24. Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics* **2012**, *90*, 891–904.
25. Mlinarić, A.; Horvat, M.; Šupak Smolčić, V. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia medica* **2017**, *27*, 447–452.
26. ICLR. ICLR 2026 Reviewer Guide. Conference Website, 2026. Accessed March 23, 2026.
27. NeurIPS 2026 Communication Chairs. What's New for the Position Paper Track at NeurIPS 2026. NeurIPS Blog, 2026. Accessed April 7, 2026.
28. NeurIPS 2026 Position Paper Track. NeurIPS 2026 Call for Position Papers. Conference Website, 2026. Accessed April 7, 2026.
29. NeurIPS 2025 Position Paper Track. Call for Position Papers 2025. Conference Website, 2025. Accessed March 23, 2026.
30. NeurIPS 2025 Position Paper Track. NeurIPS 2025 Position Paper Track FAQ. Conference Website, 2025. Accessed March 23, 2026.
31. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Harness Engineering for Language Agents: The Harness Layer as Control, Agency, and Runtime **2026**.
32. Liu, R.; Shah, N.B. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622* **2023**.
33. Xu, H.; Yue, L.; Ouyang, C.; Zheng, L.; Pan, S.; Di, S.; Zhang, M.L. FactReview: Evidence-Grounded Reviews with Literature Positioning and Execution-Based Claim Verification. *arXiv preprint arXiv:2604.04074* **2026**.
34. Wei, Q.; Holt, S.; Yang, J.; Wulfmeier, M.; van der Schaar, M. The ai imperative: Scaling high-quality peer review in machine learning. *arXiv preprint arXiv:2506.08134* **2025**.
35. Tran, D.; Valtchanov, A.; Ganapathy, K.; Feng, R.; Slud, E.; Goldblum, M.; Goldstein, T. Analyzing the Machine Learning Conference Review Process. *arXiv preprint arXiv:2011.12919* **2020**.
36. Cortes, C.; Lawrence, N.D. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774* **2021**.
37. NeurIPS 2021 Program Chairs. The NeurIPS 2021 Consistency Experiment. NeurIPS Blog, 2021. Accessed March 23, 2026.
38. Barnett, A.; Allen, L.; Aldcroft, A.; Lash, T.L.; McCreanor, V. Examining uncertainty in journal peer reviewers' recommendations: a cross-sectional study. *Royal Society Open Science* **2024**, *11*.
39. Goldberg, A.; Stelmakh, I.; Cho, K.; Oh, A.; Agarwal, A.; Belgrave, D.; Shah, N.B. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *PloS one* **2025**, *20*, e0320444.
40. Aczel, B.; Barwich, A.S.; Diekman, A.B.; Fishbach, A.; Goldstone, R.L.; Gomez, P.; Gundersen, O.E.; von Hippel, P.T.; Holcombe, A.O.; Lewandowsky, S.; et al. The present and future of peer review: Ideas, interventions, and evidence. *Proceedings of the National Academy of Sciences* **2025**, *122*, e2401232121.
41. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Human-AI productivity claims should be reported as time-to-acceptance under explicit acceptance tests, 2026.

 doi:10.20944/preprints202604.0797.v1

42. Transactions on Machine Learning Research. Acceptance Criteria. Journal Website, 2026. Accessed March 23, 2026.
43. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. OpenClaw as Language Infrastructure: A Case-Centered Survey of a Public Agent Ecosystem in the Wild **2026**.
44. Li, S.; Fan, L.; Lin, Y.; Li, Z.; Wei, X.; Ni, S.; Alinejad-Rokny, H.; Yang, M. Automatic paper reviewing with heterogeneous graph reasoning over llm-simulated reviewer-author debates. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 31717–31725.
45. Kim, J.; Lee, Y.; Lee, S. Position: The AI Conference Peer Review Crisis Demands Author Feedback and Reviewer Rewards. OpenReview, 2025. Accessed March 23, 2026.
46. Su, W. You are the best reviewer of your own papers: An owner-assisted scoring mechanism. *Advances in Neural Information Processing Systems* **2021**, *34*, 27929–27939.
47. Su, B.; Zhang, J.; Collina, N.; Yan, Y.; Li, D.; Cho, K.; Fan, J.; Roth, A.; Su, W. The ICML 2023 ranking experiment: Examining author self-assessment in ML/AI peer review. *Journal of the American Statistical Association* **2025**, pp. 1–12.
48. Transactions on Machine Learning Research. Transactions on Machine Learning Research. Journal Website, 2026. Accessed March 23, 2026.
49. ACL Rolling Review. ACL Rolling Review. Platform Website, 2026. Accessed March 23, 2026.
50. NeurIPS/ICLR/ICML Journal-to-Conference Track Oversight Committee. The NeurIPS/ICLR/ICML Journal-to-Conference Track. Conference Website, 2026. Accessed March 23, 2026.
51. NeurIPS 2026 Communication Chairs. Refining the Review Cycle: NeurIPS 2026 Area Chair Pilot. NeurIPS Blog, 2026. Accessed March 23, 2026.
52. NeurIPS 2026. Main Track Handbook 2026. Conference Website, 2026. Accessed March 23, 2026.
53. NeurIPS 2026 Communication Chairs. Introducing the Evaluations & Datasets Track at NeurIPS 2026. NeurIPS Blog, 2026. Accessed March 23, 2026.
54. ICLR. ICLR 2026 Author Guide. Conference Website, 2026. Accessed March 23, 2026.
55. ICLR 2026 Program Chairs. ICLR 2026 Response to LLM-Generated Papers and Reviews. ICLR Blog, 2025. Accessed March 23, 2026.
56. ICML. ICML 2026 Peer Review FAQ. Conference Website, 2026. Accessed March 23, 2026.
57. Agarwal, A.; Dudik, M.; Li, S.; Jaggi, M.; Shah, N.B.; Gorman, K.; Kamath, G. On Violations of LLM Review Policies. ICML Blog, 2026. Accessed March 23, 2026.
58. Zhang, T.M.; Abernethy, N.F. Reviewing scientific papers for critical problems with reasoning llms: Baseline approaches and automatic evaluation. *arXiv preprint arXiv:2505.23824* **2025**.
59. OpenReview. About OpenReview. Website, 2026. Accessed April 7, 2026.
60. OpenReview. openreview-expertise: Expertise Modeling for the OpenReview Matching System. GitHub Repository, 2026. Accessed April 7, 2026.
61. Yu, J.; Ding, Z.; Tan, J.; Luo, K.; Weng, Z.; Gong, C.; Zeng, L.; Cui, R.; Han, C.; Sun, Q.; et al. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 10164–10184.
62. ICLR 2024. ICLR 2024 Fact Sheet. Conference Fact Sheet, 2024. Accessed April 7, 2026.
63. ICLR 2021. ICLR 2021 Fact Sheet. Conference Fact Sheet, 2021. Accessed April 9, 2026.
64. ICLR 2022. ICLR 2022 Fact Sheet. Conference Fact Sheet, 2022. Accessed April 9, 2026.
65. ICLR 2023. ICLR 2023 Fact Sheet. Conference Fact Sheet, 2023. Accessed April 7, 2026.
66. OpenAccept. ICLR Acceptance Rates and Submission Stats. Historical Statistics Website, 2026. Accessed April 8, 2026.
67. OpenAccept. ICML Acceptance Rates and Submission Stats. Historical Statistics Website, 2026. Accessed April 7, 2026.
68. OpenAccept. NeurIPS Acceptance Rates and Submission Stats. Historical Statistics Website, 2026. Accessed April 8, 2026.
69. ICLR 2021 Program Chairs. The ICLR 2021 Reviewing Process and Accepted Papers. ICLR Blog / Medium, 2021. Accessed April 9, 2026.
70. ICML 2021. ICML 2021 Reviewers. Conference Website, 2021. Accessed April 9, 2026.
71. ICML 2022. ICML 2022 Reviewers. Conference Website, 2022. Accessed April 9, 2026.
72. ICML 2023. ICML 2023 Reviewers. Conference Website, 2023. Accessed April 7, 2026.
73. ICML 2024. ICML 2024 Fact Sheet. Conference Fact Sheet, 2024. Accessed April 7, 2026.
74. ICML 2025. ICML 2025 Fact Sheet. Conference Fact Sheet, 2025. Accessed April 7, 2026.

 doi:10.20944/preprints202604.0797.v1

75. ICML 2025. ICML 2025 Program Committee. Conference Website, 2025. Accessed April 9, 2026.
76. ICML 2021. 2021 ICML Organizing Committee. Conference Website, 2021. Accessed April 9, 2026.
77. ICML 2022. 2022 ICML Organizing Committee. Conference Website, 2022. Accessed April 9, 2026.
78. ICML 2023. 2023 ICML Organizing Committee. Conference Website, 2023. Accessed April 7, 2026.
79. ICML 2024. 2024 ICML Organizing Committee. Conference Website, 2024. Accessed April 7, 2026.
80. ICML 2025. 2025 ICML Organizing Committee. Conference Website, 2025. Accessed April 7, 2026.
81. NeurIPS 2022. NeurIPS 2022 Fact Sheet. Conference Fact Sheet, 2022. Accessed April 9, 2026.
82. NeurIPS 2023. NeurIPS 2023 Fact Sheet. Conference Fact Sheet, 2023. Accessed April 7, 2026.
83. NeurIPS 2024. NeurIPS 2024 Fact Sheet. Conference Fact Sheet, 2024. Accessed April 7, 2026.
84. NeurIPS 2023. 2023 Organizing Committee. Conference Website, 2023. Accessed April 7, 2026.
85. NeurIPS 2024. 2024 Organizing Committee. Conference Website, 2024. Accessed April 7, 2026.
86. NeurIPS 2025. 2025 Organizing Committee. Conference Website, 2025. Accessed April 7, 2026.
87. ICLR 2019. ICLR 2019 Area Chairs. Conference Website, 2019. Accessed April 9, 2026.
88. ICLR 2019. ICLR 2019 Committees. Conference Website, 2019. Accessed April 9, 2026.
89. ICLR 2020 Program Chairs. #OurHatata: The Reviewing Process and Research Shaping ICLR in 2020. ICLR Blog / Medium, 2020. Accessed April 9, 2026.
90. ICLR 2020. ICLR 2020 Committees. Conference Website, 2020. Accessed April 9, 2026.
91. ICLR 2021. ICLR 2021 Committees. Conference Website, 2021. Accessed April 9, 2026.
92. ICLR 2022. ICLR 2022 Committees. Conference Website, 2022. Accessed April 9, 2026.
93. ICML 2019. ICML 2019 Area Chairs. Conference Website, 2019. Accessed April 9, 2026.
94. ICML 2020. 2020 ICML Organizing Committee. Conference Website, 2020. Accessed April 9, 2026.
95. ICML 2024. ICML 2024 Reviewers. Conference Website, 2024. Accessed April 9, 2026.
96. NeurIPS 2020 Program Chairs. What We Learned from the NeurIPS 2020 Reviewing Process. NeurIPS Blog / Medium, 2020. Accessed April 9, 2026.
97. OpenReview. OpenReview NeurIPS 2021 Summary Report. OpenReview Documentation, 2021. Accessed April 9, 2026.
98. NeurIPS 2020. 2020 Organizing Committee. Conference Website, 2020. Accessed April 9, 2026.
99. He, C.; Zhou, X.; Wang, D.; Yu, X.; Xiao, L.; Li, L.; Xu, H.; Liu, W.; Miao, C. KG4ESG: The ESG Knowledge Graph Atlas **2026**.
100. He, C.; Zhou, X.; Yu, X.; Zhang, L.; Zhang, Y.; Wu, Y.; Xiao, L.; Li, L.; Wang, D.; Xu, H.; et al. SSKG Hub: An Expert-Guided Platform for LLM-Empowered Sustainability Standards Knowledge Graphs. *arXiv preprint arXiv:2603.00669* **2026**.
101. Su, B.; Collina, N.; Wen, G.; Li, D.; Cho, K.; Fan, J.; Zhao, B.; Su, W. How to Find Fantastic AI Papers: Self-Rankings as a Powerful Predictor of Scientific Impact Beyond Peer Review. *arXiv preprint arXiv:2510.02143* **2025**.
102. Agarwal, A.; Dudik, M.; Li, S.; Jaggi, M. What's New in ICML 2026 Peer Review. ICML Blog, 2026. Accessed March 23, 2026.
103. Su, W.; Su, B. Introducing ICML 2026 Policy for Self-Ranking in Reviews. ICML Blog, 2026. Accessed March 23, 2026.
104. Davies, C.; Ingram, H. Sceptics and champions: participant insights on the use of partial randomization to allocate research culture funding. *Research Evaluation* **2025**, *34*, rvaf006.
105. Stafford, T.; Rombach, I.; Hind, D.; Mateen, B.; Woods, H.B.; Dimario, M.; Wilsdon, J. Where next for partial randomisation of research funding? The feasibility of RCTs and alternatives. *Wellcome open research* **2024**, *8*, 309.
106. Zhou, R.; Chen, L.; Yu, K. Is LLM a reliable reviewer? A comprehensive evaluation of LLM on automatic paper reviewing tasks. In Proceedings of the Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024), 2024, pp. 9340–9351.
107. Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D.Y.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D.S.; Yin, Y.; et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* **2024**, *1*, AIoa2400196.
108. Chen, S.; Zhong, S.; Brumby, D.P.; Cox, A.L. What happens when reviewers receive AI feedback in their reviews? *arXiv preprint arXiv:2602.13817* **2026**.
109. Zhuang, Z.; Chen, J.; Xu, H.; Jiang, Y.; Lin, J. Large language models for automated scholarly paper review: A survey. *Information Fusion* **2025**, *124*, 103332.

 doi:10.20944/preprints202604.0797.v1

110. He, C.; Zhou, X.; Wu, Y.; Yu, X.; Zhang, Y.; Zhang, L.; Wang, D.; Lyu, S.; Xu, H.; Xiaoqiao, W.; et al. Esgenius: Benchmarking llms on environmental, social, and governance (esg) and sustainability knowledge. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 14623–14664.
111. Zhang, L.; Zhou, X.; He, C.; Wang, D.; Wu, Y.; Xu, H.; Liu, W.; Miao, C. Mmesgbench: Pioneering multimodal understanding and complex reasoning benchmark for esg tasks. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 12829–12836.
112. OpenReview. OpenReview API v2 Documentation. Documentation, 2024. Accessed April 7, 2026.
113. OpenReview. openreview-py: Official Python Client Library for the OpenReview API. GitHub Repository, 2026. Accessed April 7, 2026.
114. GROBID. Introduction - GROBID Documentation. Documentation, 2026. Accessed April 7, 2026.
115. OpenAlex. API Overview - OpenAlex Developers. API Documentation, 2026. Accessed April 7, 2026.
116. Semantic Scholar. Semantic Scholar Academic Graph API. API Documentation, 2026. Accessed April 7, 2026.