POSITION: WANT BETTER ML REVIEWS? STOP ASKING NICELY AND START INCENTIVIZING WITH A CREDIT SYSTEM

Anonymous authors
Paper under double-blind review

ABSTRACT

With soaring submission counts, stricter reciprocity review policies, widespread adoption of platforms like OpenReview, and without the offsetting pressure of publication fees, the machine learning (ML) community has one of the largest scholarly presences among all scientific fields. And yet, almost *everyone* has *many* unpleasant things to share about their review experience. Worse, there is little public space to seriously discuss — let alone debate — what makes a review system effective or how it might be improved. In this position paper, we expand our discussion from the two core problems: *How can we reasonably limit the number of submissions?* and *How can we incentivize good and discourage bad review practices?* We first assess the strengths and shortcomings of existing attempts to address such problems. Specifically, we present five takes on some popular conference mechanisms and propose two alternative designs for improvement. Our general position is that meaningful improvement in ML peer review won't come from polite best-practice suggestions tucked into Calls for Papers or Reviewer Guidelines — it requires **enforceable yet fine-grained procedural safeguards** paired with **a currency-like credit system (what we call *OpenReview Points*)**. ML practitioners can "earn" such points by contributing good review practices, and "spend" across one or multiple major conferences to redeem different kinds of "perks" — such as complimentary registration or the right to request additional review resources.

1 INTRODUCTION

This position paper argues that peer review in machine learning (ML) is unlikely to improve through polite requests or optimistic guidance tucked into Calls for Papers or Reviewer Guidelines. Fine-grained yet enforceable procedural guardrails, combined with a spendable, across-conference credit system, are almost mandatory for a sustainable review ecosystem.

Machine learning has scaled faster than nearly any other scientific field in both volume and visibility. We now have tens of thousands of paper submissions to a single conference,[1] open-access platforms like OpenReview that support interactive discussions, and increasingly reciprocal reviewing obligations to match supply with demand. On paper, the ML community has everything it needs to sustain a robust yet pleasant peer-review pipeline: we have the largest scholarly presence of any scientific field and the most modern review technology, all without the typical bottlenecks of paywalls, publication fees, or expensive memberships. However, the lived reality often feels far less functional. From cryptic or dismissive reviews to wildly inconsistent standards, frustrations with the review process are nearly universal — voiced by PhD students, seasoned professors, and industry researchers alike. Worse, there is little to no structured way to hold bad actors accountable, as well as incentives to encourage good actors to go the extra mile.

In this position paper, we expand our discussion of the two core challenges we have identified:

1. ***How can we reasonably limit the number of submissions?***

[1]The most recent NeurIPS 2025 has 21,575 valid paper submissions to the main conference alone.

2. ***How can we incentivize good and discourage bad review practices?***

We first lay the background on why these two issues are the root causes of much unpleasantness in ML review. Then, we assess some existing attempts to mitigate such issues as implemented in several ML conferences. We present our takes on such measures and, finally, propose two new mechanisms: fine-grained procedural safeguards that could be enforced at scale; and a currency-like incentive ecosystem based on something we called *"OpenReview Points"* — which would let researchers "earn" and "spend" their reviewing efforts in tangible ways across all major conferences and review cycles. We believe such mechanisms would have a fair chance of addressing many of the aforementioned shortcomings effectively and, more importantly, are flexible enough to allow each conference to adopt its own variants. We then present many alternative views to our proposed mechanisms, where we discuss how such views are valid (or not) and how our proposed mechanism shall be able to take such concerns into consideration. We conclude our paper by presenting a *Recommended Practice* section, which outlines our vision on how the first few conferences adopting a similar credit-based system should proceed and what aspects should be considered cautiously.

We faithfully emphasize that our goal is not to perfect ML peer review — as it would be unfaithful and condescending for anyone to claim that — but to make its failures rarer, less painful, and most importantly: more accountable and sustainable.

2 Root Causes

2.1 Overblown number of submissions causes all kinds of challenges.

We believe it is common knowledge that ML conferences typically receive an overblown amount of submissions (Kim et al., 2025; Yang, 2025). Naturally, this causes all kinds of practical challenges. From a manpower perspective, more submissions directly mean greater demand for reviewers and Area Chairs (ACs), which translates to a heavier workload for Senior Area Chairs (SACs) and, eventually, Program Chars (PCs). With such great pressure on every aspect of the conference review system, the results are predictable: thinner attention per paper, more rushed triage, and greater variance in both review quality and decision outcomes.

Moreover, practicality-wise, most ML conferences often guarantee the right to presentation exposure once a paper is accepted. This makes physical capacity restrictions come into play, directly imposing an upper limit on how many papers can be accepted. Many borderline or acceptance-inclined papers might be ruled out purely based on capacity constraints — a role often delegated to SACs. But considering the number of submissions versus the number of SACs,[2] this kind of assignment is, by design, unreasonable and unsustainable, as no SAC has the luxury or appetite to go through the content and review record of that many papers. In practice, this pressure incentivizes shortcutting — e.g., relying more heavily on numerical scores or other similar quick heuristics — which further amplifies randomness and weakens accountability. In fact, we have seen many SACs publicly pushing against such "force rejection for capacity" practices, as exemplified by LinkedIn posts from NeurIPS SACs Ahmad Beirami and Atlas Wang.

2.2 Lack of oversight, feedback loop, and incentive towards good and bad actors.

While reviewers, ACs, and SACs certainly have the right to provide feedback on their assigned submissions, ML conferences lack proper oversight and feedback loops for such actors. A reviewer, AC, or SAC can essentially engage in highly discretionary actions, so long as those actions are not extreme enough to trigger a desk rejection or other formal intervention (e.g., not performing the assigned review duty at all). This ecosystem leaves actors with no channel to learn how to become better, let alone any real incentive to go the extra mile.

As a result of the general lack of oversight, there is little to no systematic calibration across reviewers, limited visibility into meta-review quality, and minimal recognition for consistently careful

[2]Per `https://media.neurips.cc/Conferences/NeurIPS2024/NeurIPS2024-Fact_Sheet.pdf`, we have 195 main conference SACs but 15,671 submissions at NeurIPS 2024, making the workload roughly 80 papers per SAC.

work. Conversely, low-effort or unconstructive behavior often carries no consequences; as even with escalation, there are no rules for corresponding punishment unless their action is on the most extreme end. Without routinized feedback, transparent metrics, or positive incentives, the system neither rewards exemplary stewardship nor deters poor practices. We argue that such a lack of oversight and incentive makes the overall quality drift toward the lowest-effort equilibrium, where helpful practices like internal-review discussions and thorough AC investigations rarely happen, as they are not incentivized to go those "extra miles."

3 Our Takes

3.1 Soft and hard submission caps offer limited help.

With the growing research body of the ML community (Yang, 2025; Kim et al., 2025) and with AI-assisted research becoming more accessible[3] (Eger et al., 2025), the volume of submissions continues to grow at a pace that far outstrips the community's reviewing capacity. This escalation naturally prompts discussions around mechanisms for curbing submission rates and maintaining a manageable reviewing load, where submission caps are often proposed as one of the most direct ways to reduce such volume.

We argue that submission caps — whether "soft" (e.g., mandatory reciprocal review over X submissions) or "hard" (e.g., strict per-author quotas on how many papers can be submitted) — provide, at best, marginal relief. The core problem is not that a small set of "hyper-prolific" lead authors are personally flooding the system with new submissions (Yang, 2025), but that there is essentially no downside to submitting unready manuscripts or endlessly recycling previously rejected work with critical flaws. We argue that, in practice, per-author caps mostly trim auxiliary authors from the byline so that teams can fit under the quota; they do little to stop the same lead authors from submitting the same number of papers — worthy or not — or from repeatedly resubmitting flawed work. In other words, we argue that **submission caps mostly change who gets listed on a paper, rather than whether the paper is submitted.** This is because it would be unlikely for a team to postpone the submission of a paper simply because an auxiliary is hitting the caps. Thus, until there is a genuine negative incentive that discourages unlimited resubmission and rewards restraint, submission caps can only nibble at the edges of the volume problem, instead of addressing its core.

3.2 Irresponsible Reviewers Care Most About Their Own Works — So Asking Nicely Might Not Be Helpful

The seemingly widespread presence of bad or irresponsible review practices is rooted in the lack of accountability built into current conference mechanisms. Until very recently, most ML conferences enforced no retaliatory punishment against irresponsible reviewers — leaving bad practices essentially unchecked. **From a reviewer's perspective, the only thing that matters is their own current or future submissions. Therefore, to effectively discourage irresponsible reviewing, some form of penalty must be enforced at the submission end.** Otherwise, conferences have no real leverage and can only resort to asking nicely in Calls for Papers or Reviewer Guidelines; and despite the existence of some extremely thoughtful guidelines like the ARR Reviewer Guideline,[4] their effectiveness remains to be desired.

Recently, starting with CVPR 2025, several ML conferences have adopted what is essentially a retaliatory desk-rejection policy targeting irresponsible reviewers. At CVPR 2025, its Area Chairs (ACs) *"identified a number of highly irresponsible reviewers, those who either abandoned the review process entirely or submitted egregiously low-quality reviews, including some generated by large language models"* and ultimately issued desk rejections for 19 otherwise accepted papers involving those reviewers.[5].

While this act marks a meaningful start to enforcing hard procedural guardrails to protect review quality and integrity. We argue that retaliatory procedures as harsh as desk rejection can only offer

[3] One direct piece of evidence of this might be how an AI scientist is able to get a paper accepted at the main conference track of ACL 2025. See this blog and (Zhou & Arel, 2025) for details.

[4] https://aclrollingreview.org/reviewerguidelines

[5] https://x.com/CVPR/status/1894853624200863958

marginal benefits to the conference at large, as only a few bad actors would be extreme enough to blatantly ignore direct instructions.

3.3 Retaliatory Desk Rejection is Useful, but it Lacks Granularity and Cannot Achieve Influence at Scale

Given the precedent set by CVPR 2025, many conferences (including ICML and NeurIPS 2025) have begun adopting similar desk rejection policies targeting ultra-irresponsible reviewers. This is a step in the right direction. But desk rejection, by nature, is a blunt instrument: it's too harsh to apply broadly and can only be reasonably used for the most extreme violations with verifiable signals. In CVPR's case, it was mostly reserved for reviewers who outright abandoned their review duties — a clean, verifiable breach that leaves no room for ambiguity.

Unfortunately, the vast majority of reviewer problems in ML are much more subtle than complete negligence. Irresponsible reviews can manifest in many forms: from thoughtless boilerplate complaints like "no theory" or "needs more experiments" applied indiscriminately to every submission, to a gross misunderstanding of basic facts and refusal to reconsider, to a raised concern with no concrete support, or even the famous "Who is Adam?"[6] These reviews are much harder to police, but no less damaging.

We argue that desk rejection is too coarse a penalty to handle the long tail of poor reviewing behaviors that fall short of full abandonment. If we want meaningful deterrents at scale, we need a system that applies graduated, proportional penalties — not just all-or-nothing rulings. The fact that the CVPR 2025 procedure only results in 19 desk rejections is clear evidence that the irresponsible review issue in the ML community is far from being resolved by merely adopting this retaliatory desk rejection policy alone; more enforceable yet fine-grained procedural safeguards are necessary to handle the wide spectrum of irresponsible review practices.

3.4 100% Mandatory Reciprocal Reviewer Recruit is a Slow-Acting Poison

To keep up with rising submission counts, many ML conferences — starting with the most recent EMNLP 2025 (ARR May) — now rely on 100% reciprocal reviewer recruitment: every eligible author[7] must also review, except in extreme circumstances like parental leave. On paper, this sounds fair: if one wants to publish, one should contribute to the review pool. But in practice, it's a slow-acting poison and done so at the cost of review quality.

The policy assumes that all eligible authors of every submission are both capable and willing to provide thoughtful reviews. That is simply not true under many ML literature. Many eligible authors may have played only auxiliary roles or contributed as expert consultants on highly specialized components. They are therefore ill-suited to reviewing general-purpose ML submissions. Forcing mandatory review duties on such contributors creates a predictable outcome: low-effort reviews written solely to avoid desk rejections.

Worse, mandatory review removes the ability for people to decline, even when they know they cannot meaningfully contribute due to sensible (but non-medical-like) emergencies. Once in the reviewer pool, conferences often allow very limited flexibility for exemption. For instance, AAAI 2026 instructed their reviewers to "do your best" even if the assigned paper is outside their area of expertise. We argue that such enforced cultures would likely result in a series of rushed, templatized, or often disengaged reviews. It is worth noting that many later conferences are likely on the same page with us (in terms realizing the negative side of mandatory reciprocal review): starting from ARR July 2025, technically qualified authors may request an exemption from review duty on a case-by-case basis if they find themselves lacking the relevant expertise.[8]

We find it ironic that a mechanism designed to distribute the workload ends up degrading its quality. However, if everyone could opt out with no consequences, the system would collapse under the sheer volume of submissions. So, what is the reasonable middle ground — a way to allow reasonable

[6] `https://x.com/2prime_PKU/status/1948549824594485696`

[7] Where such eligibility is often determine by prior publication records, such as number of published works at A* or similar conferences.

[8] `https://aclrollingreview.org/exemptions2025`

opt-outs while still holding authors accountable for their share of the reviewing load? We, again, explore one such compromise in Section 4: a points-based incentive system that rewards good-faith reviewing and allows reviewers to "spend" those points to defer their reviewing obligations as one sees fit.

3.5 Helpful Implicit Expectations of Actors are Often Never Met

Conference processes implicitly assume that reviewers, ACs, and SACs will self-initiate best practices — such as timely calibration, substantive internal discussions, careful revision after rebuttals, and principled follow-ups — to ensure that informed decisions are made for each submission. **In reality, helpful practices like internal reviewer discussions almost never happen effectively because no one is incentivized to "go the extra mile."**

There is little real recognition, no tangible credit, and only minimal accountability for the extra coordination and time that these practices require; under deadline pressure, the rational response is to aim for the minimum viable effort. As a result, helpful practices like internal reviewer discussions become perfunctory or are entirely absent. Naturally, the decision quality then becomes noisier — not for lack of guidance, but for lack of aligned incentives to make the guidance actually happen.

4 Our Proposal: Fine-Grained Procedural Guardrails with a Currency-Like Incentive System

To make meaningful progress in peer review reform, we argue that two ingredients are essential: **enforceable procedural safeguards at different granularities**, and **an incentive structure that rewards good-faith participation while offering flexibility.** We propose a system based on a community-wide, cross-conference-supported economy called "OpenReview Points" — mainly for the mainstreamness of OpenReview and its good position to keep track of such balance.

This section outlines the basic principles of such a system, discusses potential enforcement strategies, and explores the feasibility of a conference-wide credit market that could finally provide conferences organizers both the "stick" and the "carrot" they currently lack.

4.1 OpenReview Points: A Currency-Like Economy Enabling Flexible Options

The current review ecosystem operates on the honor system — a reviewer is expected to perform review duties diligently and hope that others will do so as well. However, we argue that optimistic hoping is not a system. To install accountability, we propose currency-like credit system, giving contributors to the review pipeline something to earn, spend, and track.

Under our proposal, ML practitioners would accumulate OpenReview Points based on their contributions to the community. For instance — in the context of reviewing — completing a standard review might earn 1 point, helping with an emergency review might earn 2, and being recognized as an "outstanding reviewer" could grant an additional 3 points.[9] Once earned, OpenReview Points could be spent to gain access to certain "perks" and privileges. For example:

- A reviewer can spend 5 points to opt out of an assigned review duty.
- An author can spend 10 points to exempt a co-author from their reciprocal reviewing obligation.
- An author can spend 50 points to request an additional expert reviewer in the case of a highly controversial or borderline decision. In the meantime, a reviewer/AC can take this job and earn those 50 points.
- An author can spend 100 points to redeem free registration.

This economy introduces direct incentives: if one contributes meaningfully, one gains flexibility and optionality. If one does not, one's publishing privileges will begin to shrink. It also gives conference panels more space to experiment with different policies and enforcement harshness, without relying on blunt-force policies like universal reciprocal reviewing or desk rejections.

[9]We emphasize that all point values mentioned in this section are intuitively assigned for hypothetical purposes. A real OpenReview Point-based economy would require significantly more sophisticated balancing, subject to each conference's own preferences. More on such specifications in Section 6.

For instance, much of our work argues that there is a lack of incentive for actors to "go the extra mile," even when such effort can be immensely helpful (e.g., as anecdotally demonstrated in Appendix C). With point incentives, however, such "extra miles" can be encouraged: reviewers shall become more willing to initiate and engage in internal discussion, and ACs shall become more willing to investigate — simply because exemplary actions can now be rewarded. We can push this further by introducing targeted awards and penalties to shape community behavior. As discussed in Section 2.2, the lack of a reviewer feedback loop can potentially be mitigated by awarding points to authors who provide detailed reviewer feedback that reviewers may consult to improve their future practices. Similarly, as noted in Section 2.1, many reviewing issues stem from inflated submission counts. **One way to mitigate this is to require a small and refundable "submission fee" — e.g., 10 OpenReview Points — per paper.** If the paper is accepted or meets a reasonable "fair attempt" bar, the points are refunded; otherwise, they are forfeited. This soft deterrent discourages unready submissions by linking low-quality or premature work to a corresponding reduction in future publication privileges.

To be clear, we are not arguing for the enforcement of any specific rule — whether that is charging a submission fee in points, or allowing opt-outs from reviewing, or more. Rather, we argue that a currency-like system would grant every participant in the ML community far greater flexibility in how they interact with the review process. **While we fully expect friction or disagreement regarding any particular rule or redemption policy, we believe it would be difficult to argue against the utility of having a credit-based system at all.** Since it makes sense for different conferences to carve out their own rules to cater to their own communities.

4.2 Voting-Based Penalties May Introduce False Positives — And That's Acceptable, Because We Get Fine-Grained Enforcements In Return

To discourage low-effort or malicious reviewing, we propose allowing area chairs and fellow reviewers to flag irresponsible reviewer behavior with much greater flexibility and finer granularity. Of course, actions that are verifiably egregious — such as completely missing reviews or openly posting LLM-generated content — should trigger the harshest penalties (e.g., desk rejection), as these cases are binary, easy to verify, and largely undisputed.

But most bad reviewing practices do not look like that. They are far more subtle: templatized one-liners like "no theory" or "needs more experiments" applied indiscriminately, vague dismissals without justification, raising a plethora of out-of-scope questions, or reviews that clearly misunderstand the paper's core contributions and never bother to revise... **These cases are harder to catch algorithmically and rarely rise to the level where desk rejection is justifiable. Yet, they are widespread and deeply harmful.** As authors are often forced to invest time and energy to address these reviews, which is often to little effect if no oversights are placed upon the reviewers' end.

This is where a voting-based penalty system can help. For instance, if the authors report a reviewer, and that reviewer's peers on the paper — along with the area chair — unanimously agree that a review is deemed unacceptably low in quality, that reviewer could receive penalties ranging from a warning to various levels of point deduction.

We recognize that such systems introduce the possibility of false positives. But in the meantime, we argue that it is largely acceptable: First, if safeguards such as unanimous agreement, AC confirmation, and an appeal mechanism are in place, we believe the practical false positive rate can be kept low. Secondly, since point deduction is a much less extreme act compared to penalties like desk rejection, even a false case is unlikely to result in severe and immediate impacts of irrecoverable harm. More importantly, while our proposed system is far from perfect, the current system basically has the opposite problem: it has a nearly 0% true positive rate — since no matter how badly the reviewer behaves, as long as it is not at an automatically verifiable level of atrociousness, there are zero consequences. The status quo is clearly worse in a comparative sense.

No penalty system will ever be perfect, but the absence of one guaranties stagnation. We argue that a small risk of overcorrection is a worthwhile price to pay for finally holding the peer-review process to a higher standard. An even lower risk alternative is to award credits to AC/reviewers for providing detailed feedback on their peer reviews. this would enable a positive feedback loop where actors have channels to learn and become better versions of themselves. We discuss more about such specific recommended practices and our considerations behind them in Section 6.

5 Alternative Views

While we advocate for enforceable procedural safeguards and a currency-like incentive system, we recognize that not everyone will agree with this approach. Below, we discuss several alternative perspectives and respond to their concerns.

5.1 "Peer review shouldn't be gamified."

A common objection is that introducing a credit system risks gamifying the review process — turning what should be a scholarly, community-driven responsibility into a transactional system. While we sympathize with this concern, our counterpoint is simple: peer review is already governed by incentives — such as reviewing others' submissions in turn for having one's own work properly reviewed — but these incentives are just poorly aligned and implicitly stated.

Researchers submit to conference because they care deeply about getting their own work accepted; yet, they are often disincentivized from reviewing carefully and consistently. A credit system does not create incentives out of thin air — it simply formalizes them and aligns them with the broader health of the ecosystem.

5.2 "Voting-based penalties will be abused or politicized"

Another concern is that a flagging or voting-based penalty system could be misused — weaponized in borderline cases or influenced by interpersonal bias. We agree that any enforcement mechanism needs guardrails. That's why we require close-unanimous (if not totally unanimous) agreement from all other participating reviewers on the same paper and area chair confirmation before any penalty is issued. False positives are not impossible, but they are rare and correctable when hedged with these guardrails, and their damage — some point deduction — is unlikely to cause irrecoverable harm like immediate desk rejections and submission bans. The alternative — a system that allows actors to act with no accountability whatsoever — is far more damaging in the long run.

5.3 "A point system favors the privileged actors, and can result in bad things like losing quality reviewers."

Some may argue that a credit system will disproportionately benefit researchers with more time, institutional support, or prior connections — allowing them to "buy" their way out of responsibilities (e.g., being exempt from review duties) while leaving others to "pick up the slack." This is a fair concern. But in our design, points are earned through labor, not status. There is no "premium tier of citizen," only accumulated contributions through hard work. While it is still true that researchers with strong support will likely have more opportunities to contribute — as they are not otherwise occupied by some chores — their "surplus contributions" are still a net gain to the community.

We also note that, while exemptions from review duties might indeed result in losing reviewers, one thing to consider is whether those who are willing to pay a high price to be exempted are producing quality reviews (if kept by force), and whether they are likely to have the bandwidth to stay engaged with the authors. We argue that a better alternative might be to jut let them be exempted, and utilize the collected points to incentivize reviewers who do have the bandwidth and motivation in this particular conference cycle.

On the same note, one thing we would strongly advocate is **mobilizing researchers who are not main authors to participate more in the review pipeline**, as they likely have better bandwidth (since they are not under the pressure of author deadlines), and their reviews will not be as affected by feedback on their own submitted work. Under the current system, there is little incentive for researchers to do so, as most reviewers are recruited by mandatory reciprocity, which no longer applies without being an author. Our proposed credit system might provide them with a strong incentive to participate, as they can earn points to enrich their publication privileges; and specifically, have the option to spend such points to be exempt from reviewer duties when they are submitting lead-authored work, granting themselves wider bandwidth as authors when they are under rebuttal pressure.

5.4 "ALL OF THIS SOUNDS TOO BUREAUCRATIC."

Some may worry that procedural enforcement, tracking points, and adjudicating review quality will introduce too much bureaucracy into the process. We, again, see where this concern is coming from. However, conferences already invest massive effort coordinating thousands of reviews and rebuttals; we are simply proposing mechanisms to make those efforts fairer, more consistent, and more sustainable over time. While we do agree that the full form of our credit system can be too heavy to be implemented at once, we argue a gradual roll out of changes can be rather "soft landing" to existing community members.

5.5 "A CROSS-CONFERENCE RECIPROCITY MUST EXIST FIRST"

One clear and fair criticism of our credit system is that, for it to work to its full potential, multiple major conferences must adopt it. Granted that conferences like ICML, NeurIPS, and ICLR almost never work together closely, we recognize that such a prerequisite can be difficult to achieve. However, we argue that there are ML conferences well-positioned to adopt such practices: e.g., the ARR series of conferences has long implemented cross-conference measures (e.g., submission bans from the next ARR cycle), as experimented with in EMNLP 2025,[10] making them more openminded to adopting other similar cross-conference measures. Further, even if the credit system is per-conference, it can still function at a level that is better than nothing; it is just that features requiring accumulated effort may be harder to activate and experiment with.

5.6 "WHAT ABOUT EARLY-CAREER RESEARCHERS?"

Much like how companies and games handle the onboarding process for novice players and new employees, a reasonable expectation is that the point-hosting platform shall grant a baseline amount of credits to such first-time contributors, perhaps also with a protection period (much like how new hires cannot be directly placed into a Focus/PIP pipeline in tech companies), so that they have enough time and capital for trial-and-error.

Seeking endorsement from a seasoned scholar (much like how arXiv operates) is another viable option, though it might be more unfriendly to researchers from underrepresented backgrounds. For such contributors, a platform might run routine workshops, where participating and passing such workshop education shall grant these contributors such baseline credits.

6 RECOMMENDED PRACTICES

As emphasized throughout this position paper, our goal is not to promote a single, prescriptive rulebook that every conference must follow, but to advocate for a flexible framework that can adapt to different conference idiosyncrasies. Every policy comes with its own trade-offs — it is, ultimately, an art of compromise to decide which set of policies to adopt. **Under our credit system, conference panels and authors are akin to store owners and customers: the panels decide what goods are offered and at what price, while the customers decide where and how they wish to spend their money.**

However, we fully recognize that without concrete discussion of how such a system might actually operate, there will naturally be resistance — and, worse, chaos if adopted without proper consideration. Thus, this section serves as practical guidance from us authors on how the first few conferences adopting a credit-like system might proceed. We also leave our answers to many frequently asked questions to Appendix B

6.1 HEAVY ON EXISTING PERKS

We believe that early adopters should anchor their credit-like system around existing perks already offered in current ML conferences (e.g., complimentary registration, emergency reviewer invitations), rather than immediately introducing entirely new perks. Staying with existing perks provides

[10] `https://2025.emnlp.org/reviewer-policies/`

two immediate benefits: 1) Because these perks are already part of established workflows, redistributing them according to the point-based system (e.g., ranking reviewers by awarded points per conference cycle rather than relying on an AC's subjective judgment) keeps overall impact bounded. If the new distribution turns out problematic, its effects are still confined to the known scope of these already-tested perks; whereas brand-new perks introduce unknown risks. 2) We can directly compare key metrics between credit-system conferences and their historical data. Any improvement or decline is then more likely attributable to the credit system (or its specific implementation), rather than being confounded by the introduction of new perks. Using the same set of perks but with credit-based allocation gives us an opportunity to collect data on how the credit system behaves in practice. Such data shall serve as the baseline to test out different new perks.

6.2 Gentle and gradual rollout of new perks, potentially with one-off tests

When launching new perks, it is best to roll them out gently and gradually rather than all at once. This approach offers a clean testbed to monitor each perk's contribution and reduces the information load on all involved parties, who will need time to adjust.

Observant readers may notice that some of our proposed policies — e.g., free registration, the right to request additional reviewers, or refundable submission fees discussed in Section 4 — require a relatively long-term accumulation of points to become useful. Researchers would need to earn (or be penalized) enough points before crossing the thresholds of these services, stretching the evaluation horizon.

A simple way to expedite early evaluation is to introduce one-off tests. For example, in addition to point awards, conferences might grant top point-earners a one-off right to request an additional reviewer to help resolve borderline cases, with this right expiring at the end of the conference cycle. This allows organizers to directly observe whether such redeemable incentives meaningfully help — and to what extent these improvements propagate through the reviewer–AC pipeline. This kind of fast feedback might help conference organizers trim unhelpful policies quickly and expedite a faster iterations of rule sets.

6.3 Determining point values for contributions and perks

One reason we did not specify exact point values for different contributions (e.g., how many points an emergency review should yield) is that we currently lack the empirical data needed to set these responsibly — such as the typical number of emergency review requests in a conference, how often these requests are fulfilled, and how their workload compares to regular reviews. As a general guideline, however, we believe it is reasonable to treat the completion of one regular review duty as the base "unit price" of this ecosystem.

For the sake of discussion, suppose one regular review yields 1 OpenReview Point. Rather than arguing directly about how many points each contribution "should" receive, we propose working backwards from the value of the *perks* the system must support. For example, we can roughly estimate the point cost of a free registration by (1) checking how many free registrations a conference can realistically offer,[11] (2) determining what fraction of the reviewer population this corresponds to (e.g., the top 5-10%), and (3) computing the total point budget issued in that conference cycle — roughly, the number of submissions times the number of required reviews, plus bonus points (e.g., awards) and points collected via exemption fees.

With this budget in hand, we can set the point cost of a free registration so that only the intended contributor percentile can redeem it under plausible participation patterns.

Once the point cost of a perk is anchored, the point values of contributions should be calibrated relative to it. For instance, if completing one's regular review duty plus two emergency reviews typically places a reviewer in the top percentile eligible for free registration, then the point-value of an emergency review should be set accordingly. While this may still leave questions like why two emergency reviews, rather than three or five, should map to this threshold, working from anchored perks at least constrains point allocations to a meaningful, narrow range.

[11] For instance, NeurIPS 2025 offers free registration to top reviewers, which amounts to around 1,900+ reviewers. Similar travel grants and volunteer oppotunities are also offered at most ML conferences.

In other words, point-value determination begins by anchoring perk values to the scale of the economy; contribution-level point allocations then follow to maintain internal consistency rather than being chosen in isolation.

6.4 TRACK KEY METRICS AND PUBLICIZE SUCH STATISTICS

Finally, for a credit system to have a lasting impact, conferences must make informed decisions about which rules to adopt and at what point-values. Such decisions require cross-conference consistency. If, for example, NeurIPS values its perks at 10x ICLR's level for no meaningful reason, the ecosystem loses the interoperability we envision. Thus, each conference should monitor key metrics and publish these statistics as part of their post-conference fact sheets.

For example: If a new rule is implemented, do we observe increased interaction among reviewers and ACs? Do ACs report that these additional exchanges help them make more confident decisions? These statistics and reports will form the foundation for iterating toward a better implementation of the credit system and will also serve as a strong signal — even an advertisement — encouraging more conferences to adopt a shared credit currency.

7 CONCLUSION AND LIMITATIONS

In summary, we present a flexible credit-based framework aimed at improving accountability, aligned incentives, and procedural fairness in the ML peer-review pipeline. While we believe the system offers practical value and a principled foundation for future experimentation, it is equally important to acknowledge its current limitations and the challenges in evaluating such proposals.

Lack of numerical support and simulated experiments Our work lacks numerical results, but we believe discussions about review mechanisms are most meaningful in the hypothetical space — since there is no way to rewind history and A/B-test two parallel conference processes. Likewise, LLM-powered simulations add little value: too many layers of prompt, decoding, and model-specific variance compound, making such outcomes highly sensitive and difficult to trust. For instance, even if we scraped full ICLR discussion logs, what would we meaningfully do with them?

- Prompt LLM "reviewers" to apply our credit system and judge their reviews with another LLM?
- Ask roleplaying LLMs to simulate reviewer discussions and treat their frequency as a signal?
- Simulate multi-conference point accumulation and claim reviewer-load reduction?

We believe most readers would view such simulations as overly fragile and of limited interpretability. A further complication is that many key metrics are not public. Useful signals such as the number of internal reviewer discussions or reviewer-profile dynamics are unavailable. Even crude simulations — e.g., prompting LLMs to decide whether submissions fall below a "fair attempt bar" and deducting points — quickly become hand-wavy; as changing a threshold or banning simulated authors from a "next conference" would trivially reduce load, but such results would be meaningless without real baselines.

That said, we recognize the desire for some anchoring to real conferences. **Thus, we share three case studies — drawn from similar top ML venues where we served as reviewers — in Appendix C.** While works without statistical evaluation may face a higher bar for ICLR main track,[12] we believe our contribution fits ICLR's call for *"how we can improve the ways that we conduct and evaluate machine learning research"* and offers tangible value to the community — even if it improves the review pipeline only modestly.

Position paper submitted to ICLR main track Although this paper is written in a typical position-paper style, ICLR does not have a dedicated track for such works. We have confirmed with the PCs that *"position papers can be published at ICLR if they have sufficient novelty and value for the ICLR community,"* as demonstrated by prior accepted work (Ngo et al., 2024). We argue that ICLR is the most appropriate venue among the ICML/NeurIPS/ICLR trifecta — both in progressiveness and in its community's willingness to engage with systemic issues that affect all researchers. **We ask for reviewers' open-mindedness and invite active discussion.**

[12] `https://openreview.net/forum?id=fh8EYKFKns` is one accepted example.

Ethics Statement

Granted our work operates in a hypothetical space and promotes no particular ML methods or datasets, it is our honest analysis that it requires no further ethical consideration. We have carefully read the ICLR Code of Ethics and pledge to adhere to it.

The Use of Large Language Models (LLMs)

We would like to disclose that part of the writing of this paper was polished by a language model, though a human researcher is there to verify that the final output is true to the researcher's opinion.

References

Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman 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.

Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, et al. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *arXiv preprint arXiv:2502.05151*, 2025.

Armen Yuri Gasparyan, Alexey N Gerasimov, Alexander A Voronov, and George D Kitas. Rewarding peer reviewers: maintaining the integrity of science communication. *Journal of Korean medical science*, 30(4):360–364, 2015.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *PloS one*, 20(4):e0320444, 2025.

Steven Jecmen, Nihar B Shah, Fei Fang, and Leman Akoglu. On the detection of reviewer-author collusion rings from paper bidding. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL `https://openreview.net/forum?id=l8QemUZaIA`.

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

David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pp. 500–509, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936097. doi: 10.1145/1281192.1281247. URL `https://doi.org/10.1145/1281192.1281247`.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=fh8EYKFKns`.

Vishisht Rao, Justin Payan, Andrew McCallum, and Nihar B. Shah. Ml researchers support openness in peer review but are concerned about resubmission bias, 2025. URL `https://arxiv.org/abs/2511.23439`.

Anna Rogers and Isabelle Augenstein. What can we do to improve peer review in NLP? In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1256–1262, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.112. URL https://aclanthology.org/2020.findings-emnlp.112/.

Buxin Su, Jiayao Zhang, Natalie Collina, Yuling Yan, Didong Li, Kyunghyun Cho, Jianqing Fan, Aaron Roth, and Weijie Su. The icml 2023 ranking experiment: Examining author self-assessment in ml/ai peer review. *Journal of the American Statistical Association*, 0(0):1–12, 2025. doi: 10.1080/01621459.2025.2510006. URL https://doi.org/10.1080/01621459.2025.2510006.

Weijie J Su. You are the best reviewer of your own papers: An owner-assisted scoring mechanism. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=xmx5rE9QP7R.

Jibang Wu, Haifeng Xu, Yifan Guo, and Weijie Su. A truth serum for eliciting self-evaluations in scientific reviews. *arXiv preprint arXiv:2306.11154*, 2023.

Jing Yang. Position: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=gnyqRarPzW.

Yichi Zhang, Fang-Yi Yu, Grant Schoenebeck, and David Kempe. A system-level analysis of conference peer review. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 1041–1080, 2022.

Andy Zhou and Ron Arel. Tempest: Autonomous multi-turn jailbreaking of large language models with tree search. *arXiv preprint arXiv:2503.10619*, 2025.

A Related Works

Proposal of conference mechanisms for better ML review quality To the best of our knowledge, few published works have addressed **how to improve ML review quality by adjusting conference mechanisms.** The closest work to ours might be Kim et al. (2025), where the authors advocate establishing a feedback loop to reviewers and promoting reviewer rewards — two points that we also share. Specifically, Kim et al. (2025) highlights that if we allow authors to rate reviewers, those ratings will almost always be heavily influenced by the specific strengths and weaknesses (and, by extension, the scores) listed by the reviewer (Goldberg et al., 2025). Reviewers who fairly rate papers negatively may be subjected to unfair retaliatory ratings from authors. To address this, Kim et al. (2025) suggests a two-stage reveal: authors first read only the reviewer-written summary and strengths and provide a rating, where the weaknesses are then visible. We argue that this system might work to some degree, but in reality, much of a review's quality is determined by whether the highlighted weaknesses are sound and well supported. Reviewing reviewers without seeing these details would likely produce a lot of noise, and reviewers would be incentivized to write vague summaries that lack sharp substance. As for reviewer rewards, Kim et al. (2025) mostly argues for vanity perks like digital badges (e.g., ones similar to the "Pull Shark" badge on GitHub). We argue that such vanity-only perks would have much less influence than the review, submission, and cost-influencing perks we propose here.

Another piece of related work is the Isotonic Mechanism score pioneered by Su (2021) and its follow-up works like Wu et al. (2023); Su et al. (2025), which survey authors (with multiple submissions) and ask them to rank their submitted papers. A score is thereby calculated and compared with the mean of reviewers' raw scores. Should these two scores exhibit too drastic a gap, it may trigger AC intervention with an additional reviewer request, etc. We note that this work is, by large, orthogonal to ours, as it is yet another safeguard one can implement under our proposed credit system. The two works overlap in the sense that some countermeasures do show resemblance (e.g., requesting an additional reviewer).

Two pieces of relatively early but directly related work are Rogers & Augenstein (2020) and Zhang et al. (2022). Rogers & Augenstein (2020) discusses why incentives clash under a peer-review context and outlines many potential proposals (e.g., better review–paper matching, more tracks, abolishing "score-based" feedback, track-specific review formats, etc.). Similarly, Zhang et al. (2022) also investigates various policies but focuses on modeling why resubmission is so prevalent despite many works eventually getting accepted at a top venue. The main difference between these works and ours is that they do not propose a unified recipe (e.g., a credit system) as a means to address various issues; instead, custom solutions are typically proposed and discussed for each problem or edge case.

Conference Review Statistics There are a few works like Yang (2025); Cortes & Lawrence (2021); Beygelzimer et al. (2023); Goldberg et al. (2025) that collect real statistics and conduct controlled experiments from past conferences. While such works typically do not propose mechanism-based solutions, their numerical presentations help illustrate the scale and muddiness of current conferences; yet, such controlled exploration shall offer us insight into the practical dynamics of a particular mechanism design.

Boarder Relevant Art Last, outside the machine learning community, we have Gasparyan et al. (2015) analyzing peer review incentives under a mainly medical-focused context. The main argument of this work is *"none of these (financial or nonfinancial incentives) is proven effective* ***on its own****"*; however, the authors envisioned *"a strategy of combined rewards and credits for the reviewers' creative contributions seems a workable solution."*

Our work makes essentially the same argument, though framed under a point-based system (and of course, with more ML-specific flavors). Our system supports many of the typical financial and non-financial incentives mentioned in Gasparyan et al. (2015). For instance, several exemplary policies we discuss in Section 4.1 range from incentives for reviewers to write higher-quality and more timely reviews, to deterrents for authors submitting unready work, to non-financial privileges such as the right to be exempt from review assignments, and even financial compensation like free registration. We believe it is fair to say that we are not advocating for any particular type of incentive, but rather a collection of them. It just so happens that we unify them under a credit-based system, allowing their impact to last beyond a single conference.

Another point specific to Gasparyan et al. (2015) is that many reviewers dismiss incentives such as paper purchase discounts or free publication access, since their institutions already pay for publisher subscriptions, making such incentives mostly relevant to members with non-academic backgrounds. In our proposal, however, conference hosts can offer services that no institution would purchase in bulk (e.g., registration), or even privileges that money cannot buy (e.g., the right to request additional review resources). While we do not claim these incentives are inherently better — as one man's vulgarity is another's lyric — we do believe our framework offers a flexible way to combine different incentives to suit diverse needs, effectively pushing forward a system that Gasparyan et al. (2015) envisioned.

Outside the effectiveness of a certain incentive system, much of a conference's experience also depends on many other aspects, such as the reduction of collusion rings (Jecmen et al., 2025), finding better reviewer–paper mappings (Mimno & McCallum, 2007), determining the level of openness (Rao et al., 2025), or the use of LLMs for reviewers (Liu & Shah, 2023). We refer users to such works to build a deeper understanding of these important topics. We recommend readers refer to Kim et al. (2025) for an overview of such works.

B Frequently Asked Questions

Who governs this? We believe it is best to have a bank-like entity to govern point storage. Neutral platforms like OpenReview could serve as an ideal medium for tracking how many points each author has accumulated.

How do we prevent bad behavior (fraud, point farming, unfair allocation, etc.)? No mechanism can fully eliminate misconduct. Even real-world economies with tangible consequences face persistent bad actors. It would be disingenuous for us — or anyone — to claim otherwise. However, any thriving economy depends on aligning the interests of its participants. Conference organizers and authors share the goal of obtaining high-quality reviews and meaningful discussion, while reviewers (in the worst case) may simply want to earn points.

To bridge these goals, additional points can be awarded for high-quality reviews, and penalties applied for low-quality ones. We can also limit the total review capacity per author to discourage "quick review for basic point farming." Enforcement can be driven by an AC + reviewer voting mechanism, as discussed in various parts of our paper. We believe that these enforceable safeguards would be effective in mitigating typical point-farming behaviors where review quality is severely compromised.

Would Y operation be allowed (e.g., point transfer)? Point transfers should be rare and limited, as the foundation of this system is that contributors must personally engage in community service to earn points. However, some team-based privileges make sense. For instance, authors of the same paper could collectively "purchase" certain "products" — e.g., request additional review resources for borderline cases, or exempt a non-expert coauthor from review duties.

In line with our goal of proposing a framework rather than a detailed rulebook, we find it reasonable to forbid person-to-person transfers while allowing team-based purchases or awards, depending on conference preferences.

Would Z infra/consensus be needed? It is fair to note that certain infrastructure and shared consensus would be required to operationalize our credit system. At minimum, platforms like OpenReview should provide:

- A balance-tracking system to record point accumulation per author.
- A limited point-transfer interface to implement awards or penalties.

While such infrastructure is necessary, we emphasize that it is a minimal requirement. The "legislation" and enforcement of specific policies would still rest with individual conference panels. As one man's vulgarity is another's lyric, there will never be a perfect policy — only carefully considered trade-offs. Our framework respects the diverse needs of the community and supports them with great flexibility.

C THREE CASE STUDIES WHERE WE SERVE AS THE REVIEWERS

While our position paper, unfortunately, lacks real conference data to support why our proposed framework would be helpful, we believe there is even less reason to run an LLM-roleplaying simulation. We understand the perspective that having some anchors to real conferences is preferred. Here, we share three case studies — all from similar top ML conferences — where we serve as reviewers.

C.1 CASE 1: REVIEWERS CRITICIZING MATTERS OUTSIDE THE PAPER'S SCOPE.

In this case, we observed that another reviewer criticized the submitted work for reasons clearly outside its intended scope. We therefore raised our concern to the AC:

Internal comment to PC/SAC/AC

This message is set to be only visible to PC, SAC, AC, and the authors.
I want to disclose that I find reviewer `A`'s evaluation of this paper quite unreasonable. This reviewer writes:

- Certain methods, such as `method type`, demonstrate limited effectiveness in `setting`, which may restrict their practical deployment.
- The paper points out that many of the `an important task` methods tested are essentially extensions of existing models adapted for `another important task`, such as `a famous method`.

It doesn't make much sense to cite the low performance of certain featured methods as weaknesses of a dataset-proposing/benchmark paper. It is not the authors' problem if an established method underperforms. Instead, the point of benchmarking is precisely to show when a method would fail. Many benchmark works have done this — `some examples` — and it is beyond comprehensible why this is considered a weakness.
Another criticism from reviewer `A` is:

- The tasks within the benchmark may not capture all possible real-world application scenarios, possibly overlooking specific needs within certain domains.

This, in my opinion, is a boilerplate concern that can be said for literally *any* dataset. While I do agree that the proposed dataset does not capture some important `task` scenarios — `some examples` — criticizing it for "not capturing all possible real-world applications" crosses the line and feels borderline hostile. This is akin to criticizing a method paper for not evaluating on every possible dataset.
I recommend the AC to either disregard `A`'s review or consider encouraging the reviewer to revisit the evaluation.

This paper was ultimately accepted. This anecdote shows that without inner-reviewer analysis, simple rule-based policies such as "inactive $\rightarrow$ desk rejection" fail to capture cases of severely low-quality reviews. Finer-grained measures must be practiced to ensure that positive impact scales broadly, rather than being limited to a few desk-rejected papers.

C.2 CASE 2: REVIEWERS ASKING FOR PARTICULAR EXPERIMENTS AFTER THE REBUTTAL DEADLINE.

In a top ML conference where the exchange between authors and reviewers are limited to certain time window, we had a split decision situation where the non/late-responding reviewers are not supportive of the submission. As the AC is calling for consensus, we jumped in and asked:

> **Internal reviewer discussion**
>
> I skimmed over the two negative reviews of this work and found merits in many of the reviewer-raised points. However, I also find the authors' rebuttal to be proper in many regards — especially when the raised concern demands a clarification-like answer.
> It looks like the two negative reviewers have yet to address the authors' rebuttal in a meaningful way (only acks are issued, cmiiw). So, to reach a consensus, I believe it would be helpful if the two reviewers could elaborate a bit on their leftover concerns. I am happy to set aside some time to discuss such leftover issues from my perspective.

Essentially, the two negative reviewers believed that certain experiments were missing — one of which can be seen as a combination of two existing methods, and another as a specific investigatory study of the author-proposed method. While we find such suggestions to have merit, we believe they were not raised appropriately from a procedural standpoint:

> **Internal reviewer discussion**
>
> I appreciate `A`'s detailed response and updated review. I believe **`A` (as well as `B`)'s main concerns regarding `ProposedMethod` vs. `PriorWork1` + `PriorWork2` are legitimate and sound.** That being said, I am always the kind of reviewer who is "more in the authors' shoes"—for lack of better words—and I would like to present two alternative arguments regarding this concern.
> First, I believe experiment-comparison requests that touch on *combinations of existing works* should be cautiously brought up. Many methods can be combined, but their combinations typically require a number of discretionary design decisions, and it is often unlikely for authors to feature the exact combination a reviewer has in mind. In this case, `ProposedMethod` proposes a paradigm of `reducted`, where the scope of eligible combinations is wide. Thus, in my opinion, **if reviewers are specifically interested in the comparison `ProposedMethod` vs. `PriorWork1` + `PriorWork2`, such a request should be made *explicitly* before the rebuttal deadline, rather than mentioned in hindsight when the authors have no channel to address it.**
>
> From the look of it, the `ProposedMethod` authors submitted their initial rebuttal on `an early date`, which is `[redacted]` days after the review post. However, only `A` engaged substantively on `a late date`. I must note that this year's `BigConferenceName` requires only two rounds of exchange, yet only 2/4 reviewers provided those to the authors, with all engaged reviewers leaning positive. **For such reasons, while I am also interested in this comparative result and agree with `A`'s analysis, I do not believe we can use it against the authors (at least not as a singular veto reason), as the request was not properly raised from a procedural standpoint.** Imagine we were submitting a paper where reviewers were largely non-responsive, and the paper was then rejected for missing an experiment that was never explicitly asked for—it would be hard not to feel that is unfair. While I understand we all have different priorities and may have limited bandwidth for various reasons, we need to properly compensate authors when such situations occur.
>
> (I know it is uncommon for a reviewer to argue on behalf of authors, but I always do so when it is warranted. The AC is welcome to confirm that this advocacy is from me for good reasons and not the result of any collusion.)

Later, we and the two reviewers exchanged more than five comments in total, which helped the AC reach a favorable decision. This anecdote shows that many reviewers are willing to engage in internal discussions, and such exchanges are profoundly helpful in deciding borderline papers — they simply never had the "push" to initiate such discussion voluntarily. We argue that our credit-based incentive system could help elicit more of these productive discussions.

Another takeaway from this exchange is the importance of having enforceable safeguards (e.g., reply deadlines), as otherwise authors may have no channel to meaningfully rebut at all. Attaching rewards and penalties to such actions is also crucial to ensure procedural fairness. While we respect and appreciate 'A' and 'B' for engaging our discussion, the fact that they were ok with not replying

/ late replying authors is because there is virtually no penalty to them (as their "misconduct" would be too minute for desk rejection, yet the conference organizer has no other finer-grained tools) — something our credit system would help.

C.3 CASE 3: AUTHORS PRESENT UNSUPPORTIVE RESULTS WHILE CLAIMING OTHERWISE.

In this case, we found that the authors were presenting unsupportive results to one reviewer while verbally claiming the opposite — so we stepped in:

> **Internal reviewer discussion**
>
> As another reviewer, I find the reading on `ProblemX` tricky. The goal of `Task` is to have the `[reducted]`.
> In `ProblemX`, when the `component` is trained only on `dataset1`, the `dataset2` accuracy drops quite significantly — a disadvantaged result. Once the `component` is trained on both `dataset1` and `dataset2`:
>
> - The `dataset1` performance improves `a small number` over the `a baseline`, but this also comes at the cost of generating many more tokens than the `a baseline`.
> - `[redacted as it is too technical]`
> - In the end, the system shows only a `small number` accuracy gain on the task it was specifically trained on, while `doing something at a higher cost`. I think the added experiment makes the work more negative (though I appreciate the transparency), as it feels like the proposed `component` is very task-specific and does not generalize well.
>
> (This message is set to be not visible to authors so that we can have a discussion with supposedly no bias, but I am happy to adjust the visibility should you want a more public discussion.)

The reviewer exchanged views with us and we reached an agreement. The paper was ultimately rejected, with the AC explicitly citing our reasoning. Specifically, this other reviewer noted:

> **Internal reviewer discussion**
>
> (BTW, you're among the very few reviewers who respond to other review comments. I truly admire this level of dedication.)

This suggests that internal reviewer discussion is indeed rare, even though we lack direct statistical evidence.

We hope these three anecdotal case studies illustrate how even a small component of our proposed framework — encouraging more internal reviewer discussion — can meaningfully facilitate paper evaluation. While we acknowledge our bias, we believe that in all three cases, the key factor behind each paper's acceptance or rejection was our initiative in starting those discussions. This suggests that most reviewers *want* to contribute and most ACs will take such discussions seriously; they simply need a small push to take the initiative — a push that our point-based incentive could very well provide.