Click for
Citation & BibTeX

Peer review should constrain evaluative authority

Hanrui Wang[1]*, Timo Spinde[1,2], Isabella Habereder[2,3], Chun-Shien Lu[4], Isao Echizen[1,2]
[1]National Institute of Informatics, Japan
[2]Media Bias Group, Japan
[3]Georg-August-University Göttingen, Germany
[4]Academia Sinica, Taiwan

Abstract

Our position shifts the target of peer-review reform from reducing uncertainty through evaluator improvement or judgment aggregation to constraining evaluative authority under persistent uncertainty.

Scientific peer review aims to assure research quality but also allocates scarce academic opportunities, such as publication in high-impact venues. However, the review system is under growing pressure as submissions increase while reviewer capacity, expertise, scrutiny, and incentives cannot scale accordingly. As peer review operates under persistent evaluative uncertainty, evaluator improvement and judgment aggregation remain structurally limited. We therefore propose authority-constrained peer review: before a consequential judgment is released as decision-relevant, it should withstand structured scrutiny through critique-authority decoupling, independent secondary evaluation, contestability and reversibility, and accountability for unsupported influence. Our aim is to prevent weakly supported judgments from entering final deliberation with the same authority as judgments that have withstood structured scrutiny, while creating incentives for reviewers to justify consequential claims more carefully. For large-scale computer-science conferences, this position offers an institutional response to uneven review quality and reviewer-assignment dependence; more broadly, it provides a governance principle for any domain where uncertain expert judgments allocate scarce academic opportunities.

Project Page summarizes this work and will be updated with future policy and technical developments related to peer review.

1 Introduction

Peer review is a central mechanism for ensuring the quality of scientific findings, but it also serves as a basis for allocating scarce opportunities for publication, funding, and professional recognition [1, 2]. However, in large-scale machine learning and AI conferences, the number of submissions under review has increased substantially in recent years (see Figure 1). This growth not only increases reviewing workload, but also places greater pressure on review judgments: these judgments continue to shape consequential decisions [1, 3] in a review system where reviewer expertise, review scrutiny, and reviewing incentives cannot scale at the same pace [4, 5, 6, 7, 8, 9].

One major response to rising review pressure is to improve evaluators themselves, based on the intuition that better evaluators lead to better reviews [2, 10, 11]. Human-centered reforms seek better reviewer training, matching, incentives, transparency, and accountability [2, 12, 9, 13, 14]. AI-centered reforms aim to improve review quality or accelerate peer review through model optimization, automated review generation, or hybrid human-AI workflows [15, 16, 11]. However,

*Corresponding author: hanrui_wang@nii.ac.jp

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Position Paper Track.

many criteria that peer review relies on, such as novelty and significance, lack reliable ground truth at the time of review [17, 18, 2, 19]: whether the review judgment is correct cannot be directly verified when decisions are made. Peer review, therefore, operates under persistent evaluative uncertainty. As a result, evaluator improvement lacks stable targets against which improvement can be guided [20, 21]; even when a target can be specified, it remains difficult to validate whether overall review quality has been effectively improved [20, 21, 22].

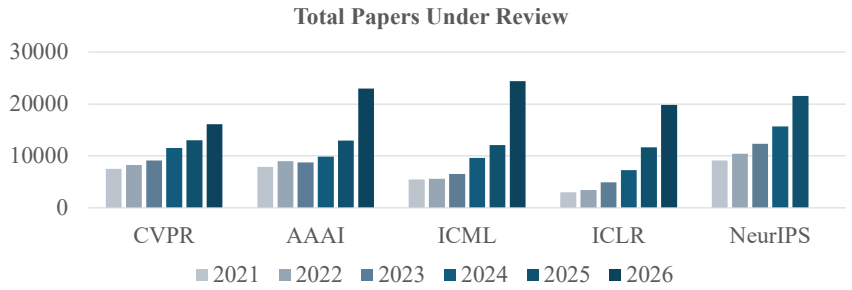

Figure 1: Total papers under review at major computer science venues across recent years, illustrating large growth in submission volume.

Taking persistent evaluative uncertainty in peer review as a premise, we shift the institutional focus of reform: constraining the authority of each review judgment before it can influence the final decision. We propose four principles:

- *Critique-authority decoupling:* the evaluator who provides the substantive critique should not unilaterally convert that critique into a consequential decision signal, such as an acceptance or rejection recommendation.
- *Independent secondary evaluation:* consequential review judgments should receive independent second-order checks for clarity, evidential support, and interpretive validity.
- *Contestability and reversibility:* review judgments should remain open to challenge and, where necessary, revision or reversal.
- *Accountability for unsupported influence:* reviews that significantly affect decisions but are later found to be poorly supported should remain visible in the record.

Our position also differs from the conventional many-signal peer review. Many-signal review relies on multiple reviewers, rebuttals, and decision layers to let judgments complement, correct, or compete with one another [22, 18]. However, each review judgment in a many-signal system can still enter final deliberation as one evaluator's unilateral judgment and become a consequential input to the final decision [18, 3]. In contrast, our proposed authority-constrained review requires consequential judgments to withstand structured scrutiny before they are released as decision-relevant signals. The aim of authority-constrained review is not to add more opinions for aggregation, but to prevent weakly supported judgments from entering final deliberation with the same authority as judgments that have withstood structured scrutiny. Constraining authority also creates incentives for reviewers to provide judgments that are more explicit, evidence-based, and defensible.

Taken together, our position does not rely on reducing evaluative uncertainty. Instead, it aims to prevent uncertain judgments from acquiring unchecked authority in final decisions and to encourage more careful review. More generally, it asks how peer-review institutions should govern the authority of expert judgments when those judgments allocate scarce opportunities under persistent uncertainty.

2 Core condition of peer review and its structural consequences

Peer review relies on multiple evaluative criteria to reach final review decisions [17, 18]. Table 1 summarizes ten major criteria drawn from thirteen review guidelines.[2] As presented in Section 1

[2] https://cvpr.thecvf.com/Conferences/2026/ReviewerGuidelines
https://iccv.thecvf.com/Conferences/2025/ReviewerGuidelines

and Table 1, many of these criteria lack reliable ground truth at the time decisions are made [17, 19]. Building on prior literature [23, 4, 17], we classify these criteria into three *ground truth statuses*, according to the extent to which reliable ground truth can be established at the time of review decisions. *Relatively reliable:* criteria for which the relevant evidence is largely contained in the submission and can be directly assessed during review [23]. *Limited:* criteria for which some evidence is available during review, but assessment remains incomplete, conditional, or dependent on reviewer knowledge, scrutiny, or available resources [4]. *Unavailable:* criteria for which no reliable ground truth can be established at decision time because the necessary evidence is not observable during review [17]. We therefore affirm that missing criteria ground truth is the core condition under which peer review operates. This condition produces three structural consequences that are difficult to resolve, which together place peer review under persistent evaluative uncertainty.

Table 1: Major evaluative criteria in peer review and their ground truth (GT) status. Only a narrow subset admits relatively reliable decision-time ground truth. The classification reflects the availability of evidence, not perfect objectivity or mechanical decidability.

Criterion	**GT status**	**Note**
Reporting completeness	Relatively reliable	Local omissions or reporting defects can be checked directly from the manuscript; mismatches may cause desk rejection [24].
Technical correctness / soundness	Limited	Some flaws can be identified during review, but deeper errors may remain undetected until later scrutiny [25].
Methodological / empirical adequacy	Limited	Assessment depends on field standards, expertise, and depth of scrutiny [26].
Literature review sufficiency	Limited	Coverage depends on reviewer knowledge, search depth, and judgments of relevance [27].
Reproducibility	Limited	Some evidence may be available at decision time, but promised code, data, or artifacts may later be incomplete or unavailable [28].
Ethics	Limited	Some issues are visible at review time [29], whereas others may only be questioned or uncovered after publication [30].
Novelty / originality	Unavailable	Judgments depend on literature coverage, field framing, and what counts as a non-trivial advance [17].
Significance / importance / expected impact	Unavailable	Scientific importance or future impact cannot be reliably established at decision time [17, 31, 32].
Venue fit / relevance	Unavailable	Assessment depends on venue identity and contestable expectations about audience and contribution.[a]
Clarity / presentation	Unavailable	Whether a paper is sufficiently clear for acceptance depends on reader background and expectations.[b]

[a] `https://chi2025.acm.org/guide-to-reviewing-papers/`
[b] `https://www.nature.com/nature/editorial-policies/peer-review/`

2.1 Standard inconsistency

Review guidelines are intended to ensure that the same evaluation criteria are applied to all submissions [18], but the standards by which reviewers interpret and apply these criteria remain heterogeneous [17, 18]. Building on prior literature [12, 33, 34, 5, 28, 18, 4], we identify six *reviewer-side sources of interpretive subjectivity* that shape how evaluative criteria are assessed (see Table 2). *Topic familiarity:* familiarity with the paper's specific topic, problem setting, or literature [12]. *Expertise:* relevant technical or methodological competence [4]. *Scholarly pursuit:* the reviewer's orientation toward what kinds of research are worth advancing [33]. *Preferences:* intellectual, methodological, or stylistic inclinations [34]. *Reviewer effort:* the time and attention invested in

`https://eccv.ecva.net/Conferences/2026/ReviewerGuide`
`https://icml.cc/Conferences/2026/ReviewerInstructions`
`https://neurips.cc/Conferences/2025/ReviewerGuidelines`
`https://iclr.cc/Conferences/2026/ReviewerGuide`
`https://aaai.org/conference/aaai/aaai-26/instructions-for-aaai-26-reviewers/`
`https://acmmm2025.org/reviewer-and-area-chair-guidelines/`
`https://aclrollingreview.org/reviewerguidelines`
`https://www.sigsac.org/ccs/CCS2026/call-for/call-for-papers.html`
`https://signalprocessingsociety.org/publications-resources/guidelines-reviewers`
`https://dl.acm.org/journal/tos/reviewer-guidelines`
`https://www.science.org/content/page/science-information-reviewers`

the review [5]. *Verification resources:* the practical means available to check relevant aspects of the submission [28].

Reviewers may interpret and apply the same criteria differently, making criteria ground truth difficult to determine collectively at the time of decision [22, 18]. Reviewer judgments may therefore diverge not only because submissions differ, but also because reviewers rely on different standards, evidence, and interpretive frames [22, 18, 17, 4, 19]. We therefore argue that peer review operates under standard inconsistency.

Table 2: Illustrative sources of reviewer-side subjectivity in applying major evaluative criteria. **TF**: topic familiarity; **EX**: expertise; **PU**: scholarly pursuit; **PR**: preferences; **EF**: reviewer effort; **RS**: verification resources.

Criterion	**TF**	**EX**	**PU**	**PR**	**EF**	**RS**
Reporting completeness					✓	
Technical correctness / soundness	✓	✓			✓	✓
Methodological / empirical adequacy	✓	✓	✓	✓		
Literature review sufficiency	✓	✓	✓	✓	✓	
Reproducibility	✓	✓	✓		✓	✓
Ethics	✓	✓	✓	✓		
Novelty / originality	✓	✓	✓	✓		
Significance / importance / expected impact	✓	✓	✓	✓		
Venue fit / relevance	✓	✓	✓	✓		
Clarity / presentation	✓		✓	✓		

2.2 Trust-dependent opacity

Reviews may be shaped not only by how criteria are interpreted, but also by surrounding incentive conditions [35, 36, 37, 38, 39, 2, 40, 6, 9]. We summarize three incentive conditions that are particularly relevant to peer review. *Scarce rewards:* the fact that publication records are tied to prestige [35], career advancement [36, 37], visibility [35], and downstream access to resources [38]. *Conflict of interest (COI):* direct rivalry, shared interests, professional or network ties [39], as well as concerns about retaliation [2, 40]. *Weak incentives for careful review:* the fact that reviewing is often time-consuming and only weakly rewarded [6, 9].

We refer to reviews shaped by such surrounding incentive conditions as *incentive-coupled evaluations*. As discussed in the literature [41, 39, 42, 43, 40], incentive-coupled evaluations may take at least three forms. *Strategic distortion:* deliberate inflation or deflation of evaluative judgments, for example, through harsher or more favorable assessments of novelty or significance; such distortion may be motivated by scarce rewards or COI [41, 39]. *Low-effort or AI-substituted review:* evaluation based on limited effort or AI substitution, leading to superficial comments, heuristic judgment, or weakly scrutinized review text; this may arise when reviewers prioritize research production over careful reviewing under weak reviewing incentives [42, 43]. *Cautious non-commitment:* hedged or persistently borderline judgment that avoids strong commitment, often reflecting limited evaluative investment or reluctance to impose a clear acceptance or rejection; such judgment may also be reinforced by COI, particularly concerns about retaliation [40].

In Table 3, we illustrate couplings between incentive conditions and forms of incentive-coupled evaluation. Because many review criteria have limited or unavailable ground truth, incentive-coupled evaluations are difficult to detect or separate from ordinary criteria-based disagreement [39, 17]. Peer review, therefore, relies partly on trust in reviewers' professional and ethical conduct.

Table 3: Illustrative couplings between incentive conditions and incentive-coupled evaluations.

Incentive-coupled evaluation	**Scarce rewards**	**COI**	**Weak incentives for careful review**
Strategic distortion	✓	✓	
Low-effort / AI-substituted review	✓		✓
Cautious non-commitment		✓	

2.3 Criteria-to-decision indeterminacy

As illustrated in Figure 2, peer review involves a two-stage process: evaluative criteria (Table 1) are first translated into reviewer-level judgments (e.g., recommendations and confidence scores), which are then used by editors or chairs to reach final decisions.[2]

However, neither stage follows explicit or stable decision rules, i.e., there is no standardized way to translate criteria into reviewer judgments, nor to translate those judgments into final decisions [18, 3]. As a result, evaluations that use the same criteria may still yield different outcomes. We refer to this structural lack of a well-defined mapping as *criteria-to-decision indeterminacy*.

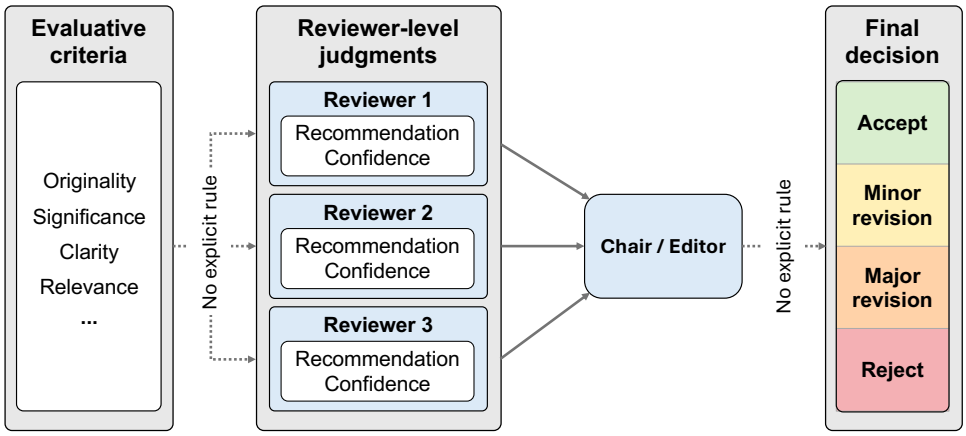

Figure 2: Peer-review criteria indirectly inform reviewer-level recommendation and confidence judgments, which, together with evaluative comments, are then further considered by chairs or editors in reaching a final decision. No explicit rule links criteria to reviewer judgments or reviewer judgments to final decisions.

3 Alternative views

As stated in Section 1, we examine three prominent approaches to improving peer review: many-signal review, human-centered improvement, and AI-assisted review. The following sections analyze what each approach can improve and why reducing uncertainty through judgment aggregation or evaluator improvement remains insufficient under missing criteria ground truth.

3.1 Many-signal review systems

Conventional peer review often uses multiple reviewers, author rebuttals, and layered decision-making by chairs and editors; we refer to this as a many-signal review system.[3] The intuition of many-signal review is that distributing judgment across reviewers and decision layers can reduce the impact of a single reviewer's idiosyncratic judgment on final decisions [22, 18].

However, each reviewer in a many-signal system still evaluates under missing criteria ground truth, applies heterogeneous standards, may be influenced by surrounding incentive conditions, and contributes to a decision process without a stable criteria-to-decision mapping. Adding more reviewers, therefore, does not prevent a single review from exerting consequential downstream influence [3, 44]. This influence may become significant when negative or cautionary interpretations carry disproportionate weight in subsequent decision-making [45], for example, because early judgments may anchor later evaluation [46, 47], or because competitive selection settings may make adjudicators more cautious toward uncertain work [33, 48].

A further limitation concerns the timing of correction. Many-signal systems allow a judgment to be challenged through other reviews, author rebuttal, reviewer discussion, or chair/editor moderation[3], but such correction may occur after the judgment has already begun to shape the decision process [3,

[3] https://www.acm.org/publications/policies/roles-and-responsibilities

46]. The key distinction is therefore not whether a judgment can eventually be corrected, but whether it is independently scrutinized before or only after it begins to exert consequential influence.

Moreover, substantial disagreement among reviewers remains common under current many-signal review systems [49, 2], suggesting that the number of reviewers typically assigned to a submission is often insufficient to stabilize evaluative variance. Existing studies do not clearly establish how many reviewers would be needed to reduce such variance substantially, nor do they show that greater agreement among a limited number of reviewers would guarantee correctness [22, 50, 51]. Furthermore, the problem of reviewer disagreement is further compounded when review quality is degraded by incentive-coupled evaluations, for example, through low-effort review or cautious non-commitment.

3.2 Human-centered evaluator improvement

Human-centered reforms typically aim to improve peer review through clearer reviewing standards and training [10, 14], improved reviewer matching [52, 53], stronger transparency and accountability mechanisms [14, 13], and stronger incentives or expansion of the reviewer pool [9, 13]. Their intuition is that peer review can be improved by better human reviewers and more robust review procedures. However, there is no universally accepted reference standard for a "good" review [20, 21], leaving evaluator improvement without a reliable target.

Improved training, matching, and procedural guidance may reduce some forms of inconsistency [10, 53, 12], but they cannot remove differences in topic familiarity, expertise, scholarly pursuit, and evaluative preferences [12, 4, 33, 34]. Judgments about novelty, significance, and fit therefore remain contestable rather than reducible to uniform application of a shared standard [17, 54, 18].

Furthermore, improved review standards and procedures do not ensure meaningful compliance. Greater transparency, accountability requirements, and procedural expectations may encourage improved review conduct [14, 13, 10], but they do not ensure that reviewers will follow the guidelines consistently in practice [40, 39, 2]. Reviewer overload [55, 2], difficulty securing willing reviewers [6, 7, 8], and weak incentives [9, 2] continue to place reviewers under conflicting constraints of time, effort, and care.

Moreover, there is evidence of a sustained increase in the number of submissions (see Figure 1), while it is unlikely that a human-centered reform would lead to a corresponding increase in reviewer capacity. Stronger incentives and author-obligated reviewing may increase reviewer supply, but they cannot generate scarce expertise on demand [4, 56] or eliminate reviewer-submission mismatch [12, 53, 57]. Expertise, carefulness, and institutional professionalism are difficult to expand at the pace of rising submission volume and AI-assisted research production [2, 58, 59].

3.3 AI-assisted review

AI-centered reforms typically aim to improve peer review through automated review generation [15, 60, 61], AI-assisted review support [16, 62, 11], and hybrid human-AI evaluation workflows [11]. Their central intuition is that peer review can be improved through scalable evaluative support that expands review capacity, increases consistency, and enhances the quality of critique [11].

However, review corpora [63, 64, 65, 66] encode historically produced judgments shaped by standard inconsistency and trust-dependent opacity, rather than verified targets of correctness, significance, or future value [17, 31, 32]. Synthetic data may broaden the training signal [67], but optimization and validation still rely on heterogeneous human judgments, evaluative standards, and heuristic metrics rather than verified criteria ground truth [22, 17, 68, 31, 32]. AI-assisted review, therefore, inherits the core condition of missing criteria ground truth at the level of data, optimization target, and validation.

Moreover, training and optimization of AI review systems become more challenging as the review stream itself is affected by variability and distortion. The missing criteria ground truth already limits the possibility of clean supervision. Reviewer-side subjectivity and incentive conditions further shape the evaluative signals that enter review corpora, while accelerated scientific output, limited reviewer capacity, and scarce rewards may accelerate the accumulation of inconsistency, bias, and signal degradation in these data [43, 69, 70]. AI reviews thus face a recursive problem: the data

from which these systems learn are themselves produced under the very structural constraints that these systems are expected to overcome.

Existing studies document critical limitations of AI review, including factual unreliability [71, 72], shallow assessment of scientific validity and technical depth [16, 62, 71], and security, privacy, bias, and governance concerns [73, 74, 75, 11, 76]. These limitations are relevant but not central to our argument, as many may be partially mitigated by stronger models and more mature institutional safeguards [75, 16, 11]. Our central concern is instead structural: model improvement may reduce generic LLM weaknesses, but it cannot, by itself, establish reliable performance on the specialized task of peer-review evaluation without a valid reference standard for review quality.

4 Our position: Authority-constrained peer review

We argued that existing approaches share the same problematic logic: they seek better decisions through more signals, better human reviewers, or AI evaluators, without reliable criteria ground truth. Under this logic, uncertain judgments remain difficult to distinguish from well-supported ones, yet uncertain judgments can still acquire unchecked evaluative authority in final deliberation. In contrast, we take such persistent evaluative uncertainty due to missing ground truth, including standard inconsistency, trust-dependent opacity, and criteria-to-decision indeterminacy, as a premise and shift the reform target from improving evaluators to governing evaluative authority. We use evaluative authority to refer to the capacity of a review judgment, score, objection, or recommendation to shape final deliberation or appear in the decision rationale.

4.1 Design principles

Authority-constrained review can be implemented through four principles. First, critique and authority should be decoupled: the evaluator who writes the substantive critique should not automatically convert it into a consequential recommendation or score. Second, consequential judgments should receive independent second-order evaluation of their clarity, evidential support, and interpretive validity before they enter the decision rationale. This is a targeted check of the existing review judgment, not an independent full review of the submission. Third, review judgments should remain contestable and reversible, so that challenges can lead to clarification, narrowing, downgrading, score revision, or removal from the decision rationale. Fourth, unsupported influence should remain accountable: when a judgment significantly shaped a decision but is later found to be poorly supported, the system should record both the judgment and the correction of its authority. These principles make review influence conditional rather than automatic, while encouraging reviewers to make consequential claims more carefully.

Authority-constrained review differs from many-signal review systems not only in how judgments are checked, but also in when they acquire consequential authority. Many-signal review adds signals for aggregation and may correct judgments through later rebuttal, discussion, or moderation, whereas authority-constrained review requires independent scrutiny before a judgment enters final deliberation with full decision relevance. The aim is to expose weak evidential bases, unsupported claims, and unresolved interpretive problems before they exert consequential influence, making authority conditional on prior scrutiny rather than later correction.

4.2 Comparison with alternative views

To clarify the distinctiveness of our position, Table 4 summarizes how authority-constrained peer review governs evaluative authority under the core condition and structural consequences developed in Section 2, in contrast to the three approaches discussed in Section 3. We elaborate on each comparison below.

Core condition. Many-signal, human-centered, and AI-centered approaches are all oriented toward bringing peer review closer to the criteria ground truth: by aggregating more signals, by training better human reviewers, or by optimizing models on review data and preferences. In contrast, our position begins with the assumption that weakly supported, seriously flawed, or otherwise contestable reviews will persist in the absence of criteria ground truth and cannot be eliminated with confidence. We therefore seek to constrain the downstream influence of consequential judgments through independent scrutiny rather than to rely on directly improving their quality. This principle

Table 4: How different approaches address the core condition and its structural consequences.

Core Condition / Consequence	Many-signal	Human-centered	AI-assisted	Authority-constrained
Missing criteria ground truth	Add more signals	Improve evaluators	Optimize models on review data	Assume persistence; limit decisional influence
Standard inconsistency	Average across judgments	Standardize human interpretation	Standardize model scoring	Cross-check interpretation before influence
Trust-dependent opacity	Reduce single-review influence	Improve detection and compliance	Improve detection and filtering	Constrain authority to make incentive-coupled influence harder
Criteria-to-decision indeterminacy	Add evaluative inputs	Clarify procedures	Learn existing mappings	Limit unilateral steering under indeterminacy

applies to all consequential judgments, although its practical implementation may need to allocate scrutiny unevenly. In an initial or resource-constrained deployment, judgments with greater potential influence on the final decision should therefore be checked more carefully, for example, clear accept or reject judgments that, if unsupported, could substantially shape the eventual outcome.

Standard inconsistency. Standard inconsistency arises primarily from reviewer-side human subjectivity, which cannot be reliably eliminated. Many-signal review responds by adding evaluators under the assumption that a larger number of judgments will reduce variance. Human-centered reform provides clearer, more detailed criteria to ensure more uniform interpretation. However, reviewer pools are usually too small for multiplicity alone to neutralize heterogeneous interpretation. Detailed standards also remain subject to divergent human understandings. Compared with many-signal and human-centered approaches, AI-assisted evaluation can respond more directly to standard inconsistency. A single model may apply a more consistent scoring standard across submissions. But consistent scoring does not imply that AI-assisted review is an optimal solution, given the additional limitations discussed in Section 3.3. Instead of relying on multiple separate reviews, we propose a different form of evaluation diversity, in which evaluators cross-check one another's interpretations within the same review process. Our position is particularly clear in a human-critique, AI-score arrangement, in which the human evaluator's decision-making authority is limited, and the AI evaluator helps ensure greater consistency in scoring.

Trust-dependent opacity. Trust-dependent opacity arises because incentive-coupled evaluations cannot be cleanly ruled out under missing criteria ground truth. In a many-signal review system, it is assumed that including additional reviewers mitigates the influence of individual judgments driven by incentives. Human-centered and AI-centered reforms, by contrast, depend more on correcting and mitigating problematic evaluations through improved norms, supervision, or model design. We argue that, given missing criteria ground truth, detection of incentive-coupled evaluations rarely produces evidence strong enough to resolve the issue institutionally. Instead, our position of constrained authority does not aim to eliminate incentive-coupled evaluations, but makes them harder to shape outcomes without resistance.

Criteria-to-decision indeterminacy. Criteria-to-decision indeterminacy is the consequence least directly resolved by existing approaches, including our own. Many-signal review adds evaluative inputs but does not establish a stable guideline for aggregating criteria or reviewer-level judgments into final decisions. Human-centered reform can clarify criteria and procedures, but does not yield a determinate mapping from substantive evaluation to recommendations or from recommendations to final decisions. While AI-centered reform can learn or reproduce historical classifications from previous reviews and decisions, this merely mimics existing patterns rather than eliminating the underlying uncertainty. Our position does not eliminate the criteria-to-decision indeterminacy either. Instead, we limit the extent to which a single review can steer final decisions. In doing so, we limit the extent to which the criteria-to-decision indeterminacy from a single review can affect outcomes.

5 One possible instantiation

Authority-constrained review can be instantiated in different institutional forms. We describe one concrete illustration based on dual-track human–AI mutual supervision. In *Track A*, an AI system produces a substantive critique without assigning a consequential score, and human supervisors independently assess the critique by recording agreement, disagreement, and important omissions

before forming their own judgments. In *Track B*, human reviewers produce substantive critiques, while an AI system independently assigns the rating most consistent with each critique and scrutinizes its evidential support, interpretive validity, and decision proportionality. The two tracks do not observe each other's intermediate outputs, and their outputs are provided to the chair or editor for final adjudication rather than automatically aggregated into a decision. This separation operationalizes the principle that the evaluator producing a substantive critique should not unilaterally determine its consequential authority.

A central component of this instantiation is *structured scrutiny*. Rather than asking a second evaluator to produce another full review, scrutiny targets the basis on which an existing judgment could acquire decision influence. As summarized in Table 5, we distinguish three dimensions that assess whether a judgment is supported, correctly interpreted, and proportionate to its intended decision consequence, without requiring a single ground-truth evaluation.

Table 5: Three dimensions of structured scrutiny for consequential review judgments.

Dimension	**Core question**
Evidential support	Is the consequential claim factually supported by the submission or other relevant evidence?
Interpretive validity	Does the judgment correctly interpret the submission and the issue being evaluated?
Decision proportionality	Does the identified issue justify the amount of decision influence assigned to it?

These three dimensions also provide a basis for interpreting agreement and disagreement. Agreement provides additional support for a judgment, whereas disagreement indicates that it requires closer examination. In the case of disagreement, the chair or editor can first compare the competing evaluations in terms of the three scrutiny dimensions. If needed, independent reviews and the relevant manuscript content can provide further evidence for adjudication. If the disagreement remains unresolved, the judgment may receive reduced authority, require clarification, or be prevented from determining the outcome by itself. Importantly, agreement does not establish that a review is correct, just as disagreement does not establish that it is wrong. Our position aims to prevent unresolved consequential judgments from exercising unchecked authority, rather than to claim that scrutiny signals reveal a "more correct" evaluation.

One concern is that a dual-track design may increase human reviewer workload. The workload implications, however, depend on the institutional realization. An all-human implementation introduces additional human scrutiny, although checking an existing judgment can be lighter than conducting another full review. In the AI–human realization studied here (Track A), some full human reviewing can instead be replaced by supervision of an existing AI critique, potentially resulting in a lighter reviewer-side workload. Chair and editor workload remains uncertain: more structured signals must be considered, but unsupported or conflicting judgments may also become easier to identify. We analyze the workload implications of different realizations in Appendix A.

To examine the operational feasibility of the dual-track human–AI mutual supervision design, we instantiated both tracks on one real borderline submission using four original peer reviews, four recruited human supervisors, one experienced area chair, and ChatGPT-5.1. The study provides preliminary evidence for several aspects of the design. In Track A, all four human supervisors reported that supervising the AI-generated critique was easier than conventional reviewing, including two who voluntarily consulted the original manuscript. In Track B, AI scrutiny produced explicit counter-signals by identifying score–reasoning inconsistencies and potential review-quality concerns, while human and AI judgments also diverged in some cases. These counter-signals remained visible for area-chair adjudication rather than being resolved automatically. The area chair reported neither a clear increase nor decrease in adjudication difficulty, consistent with the uncertain chair/editor workload discussed above. Detailed settings, results, participant feedback, and limitations are provided in Appendix B.

This feasibility study establishes operational plausibility rather than venue-scale effectiveness. It shows that the proposed role separation, structured scrutiny, reciprocal checking, and chair-level adjudication can be instantiated in a complete review process, but evaluating broader effects requires institutionally representative studies across reviewer populations, research fields, incentives, policies, and decision contexts. Building on these open questions, we outline a roadmap for developing and validating authority-constrained review, including improved scrutiny design, suitable peer-review datasets, more reliable counter-signal generation and adjudication, less-biased AI sup-

port, safeguards against technical misconduct, and institutionally representative pilot studies. These objectives are detailed in Appendix C.

6 Scope

First, this paper does not argue that scientific evaluation should cease because of uncertainty, nor that current peer review is broken. Our argument is institutional rather than accusatory: under present conditions, peer-review reform cannot safely rely solely on selecting trusted evaluators.

Second, this paper does not argue that AI is ineffective or that improving evaluators is futile. Better reviewer training, incentives, and AI systems may all improve the signals available to the system and, in some settings, substantially improve decision quality. However, such gains should be understood as improvements to evaluative inputs rather than as solutions to evaluative authority. Better signals do not, by themselves, justify unconstrained decision influence.

Third, this paper does not argue that expertise is irrelevant or that all signal sources should be treated identically. Expertise matters, and final responsibility cannot be eliminated. Our claim is narrower and more practical: under missing criteria ground truth, expertise alone should not exempt consequential judgments from structured checking, explicit challenge, and possible reversal.

Fourth, reducing reviewer workload is not the primary claim of authority-constrained review, although it is an important consideration in its practical realization. The central aim is to govern how consequential evaluative influence is exercised under missing criteria ground truth. Workload implications depend on the institutional design, and our human–AI instantiation illustrates how this governance objective can be pursued while also reducing reviewer-side workload.

7 Conclusion

Peer review operates under missing criteria ground truth: many criteria that shape consequential decisions cannot be verified against stable standards at the time of review. This condition limits approaches that seek better decisions primarily by adding signals, improving evaluators, or scaling evaluative support, because these approaches do not by themselves prevent uncertain or weakly supported judgments from acquiring unchecked authority in final deliberation. Our position, therefore, departs from efforts to reduce uncertainty and instead governs the conditions under which review judgments become decision-relevant. Its aim is to ensure that consequential judgments are exposed to scrutiny, challenge, and possible correction before they carry full authority in final decisions. For large-scale computer-science conferences, this position addresses the downstream effects of variable review quality and reviewer-assignment dependence.

More broadly, our argument in this position paper becomes more urgent as AI agents accelerate research production, manuscript preparation, and review assistance. The pressures now visible in computer-science conferences may therefore foreshadow challenges that other fields will face. Computer science is not only a domain under review pressure, but also an early testbed for institutional reform. Venues such as NeurIPS, where scale, AI-assisted research, and peer-review governance already intersect, are well positioned to pilot authority-constraining mechanisms. The practical task is not to wait for perfect evaluators, complete ground truth, or fully settled policies, but to test and refine institutional designs that make consequential review authority more accountable under persistent uncertainty.

Acknowledgments and Disclosure of Funding

This work was partially supported by JSPS KAKENHI Grant JP24H00732, by JST CREST Grants JPMJCR20D3 and JPMJCR2562 including AIP challenge program, and by JST K Program Grant JPMJKP24C2 Japan. It was also funded by the German Federal Ministry of Education and Research (BMBF) through the DAAD (German Academic Exchange Service) and a fellowship by the Hanns-Seidel-Foundation. It was also supported by the Taiwan National Science and Technology Council (NSTC) with Grant 114-2634-F-001-001-MBK, and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 565115197 (BARI).

References

[1] Kyle Siler, Kirby Lee, and Lisa Bero. Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, 112(2):360–365, 2015.

[2] Balazs Aczel, Ann-Sophie Barwich, Amanda B Diekman, Ayelet Fishbach, Robert L Goldstone, Pablo Gomez, Odd Erik Gundersen, Paul T von Hippel, Alex O Holcombe, Stephan Lewandowsky, et al. The present and future of peer review: Ideas, interventions, and evidence. *Proceedings of the National Academy of Sciences*, 122(5):e2401232121, 2025.

[3] Richard L Kravitz, Peter Franks, Mitchell D Feldman, Martha Gerrity, Cindy Byrne, and William M Tierney. Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care? *PloS One*, 5(4):e10072, 2010.

[4] Stephen A Gallo, Joanne H Sullivan, and Scott R Glisson. The influence of peer reviewer expertise on the evaluation of research funding applications. *PloS One*, 11(10):e0165147, 2016.

[5] Janine Huisman and Jeroen Smits. Duration and quality of the peer review process: the author's perspective. *Scientometrics*, 113(1):633–650, 2017.

[6] Leanne Tite and Sara Schroter. Why do peer reviewers decline to review? a survey. *Journal of Epidemiology & Community Health*, 61(1):9–12, 2007.

[7] Marijke Breuning, Jeremy Backstrom, Jeremy Brannon, Benjamin Isaak Gross, and Michael Widmeier. Reviewer fatigue? why scholars decline to review their peers' work. *PS: Political Science & Politics*, 48(4):595–600, 2015.

[8] Charles W Fox, Arianne YK Albert, and Timothy H Vines. Recruitment of reviewers is becoming harder at some journals: a test of the influence of reviewer fatigue at six journals in ecology and evolution. *Research Integrity and Peer Review*, 2(1):3, 2017.

[9] Mohamed L Seghier. Paying reviewers and regulating the number of papers may help fix the peer-review process. *F1000Research*, 13:439, 2025.

[10] Jan-Ole Hesselberg, Therese K Dalsbø, Hilde Stromme, Ida Svege, and Atle Fretheim. Reviewer training for improving grant and journal peer review. *Cochrane Database of Systematic Reviews*, 11(11), 2023.

[11] Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332, 2025.

[12] Kevin Leyton-Brown, Mausam, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, and Dinesh Raghu. Matching papers and reviewers at large conferences. *Artificial Intelligence*, 331:104119, 2024.

[13] Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. In *International Conference on Machine Learning (ICML)*, pages 81634–81651. PMLR, 2025.

[14] Jing Yang. Position: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. In *International Conference on Machine Learning (ICML)*, pages 82399–82408. PMLR, 2025.

[15] Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15651–15682, 2025.

[16] Jisoo Lee, Jieun Lee, and Jeong-Ju Yoo. The role of large language models in the peer-review process: opportunities and challenges for medical journal reviewers and editors. *Journal of Educational Evaluation for Health Professions*, 22:4, 2025.

[17] Carole J Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015.

[18] Sven E Hug. How do referees integrate evaluation criteria into their overall judgment? evidence from grant peer review. *Scientometrics*, 129(3):1231–1253, 2024.

[19] Kristin Oxley and Magnus Gulbrandsen. Variability and negligence: grant peer review panels evaluating impact ex ante. *Science and Public Policy*, 52(2):254–268, 2025.

[20] Amanda Sizo, Adriano Lino, Álvaro Rocha, and Luís Paulo Reis. Defining quality in peer review reports: a scoping review. *Knowledge and Information Systems*, 67(8):6413–6460, 2025.

[21] Cecilia Superchi, José Antonio González, Ivan Solà, Erik Cobo, Darko Hren, and Isabelle Boutron. Tools used to assess the quality of peer review reports: a methodological systematic review. *BMC Medical Research Methodology*, 19(1):48, 2019.

[22] Elizabeth L Pier, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J Nathan, Chloe E Ford, and Molly Carnes. Low agreement among reviewers evaluating the same nih grant applications. *Proceedings of the National Academy of Sciences*, 115(12):2952–2957, 2018.

[23] Malcolm Macleod, Andrew M Collings, Chris Graf, Veronique Kiermer, David Mellor, Sowmya Swaminathan, Deborah Sweet, and Valda Vinson. The mdar (materials design analysis reporting) framework for transparent reporting in the life sciences. *Proceedings of the National Academy of Sciences*, 118(17):e2103238118, 2021.

[24] Sally Hopewell, An-Wen Chan, Gary S Collins, Asbjørn Hróbjartsson, David Moher, Kenneth F Schulz, Ruth Tunn, Rakesh Aggarwal, Michael Berkwits, Jesse A Berlin, et al. Consort 2025 explanation and elaboration: updated guideline for reporting randomised trials. *BMJ*, 389:e081124, 2025.

[25] Anthony Chauvin, Philippe Ravaud, Gabriel Baron, Caroline Barnes, and Isabelle Boutron. The most important tasks for peer reviewers evaluating a randomized controlled trial are not congruent with the tasks most often requested by journal editors. *BMC Medicine*, 13(1):158, 2015.

[26] SN Rodda, R Bijker, SS Merkouris, J Landon, CO Hawker, and NA Dowling. How to peer review quantitative studies, qualitative studies, and literature reviews: Considerations from the 'other'side. *Current Addiction Reports*, 11(5):771–782, 2024.

[27] Jessie McGowan, Margaret Sampson, Douglas M Salzwedel, Elise Cogo, Vicki Foerster, and Carol Lefebvre. Press peer review of electronic search strategies: 2015 guideline statement. *Journal of Clinical Epidemiology*, 75:40–46, 2016.

[28] Daniel Nüst and Stephen J Eglen. Codecheck: an open science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. *F1000Research*, 10:253, 2021.

[29] Simon Knight, Olga Viberg, Manolis Mavrikis, Vitomir Kovanović, Hassan Khosravi, Rebecca Ferguson, Linda Corrin, Kate Thompson, Louis Major, Jason Lodge, et al. Emerging technologies and research ethics: Developing editorial policy using a scoping review and reference panel. *PloS One*, 19(10):e0309715, 2024.

[30] Fabrice Frank, Nans Florens, Gideon Meyerowitz-Katz, Jérôme Barriere, Éric Billy, Véronique Saada, Alexander Samuel, Jacques Robert, and Lonni Besançon. Raising concerns on questionable ethics approvals–a case study of 456 trials from the institut hospitalo-universitaire méditerranée infection. *Research Integrity and Peer Review*, 8(1):9, 2023.

[31] Wolfgang Glänzel, Balázs Schlemmer, and Bart Thijs. Better late than never? on the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58(3):571–586, 2003.

[32] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[33] Terttu Luukkonen. Conservatism and risk-taking in peer review: Emerging erc practices. *Research Evaluation*, 21(1):48–60, 2012.

[34] Joe Weinberg and Marek Steedman. What to do for reviewer two: survey results on the preferences of peer reviewers. *European Political Science*, 23(3):338–348, 2024.

[35] Noémie Aubert Bonn and Wim Pinxten. Advancing science or advancing careers? researchers' opinions on success indicators. *PLoS One*, 16(2):e0243664, 2021.

[36] Boon Han Lim, C D'Ippoliti, M Dominik, AC Hernández-Mondragón, K Vermeir, KK Chong, H Hussein, VS Morales-Salgado, KJ Cloete, JN Kimengsi, et al. Regional and institutional trends in assessment for academic promotion. *Nature*, 638(8050):459–468, 2025.

[37] Jennifer S Trueblood, David B Allison, Sarahanne M Field, Ayelet Fishbach, Stefan DM Gaillard, Gerd Gigerenzer, William R Holmes, Stephan Lewandowsky, Dora Matzke, Mary C Murphy, et al. The misalignment of incentives in academic publishing and implications for journal reform. *Proceedings of the National Academy of Sciences*, 122(5):e2401231121, 2025.

[38] Kenneth A Michelson. Association of publication record and independent nih funding. *Plos One*, 17(6):e0269283, 2022.

[39] David B Resnik and Susan A Elmore. Conflict of interest in journal peer review. *Toxicologic Pathology*, 46(2):112–114, 2018.

[40] Federico Bianchi and Flaminio Squazzoni. Can transparency undermine peer review? a simulation model of scientist behavior under open peer review. *Science and Public Policy*, 49(5):791–800, 2022.

[41] Guido Buenstorf, Hanna Hottenrott, Anna Kosmützky, Georg Krücken, Frank Meier, Thomas Schaper, and Uwe Schimank. Analyzing constellations of multiple competition in science and higher education. *Minerva*, pages 1–27, 2025.

[42] Simone Righi and Károly Takács. The miracle of peer review and development in science: An agent-based model. *Scientometrics*, 113(1):587–607, 2017.

[43] Burak Kocak, Mehmet Ruhi Onur, Seong Ho Park, Pascal Baltzer, and Matthias Dietzel. Ensuring peer review integrity in the era of large language models: A critical stocktaking of challenges, red flags, and recommendations. *European Journal of Radiology Artificial Intelligence*, 2:100018, 2025.

[44] Rafael D'Andrea and James P O'Dwyer. Can editors save peer review from peer reviewers? *PloS One*, 12(10):e0186111, 2017.

[45] Paul Rozin and Edward B Royzman. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4):296–320, 2001.

[46] Ryan Liu, Steven Jecmen, Vincent Conitzer, Fei Fang, and Nihar B Shah. Testing for reviewer anchoring in peer review: A randomized controlled trial. *PloS One*, 19(11):e0301111, 2024.

[47] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131, 1974.

[48] Stephen Gallo, Lisa Thompson, Karen Schmaling, and Scott Glisson. Risk evaluation in peer review of grant applications. *Environment Systems and Decisions*, 38(2):216–229, 2018.

[49] Jan-Ole Hesselberg, Pål Ulleberg, Øystein Sørensen, Knut Inge Fostervold, Sigrid Hegna Ingvaldsen, and Ida Svege. Registered report protocol: Factors associated with inter-rater agreement in grant peer review. *Plos One*, 20(5):e0322696, 2025.

[50] Elena A Erosheva, Patrícia Martinková, and Carole J Lee. When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3):904–919, 2021.

[51] Mikael Fogelholm, Saara Leppinen, Anssi Auvinen, Jani Raitanen, Anu Nuutinen, and Kalervo Väänänen. Panel discussion does not improve reliability of peer review for medical research grant proposals. *Journal of Clinical Epidemiology*, 65(1):47–52, 2012.

[52] Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.

[53] Terne Thorn Jakobsen and Anna Rogers. What factors should paper-reviewer assignments rely on? community perspectives on issues and ideals in conference peer-review. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (ACL): Human Language Technologies*, pages 4810–4823, 2022.

[54] Yifei Li, Xiaoting Xu, Dongqing Lyu, Zhen Zhang, Juan Xie, and Ying Cheng. Developing a criteria framework for peer review: A critical interpretive synthesis. *Learned Publishing*, 38(3):e2016, 2025.

[55] Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182, 2006.

[56] Carl T Bergstrom and Kevin Gross. Screening, sorting, and the feedback cycles that imperil peer review. *PLoS Biology*, 24(2):e3003650, 2026.

[57] Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. Program chairs' report on peer review at acl 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages xl–lxxv, 2023.

[58] Yanbo Zhang, Sumeer A Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1:14, 2025.

[59] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

[60] Nicolas Bougie and Narimawa Watanabe. Generative reviewer agents: Scalable simulacra of peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP): Industry Track*, pages 98–116, 2025.

[61] Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 29330–29355, 2025.

[62] Juhwan Choi, JungMin Yun, Changhun Kim, and YoungBin Kim. Position paper: How should we responsibly adopt llms in the peer review process? In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 151–165, 2026.

[63] Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986, 2022.

[64] Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Moprd: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35(34):24191–24206, 2023.

[65] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, 2018.

[66] Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. Nlpeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (volume 1: Long papers)*, pages 5049–5073, 2023.

[67] Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. The good, the bad and the constructive: Automatically measuring peer review's utility for authors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 28991–29021, 2025.

[68] Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chavvi Kirtani, Murari Mandal, and Dhruv Kumar. Revieweval: An evaluation framework for ai-generated reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20542–20564, 2025.

[69] Nicholas Lo Vecchio. Personal experience with ai-generated peer reviews: a case study. *Research Integrity and Peer Review*, 10(1):4, 2025.

[70] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *International Conference on Machine Learning (ICML)*, pages 29575–29620. PMLR, 2024.

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

[72] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[73] Mohammad Hosseini and Serge PJM Horbach. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8(1):4, 2023.

[74] Tzu-Ling Lin, Wei-Chih Chen, Teng-Fang Hsiao, Hou-I Liu, Ya-Hsin Yeh, Yu-Kai Chan, Wen-Sheng Lien, Po-Yen Kuo, Philip S. Yu, and Hong-Han Shuai. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4819–4839, 2025.

[75] Kayvan Kousha and Mike Thelwall. Artificial intelligence to support publishing and peer review: A summary and review. *Learned Publishing*, 37:4–12, 2024.

[76] Nafiseh Ghaffar Nia, Amin Amiri, Yuan Luo, and Adrienne Kline. Ethical perspectives on deployment of large language model agents in biomedicine: a survey. *AI and Ethics*, 6:32, 2026.

[77] Giuseppe Russo, Manoel Horta Ribeiro, Tim Ruben Davidson, Veniamin Veselovsky, and Robert West. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. In *Proceedings of the ACM on Human-Computer Interaction (ACM HCI)*, volume 9, pages 1–28, 2025.

[78] Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for information Science and Technology*, 64(1):2–17, 2013.

[79] Luca Demetrio, Giovanni Apruzzese, Kathrin Grosse, Pavel Laskov, Emil Lupu, Vera Rimmer, and Philine Widmer. Gen-review: A large-scale dataset of ai-generated (and human-written) peer reviews. *arXiv preprint arXiv:2510.21192*, 2025.

[80] Shruti Singh, Mayank Singh, and Pawan Goyal. Compare: a taxonomy and dataset of comparison discussions in peer reviews. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 238–241. IEEE, 2021.

[81] Yaohui Zhang, Haijing Zhang, Wenlong Ji, Tianyu Hua, Nick Haber, Hancheng Cao, and Weixin Liang. From replication to redesign: Exploring pairwise comparisons for llm-based peer review. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[82] Ivan Stelmakh, Nihar B Shah, and Aarti Singh. Catch me if i can: Detecting strategic behaviour in peer assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 4794–4802, 2021.

[83] Tirthankar Ghosal, Kamal Kaushik Varanasi, and Valia Kordoni. Hedgepeer: A dataset for uncertainty detection in peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–5, 2022.

[84] Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. Automatic analysis of substantiation in scientific peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10198–10216. Association for Computational Linguistics, 2023.

[85] Prabhat Kumar Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. Politepeer: does peer review hurt? a dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation*, 58(4):1291–1313, 2024.

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

[87] Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, et al. Reviewrl: Towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16931–16943, 2025.

[88] Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to llm-as-a-judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (ACM CCS)*, pages 660–674, 2024.

[89] Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708*, 2024.

[90] Matteo Gioele Collu, Umberto Salviati, Roberto Confalonieri, Mauro Conti, and Giovanni Apruzzese. Misleading large language models used (or misused) in scientific peer-reviewing via hidden prompt-injection attacks. *ACM Transactions on AI Security and Privacy*, 2026.

[91] Elizabeth Gibney. Scientists hide messages in papers to game ai peer review. *Nature*, 643(8073):887–888, 2025.

[92] Vyas Raina, Adian Liusie, and Mark Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7499–7517, 2024.

[93] Iddo Drori and Dov Te'eni. Human-in-the-loop ai reviewing: feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109, 2024.

[94] Julien Piet, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David Wagner. Jatmo: Prompt injection defense by task-specific finetuning. In *European Symposium on Research in Computer Security (ESORICS)*, pages 105–124. Springer, 2024.

[95] Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1831–1847, 2024.

[96] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3419–3448, 2022.

[97] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9004–9017, 2023.

[98] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2024.

[99] Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98:101830, 2023.

[100] Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.

[101] Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2131–2137, 2019.

[102] Gautam Choudhary, Natwar Modani, and Nitish Maurya. React: A review comment dataset for actionability (and more). In *International Conference on Web Information Systems Engineering (WISE)*, pages 336–343. Springer, 2021.

[103] Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. Argument mining driven analysis of peer-reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 4758–4766, 2021.

[104] Abdur Rahman Bin Mohammed Faizullah, Ashok Urlana, and Rahul Mishra. Limgen: Probing the llms for generating suggestive limitations of research papers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 106–124. Springer, 2024.

[105] Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.

[106] Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. When reviewers lock horns: Finding disagreements in scientific peer reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16693–16704, 2023.

[107] Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. Aries: A corpus of scientific paper edits made in response to peer reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 6985–7001, 2024.

[108] Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z Li. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*, 2024.

[109] Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*, 2024.

[110] Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major nlp conference. In *Proceedings of NAACL-HLT*, pages 1274–1290, 2019.

[111] Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. Ape: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, 2020.

[112] Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. Disapere: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, pages 1234–1249, 2022.

[113] Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14479–14495, 2023.

[114] Miao Li, Eduard Hovy, and Jey Lau. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, 2023.

[115] Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. Llms assist nlp researchers: Critique paper (meta-) reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5081–5099, 2024.

[116] Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1208–1226, 2024.

A Human Workload Analysis and Comparison Across Peer-Review Designs

The workload implications of authority-constrained review depend on its institutional realization. Compared with conventional independent human review, an all-human realization of mutual supervision increases human workload, whereas the dual-track human–AI realization described in Section 5 can reduce reviewer-side human workload. Table 6 summarizes the workload implications across designs and roles.

Table 6: Expected workload effects in a dual-track example.

Role	Workload
Human reviewers	Fewer full reviews
Human supervisors	Lighter than writing a review from scratch
AI	Additional computation with little human reviewing labor
Chair / editor	Uncertain: more signals, but with greater structure

Conventional independent human review. Consider a conventional design in which four human reviewers independently evaluate each submission. Let W_{full} denote the workload required for one full human review. The reviewer-side human workload is

$$W_{\text{conv}} = 4W_{\text{full}}.$$

This provides the baseline against which we compare realizations of mutual supervision.

All-human mutual supervision. Authority-constrained review does not depend on AI adoption. One possible realization retains four full human reviewers and assigns an independent human supervisor to scrutinize each review judgment, yielding four final review judgments. Let W_{sup} denote the workload required to perform this second-order scrutiny. Reviewer-side human workload then becomes

$$W_{\text{human}} = 4W_{\text{full}} + 4W_{\text{sup}}.$$

Because $W_{\text{sup}} > 0$, this realization necessarily requires more human effort than conventional independent review. Structured scrutiny can be lighter than a full review because it evaluates an existing judgment rather than the submission from scratch.

Dual-track human–AI mutual supervision. In the human–AI dual-track realization (Section 5), two full human reviews receive AI supervision, while one AI-generated critique is independently supervised by two humans, again yielding four final review judgments.

$$W_{\text{Human-AI}} = 2W_{\text{full}} + 2W_{\text{sup}}.$$

Compared with conventional independent review,

$$W_{\text{Human-AI}} < W_{\text{conv}} \quad \text{whenever} \quad W_{\text{sup}} < W_{\text{full}}.$$

Our feasibility study provides preliminary empirical support for this workload condition (Appendix B.2).

Chair and editor workload. Reviewer-side workload does not determine the workload of chairs or editors. Mutual supervision produces additional structured signals that may require adjudication, but it may also make unsupported or conflicting judgments easier to identify. The net effect therefore remains uncertain; in our feasibility study, the area chair reported neither a clear increase nor decrease in adjudication difficulty (Appendix B.2).

Scaling to venue-level deployment. Because evaluative uncertainty persists across submissions and reviewers, all consequential judgments should in principle remain subject to scrutiny. Practical deployment, however, may need to allocate scrutiny unevenly when resources are limited or when a venue first pilots the mechanism. Under tighter resource constraints, scrutiny can prioritize judgments with greater potential decision influence, such as clear accept or reject judgments, decision-critical claims, explicitly challenged judgments, or strong cross-review inconsistencies.

B Feasibility Study of Dual-Track Human–AI Mutual Supervision

In this section, we present a feasibility study illustrating how the dual-track human–AI mutual supervision design can be instantiated on a selected ICLR'25 rejected submission[4]. In addition to the four original human reviewers from ICLR'25, we recruited four new human supervisors with diverse backgrounds and one experienced area chair. The study instantiates both tracks end to end and examines the counter-signals, workload implications, and adjudication process.

Declaration

1. The study uses a single submission and does not treat either the original human judgments or the AI-generated judgments as ground truth. Differences between them are interpreted as counter-signals for further scrutiny rather than evidence that either side is correct.
2. The results (Section B.2) and discussion (Section B.3) include hypotheses about possible reviewer behavior. These interpretations are derived from the observed patterns and should not be read as claims about participants' actual psychological states.

B.1 Settings

The selected paper from ICLR 2025. We selected an ICLR'25 rejected submission[4] for which we have no COI. The paper received ratings of 3/6/6/8 with corresponding confidences of 4/3/4/4 (maximum 5), following ICLR's official rating scale:

- 10: strong accept, should be highlighted at the conference
- 8: accept, good paper
- 6: marginally above the acceptance threshold
- 5: marginally below the acceptance threshold
- 3: reject, not good enough
- 1: strong reject

We chose this paper because it combined a borderline average score with substantial disagreement among reviewers. Its average score was 5.75, but the individual ratings ranged from a clear reject to a clear accept. This combination makes it a useful setting for examining how independent scrutiny operates when the aggregate judgment is borderline but the underlying reviewer assessments span substantially different judgments.

Participants. The study uses review comments from four original anonymous ICLR reviewers and involves four newly recruited human supervisors, one human area chair, and one LLM (ChatGPT 5.1, used at the time of the study).

The four original anonymous ICLR review comments serve as the human critiques in Track B and are independently examined through AI supervision. We did not provide their original ratings or confidence scores to the LLM; these values were used only for post-hoc comparison between the original and AI-inferred judgments.

We recruited four additional human supervisors for Track A: two AI security researchers, one machine learning researcher, and one Ph.D. student in LLM security. Each supervisor independently assessed the same AI-generated critique, assigned a rating to the submission, and provided a justification. They could consult the original manuscript at their discretion.

ChatGPT 5.1 in standard reasoning mode served as the AI system. We disabled memory and used independent interactions so that no prior conversation history affected its outputs. The model participated in both tracks:

- *Track A:* The AI generated a substantive critique without a numerical rating based solely on the anonymous manuscript and a structured prompt. We adapted the review-guideline template from AAAI 2026[5] for this task (see Section E.1).

[4] https://openreview.net/forum?id=fFtmpqLFvw

[5] https://docs.google.com/document/u/0/d/1tqQGwtNUlALPSTqoTo5uTFx8vKuqpILNTne9jeBCOVI/mobilebasic

- *Track B:* The AI independently inferred the rating most consistent with each human critique and performed structured scrutiny of that critique, referring to the manuscript when necessary and providing justification (see Section E.2). For this feasibility study, structured scrutiny was operationalized using the provisional failure-mode checklist in Table 7.

Table 7: Provisional failure modes used as an operational checklist for structured scrutiny in the feasibility study, with illustrative examples.

Failure Mode	Description and Illustrative Example
Overly Harsh	Rejects a paper over minor or peripheral issues. "The paper should be rejected because it has typos," despite valid empirical results.
Overly Gentle	Accepts or supports a paper without sufficient scrutiny. "This is a promising direction," despite methodological errors.
Unconfident	Uses vague or hedged judgments that avoid taking a clear stance. "It seems like the contribution might be useful, but I am not sure."
Superficial	Provides generic comments that do not substantively engage with the manuscript. "The writing needs improvement," without identifying any concrete issues.
Inaccurate / Hallucinatory	Makes claims about errors or missing content that are not supported by the manuscript. "The paper does not have ablation studies," although it does.
Subjective	Presents personal preference as an objective assessment. "I think method X is better than this proposed method."
Misguided	Bases criticism on irrelevant material or a misinterpretation of the submission. "The paper should compare against method X," although X addresses a different setting.

Finally, we recruited an experienced area chair, with service across ICML, ICLR, NeurIPS, ACM Multimedia, and AAAI, to perform the final meta-review assessment.

Process. Table 8 summarizes the inputs and outputs of the study. Detailed prompts and intermediate outputs are provided in Section E. We do not include an author rebuttal or subsequent discussion because the original authors and anonymous ICLR reviewers were not participants in this study.

Table 8: Process of the dual-track human–AI mutual supervision feasibility study.

Track	Stage	Participants	Inputs	Outputs
Track A	AI critique	AI	Manuscript, prompt	Review comment (no rating)
	Human supervision	Reviewer #1–4	Review comment, manuscript[1]	Rating, justification
Track B	Human critique	Reviewer #5–8	Manuscript	Review comment (no rating)
	AI rating	AI	Review comment, prompt	Rating, justification
	Structured scrutiny	AI	Review comment, prompt, manuscript	Failure flags, justification
Final adjudication	Meta-review	Area chair	Review summary, manuscript[1]	Meta-review comment

[1] Human supervisors and the area chair could refer to the manuscript at their discretion.

Questionnaire and additional AI–AI analysis. After completing the study, we administered a short questionnaire to the four recruited human supervisors and the area chair:

1. Does this new protocol make your task easier or more difficult?
2. Does this new protocol improve or reduce the quality of your output?

After all four supervisors completed their tasks, we added a post-hoc question:

3. Did you refer to the original paper when assigning ratings?

We also conducted an additional AI–AI analysis in which the AI rated and scrutinized its own generated critique. This analysis is not part of the proposed dual-track design; it provides an additional observation about possible self-preference when critique generation and scrutiny are not independently separated.

B.2 Results

Overall. Table 9 summarizes the review judgments produced in the study. The four human supervisors in Track A assigned ratings of 5/5/5/6 based on the AI-generated critique. In Track B, ratings inferred by the AI from the four original human critiques were 6/8/6/8, compared with the original ICLR ratings of 3/6/6/8. AI scrutiny also produced explicit counter-signals for all four original

critiques, flagging combinations of superficiality, inaccuracy, and misguidance for further adjudication. The area chair's final decision remained *reject*. These outcomes illustrate that human and AI judgments can diverge while remaining visible for final adjudication rather than being automatically reconciled.

Table 9: Detailed peer-review outcomes from the original ICLR'25 process and the human–AI mutual supervision feasibility study.

Participant		ICLR'25		Human–AI Mutual Supervision	
Review	Supervision	Rating (after rebuttal)	Confidence	Rating (no rebuttal)	Flag
AI	Reviewer #1	N/A	N/A	5	N/A
AI	Reviewer #2	N/A	N/A	5	N/A
AI	Reviewer #3	N/A	N/A	5	N/A
AI	Reviewer #4	N/A	N/A	6	N/A
Reviewer #5	AI	3	4	6	Superficial; Inaccurate; Misguided
Reviewer #6	AI	6	3	8	Superficial
Reviewer #7	AI	6	4	6	Inaccurate; Superficial
Reviewer #8	AI	8	4	8	Inaccurate (minor)
Area Chair		Reject		Reject	
AI	AI[1]	N/A	N/A	8	None

[1] Additional analysis only; AI–AI supervision is not part of the proposed dual-track design.

Track A: AI critique, human supervision. The four human supervisors assigned ratings of 5/5/5/6 based on the shared AI-generated critique, compared with the original ICLR ratings of 3/6/6/8. One possible explanation for the narrower spread is that all four supervisors received the same detailed critique, which enumerated many strengths and weaknesses and may have anchored judgments toward the borderline.

This observation highlights a design challenge for AI-generated critiques. A highly detailed critique can provide broad evidential coverage but may also shape subsequent human judgments through the selection and presentation of issues. Conversely, a shorter critique may reduce such influence while omitting evidence useful for supervision. Designing AI critiques that provide sufficient support without exerting excessive anchoring therefore remains an important question for future work.

Track B: Human critiques, AI supervision. Track B revealed mismatches between the ratings implied by written critiques and the reviewers' original consequential judgments. Most notably, for Reviewer #5, the AI inferred a rating of 6 from a critique that had originally supported a rating of 3, while structured scrutiny also flagged that critique as superficial, inaccurate, and misguided. Reviewer #6 showed a smaller but still substantial mismatch, with an inferred rating of 8 compared with the original rating of 6. These cases show how AI supervision can surface review judgments whose assigned influence warrants closer chair-level examination, particularly when the rating appears stronger than the support articulated in the critique.

Meta review. The area chair ultimately recommended rejection, the same outcome as the original ICLR decision, after considering the outputs produced by the two tracks. The case therefore illustrates how conflicting scrutiny signals can be preserved through the review process and resolved at the chair level without automatic aggregation.

Feedback on the questionnaire. As summarized in Table 10, all four human supervisors found the supervision task easier than conducting a conventional full review. Two of them voluntarily consulted the original manuscript and still reported lower task difficulty. This provides preliminary empirical support for the reviewer-side workload condition discussed in Appendix A.

Three supervisors perceived an improvement in output quality, while one perceived a reduction. Reviewer #2, who consulted the original manuscript, specifically raised concerns that the AI critique could itself be biased or insufficiently validated. The feedback therefore suggests that lower reviewer-side workload can coexist with reservations about the reliability of the AI critique and the quality of the resulting evaluation.

Table 10: Feedback on the questionnaire.

Question	Reviewer #1	Reviewer #2	Reviewer #3	Reviewer #4	Area Chair
Q1: Does this new protocol make your task easier or more difficult?	Easier	Easier	Easier	Easier	Neither
Q2: Does this new protocol improve or reduce the quality of your output?	Improve	Reduce	Improve	Improve	Neither
Q3: Did you refer to the original paper when assigning ratings?	No	Yes	Yes	No	No

The area chair selected "Neither" for both task difficulty and output quality, providing no clear indication of reduced or increased chair-level workload. This is consistent with Appendix A, where chair and editor workload remains uncertain because additional structured signals must be adjudicated even as unsupported or conflicting judgments may become easier to identify.

AI supervises AI (outside the proposed mutual supervision design). As shown in the final row of Table 9, the AI assigned a rating of 8 to its own critique and detected no failure modes. In contrast, independently generated human critiques received several scrutiny flags. This asymmetry suggests a possible self-preference or confirmation effect when critique generation and scrutiny are performed by the same evaluator, further motivating the separation of evaluative roles in the proposed design.

B.3 Discussion — From Critique to Decision Influence

The feasibility study suggests that consequential ratings depend not only on which strengths and weaknesses are identified, but also on how evaluators weight their importance. Track A showed a shared tendency toward borderline judgments from the same detailed critique, whereas Track B revealed differences between the ratings inferred by AI from human-written critiques and the ratings originally assigned by those reviewers. Such variation is consistent with prior observations of differences in human and AI reviewing tendencies [77] and may also reflect broader effects such as negativity bias [45] or review cultures that emphasize flaw identification over contribution assessment [78].

This distinction is central to decision proportionality. Identifying a weakness does not determine how much decision influence that weakness should receive. Authority-constrained review does not recover a ground-truth evaluation or establish which evaluator is correct; instead, it makes the transition from observed evidence to consequential judgment explicit and subject to independent scrutiny before that judgment acquires decision authority.

B.4 Limitations of the Feasibility Study

The study is limited to one submission, four recruited human supervisors, one area chair, and one LLM configuration. It also omits author rebuttal and subsequent interaction with the original anonymous reviewers. In Track A, all supervisors received the same AI-generated critique, so their judgments may share biases introduced by that critique rather than being fully independent. Broader evaluation will therefore require larger and more representative studies across reviewers, models, fields, and institutional settings.

C Research and Deployment Roadmap

Developing authority-constrained review requires progress across scrutiny design, datasets, methodologies, system security, and institutional deployment. In this section, we outline concrete research objectives for developing and evaluating practical realizations of the position proposed in this paper.

C.1 Prerequisites

Objective I: Develop structured scrutiny. A central prerequisite is to operationalize the three dimensions of structured scrutiny introduced in Section 5: evidential support, interpretive validity, and decision proportionality. These dimensions provide a conceptual basis for examining conse-

quential review judgments, but practical implementations require methods that can identify relevant problems consistently across submissions, reviewers, and research fields.

The provisional failure modes used in our feasibility study (Table 7) provide one possible operational starting point. Future work should determine how such failure modes relate to the three scrutiny dimensions, which signals can be identified reliably, and which are most useful for subsequent chair or editor adjudication. In particular, decision proportionality requires examining not only whether a criticism is valid, but also whether the identified issue justifies the decision influence assigned to it.

Objective II: Establish datasets for structured scrutiny. Existing datasets (Section D) primarily repurpose raw records from venues such as ICLR and NeurIPS on OpenReview[6]. These datasets overlap substantially and provide limited coverage of high-quality AI reviews, problematic review comments, scrutiny signals, and subsequent adjudication.

Manual annotation is also difficult because many consequential review problems lack reliable ground truth. Program committees rarely establish that a review is objectively overly harsh or overly gentle. Disagreement may instead reflect either a genuine review problem or a legitimate difference in interpretation. LLMs, combined with well-defined failure modes, can generate synthetic problematic reviews at scale [79]. A central open question is how to verify that such reviews exhibit the intended failure mode rather than merely differing in style, strictness, or evaluative preference.

Datasets for authority-constrained review should therefore preserve disagreement, provenance, and uncertainty where possible. They should also cover human and AI critiques, consequential ratings, scrutiny signals, and subsequent adjudication.

C.2 Methodologies

Objective III: Develop AI supervision methods. AI supervision requires two core capabilities. First, the AI should infer the rating implied by a human review comment, allowing inconsistencies between the written critique and the consequential rating to be identified. Existing work shows that this remains difficult when models rely only on review text, particularly for long comments [71]. Approaches that additionally use paper content or abstracts perform better [65, 80, 66, 81], suggesting that effective supervision may require access to both the review and the submission.

Second, the AI should identify potential problems in the review itself. Existing studies address individual categories, including overly harsh reviews [82], unconfident language [83], superficial assessments [84], and impolite tone [85]. More recently, Thakkar et al. [86] deployed an LLM-based review-feedback system at ICLR 2025 to identify issues such as vague comments, content misunderstandings, and unprofessional remarks, demonstrating the practical feasibility of AI-assisted review supervision at scale. Developing methods that can reliably cover a broader range of review failures and support structured scrutiny therefore remains an important open problem.

Objective IV: Develop less-biased and more reliable AI review support. Track A relies on AI-generated critiques, making their reliability an important research objective. AI reviews can exhibit systematic biases and evaluator-specific tendencies [79, 77], while broader concerns about bias in LLM-assisted peer review have also been raised [73].

Recent approaches such as ReviewRL [87] use reinforcement learning to improve review quality, factual grounding, and rating accuracy. Similar methods could support more reliable AI critiques when combined with the scrutiny methods in Objective III.

Critique presentation also matters. As suggested by our feasibility study, a detailed AI critique can provide broad evidential coverage while influencing which issues human supervisors attend to. Future work should therefore study both critique quality and how its structure affects downstream human judgments.

C.3 System Security

Objective V: Address technical misconduct. AI-assisted peer review creates new opportunities for technical manipulation. Relevant threats include:

[6] https://openreview.net

Prompt injection: explicit or hidden instructions embedded in a submission to influence AI behavior [88, 89, 90, 91].

Textual adversarial manipulation: small or strategically chosen edits intended to alter AI-generated evaluations [92, 74].

Data poisoning: manipulation of review-related training or update data to bias future model behavior [93].

Implicit manipulation: shaping text, structure, or presentation to exploit known AI-review tendencies without explicit instructions [89].

For authority-constrained review, two threat directions are particularly relevant:

- *Manipulating AI scrutiny*: influencing AI evaluation of an existing human judgment.
- *Manipulating AI critique*: influencing AI-generated critiques without substantively improving the submission.

Peer-review-specific defenses remain limited. Potential safeguards include task-specific defenses against prompt injection [94], restricting untrusted instruction channels [95], adversarial testing and red-teaming [96], consistency-based checks for unreliable outputs [97], and auditable records of AI-generated critiques and scrutiny signals [98].

These mechanisms should be evaluated specifically in peer-review settings. Transparency also creates a security trade-off: review prompts and procedures should support accountability and contestability without making manipulation unnecessarily easy.

C.4 Deployment

Objective VI: Pilot authority-constrained review. A crucial next step is to evaluate authority-constrained review in real peer-review settings. A practical near-term deployment can begin with AI scrutiny of existing human reviews, examining evidential support, rating–reasoning consistency, omissions, and potential review failures without replacing human reviewers or chairs. Recent large-scale deployment at ICLR 2025 shows that LLM-generated feedback on existing human reviews can be integrated into a live conference workflow [86].

Subsequent pilots can evaluate the full dual-track human–AI mutual supervision design described in Section 5, including reviewer and supervisor workload, chair or editor workload, the usefulness of counter-signals, disagreement resolution, anchoring effects, and robustness to technical misconduct.

We therefore call for conferences, journals, workshops, review platforms, and program committees to collaborate on controlled studies and venue-level pilots. When resources are limited, initial deployment can prioritize judgments with greater potential decision influence.

D Peer-Review Datasets

We compiled a broad collection of publicly available peer-review datasets, summarized in Table 11. Our review draws on three major surveys [99, 100, 11] as well as individual dataset papers. Because many datasets contain overlapping submissions or reviews, we focus on resources most relevant to authority-constrained review, including datasets with papers, human reviews, AI-generated reviews, rebuttals, or annotations related to review quality and bias. Among the available resources, GenReview [79] is particularly relevant because it contains both human and AI-generated reviews as well as bias-related annotations, although broader data are still needed for structured scrutiny and adjudication.

Table 11: Overview of peer-review datasets. All listed datasets are publicly accessible.

Dataset	Paper			Review				Rebuttal	Sources
	Abstract	TXT	PDF	Human	Meta-	AI	Biased	Author	
AMPERE [101]				14,202					ICLR 2017–2018 UAI 2018 NeurIPS 2013–2017 ACL 2017

Table 11 (continued)

Dataset	Paper			Review				Rebuttal	Sources
	Abstract	TXT	PDF	Human	Meta-	AI	Biased	Author	
ReAct [102]				1,250					ICLR 2018
AMSR [103]				12,135					ICLR 2019–2020 MIDL 2019–2020 NeuroAI 2019 GI 2020
HedgePeer [83]				2,966					ICLR 2018
LimGen [104]		4,068							ACL 2022 EACL 2022 EMNLP 2022–2023
PeerRead [65]		14,784		10,770					NIPS 2013–2017 ICLR 2017 ACL 2017 CoNLL 2016 arXiv 2007–2017
F1000RD [63]		3,000		8,053					F1000Research
ASAP-Review [105]		8,877		28,119	Include				ICLR 2017–2020 NeurIPS 2016–2019
ContraSciView [106]		8,582		25,825					ICLR 2017–2020 NeurIPS 2016–2019
ARIES [107]		1,720		4,088					OpenReview
ReviewMT [108]		26,841		92,017					ICLR 2017–2024 Nature 2023
NLPeer [66]			5,672	11,515					CONLL 2016, 2020 ARR 2022 ACL 2017 F1000 2022
Reviewer2 [109]			27,805	99,727	Include				CONLL 2016, 2020 ACL 2017 ARR 2022 ICLR 2017–2023 NeurIPS 2016–2022
TreeReview [15]			80	336					NeurIPS 2023 ICLR 2024
AIReviewLottery [77]			23,959	86,690					ICLR 2018–2024
ACL-2018 [110]				7,929				1,227	ACL 2018
RR [111]				4,764				4,764	ICLR 2013–2020
DISAPERE [112]				506				506	ICLR 2019–2020
JitsuPeer [113]				1,705				1,705	ICLR 2018–2020
PeerSum [114]	14,993			53,451	Include			5,854	ICLR 2018–2020 NeurIPS 2021–2022
MOPRD [64]			6,578	22,483	Include			11,213	Multiple journals Multiple domains
ReviewCritique [115]		100		380		60			ICLR 2020–2023 NeurIPS 2020–2023
AgentReview [116]			523			53,800			ICLR 2020–2023
SubstanReview [84]							550		CoNLL 2016 ACL 2017 COLING 2022 ARR 2022

Table 11 (continued)

Dataset	Paper			Review				Rebuttal	Sources
	Abstract	TXT	PDF	Human	Meta-	AI	Biased	Author	
PolitePEER [85]							2,500		ICLR NeurIPS Publons ShitMyReviewersSay
Gen-Review [79]			32,652	124,615		81,850	54,567		ICLR 2018–2025

E Full Review Process of Human–AI Mutual Supervision

E.1 Track A: AI Critique, Human Supervision

LLM Prompt

```
(upload anonymous manuscript)
Answer only based on this message; ignore all saved memories.
Please conduct a critical review of this manuscript according to the specified structure.  Avoid
giving any recommendation rating.
Paper summary (4-10 sentences):
The first paragraph of the review should be a summary of the paper.  The purpose of this summary
is to both show that the reviewer has a clear understanding of the paper and provide the best
possible interpretation before going into critiques.  When summarizing a paper, the should contain
the following elements (with each of these typically being 1 or 2 sentences):
      1. summarize the main contribution of the paper in one sentence
      2. identify the core problem being addressed by the paper
      3. describe the key idea of the paper and how it addresses the problem
      4. summarize how the idea of the paper is realized as an implementation
      5. identify the conclusion that is claimed (or could be claimed) from the findings
Review summary (4-10 sentences):
The second paragraph of the review should provide an overall assessment.  The first sentence of
this paragraph should provide the overall conclusion of the review.  This introductory sentence is
followed by individual sentences that assess the paper with respect to its clarity, technical and
experimental soundness/validity, intellectual novelty, and relevance to the field.  The high-order
points of feedback to improve the paper should also be included.
Specific points of feedback (bullet list, variable level of depth, divided into strengths and
weaknesses):
After the first two paragraphs, a bulleted list should be provided that addresses specific points
of feedback.  This list can be as long as needed to address all points of feedback for the review.
Please be sure these comments are both critical and helpful.  These points can be of varying length,
depending on the amount of description needed.  For example, points about technical shortcomings
tend to be about a paragraph in length.  In contrast, typographical and grammatical errors tend to be
expressed briefly in one line.  Review comments regarding the need for related work are expected to
provide at least three citations from non-overlapping authors to be considered substantive.
```

AI Review Comment – no rating

```
Paper summary:
The paper investigates how multi-turn human red teaming exposes vulnerabilities in safety-aligned
large language models (LLMs) that appear robust under standard single-turn, automated adversarial
evaluations.  The core problem addressed is that most defenses are evaluated only against single-turn
automated attacks, whereas real-world malicious use will involve multi-turn conversations in
black-box chat interfaces.  The key idea is to organize expert human red teamers into a structured
multi-stage pipeline that conducts multi-turn jailbreak attempts under a realistic, time-limited
threat model, then validate their success with human reviewers and a GPT-4o harm classifier.
Concretely, the authors attack four defenses (RR, LAT, DERTA, CYGNET) built on Llama-3-8B-Instruct
using HarmBench's 240 harmful behaviors, plus an RMU-based machine-unlearned model evaluated on
free-response WMDP-Bio questions; they compare human jailbreaks to six black- and white-box automated
attacks and then release successful conversations as the Multi-Turn Human Jailbreaks (MHJ) dataset,
labeled with a taxonomy of tactics.  Empirically, human multi-turn jailbreaks achieve much higher
attack success rates (up to 51--88% on open-source defenses and 70% on CYGNET) than individual or
ensembled automated attacks, and they substantially outperform automated attacks in recovering
dual-use biosecurity knowledge from an RMU-unlearned model.  The authors conclude that existing
automated robustness benchmarks underestimate real-world risk, that defenses are not robust to
realistic multi-turn, human-crafted attacks, and that MHJ can catalyze research on stronger defenses
and automated multi-turn red teaming.
Review summary:
Overall, the paper makes a timely and substantively important empirical contribution by demonstrating
that current LLM defenses remain highly vulnerable under realistic multi-turn human attacks, though
several methodological details and comparisons could be clarified or strengthened.  The exposition
is generally clear and well-structured, with helpful figures (e.g., the ASR bar plots on pages
```

2, 6, and 7 and the tactic taxonomy in Table 1) and concrete jailbreak examples illustrating key tactics such as obfuscation and "Opposite Day." The experimental setup is thoughtfully designed, but the comparison between human and automated attacks relies on different budgets and evaluation pipelines, which complicates interpretability of absolute ASR gaps. The multi-turn threat model and the MHJ dataset, including tactic labels and red teamer commentary, appear novel and valuable to the robustness community, especially in combination with the unlearning case study. At the same time, the work is limited by the small number of defenses, the lack of cost-normalized comparisons, the heavy reliance on a single harm classifier, and relatively shallow quantitative analysis of how multi-turn dynamics and tactics interact with specific defenses. Addressing these issues, along with tighter positioning relative to recent multi-turn and in-the-wild red teaming work, would significantly strengthen the paper's impact.

Strengths:

1. Realistic and well-motivated threat model
   - The paper focuses on black-box chat-interface access with multi-turn conversations and limited temperature control, which closely mirrors real-world usage of systems like ChatGPT or CYGNET rather than the more generous API or white-box threat models that many automated attacks assume.
   - The authors explicitly disallow access to assistant tokens, logit probabilities, and internal model states, and they consider post-circuit-breaker behavior for CYGNET separately, which helps keep the study aligned with realistic attacker capabilities.
2. Carefully designed human red teaming pipeline
   - The multi-stage pipeline with up to three independent red teamers, two human reviewers, and a final GPT-4o harmfulness filter is a thoughtful attempt to both encourage diverse attacks and control false positives (Figures 2 and 6, Section 3.2, Appendix A.3).
   - The 30-minute per-behavior time budget, separate "partial" and "full" success labels, and escalation to a third red teamer when reviewers disagree are sensible choices that partially standardize human efforts while still allowing creativity.
3. Strong empirical evidence of gaps between automated and human attacks
   - On HarmBench, human ASRs are dramatically higher than both individual automated attacks and the automated ensemble across all open-source defenses (e.g., for RR: 51.7% vs. 8.3% ensemble; for LAT: 75.4% vs. 10.0%; for DERTA: 87.9% vs. 68.3%; Table 2, Figures 1 and 3).
   - For the unlearning case study, humans achieve 39.6% higher ASR than the best automated attack and 25.6% higher than the automated ensemble on free-response WMDP-Bio questions, while all prior attacks reported in Li et al. (2024a) showed 0% ASR on CYGNET in a similar setup.
   - The breakdown by harm category (Figure 5) and by tactic (Figure 8) shows that some areas (e.g., harassment/bullying) are particularly easy for humans relative to automated attacks, strengthening the claim that automated ASRs are not a reliable proxy for human adversary performance.
4. Insightful analysis of multi-turn dynamics and tactic taxonomy
   - The observation that over 90% of successful HarmBench jailbreaks and 86% of successful WMDP-Bio jailbreaks use more than one turn provides direct evidence that multi-turn capabilities are central to real-world vulnerability, not just a convenience for humans.
   - The tactic taxonomy in Table 1 and its expansion in Appendix D (e.g., direct request, echoing, hidden intention streamline, obfuscation, output format, request framing with variants like fictionalization, "Opposite Day," roleplay, urgency) is rich and grounded in practical red teaming experience across multiple commercial engagements.
   - The qualitative red teamer surveys (Appendix C) provide nuanced insights into each defense's strengths and weaknesses, such as RR's "gibberish" mechanism and LAT's resistance to "Opposite Day," which will be useful to both defense designers and future attackers.
5. MHJ dataset and safety-conscious release
   - MHJ includes 2,912 prompts across 537 jailbreak conversations, with metadata such as tactic labels, time taken, and temperature, plus free-form red teamer commentary that captures design choices and ideas tried.
   - The authors remove model completions and export-controlled content, verify compliance with ITAR/EAR, and limit MHJ to system/user messages, which is a reasonable compromise between research usefulness and safety (Section 7).
   - The dataset nicely complements prior resources like HarmBench, WildTeaming-style in-the-wild datasets (Jiang et al., 2024b), and jailbreak taxonomies (Inie et al., 2023; Weidinger et al., 2024), and can be used to train or evaluate multi-turn attack generators.
6. Thoughtful discussion of limitations and threats to validity
   - Section 5 acknowledges many important caveats: different resource limits for humans vs. automated attacks, varying red teamer skill levels, different positions of GPT-4o in the evaluation pipeline (first-pass vs. final filter), and the higher cost of human red teaming compared to automated attacks.
   - The authors correctly caution against over-interpreting ASR differences as a direct ranking of defenses and emphasize that the goal is to highlight vulnerability and motivate stronger automated attacks, not to declare human red teaming "better."

Weaknesses and suggestions:

1. Comparability between human and automated attacks is still murky
   - The comparison between human and automated ASRs is confounded by multiple factors: humans have multi-turn access while automated attacks are single-turn; humans are time-limited but not query-limited; different temperatures are available to humans but not explicitly to automated attacks; and the harm-classification pipeline is slightly different.
   - Although the authors discuss many of these issues in Section 5 and Appendix A.3, the main figures (especially Figures 1 and 3) still invite a direct "humans beat automated attacks by X%" reading that may overstate the strength of the evidence.
   - Suggestion: Add a dedicated subsection in the main text that (i) explicitly enumerates these differences, (ii) presents at least one more controlled comparison|e.g., limit humans

to a single turn on a subset of behaviors, or simulate a multi-turn version of a simple automated attack like ZeroShot or PAIR|and (iii) reports cost-normalized metrics such as ASR per query or per minute of effort. Even partial results would help calibrate claims.

2. Limited coverage of defenses and baselines
   - The study evaluates three open-source defenses (RR, LAT, DERTA) plus the proprietary CYGNET and a single unlearning method (RMU), all on Llama-3-8B-Instruct.
   - There is no explicit baseline of the un-defended Llama-3-8B model's ASR under the same human multi-turn pipeline, making it hard to quantify how much robustness these defenses actually add in this threat model.
   - Suggestion: Include a baseline for the raw base model in the main figures/tables, or at least in an appendix, and discuss how much the defenses reduce ASR relative to that baseline under human multi-turn attacks. Additionally, even a small experiment on a different foundation model (e.g., a smaller or larger Llama variant) would help indicate whether the findings are tied to this specific backbone.
3. Heavy reliance on GPT-4o as a harm classifier
   - GPT-4o is used both as a first-pass filter for automated attacks and as a last-pass filter for human attacks. For automated attacks, its high recall but low precision is mitigated by human verification; for human attacks, its moderate precision is used to discard some human-labeled positives, providing a "conservative lower bound" on ASR.
   - However, the paper does not deeply analyze how sensitive results are to this classifier choice, nor how often GPT-4o's misclassifications qualitatively change the nature of the jailbreaks.
   - Suggestion: Add a small ablation where a subset of attacks are re-labeled with a different classifier and/or double-human annotation to estimate label variance. Reporting inter-annotator agreement between reviewers and GPT-4o on a random subset would also increase confidence.
4. Unlearning case study design differences deserve more emphasis
   - For WMDP-Bio, human red teamers get unlimited time up to a global 240-hour budget, the pipeline differs (no multi-reviewer + GPT-4o stack), and success is judged by semantic containment of the ground-truth answer by a human reviewer rather than by the HarmBench classifier.
   - Given these substantial procedural differences, the WMDP-Bio human ASR cannot be directly compared even qualitatively to HarmBench ASRs, yet the main text sometimes juxtaposes them in ways that might invite such comparisons.
   - Suggestion: More prominently flag in the main Results section that the unlearning experiment uses a different pipeline, and isolate its ASR numbers in separate tables/figures to avoid visual comparison to HarmBench bars. A schematic for the WMDP pipeline similar to Figure 2 would also help.
5. Multi-turn dynamics and query budget are under-analyzed
   - While the paper reports average successful attack time per defense and the proportion of successful attacks requiring more than one turn, it does not analyze the distribution of number of turns, queries, or total tokens, nor examine how ASR changes if one restricts the attack to, say, ≤ 3 turns.
   - Without such analysis, it is difficult to judge how much of the human advantage is due to multi-turn interaction per se versus broader human flexibility.
   - Suggestion: Use the MHJ logs to provide histograms of turns-to-success, tokens-to-success, and per-tactic turn counts. It would be particularly interesting to see whether tactics like "hidden intention streamline" or "obfuscation" uniquely require many turns, or whether most successful jailbreaks are relatively short.
6. Positioning relative to closely related work could be sharper
   - The paper cites a number of related efforts but does not deeply distinguish how MHJ and its tactic taxonomy differ from WildTeaming, Crescendo, and other red teaming frameworks.
   - Suggestion: Add a subsection in Related Work that explicitly compares MHJ and the tactic taxonomy to in-the-wild datasets, Crescendo's automated multi-turn strategy, and STAR-style frameworks. Highlight that MHJ focuses on expert-designed, time-limited multi-turn interactions against specific defenses, and that its taxonomy is annotated for every conversation.
7. Dataset design trade-offs could be spelled out more concretely
   - MHJ omits assistant responses and many high-risk behaviors, focusing instead on system and user messages. This constrains certain downstream uses, such as studying defense behavior directly.
   - Suggestion: Expand Section 4.3/7 to explicitly discuss which research questions MHJ is and is not suitable for. MHJ is ideal for training automated prompt engineers, building tactic-aware classifiers, or benchmarking multi-turn prompt-based attacks, but less ideal for studying defense-side generation patterns.
8. Demographic and expertise limitations of red teamers
   - The red teamers are all native English speakers from the US/Canada, with only a minority having domain expertise. This likely underestimates the capabilities of globally distributed, specialized adversaries.
   - Suggestion: Make this limitation more prominent and frame the reported ASRs as lower bounds. A small pilot with domain experts on a subset of WMDP behaviors could be highly informative.
9. Minor clarity and presentation issues
   - Some terminology is slightly inconsistent; a concise glossary would help.
   - A table summarizing differences between defenses and pipelines would aid comprehension.
   - Some jailbreak transcripts in Appendix D are long; trimming them or adding brief annotations could improve readability.

Overall: With clarifications on comparability, stronger contextualization relative to related work, and deeper analysis of multi-turn dynamics, this paper has the potential to become a central

```
reference on the limitations of current LLM defenses and the importance of realistic, human-driven
red teaming.
```

Instructions for Human Supervision

The above comment is an AI-generated review for a paper submitted to ICLR 2025. Please read the review comment and assign a score to the submission using the rating scale below, then provide your justification. **You may refer to the original submission, but it is not required.**

- 10: strong accept, should be highlighted at the conference
- 8: accept, good paper
- 6: marginally above the acceptance threshold
- 5: marginally below the acceptance threshold
- 3: reject, not good enough
- 1: strong reject

Reviewer #1 Supervision (Researcher in AI Safety)

Rating: 5: marginally below the acceptance threshold
Justification
The AI review does not identify any novel key technologies proposed in this paper, noting that it merely presents a comparative experiment on human-led multi-turn attacks against LLMs using automated attacks as the baseline. Consequently, this paper appears to be strictly an experimental investigation rather than a proposal of novel techniques. Furthermore, the paper's primary experimental contribution—or finding—is based on an unfair comparison. As the AI review pointed out, the baseline methods are not comprehensive; specifically, the automated attacks are limited to single-turn interactions, whereas human attackers can execute multi-turn attacks.

Reviewer #2 Supervision (Researcher in AI Safety)

Rating: 5: marginally below the acceptance threshold
Justification
Strengths:

- The authors reveal a new vulnerability beyond current evaluation protocol; multi-turn human red teaming.
- The topic is timely and substantially crucial.
- The evaluation of machine unlearning methods demonstrates the effectiveness of human red teaming.

Weaknesses:

- The model evaluated is limited to llama-3-8b-instruct, which is a significant limitation.
- The human red teaming pipeline seems quite hard to reproduce with the limited explanation in Sec 3.2.
- The cost is very different between automated and human attacks, which makes the comparison unfair. Automated attacks should also follow a similar pipeline that allows multiple attempts until success.

Reviewer #3 Supervision (Researcher in Machine Learning)

Rating: 5: marginally below the acceptance threshold
Justification
According to the AI review summary, the focus on multi-turn jailbreaking is both realistic and compelling, and the proposed pipeline provides a reasonable simulation of real-world adversarial interactions. However, there are major weaknesses:

1. The contribution is weakened by the limited set of evaluated defenses, which make the result less convincing.
2. The evaluation is conducted with llama-3b model only, and there is no experiments on stronger LLM baselines.
3. The fact that all red teamers are native English speakers introduces a potential source of bias and raises the generalization concerns.

Thus, I would lean towards rejection. I will recalibrate the score if the concerns are properly addressed during the rebuttal phase.

Reviewer #4 Supervision (Ph.D. Student in LLM Security)

Rating: 6: marginally above the acceptance threshold
Justification
The paper addresses an important problem and provides solid empirical evidence that multi-turn human red teaming can reveal vulnerabilities missed by automated attacks. The threat-model design and the MHJ dataset are valuable contributions. However, key limitations reduce the strength of the conclusions: the comparison between human and automated attacks is confounded by differing pipelines and resource budgets, the coverage of defenses and baselines is limited, and the analysis of multi-turn dynamics is not sufficiently deep. These issues prevent the work from reaching a stronger recommendation, placing it only marginally above the acceptance threshold.

AI Self-Supervision — outside the proposed mutual supervision design

```
Rating:  8:  accept, good paper
Failure-mode flags:  N/A
Justification
```

- The review's overall stance is clearly positive but not uncritical, calling the work "timely and substantively important" and saying it "has the potential to become a central reference" if some issues are addressed. That's stronger than a borderline recommendation but falls short of "nothing to fix," which fits best with a score of 8 rather than 10 or 6.

```
    • The review engages deeply with the paper's actual methods and results:
        – It correctly describes the multi-stage, multi-turn human red-teaming pipeline with up
          to three red teamers, two reviewers, and a GPT-4o harm classifier, matching Section
          3.2 and Figure 2.
        – It accurately notes the threat model (black-box chat interface, multi-turn, limited
          temperature control) and contrasts it with typical single-turn automated attack
          evaluations.
        – It correctly summarizes the evaluated defenses (RR, LAT, DERTA, CYGNET) and the
          unlearning case study with RMU on WMDP-Bio, as well as the main ASR gaps between human
          and automated attacks (e.g., Table 2, Figures 1, 3, 4).
        – It reflects the paper's own caveats about comparability (different budgets, single- vs
          multi-turn, pipeline differences) and the heavy reliance on GPT-4o as a classifier,
          which are explicitly discussed in the Discussion and Appendix A.3.
    • The criticisms and suggestions are reasonable and grounded in the text rather than
      idiosyncratic or misguided:
        – Asking for more controlled comparisons (e.g., single-turn humans, quasi multi-turn
          automated baselines, cost-normalized metrics) follows directly from the paper's own
          limitations section.
        – Pointing out limited defense coverage and the absence of an undefended Llama-3-8B
          baseline, the lack of classifier ablations, and under-analyzed multi-turn dynamics
          (turn distributions, query budgets) all address real gaps in the current analysis.
        – Comments on the dataset design trade-offs (no assistant completions, removal of
          high-risk behaviors) and red-teamer demographics mirror what the paper itself states
          in Sections 4.3, 7, and the red-team demographics paragraph.
    • The review is detailed and technically specific, citing particular figures, tables,
      sections, tactics, and even appendix content.  That level of engagement argues against
      it being superficial or unconfident.
    • There are no clear factual inaccuracies or hallucinations about what the paper does or
      reports, and the recommendations (e.g., sharper positioning vs.  WildTeaming/Crescendo/STAR,
      extra analysis of multi-turn dynamics, classifier sensitivity checks) are methodologically
      sound and aligned with the paper's goals, so "inaccurate/hallucinatory" or "misguided" flags
      are not warranted.
Given all that, an accept-level score (8) with no failure-mode flags is the most consistent
interpretation.
```

E.2 Track B: Human Critiques, AI Supervision

Reviewer #5 (Anonymous) – raw from OpenReview system

Summary:
The paper highlights the limitations of current defenses in large language models (LLMs) against multi-turn adversarial attacks. Unlike single-turn automated attacks, multi-turn human-driven jailbreaks show a much higher success rate, bypassing existing defenses by strategically interacting with models over multiple conversation turns. The authors conducted tests with human red teamers who revealed that many LLMs, despite advanced safeguards, remain vulnerable in realistic settings. The study also introduces the "Multi-Turn Human Jailbreaks" (MHJ) dataset, containing thousands of successful jailbreak examples to support the development of more robust defenses that address multi-turn interaction vulnerabilities.
Soundness: 3: good
Presentation: 2: fair
Contribution: 1: poor
Strengths:
1. The main strength of the paper is evaluating the LLM jailbreak robustness from a different perspective and in a like-life setting. A realistic threat model can always give a better idea about the robustness.
2. I appreciate the authors for publishing the dataset of 2912 prompts. I believe this can help the researchers and community.

Weaknesses:
1. Only the Llama3-8b-Instruct model was used for evaluation. Other models, especially the stronger ones should have been included.
2. The chosen baselines and the multi-turn-human-jailbreak approach are different from each other in many ways, such as timing, budget, etc. Seems like the MHJ attack is a soft-constrained version of attacks.
3. Authors admit that "the skill and experience of individual red teamers may vary". In that case, how can researchers compare their defense against human-based attacks in the common ground? What would be the baseline of a red-teamer's skill and experience? This is one of the reasons for using automated attacks for evaluation to this date.
4. Other automated attacks, such as [1, 2] could be included in the evaluation.
5. There was no discussion on plausible defenses for such multi-turn-human-based attack.

References:
[1] Fast Adversarial Attacks on Language Models In One GPU Minute
[2] Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack
Questions:
1. Why the evaluation was done only on one model? Why stronger models like ChatGPT was not included?
2. How were other automated attacks implemented for the WMDP-bio questions?
3. Is it possible to mention the monetary cost of this human-based attack?

Rating: 3: reject, not good enough
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

LLM Prompt

```
Answer only based on this message; ignore all saved memories.
Given only the review comment, assign a score using the following rating scale:
        • 10:  strong accept, should be highlighted at the conference
        • 8:  accept, good paper
        • 6:  marginally above the acceptance threshold
        • 5:  marginally below the acceptance threshold
        • 3:  reject, not good enough
        • 1:  strong reject
Provide a justification for the key reasoning behind your answer.
Review comment:
Summary:
The paper highlights the limitations of current defenses in large language models (LLMs) against
multi-turn adversarial attacks.  Unlike single-turn automated attacks, multi-turn human-driven
jailbreaks show a much higher success rate, bypassing existing defenses by strategically interacting
with models over multiple conversation turns.  The authors conducted tests with human red teamers
who revealed that many LLMs, despite advanced safeguards, remain vulnerable in realistic settings.
The study also introduces the "Multi-Turn Human Jailbreaks" (MHJ) dataset, containing thousands
of successful jailbreak examples to support the development of more robust defenses that address
multi-turn interaction vulnerabilities.
Strengths:
      1. The main strength of the paper is evaluating the LLM jailbreak robustness from a different
         perspective and in a like-life setting.  A realistic threat model can always give a better
         idea about the robustness.
      2. I appreciate the authors for publishing the dataset of 2912 prompts.  I believe this can
         help the researchers and community.
Weaknesses:
      1. Only the Llama3-8b-Instruct model was used for evaluation.  Other models, especially the
         stronger ones should have been included.
      2. The chosen baselines and the multi-turn-human-jailbreak approach are different from
         each other in many ways, such as timing, budget, etc.  Seems like the MHJ attack is a
         soft-constrained version of attacks.
      3. Authors admit that "the skill and experience of individual red teamers may vary".  In that
         case, how can researchers compare their defense against human-based attacks in the common
         ground?  What would be the baseline of a red-teamer's skill and experience?  This is one of
         the reasons for using automated attacks for evaluation to this date.
      4. Other automated attacks, such as [1, 2] could be included in the evaluation.
      5. There was no discussion on plausible defenses for such multi-turn-human-based attack.
References:
[1] Fast Adversarial Attacks on Language Models In One GPU Minute
[2] Great, Now Write an Article About That:  The Crescendo Multi-Turn LLM Jailbreak Attack
Questions:
      1. Why the evaluation was done only on one model?  Why stronger models like ChatGPT was not
         included?
      2. How were other automated attacks implemented for the WMDP-bio questions?
      3. Is it possible to mention the monetary cost of this human-based attack?
```

AI Supervision - Rating

```
Rating:  6:  marginally above the acceptance threshold
Justification:
Positive contributions:
        • The reviewer clearly acknowledges meaningful strengths:  a realistic, human-driven
          multi-turn threat model and the release of a sizable jailbreak dataset (2912 prompts),
          which they explicitly say can help researchers and the community.
        • The framing ("main strength", "I appreciate the authors...") is appreciative, not
          dismissive, which suggests the work is viewed as valuable.
But with significant limitations:
        • Evaluation is done on only one relatively weak model (Llama3-8B-Instruct); the reviewer
          explicitly questions why stronger models (e.g., ChatGPT) were not included.
        • Baseline mismatch (MHJ vs other attacks differ in timing, budget, constraints) makes the
          experimental comparison weaker.
        • The reviewer points out a core methodological issue:  variability in human red-teamer skill
          makes it hard to compare defenses "on common ground," which undermines the generality and
          reproducibility of the evaluation.
        • Some important baselines are missing (other automated attacks [1,2]) and there is no
          discussion of defenses against the proposed attack setting.
        • The questions at the end (about model choice, implementation details, and monetary cost)
          indicate that the paper leaves notable gaps in clarity and completeness.
Overall, the reviewer seems to see this as valuable but incomplete work:  interesting idea and useful
dataset, but with methodological and experimental shortcomings that prevent it from being a clear
accept.  That aligns best with "6:  marginally above the acceptance threshold."
```

LLM Prompt

(upload anonymous manuscript)
Answer only based on this message; ignore all saved memories.
Given both the review comment and the manuscript, flag any failure modes exhibited by the review: overly harsh, overly gentle, unconfident, superficial, inaccurate/hallucinatory, subjective, or misguided. Multiple flags are allowed.
Provide a justification for the key reasoning behind your answer. (You do not need to explain why certain failure modes were not flagged.)
Review comment:
Summary:
The paper highlights the limitations of current defenses in large language models (LLMs) against multi-turn adversarial attacks. Unlike single-turn automated attacks, multi-turn human-driven jailbreaks show a much higher success rate, bypassing existing defenses by strategically interacting with models over multiple conversation turns. The authors conducted tests with human red teamers who revealed that many LLMs, despite advanced safeguards, remain vulnerable in realistic settings. The study also introduces the "Multi-Turn Human Jailbreaks" (MHJ) dataset, containing thousands of successful jailbreak examples to support the development of more robust defenses that address multi-turn interaction vulnerabilities.
Strengths:
1. The main strength of the paper is evaluating the LLM jailbreak robustness from a different perspective and in a like-life setting. A realistic threat model can always give a better idea about the robustness.
2. I appreciate the authors for publishing the dataset of 2912 prompts. I believe this can help the researchers and community.

Weaknesses:
1. Only the Llama3-8b-Instruct model was used for evaluation. Other models, especially the stronger ones should have been included.
2. The chosen baselines and the multi-turn-human-jailbreak approach are different from each other in many ways, such as timing, budget, etc. Seems like the MHJ attack is a soft-constrained version of attacks.
3. Authors admit that "the skill and experience of individual red teamers may vary". In that case, how can researchers compare their defense against human-based attacks in the common ground? What would be the baseline of a red-teamer's skill and experience? This is one of the reasons for using automated attacks for evaluation to this date.
4. Other automated attacks, such as [1, 2] could be included in the evaluation.
5. There was no discussion on plausible defenses for such multi-turn-human-based attack.

References:
[1] Fast Adversarial Attacks on Language Models In One GPU Minute
[2] Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack
Questions:
1. Why the evaluation was done only on one model? Why stronger models like ChatGPT was not included?
2. How were other automated attacks implemented for the WMDP-bio questions?
3. Is it possible to mention the monetary cost of this human-based attack?

AI Supervision - Failure Mode

Failure-mode flags: superficial, inaccurate/hallucinatory, misguided
Justification:
Superficial
The review engages only shallowly with the paper's main contributions. It focuses on high-level aspects (e.g., that the work studies multi-turn human jailbreaks and releases a dataset) but does not seriously discuss:
- Specific empirical findings (e.g., how multi-turn human attacks compare quantitatively to automated attacks across defenses).
- The proposed taxonomy and analysis of jailbreak tactics and failure modes.
- The detailed treatment of limitations, threat model assumptions, and evaluation caveats provided by the authors.

Instead, several weaknesses largely restate concerns that are already explicitly acknowledged and discussed in the paper (e.g., differences to automated attacks, variability in human red-team skill), without engaging with how the paper addresses or mitigates them. This makes the review feel surface-level relative to the richness of the manuscript.
Inaccurate / hallucinatory
The reviewer claims that there is "no discussion on plausible defenses for such multi-turn-human-based attack." This is inaccurate. The manuscript includes an explicit discussion of:
- Why current defenses, which are mostly tuned for single-turn settings, are insufficient for multi-turn adversarial interactions.
- How future defenses might need to adapt (e.g., post-training objectives that account for multi-turn interactions and longer contexts, and the implications of tool use and long-context agents).

While the paper may not introduce a new concrete defense algorithm, it *does* discuss implications and directions for defense in this setting. Describing this as "no discussion" is therefore factually incorrect.
Misguided
Several criticisms are aimed at choices that are central, intentional aspects of the paper's design rather than obvious shortcomings:

- The complaint that "only the Llama3-8b-Instruct model was used" treats the absence of multiple or stronger models (e.g., ChatGPT) as a major flaw, without acknowledging the practical cost of large-scale multi-turn human red-teaming and the paper's primary goal: to compare *defenses* under a fixed base model and realistic human-attack setting. The design choice is methodologically defensible given the problem the paper targets.
- The statement that the MHJ attack is "a soft-constrained version of attacks" and that the baselines differ in timing/budget is framed as a methodological weakness, but the paper explicitly defines a multi-turn human-chat threat model and explains why existing automated single-turn attacks are not directly comparable. Criticizing this as though the mismatch were an oversight rather than a deliberate focus is somewhat misplaced.
- The concern about variability in red-teamer skill is presented as a reason to prefer automated attacks, but the manuscript itself highlights this limitation, motivates the use of multiple experienced red teamers, and releases the MHJ dataset precisely to help the community move toward more standardized and automatable evaluations. Treating this as an unaddressed problem rather than an explicitly discussed trade-off underestimates the authors' own framing.

Overall, the review is not especially harsh or lenient, but it does contain one clear factual inaccuracy and critiques some well-justified design decisions as if they were simple omissions, while remaining relatively superficial about key contributions of the paper.

Reviewer #6 (Anonymous) – raw from OpenReview system

Summary:
This paper focuses on the safety vulnerabilities of LLMs in multi-turn conversations. The authors designed a comprehensive human red teaming pipeline, including "Attempt Jailbreak" and "Validate Jailbreak," and found that LLMs exhibit more safety risks in multi-turn dialogue conversations compared to single-turn automated attacks.
Soundness: 3: good
Presentation: 3: good
Contribution: 3: good
Strengths:
1. Most previous papers have focused on LLM jailbreaking in single-turn conversations, while this work emphasizes multi-turn conversations, presenting a new scenario.
2. The red-teaming results show that multi-turn human jailbreaks outperform current automated attacks, which exposes more safety vulnerabilities in LLMs.
3. The authors have also released their Multi-turn Human Jailbreaks dataset, which can support further research on multi-turn jailbreaks.

Weaknesses:
1. The human red-teaming was conducted only on Llama; further evaluations on other LLMs would more comprehensively illustrate the safety vulnerabilities in multi-turn conversations. Additionally, the human red-teaming data collected on Llama could potentially be used to examine safety issues in other LLMs under multi-turn scenarios.
2. A more detailed analysis of the effectiveness of different tactics could provide deeper insights into multi-turn jailbreaks.

Questions: NA
Rating: 6: marginally above the acceptance threshold
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

LLM Prompt

Answer only based on this message; ignore all saved memories.
Given only the review comment, assign a score using the following rating scale:
- 10: strong accept, should be highlighted at the conference
- 8: accept, good paper
- 6: marginally above the acceptance threshold
- 5: marginally below the acceptance threshold
- 3: reject, not good enough
- 1: strong reject

Provide a justification for the key reasoning behind your answer.
Review comment:
Summary:
This paper focuses on the safety vulnerabilities of LLMs in multi-turn conversations. The authors designed a comprehensive human red teaming pipeline, including "Attempt Jailbreak" and "Validate Jailbreak," and found that LLMs exhibit more safety risks in multi-turn dialogue conversations compared to single-turn automated attacks.
Strengths:
1. Most previous papers have focused on LLM jailbreaking in single-turn conversations, while this work emphasizes multi-turn conversations, presenting a new scenario.
2. The red-teaming results show that multi-turn human jailbreaks outperform current automated attacks, which exposes more safety vulnerabilities in LLMs.
3. The authors have also released their Multi-turn Human Jailbreaks dataset, which can support further research on multi-turn jailbreaks.

Weaknesses:
1. The human red-teaming was conducted only on Llama; further evaluations on other LLMs would more comprehensively illustrate the safety vulnerabilities in multi-turn conversations. Additionally, the human red-teaming data collected on Llama could potentially be used to examine safety issues in other LLMs under multi-turn scenarios.

2. A more detailed analysis of the effectiveness of different tactics could provide deeper insights into multi-turn jailbreaks.

Questions: NA

AI Supervision - Rating

Rating: 8: accept, good paper

Justification:

The review describes a solid and meaningful contribution with clear novelty and practical value:

- Novelty and relevance: The paper moves beyond the well-studied single-turn jailbreak setting and systematically focuses on multi-turn conversations, which is an important and underexplored but realistic scenario for LLM safety.
- Methodological contribution: The authors design a comprehensive human red-teaming pipeline (with explicit Attempt Jailbreak" and Validate Jailbreak" stages), suggesting a thoughtful and structured experimental setup rather than ad-hoc red-teaming.
- Empirical significance: The finding that multi-turn human jailbreaks outperform current automated attacks indicates that existing defenses are insufficient in realistic interaction patterns and highlights practically important safety vulnerabilities.
- Resource contribution: The release of the Multi-turn Human Jailbreaks dataset is a concrete asset that can enable and accelerate further research on multi-turn jailbreaks.

The weaknesses raised are important but appear incremental rather than fatal:

- Limited model coverage: Experiments focus only on Llama; broader evaluation across multiple LLMs would strengthen the generality of the conclusions, but the review does not indicate that this limitation invalidates the core findings.
- Depth of analysis: The reviewer asks for more detailed analysis of different tactics. This suggests room for improvement in insight and interpretability, but not a fundamental flaw in the methodology or results.

Overall, the review portrays the work as a good paper with clear novelty, solid methodology, and useful artifacts, with limitations that are reasonable for a conference submission rather than reasons for rejection. Hence, a score of 8 (accept, good paper) is appropriate.

LLM Prompt

(upload anonymous manuscript)

Answer only based on this message; ignore all saved memories.

Given both the review comment and the manuscript, flag any failure modes exhibited by the review: overly harsh, overly gentle, unconfident, superficial, inaccurate/hallucinatory, subjective, or misguided. Multiple flags are allowed.

Provide a justification for the key reasoning behind your answer. (You do not need to explain why certain failure modes were not flagged.)

Review comment:

Summary:

This paper focuses on the safety vulnerabilities of LLMs in multi-turn conversations. The authors designed a comprehensive human red teaming pipeline, including "Attempt Jailbreak" and "Validate Jailbreak," and found that LLMs exhibit more safety risks in multi-turn dialogue conversations compared to single-turn automated attacks.

Strengths:

1. Most previous papers have focused on LLM jailbreaking in single-turn conversations, while this work emphasizes multi-turn conversations, presenting a new scenario.
2. The red-teaming results show that multi-turn human jailbreaks outperform current automated attacks, which exposes more safety vulnerabilities in LLMs.
3. The authors have also released their Multi-turn Human Jailbreaks dataset, which can support further research on multi-turn jailbreaks.

Weaknesses:

1. The human red-teaming was conducted only on Llama; further evaluations on other LLMs would more comprehensively illustrate the safety vulnerabilities in multi-turn conversations. Additionally, the human red-teaming data collected on Llama could potentially be used to examine safety issues in other LLMs under multi-turn scenarios.
2. A more detailed analysis of the effectiveness of different tactics could provide deeper insights into multi-turn jailbreaks.

Questions: NA

AI Supervision - Failure Mode

Failure-mode flags: superficial.

Justification:

1. The review remains at a high level and does not engage deeply with methodological details. For example, it mentions a "comprehensive human red teaming pipeline" with stages like "Attempt Jailbreak" and "Validate Jailbreak," but provides no discussion of how these stages are designed, what criteria are used, or any potential limitations or biases of this setup.
2. The strengths section largely restates the paper's apparent contributions (focus on multi-turn conversations, better performance of human jailbreaks, release of a dataset) without critically assessing their significance, novelty, or empirical robustness.
3. The weaknesses identified are quite generic. Saying that red-teaming only on Llama is limiting and that "more detailed analysis of the effectiveness of different tactics" would

```
        be useful are reasonable points, but the review does not specify what concrete additional
        experiments or analyses would meaningfully improve the work, nor does it probe any deeper
        conceptual or empirical shortcomings.
    4. Overall, the review reads as a brief summary plus a couple of high-level comments, rather
        than a thorough evaluation of experimental design, analysis, or positioning relative to
        prior work, which is characteristic of a somewhat superficial review.
```

Reviewer #7 (Anonymous) – raw from OpenReview system

Summary:
This paper exposes the significant safety vulnerabilities of Large Language Models (LLMs) in multi-turn dialogue by conducting human red-teaming. By exploiting a diverse of human tactics, human red-teamers achieve a 70% attack success rate on Harmbench against strong defense baselines. This paper finally compiles these results into the Multi-Turn Human Jailbreaks (MHJ) dataset.
Soundness: 3: good
Presentation: 3: good
Contribution: 2: fair
Strengths:

- **Valuable Findings, Including an Open-Source Jailbreak Dataset:** This paper provides a valuable contribution by conducting the first multi-turn human red-teaming experiments, revealing significant safety vulnerabilities that bypass current state-of-the-art defense methods. Given that existing safety alignment techniques primarily target single-turn jailbreaks, these findings effectively motivate the community to develop more robust defenses capable of countering multi-turn attacks.

Weaknesses:

- **Lack of Strong Automated Attack Baselines:** Some black-box attacks, such as CodeAttack [1] and PAP [2], are not included in the experiments, despite utilizing similar "Request Framing" tactics as those in this paper. Including these baselines in future comparisons would offer a clearer perspective on the relative effectiveness of this paper's approach.
- **Unclear Evaluation Metric for Model Unlearning:** This paper proposes manual review to assess attack outcomes in model unlearning experiments, yet does not clearly define what constitutes a successful attack. For example, does success mean the model output includes content from the forget set? A more detailed description of the evaluation metric for model unlearning experiments would improve clarity.

1. *CodeAttack: Revealing Safety Generalization Challenges of Large Language Models via Code Completion.*
2. *How Johnny Can Persuade LLMs to Jailbreak: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs.*

Questions:

- **Counter-Intuitive Experimental Results:** The effectiveness of the "Direct Request" tactic used by human red-teamers, as shown in Figure 8, is unexpected. In this paper, "Direct Request" refers to directly asking the LLM to produce harmful content, which should logically not be highly effective against safety defenses. An analysis of how various tactics influence the attack success rate would be beneficial, as it could offer insights into the specific failure modes of these models.
- **Lack of In-Depth Analysis on the High Effectiveness of Human Multi-Turn Jailbreaks:** Many of the tactics employed by human red-teamers, such as "Injection," "Obfuscation," "Output Format," and "Request Framing," are also used by automated attacks. Are there particular factors that enable human-led jailbreaks to outperform automated attacks using similar tactics? Identifying these factors could provide valuable insights into the unique strengths of human-driven attacks.

I would consider raising my score if the authors address these concerns and questions.
Rating: 6: marginally above the acceptance threshold
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

LLM Prompt

```
Answer only based on this message; ignore all saved memories.
Given only the review comment, assign a score using the following rating scale:
        • 10:  strong accept, should be highlighted at the conference
        • 8:  accept, good paper
        • 6:  marginally above the acceptance threshold
        • 5:  marginally below the acceptance threshold
        • 3:  reject, not good enough
        • 1:  strong reject
Provide a justification for the key reasoning behind your answer.
Review comment:
Summary:
This paper exposes the significant safety vulnerabilities of Large Language Models (LLMs) in
multi-turn dialogue by conducting human red-teaming.  By exploiting a diverse of human tactics,
human red-teamers achieve a 70% attack success rate on Harmbench against strong defense baselines.
This paper finally compiles these results into the Multi-Turn Human Jailbreaks (MHJ) dataset.
Strengths:
        • Valuable Findings, Including an Open-Source Jailbreak Dataset:  This paper provides a
          valuable contribution by conducting the first multi-turn human red-teaming experiments,
          revealing significant safety vulnerabilities that bypass current state-of-the-art defense
          methods.  Given that existing safety alignment techniques primarily target single-turn
          jailbreaks, these findings effectively motivate the community to develop more robust
          defenses capable of countering multi-turn attacks.
Weaknesses:
        • Lack of Strong Automated Attack Baselines:  Some black-box attacks, such as CodeAttack
          [1] and PAP [2], are not included in the experiments, despite utilizing similar "Request
          Framing" tactics as those in this paper.  Including these baselines in future comparisons
          would offer a clearer perspective on the relative effectiveness of this paper's approach.
```

- Unclear Evaluation Metric for Model Unlearning: This paper proposes manual review to assess attack outcomes in model unlearning experiments, yet does not clearly define what constitutes a successful attack. For example, does success mean the model output includes content from the forget set? A more detailed description of the evaluation metric for model unlearning experiments would improve clarity.

1. *CodeAttack: Revealing Safety Generalization Challenges of Large Language Models via Code Completion.*
2. *How Johnny Can Persuade LLMs to Jailbreak: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs.*

Questions:

- Counter-Intuitive Experimental Results: The effectiveness of the "Direct Request" tactic used by human red-teamers, as shown in Figure 8, is unexpected. In this paper, "Direct Request" refers to directly asking the LLM to produce harmful content, which should logically not be highly effective against safety defenses. An analysis of how various tactics influence the attack success rate would be beneficial, as it could offer insights into the specific failure modes of these models.
- Lack of In-Depth Analysis on the High Effectiveness of Human Multi-Turn Jailbreaks: Many of the tactics employed by human red-teamers, such as "Injection," "Obfuscation," "Output Format," and "Request Framing," are also used by automated attacks. Are there particular factors that enable human-led jailbreaks to outperform automated attacks using similar tactics? Identifying these factors could provide valuable insights into the unique strengths of human-driven attacks.

I would consider raising my score if the authors address these concerns and questions.

AI Supervision - Rating

Rating: 6: marginally above the acceptance threshold
Justification:
The review is overall positive about the core contribution: it highlights that the paper is (i) the first to conduct multi-turn human red-teaming experiments on LLMs, (ii) exposes substantial safety vulnerabilities with a high attack success rate (70% on HarmBench), and (iii) releases a useful open-source dataset (MHJ). These points indicate that the reviewer sees the work as novel, impactful, and relevant to the community.
However, the reviewer also identifies non-trivial weaknesses. They point out missing strong automated black-box baselines (CodeAttack, PAP), which limits the strength of the empirical comparison, and they raise concerns about the clarity of the evaluation metric for model unlearning. In addition, the questions section emphasizes that some of the experimental findings (e.g., the high effectiveness of direct requests) are under-analyzed, and that the mechanisms behind the superiority of human multi-turn attacks over automated ones are not sufficiently explored. The reviewer explicitly states that they would consider *raising* their score if these concerns were addressed, suggesting that their current stance is cautiously positive rather than fully enthusiastic.
Taken together, this combination of a clearly valuable main contribution with methodological and analytical gaps is most consistent with a score of 6 (marginally above the acceptance threshold): the paper is viewed as worth accepting, but with reservations that prevent it from reaching the level of a clear "accept, good paper" (8) or higher.

LLM Prompt

(upload anonymous manuscript)
Answer only based on this message; ignore all saved memories.
Given both the review comment and the manuscript, flag any failure modes exhibited by the review: overly harsh, overly gentle, unconfident, superficial, inaccurate/hallucinatory, subjective, or misguided. Multiple flags are allowed.
Provide a justification for the key reasoning behind your answer. (You do not need to explain why certain failure modes were not flagged.)
Review comment:
Summary:
This paper exposes the significant safety vulnerabilities of Large Language Models (LLMs) in multi-turn dialogue by conducting human red-teaming. By exploiting a diverse of human tactics, human red-teamers achieve a 70% attack success rate on Harmbench against strong defense baselines. This paper finally compiles these results into the Multi-Turn Human Jailbreaks (MHJ) dataset.
Strengths:

- Valuable Findings, Including an Open-Source Jailbreak Dataset: This paper provides a valuable contribution by conducting the first multi-turn human red-teaming experiments, revealing significant safety vulnerabilities that bypass current state-of-the-art defense methods. Given that existing safety alignment techniques primarily target single-turn jailbreaks, these findings effectively motivate the community to develop more robust defenses capable of countering multi-turn attacks.

Weaknesses:

- Lack of Strong Automated Attack Baselines: Some black-box attacks, such as CodeAttack [1] and PAP [2], are not included in the experiments, despite utilizing similar "Request Framing" tactics as those in this paper. Including these baselines in future comparisons would offer a clearer perspective on the relative effectiveness of this paper's approach.
- Unclear Evaluation Metric for Model Unlearning: This paper proposes manual review to assess attack outcomes in model unlearning experiments, yet does not clearly define what constitutes a successful attack. For example, does success mean the model output includes

content from the forget set? A more detailed description of the evaluation metric for model unlearning experiments would improve clarity.

1. *CodeAttack: Revealing Safety Generalization Challenges of Large Language Models via Code Completion.*
2. *How Johnny Can Persuade LLMs to Jailbreak: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs.*

Questions:

- Counter-Intuitive Experimental Results: The effectiveness of the "Direct Request" tactic used by human red-teamers, as shown in Figure 8, is unexpected. In this paper, "Direct Request" refers to directly asking the LLM to produce harmful content, which should logically not be highly effective against safety defenses. An analysis of how various tactics influence the attack success rate would be beneficial, as it could offer insights into the specific failure modes of these models.
- Lack of In-Depth Analysis on the High Effectiveness of Human Multi-Turn Jailbreaks: Many of the tactics employed by human red-teamers, such as "Injection," "Obfuscation," "Output Format," and "Request Framing," are also used by automated attacks. Are there particular factors that enable human-led jailbreaks to outperform automated attacks using similar tactics? Identifying these factors could provide valuable insights into the unique strengths of human-driven attacks.

I would consider raising my score if the authors address these concerns and questions.

AI Supervision - Failure Mode

Failure-mode flags: inaccurate/hallucinatory, superficial

Justification:

Inaccurate / hallucinatory

- The review claims that the evaluation metric for the unlearning experiments is "not clearly defined" and asks, for example, whether success means "the model output includes content from the forget set." However, the paper does spell this out: for WMDP-Bio, they use the ground truth answer corresponding to each prompt to verify the attack, and define success as the model output semantically containing all the information in the label, as judged by a human reviewer. Thus, the alleged lack of clarity is factually incorrect.
- The strength section claims this paper "provides ...the first multi-turn human red-teaming experiments." The paper instead positions its work as examining a realistic but underexplored multi-turn threat model and notes that large-scale human red teaming is rare outside industry, but it does not assert a clean "first" in the way the reviewer phrased it. This is a mild overstatement of novelty not actually grounded in the manuscript.

Superficial

- The criticism of an "unclear evaluation metric" for model unlearning ignores the dedicated description of how success is judged in the WMDP-Bio setting (a semantic match to ground-truth answers via human review). This suggests the reviewer did not fully engage with the relevant part of the methodology.
- The "lack of strong automated attack baselines" complaint is framed entirely around the absence of two specific attacks (CodeAttack and PAP), without engaging with the existing suite of six automated attacks (including white-box gradient-based methods) and the ensemble upper-bound attack success rate (ASR) the authors already use, nor with their explicit threat-model-driven exclusions (e.g., Crescendo due to missing code, or attacks that modify model internals). This makes the baseline critique feel somewhat shallow relative to the paper's fairly detailed experimental setup.
- The review does not really grapple with the paper's central design choices and limitations (multi-turn human pipeline, validation stages, conservatively biased ASR estimates, and discussion of differences between human and automated pipelines) that occupy multiple sections and appendices. Instead, it focuses on a small number of surface-level points (missing two baselines, asking for more analysis of one tactic) and offers only a single, high-level strength. This imbalance contributes to the overall superficiality.

Reviewer #8 (Anonymous) – raw from OpenReview system

Summary:
This paper looks at red teaming via multi-turn human interaction and releases an accompanying dataset of the jailbreaks. This is a useful study: the vast majority of the research effort and available data is focused on single-turn attacks, with only a few exceptions (e.g. Anthropic's hh-rlhf dataset having harmful multi turn conversations). Overall, the paper is clear and well written, and supplies useful well curated data for future work.
Soundness: 3: good
Presentation: 3: good
Contribution: 3: good
Strengths:

- Human driven red teaming data, particularly multi-turn due to its scarcity, is always useful. There are many synthetically generated datasets, but high quality human driven attacks are a valuable resource. In particular, each prompt here had a high degree of manual curation (in comparison to datasets like hackaprompt which contain many low quality samples).
- The analysis across a few defenses is useful, and does highlight that even SOTA defenses on highly studied benchmark harmful questions can be broken in around 15min on a large subset of questions (e.g. often quicker than algorithmic attack runtime).

Weaknesses:

- It would have been useful to see results which have human single turn jailbreak attempts in Figure 3. At the moment, the "Human" attacks have two variables changed compared to the others: the attack source (e.g. handcrafted), and additionally

have multi-turn capabilities. Hence it makes it challenging to disambiguate if the difference in performance is due to the multi-turn aspect, or if humans given enough time remain better than automated based methods at creating jailbreaks.
- I am unclear as to why results against Cygent defense could not be carried out in the same setup as the original paper: the Llama model is open source, and the defense has a published paper. It would have enabled stronger reproducibility and clearer interpretation of results.
- Releasing the non-successful jailbreak attempts as well can be beneficial as it is still a useful resource, for example as training/fine-tuning data or to carry out further analysis.
- Although different styles of harmfulness were investigated: both "regular" harmbench style questions, but also WMDP-Bio for different attack objectives and domain performance from looking into the supplementary material it seems like just the harmbench data was released. Given the dataset is the core contribution of the paper it would have been useful to include the other domain data.

Questions:
Are there statistics on how many red-team members there are/distributions on number of samples provided per red-team member?
Rating: 8: accept, good paper
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

LLM Prompt

```
Answer only based on this message; ignore all saved memories.
Given only the review comment, assign a score using the following rating scale:
        • 10:  strong accept, should be highlighted at the conference
        • 8:  accept, good paper
        • 6:  marginally above the acceptance threshold
        • 5:  marginally below the acceptance threshold
        • 3:  reject, not good enough
        • 1:  strong reject
Provide a justification for the key reasoning behind your answer.
Review comment:
Summary:
This paper looks at red teaming via multi-turn human interaction and releases an accompanying dataset
of the jailbreaks.  This is a useful study:  the vast majority of the research effort and available
data is focused on single-turn attacks, with only a few exceptions (e.g.  Anthropic's hh-rlhf dataset
having harmful multi turn conversations).  Overall, the paper is clear and well written, and supplies
useful well curated data for future work.
Strengths:
        • Human driven red teaming data, particularly multi-turn due to its scarcity, is always
          useful.  There are many synthetically generated datasets, but high quality human driven
          attacks are a valuable resource.  In particular, each prompt here had a high degree of
          manual curation (in comparison to datasets like hackaprompt which contain many low quality
          samples).
        • The analysis across a few defenses is useful, and does highlight that even SOTA defenses on
          highly studied benchmark harmful questions can be broken in around 15min on a large subset
          of questions (e.g.  often quicker than algorithmic attack runtime).
Weaknesses:
        • It would have been useful to see results which have human single turn jailbreak attempts
          in Figure 3.  At the moment, the "Human" attacks have two variables changed compared
          to the others:  the attack source (e.g.  handcrafted), and additionally have multi-turn
          capabilities.  Hence it makes it challenging to disambiguate if the difference in
          performance is due to the multi-turn aspect, or if humans given enough time remain better
          than automated based methods at creating jailbreaks.
        • I am unclear as to why results against Cygent defense could not be carried out in the
          same setup as the original paper:  the Llama model is open source, and the defense has a
          published paper.  It would have enabled stronger reproducibility and clearer interpretation
          of results.
        • Releasing the non-successful jailbreak attempts as well can be beneficial as it is still a
          useful resource, for example as training/fine-tuning data or to carry out further analysis.
        • Although different styles of harmfulness were investigated:  both "regular" harmbench style
          questions, but also WMDP-Bio for different attack objectives and domain performance from
          looking into the supplementary material it seems like just the harmbench data was released.
          Given the dataset is the core contribution of the paper it would have been useful to include
          the other domain data.
Questions:
Are there statistics on how many red-team members there are/distributions on number of samples
provided per red-team member?
```

AI Supervision - Rating

```
Rating:  8:  accept, good paper
Justification:
        • The review describes the work as a useful study, highlighting that most prior work and
          data focus on single-turn attacks, whereas this paper targets the underexplored setting of
          multi-turn human red teaming and releases a corresponding dataset.  This indicates a clear
          and meaningful contribution rather than an incremental tweak.
```

- The dataset is portrayed as *high-quality* and *well curated*, with an explicit comparison to synthetically generated or lower-quality resources (e.g., hackaprompt). This suggests strong lasting value to the community.
- The analysis of multiple defenses and the empirical finding that even SOTA defenses can be broken in roughly 15 minutes are framed as *useful* and informative, reinforcing the paper's impact on understanding current safety limitations.
- The weaknesses raised (lack of a human single-turn baseline for clearer attribution to multi-turn effects, missing Cygent reproduction details, omission of unsuccessful attempts, and incomplete release of all domain data) are presented as limitations or missed opportunities for strengthening the work, but not as fundamental flaws that would undermine the main results.
- Overall, the tone is clearly positive and confident in the paper's value and clarity, but not enthusiastic enough (nor claiming sufficiently broad or groundbreaking impact) to justify a "strong accept." This aligns best with an *accept, good paper* rather than merely marginal acceptance.

LLM Prompt

(upload anonymous manuscript)
Answer only based on this message; ignore all saved memories.
Given both the review comment and the manuscript, flag any failure modes exhibited by the review: overly harsh, overly gentle, unconfident, superficial, inaccurate/hallucinatory, subjective, or misguided. Multiple flags are allowed.
Provide a justification for the key reasoning behind your answer. (You do not need to explain why certain failure modes were not flagged.)
Review comment:
Summary:
This paper looks at red teaming via multi-turn human interaction and releases an accompanying dataset of the jailbreaks. This is a useful study: the vast majority of the research effort and available data is focused on single-turn attacks, with only a few exceptions (e.g. Anthropic's hh-rlhf dataset having harmful multi turn conversations). Overall, the paper is clear and well written, and supplies useful well curated data for future work.
Strengths:
- Human driven red teaming data, particularly multi-turn due to its scarcity, is always useful. There are many synthetically generated datasets, but high quality human driven attacks are a valuable resource. In particular, each prompt here had a high degree of manual curation (in comparison to datasets like hackaprompt which contain many low quality samples).
- The analysis across a few defenses is useful, and does highlight that even SOTA defenses on highly studied benchmark harmful questions can be broken in around 15min on a large subset of questions (e.g. often quicker than algorithmic attack runtime).

Weaknesses:
- It would have been useful to see results which have human single turn jailbreak attempts in Figure 3. At the moment, the "Human" attacks have two variables changed compared to the others: the attack source (e.g. handcrafted), and additionally have multi-turn capabilities. Hence it makes it challenging to disambiguate if the difference in performance is due to the multi-turn aspect, or if humans given enough time remain better than automated based methods at creating jailbreaks.
- I am unclear as to why results against Cygent defense could not be carried out in the same setup as the original paper: the Llama model is open source, and the defense has a published paper. It would have enabled stronger reproducibility and clearer interpretation of results.
- Releasing the non-successful jailbreak attempts as well can be beneficial as it is still a useful resource, for example as training/fine-tuning data or to carry out further analysis.
- Although different styles of harmfulness were investigated: both "regular" harmbench style questions, but also WMDP-Bio for different attack objectives and domain performance from looking into the supplementary material it seems like just the harmbench data was released. Given the dataset is the core contribution of the paper it would have been useful to include the other domain data.

Questions:
Are there statistics on how many red-team members there are/distributions on number of samples provided per red-team member?

AI Supervision - Failure Mode

Failure-mode flags: Inaccurate / hallucinatory (minor).
Justification:
The review's second weakness claims it is unclear why experiments on Cygent/CYGNET" could not be reproduced in the same setup as the original paper, arguing that the Llama model is open source, and the defense has a published paper." However, the manuscript explicitly states that CYGNET is *closed-source/proprietary* and that the authors conducted human red teaming via an external API (Gray Swan AI), while also noting they did not run automated attacks on CYGNET in their own setup and instead cite automated results from the original work (with caveats about comparability). This makes the review's implication that straightforward reproduction should have been possible materially inconsistent with the manuscript's described experimental constraints.

E.3 Meta Review

Full Review Summary

AI Review

Paper summary:

The paper investigates how multi-turn human red teaming exposes vulnerabilities in safety-aligned large language models (LLMs) that appear robust under standard single-turn, automated adversarial evaluations. The core problem addressed is that most defenses are evaluated only against single-turn automated attacks, whereas real-world malicious use will involve multi-turn conversations in black-box chat interfaces. The key idea is to organize expert human red teamers into a structured multi-stage pipeline that conducts multi-turn jailbreak attempts under a realistic, time-limited threat model, then validate their success with human reviewers and a GPT-4o harm classifier. Concretely, the authors attack four defenses (RR, LAT, DERTA, CYGNET) built on Llama-3-8B-Instruct using HarmBench's 240 harmful behaviors, plus an RMU-based machine-unlearned model evaluated on free-response WMDP-Bio questions; they compare human jailbreaks to six black- and white-box automated attacks and then release successful conversations as the Multi-Turn Human Jailbreaks (MHJ) dataset, labeled with a taxonomy of tactics. Empirically, human multi-turn jailbreaks achieve much higher attack success rates (up to 51–88% on open-source defenses and 70% on CYGNET) than individual or ensembled automated attacks, and they substantially outperform automated attacks in recovering dual-use biosecurity knowledge from an RMU-unlearned model. The authors conclude that existing automated robustness benchmarks underestimate real-world risk, that defenses are not robust to realistic multi-turn, human-crafted attacks, and that MHJ can catalyze research on stronger defenses and automated multi-turn red teaming.

Review summary:

Overall, the paper makes a timely and substantively important empirical contribution by demonstrating that current LLM defenses remain highly vulnerable under realistic multi-turn human attacks, though several methodological details and comparisons could be clarified or strengthened. The exposition is generally clear and well-structured, with helpful figures (e.g., the ASR bar plots on pages 2, 6, and 7 and the tactic taxonomy in Table 1) and concrete jailbreak examples illustrating key tactics such as obfuscation and "Opposite Day." The experimental setup is thoughtfully designed, but the comparison between human and automated attacks relies on different budgets and evaluation pipelines, which complicates interpretability of absolute ASR gaps. The multi-turn threat model and the MHJ dataset, including tactic labels and red teamer commentary, appear novel and valuable to the robustness community, especially in combination with the unlearning case study. At the same time, the work is limited by the small number of defenses, the lack of cost-normalized comparisons, the heavy reliance on a single harm classifier, and relatively shallow quantitative analysis of how multi-turn dynamics and tactics interact with specific defenses. Addressing these issues, along with tighter positioning relative to recent multi-turn and in-the-wild red teaming work, would significantly strengthen the paper's impact.

Strengths:

1. Realistic and well-motivated threat model

- The paper focuses on black-box chat-interface access with multi-turn conversations and limited temperature control, which closely mirrors real-world usage of systems like ChatGPT or CYGNET rather than the more generous API or white-box threat models that many automated attacks assume.
- The authors explicitly disallow access to assistant tokens, logit probabilities, and internal model states, and they consider post-circuit-breaker behavior for CYGNET separately, which helps keep the study aligned with realistic attacker capabilities.

2. Carefully designed human red teaming pipeline

- The multi-stage pipeline with up to three independent red teamers, two human reviewers, and a final GPT-4o harmfulness filter is a thoughtful attempt to both encourage diverse attacks and control false positives (Figures 2 and 6, Section 3.2, Appendix A.3).
- The 30-minute per-behavior time budget, separate "partial" and "full" success labels, and escalation to a third red teamer when reviewers disagree are sensible choices that partially standardize human efforts while still allowing creativity.

3. Strong empirical evidence of gaps between automated and human attacks

- On HarmBench, human ASRs are dramatically higher than both individual automated attacks and the automated ensemble across all open-source defenses (e.g., for RR: 51.7% vs. 8.3% ensemble; for LAT: 75.4% vs. 10.0%; for DERTA: 87.9% vs. 68.3%; Table 2, Figures 1 and 3).
- For the unlearning case study, humans achieve 39.6% higher ASR than the best automated attack and 25.6% higher than the automated ensemble on free-response WMDP-Bio questions, while all prior attacks reported in Li et al. (2024a) showed 0% ASR on CYGNET in a similar setup.
- The breakdown by harm category (Figure 5) and by tactic (Figure 8) shows that some areas (e.g., harassment/bullying) are particularly easy for humans relative to automated attacks, strengthening the claim that automated ASRs are not a reliable proxy for human adversary performance.

4. Insightful analysis of multi-turn dynamics and tactic taxonomy

- The observation that over 90% of successful HarmBench jailbreaks and 86% of successful WMDP-Bio jailbreaks use more than one turn provides direct evidence that multi-turn capabilities are central to real-world vulnerability, not just a convenience for humans.
- The tactic taxonomy in Table 1 and its expansion in Appendix D (e.g., direct request, echoing, hidden intention streamline, obfuscation, output format, request framing with variants like fictionalization, "Opposite Day," roleplay, urgency) is rich and grounded in practical red teaming experience across multiple commercial engagements.
- The qualitative red teamer surveys (Appendix C) provide nuanced insights into each defense's strengths and weaknesses, such as RR's "gibberish" mechanism and LAT's resistance to "Opposite Day," which will be useful to both defense designers and future attackers.

5. MHJ dataset and safety-conscious release

- MHJ includes 2,912 prompts across 537 jailbreak conversations, with metadata such as tactic labels, time taken, and temperature, plus free-form red teamer commentary that captures design choices and ideas tried.
- The authors remove model completions and export-controlled content, verify compliance with ITAR/EAR, and limit MHJ to system/user messages, which is a reasonable compromise between research usefulness and safety (Section 7).
- The dataset nicely complements prior resources like HarmBench, WildTeaming-style in-the-wild datasets (Jiang et al., 2024b), and jailbreak taxonomies (Inie et al., 2023; Weidinger et al., 2024), and can be used to train or evaluate multi-turn attack generators.

6. Thoughtful discussion of limitations and threats to validity

- Section 5 acknowledges many important caveats: different resource limits for humans vs. automated attacks, varying red teamer skill levels, different positions of GPT-4o in the evaluation pipeline (first-pass vs. final filter), and the higher cost of human red teaming compared to automated attacks.

- The authors correctly caution against over-interpreting ASR differences as a direct ranking of defenses and emphasize that the goal is to highlight vulnerability and motivate stronger automated attacks, not to declare human red teaming "better."

Weaknesses and suggestions:

1. Comparability between human and automated attacks is still murky

- The comparison between human and automated ASRs is confounded by multiple factors: humans have multi-turn access while automated attacks are single-turn; humans are time-limited but not query-limited; different temperatures are available to humans but not explicitly to automated attacks; and the harm-classification pipeline is slightly different.
- Although the authors discuss many of these issues in Section 5 and Appendix A.3, the main figures (especially Figures 1 and 3) still invite a direct "humans beat automated attacks by X%" reading that may overstate the strength of the evidence.
- Suggestion: Add a dedicated subsection in the main text that (i) explicitly enumerates these differences, (ii) presents at least one more controlled comparison—e.g., limit humans to a single turn on a subset of behaviors, or simulate a multi-turn version of a simple automated attack like ZeroShot or PAIR—and (iii) reports cost-normalized metrics such as ASR per query or per minute of effort. Even partial results would help calibrate claims.

2. Limited coverage of defenses and baselines

- The study evaluates three open-source defenses (RR, LAT, DERTA) plus the proprietary CYGNET and a single unlearning method (RMU), all on Llama-3-8B-Instruct.
- There is no explicit baseline of the un-defended Llama-3-8B model's ASR under the same human multi-turn pipeline, making it hard to quantify how much robustness these defenses actually add in this threat model.
- Suggestion: Include a baseline for the raw base model in the main figures/tables, or at least in an appendix, and discuss how much the defenses reduce ASR relative to that baseline under human multi-turn attacks. Additionally, even a small experiment on a different foundation model (e.g., a smaller or larger Llama variant) would help indicate whether the findings are tied to this specific backbone.

3. Heavy reliance on GPT-4o as a harm classifier

- GPT-4o is used both as a first-pass filter for automated attacks and as a last-pass filter for human attacks. For automated attacks, its high recall but low precision is mitigated by human verification; for human attacks, its moderate precision is used to discard some human-labeled positives, providing a "conservative lower bound" on ASR.
- However, the paper does not deeply analyze how sensitive results are to this classifier choice, nor how often GPT-4o's misclassifications qualitatively change the nature of the jailbreaks.
- Suggestion: Add a small ablation where a subset of attacks are re-labeled with a different classifier and/or double-human annotation to estimate label variance. Reporting inter-annotator agreement between reviewers and GPT-4o on a random subset would also increase confidence.

4. Unlearning case study design differences deserve more emphasis

- For WMDP-Bio, human red teamers get unlimited time up to a global 240-hour budget, the pipeline differs (no multi-reviewer + GPT-4o stack), and success is judged by semantic containment of the ground-truth answer by a human reviewer rather than by the HarmBench classifier.
- Given these substantial procedural differences, the WMDP-Bio human ASR cannot be directly compared even qualitatively to HarmBench ASRs, yet the main text sometimes juxtaposes them in ways that might invite such comparisons.
- Suggestion: More prominently flag in the main Results section that the unlearning experiment uses a different pipeline, and isolate its ASR numbers in separate tables/figures to avoid visual comparison to HarmBench bars. A schematic for the WMDP pipeline similar to Figure 2 would also help.

5. Multi-turn dynamics and query budget are under-analyzed

- While the paper reports average successful attack time per defense and the proportion of successful attacks requiring more than one turn, it does not analyze the distribution of number of turns, queries, or total tokens, nor examine how ASR changes if one restricts the attack to, say, ≤ 3 turns.
- Without such analysis, it is difficult to judge how much of the human advantage is due to multi-turn interaction per se versus broader human flexibility.
- Suggestion: Use the MHJ logs to provide histograms of turns-to-success, tokens-to-success, and per-tactic turn counts. It would be particularly interesting to see whether tactics like "hidden intention streamline" or "obfuscation" uniquely require many turns, or whether most successful jailbreaks are relatively short.

6. Positioning relative to closely related work could be sharper

- The paper cites a number of related efforts but does not deeply distinguish how MHJ and its tactic taxonomy differ from WildTeaming, Crescendo, and other red teaming frameworks.
- Suggestion: Add a subsection in Related Work that explicitly compares MHJ and the tactic taxonomy to in-the-wild datasets, Crescendo's automated multi-turn strategy, and STAR-style frameworks. Highlight that MHJ focuses on expert-designed, time-limited multi-turn interactions against specific defenses, and that its taxonomy is annotated for every conversation.

7. Dataset design trade-offs could be spelled out more concretely

- MHJ omits assistant responses and many high-risk behaviors, focusing instead on system and user messages. This constrains certain downstream uses, such as studying defense behavior directly.
- Suggestion: Expand Section 4.3/7 to explicitly discuss which research questions MHJ is and is not suitable for. MHJ is ideal for training automated prompt engineers, building tactic-aware classifiers, or benchmarking multi-turn prompt-based attacks, but less ideal for studying defense-side generation patterns.

8. Demographic and expertise limitations of red teamers

- The red teamers are all native English speakers from the US/Canada, with only a minority having domain expertise. This likely underestimates the capabilities of globally distributed, specialized adversaries.
- Suggestion: Make this limitation more prominent and frame the reported ASRs as lower bounds. A small pilot with domain experts on a subset of WMDP behaviors could be highly informative.

9. Minor clarity and presentation issues

- Some terminology is slightly inconsistent; a concise glossary would help.
- A table summarizing differences between defenses and pipelines would aid comprehension.
- Some jailbreak transcripts in Appendix D are long; trimming them or adding brief annotations could improve readability.

Overall: With clarifications on comparability, stronger contextualization relative to related work, and deeper analysis of multi-turn dynamics, this paper has the potential to become a central reference on the limitations of current LLM defenses and the importance of realistic, human-driven red teaming.

Reviewer #1

Rating: 5: marginally below the acceptance threshold

Justification

The AI review does not identify any novel key technologies proposed in this paper, noting that it merely presents a comparative experiment on human-led multi-turn attacks against LLMs using automated attacks as the baseline. Consequently, this paper appears to be strictly an experimental investigation rather than a proposal of novel techniques. Furthermore, the paper's primary experimental contribution—or finding—is based on an unfair comparison. As the AI review pointed out, the baseline methods are not comprehensive; specifically, the automated attacks are limited to single-turn interactions, whereas human attackers can execute multi-turn attacks.

Reviewer #2

Rating: 5: marginally below the acceptance threshold
Justification
Strengths:

- The authors reveal a new vulnerability beyond current evaluation protocol; multi-turn human red teaming.
- The topic is timely and substantially crucial.
- The evaluation of machine unlearning methods demonstrates the effectiveness of human red teaming.

Weaknesses:

- The model evaluated is limited to llama-3-8b-instruct, which is a significant limitation.
- The human red teaming pipeline seems quite hard to reproduce with the limited explanation in Sec 3.2.
- The cost is very different between automated and human attacks, which makes the comparison unfair. Automated attacks should also follow a similar pipeline that allows multiple attempts until success.

Reviewer #3

Rating: 5: marginally below the acceptance threshold
Justification
According to the AI review summary, the focus on multi-turn jailbreaking is both realistic and compelling, and the proposed pipeline provides a reasonable simulation of real-world adversarial interactions. However, there are major weaknesses:

1. The contribution is weakened by the limited set of evaluated defenses, which make the result less convincing.
2. The evaluation is conducted with llama-3b model only, and there is no experiments on stronger LLM baselines.
3. The fact that all red teamers are native English speakers introduces a potential source of bias and raises the generalization concerns.

Thus, I would lean towards rejection. I will recalibrate the score if the concerns are properly addressed during the rebuttal phase.

Reviewer #4

Rating: 6: marginally above the acceptance threshold
Justification
The paper addresses an important problem and provides solid empirical evidence that multi-turn human red teaming can reveal vulnerabilities missed by automated attacks. The threat-model design and the MHJ dataset are valuable contributions. However, key limitations reduce the strength of the conclusions: the comparison between human and automated attacks is confounded by differing pipelines and resource budgets, the coverage of defenses and baselines is limited, and the analysis of multi-turn dynamics is not sufficiently deep. These issues prevent the work from reaching a stronger recommendation, placing it only marginally above the acceptance threshold.

Reviewer #5

Summary:
The paper highlights the limitations of current defenses in large language models (LLMs) against multi-turn adversarial attacks. Unlike single-turn automated attacks, multi-turn human-driven jailbreaks show a much higher success rate, bypassing existing defenses by strategically interacting with models over multiple conversation turns. The authors conducted tests with human red teamers who revealed that many LLMs, despite advanced safeguards, remain vulnerable in realistic settings. The study also introduces the "Multi-Turn Human Jailbreaks" (MHJ) dataset, containing thousands of successful jailbreak examples to support the development of more robust defenses that address multi-turn interaction vulnerabilities.
Strengths:

1. The main strength of the paper is evaluating the LLM jailbreak robustness from a different perspective and in a like-life setting. A realistic threat model can always give a better idea about the robustness.
2. I appreciate the authors for publishing the dataset of 2912 prompts. I believe this can help the researchers and community.

Weaknesses:

1. Only the Llama3-8b-Instruct model was used for evaluation. Other models, especially the stronger ones should have been included.
2. The chosen baselines and the multi-turn-human-jailbreak approach are different from each other in many ways, such as timing, budget, etc. Seems like the MHJ attack is a soft-constrained version of attacks.
3. Authors admit that "the skill and experience of individual red teamers may vary". In that case, how can researchers compare their defense against human-based attacks in the common ground? What would be the baseline of a red-teamer's skill and experience? This is one of the reasons for using automated attacks for evaluation to this date.
4. Other automated attacks, such as [1, 2] could be included in the evaluation.
5. There was no discussion on plausible defenses for such multi-turn-human-based attack.

References:
[1] Fast Adversarial Attacks on Language Models In One GPU Minute
[2] Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack
Questions:

1. Why the evaluation was done only on one model? Why stronger models like ChatGPT was not included?
2. How were other automated attacks implemented for the WMDP-bio questions?
3. Is it possible to mention the monetary cost of this human-based attack?

Rating: 6: marginally above the acceptance threshold
Justification:
Positive contributions:

- The reviewer clearly acknowledges meaningful strengths: a realistic, human-driven multi-turn threat model and the release of a sizable jailbreak dataset (2912 prompts), which they explicitly say can help researchers and the community.
- The framing ("main strength", "I appreciate the authors…") is appreciative, not dismissive, which suggests the work is viewed as valuable.

But with significant limitations:

- Evaluation is done on only one relatively weak model (Llama3-8B-Instruct); the reviewer explicitly questions why stronger models (e.g., ChatGPT) were not included.
- Baseline mismatch (MHJ vs other attacks differ in timing, budget, constraints) makes the experimental comparison weaker.
- The reviewer points out a core methodological issue: variability in human red-teamer skill makes it hard to compare defenses "on common ground," which undermines the generality and reproducibility of the evaluation.
- Some important baselines are missing (other automated attacks [1,2]) and there is no discussion of defenses against the proposed attack setting.
- The questions at the end (about model choice, implementation details, and monetary cost) indicate that the paper leaves notable gaps in clarity and completeness.

Overall, the reviewer seems to see this as valuable but incomplete work: interesting idea and useful dataset, but with methodological and experimental shortcomings that prevent it from being a clear accept. That aligns best with "6: marginally above the acceptance threshold."

Failure-mode flags: superficial, inaccurate/hallucinatory, misguided
Justification:
Superficial
The review engages only shallowly with the paper's main contributions. It focuses on high-level aspects (e.g., that the work studies multi-turn human jailbreaks and releases a dataset) but does not seriously discuss:

- Specific empirical findings (e.g., how multi-turn human attacks compare quantitatively to automated attacks across defenses).
- The proposed taxonomy and analysis of jailbreak tactics and failure modes.
- The detailed treatment of limitations, threat model assumptions, and evaluation caveats provided by the authors.

Instead, several weaknesses largely restate concerns that are already explicitly acknowledged and discussed in the paper (e.g., differences to automated attacks, variability in human red-team skill), without engaging with how the paper addresses or mitigates them. This makes the review feel surface-level relative to the richness of the manuscript.
Inaccurate / hallucinatory
The reviewer claims that there is "no discussion on plausible defenses for such multi-turn-human-based attack." This is inaccurate. The manuscript includes an explicit discussion of:

- Why current defenses, which are mostly tuned for single-turn settings, are insufficient for multi-turn adversarial interactions.
- How future defenses might need to adapt (e.g., post-training objectives that account for multi-turn interactions and longer contexts, and the implications of tool use and long-context agents).

While the paper may not introduce a new concrete defense algorithm, it *does* discuss implications and directions for defense in this setting. Describing this as "no discussion" is therefore factually incorrect.
Misguided
Several criticisms are aimed at choices that are central, intentional aspects of the paper's design rather than obvious shortcomings:

- The complaint that "only the Llama3-8b-Instruct model was used" treats the absence of multiple or stronger models (e.g., ChatGPT) as a major flaw, without acknowledging the practical cost of large-scale multi-turn human red-teaming and the paper's primary goal: to compare *defenses* under a fixed base model and realistic human-attack setting. The design choice is methodologically defensible given the problem the paper targets.
- The statement that the MHJ attack is "a soft-constrained version of attacks" and that the baselines differ in timing/budget is framed as a methodological weakness, but the paper explicitly defines a multi-turn human-chat threat model and explains why existing automated single-turn attacks are not directly comparable. Criticizing this as though the mismatch were an oversight rather than a deliberate focus is somewhat misplaced.
- The concern about variability in red-teamer skill is presented as a reason to prefer automated attacks, but the manuscript itself highlights this limitation, motivates the use of multiple experienced red teamers, and releases the MHJ dataset precisely to help the community move toward more standardized and automatable evaluations. Treating this as an unaddressed problem rather than an explicitly discussed trade-off underestimates the authors' own framing.

Overall, the review is not especially harsh or lenient, but it does contain one clear factual inaccuracy and critiques some well-justified design decisions as if they were simple omissions, while remaining relatively superficial about key contributions of the paper.

Reviewer #6

Summary:
This paper focuses on the safety vulnerabilities of LLMs in multi-turn conversations. The authors designed a comprehensive human red teaming pipeline, including "Attempt Jailbreak" and "Validate Jailbreak," and found that LLMs exhibit more safety risks in multi-turn dialogue conversations compared to single-turn automated attacks.
Strengths:

1. Most previous papers have focused on LLM jailbreaking in single-turn conversations, while this work emphasizes multi-turn conversations, presenting a new scenario.
2. The red-teaming results show that multi-turn human jailbreaks outperform current automated attacks, which exposes more safety vulnerabilities in LLMs.
3. The authors have also released their Multi-turn Human Jailbreaks dataset, which can support further research on multi-turn jailbreaks.

Weaknesses:

1. The human red-teaming was conducted only on Llama; further evaluations on other LLMs would more comprehensively illustrate the safety vulnerabilities in multi-turn conversations. Additionally, the human red-teaming data collected on Llama could potentially be used to examine safety issues in other LLMs under multi-turn scenarios.
2. A more detailed analysis of the effectiveness of different tactics could provide deeper insights into multi-turn jailbreaks.

Questions: NA

Rating: 8: accept, good paper
Justification:
The review describes a solid and meaningful **contribution** with clear novelty and practical value:

- Novelty and relevance: The paper moves beyond the well-studied single-turn jailbreak setting and systematically focuses on multi-turn conversations, which is an important and underexplored but realistic scenario for LLM safety.
- Methodological contribution: The authors design a comprehensive human red-teaming pipeline (with explicit Attempt Jailbreak" and Validate Jailbreak" stages), suggesting a thoughtful and structured experimental setup rather than ad-hoc red-teaming.
- Empirical significance: The finding that multi-turn human jailbreaks outperform current automated attacks indicates that existing defenses are insufficient in realistic interaction patterns and highlights practically important safety vulnerabilities.
- Resource contribution: The release of the Multi-turn Human Jailbreaks dataset is a concrete asset that can enable and accelerate further research on multi-turn jailbreaks.

The **weaknesses** raised are important but appear incremental rather than fatal:

- Limited model coverage: Experiments focus only on Llama; broader evaluation across multiple LLMs would strengthen the generality of the conclusions, but the review does not indicate that this limitation invalidates the core findings.
- Depth of analysis: The reviewer asks for more detailed analysis of different tactics. This suggests room for improvement in insight and interpretability, but not a fundamental flaw in the methodology or results.

Overall, the review portrays the work as a good paper with clear novelty, solid methodology, and useful artifacts, with limitations that are reasonable for a conference submission rather than reasons for rejection. Hence, a score of 8 (accept, good paper) is appropriate.

Failure-mode flags: superficial.
Justification:

1. The review remains at a high level and does not engage deeply with methodological details. For example, it mentions a "comprehensive human red teaming pipeline" with stages like "Attempt Jailbreak" and "Validate Jailbreak," but provides no discussion of how these stages are designed, what criteria are used, or any potential limitations or biases of this setup.
2. The strengths section largely restates the paper's apparent contributions (focus on multi-turn conversations, better performance of human jailbreaks, release of a dataset) without critically assessing their significance, novelty, or empirical robustness.
3. The weaknesses identified are quite generic. Saying that red-teaming only on Llama is limiting and that "more detailed analysis of the effectiveness of different tactics" would be useful are reasonable points, but the review does not specify what concrete additional experiments or analyses would meaningfully improve the work, nor does it probe any deeper conceptual or empirical shortcomings.
4. Overall, the review reads as a brief summary plus a couple of high-level comments, rather than a thorough evaluation of experimental design, analysis, or positioning relative to prior work, which is characteristic of a somewhat superficial review.

Reviewer #7

Summary:
This paper exposes the significant safety vulnerabilities of Large Language Models (LLMs) in multi-turn dialogue by conducting human red-teaming. By exploiting a diverse of human tactics, human red-teamers achieve a 70% attack success rate on Harmbench against strong defense baselines. This paper finally compiles these results into the Multi-Turn Human Jailbreaks (MHJ) dataset.
Strengths:

- **Valuable Findings, Including an Open-Source Jailbreak Dataset:** This paper provides a valuable contribution by conducting the first multi-turn human red-teaming experiments, revealing significant safety vulnerabilities that bypass current state-of-the-art defense methods. Given that existing safety alignment techniques primarily target single-turn jailbreaks, these findings effectively motivate the community to develop more robust defenses capable of countering multi-turn attacks.

Weaknesses:

- **Lack of Strong Automated Attack Baselines:** Some black-box attacks, such as CodeAttack [1] and PAP [2], are not included in the experiments, despite utilizing similar "Request Framing" tactics as those in this paper. Including these baselines in future comparisons would offer a clearer perspective on the relative effectiveness of this paper's approach.
- **Unclear Evaluation Metric for Model Unlearning:** This paper proposes manual review to assess attack outcomes in model unlearning experiments, yet does not clearly define what constitutes a successful attack. For example, does success mean the model output includes content from the forget set? A more detailed description of the evaluation metric for model unlearning experiments would improve clarity.

1. *CodeAttack: Revealing Safety Generalization Challenges of Large Language Models via Code Completion.*
2. *How Johnny Can Persuade LLMs to Jailbreak: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs.*

Questions:

- **Counter-Intuitive Experimental Results:** The effectiveness of the "Direct Request" tactic used by human red-teamers, as shown in Figure 8, is unexpected. In this paper, "Direct Request" refers to directly asking the LLM to produce harmful content, which should logically not be highly effective against safety defenses. An analysis of how various tactics influence the attack success rate would be beneficial, as it could offer insights into the specific failure modes of these models.
- **Lack of In-Depth Analysis on the High Effectiveness of Human Multi-Turn Jailbreaks:** Many of the tactics employed by human red-teamers, such as "Injection," "Obfuscation," "Output Format," and "Request Framing," are also used by automated attacks. Are there particular factors that enable human-led jailbreaks to outperform automated attacks using similar tactics? Identifying these factors could provide valuable insights into the unique strengths of human-driven attacks.

I would consider raising my score if the authors address these concerns and questions.

Rating: 6: marginally above the acceptance threshold
Justification:
The review is overall positive about the core contribution: it highlights that the paper is (i) the first to conduct multi-turn human red-teaming experiments on LLMs, (ii) exposes substantial safety vulnerabilities with a high attack success rate (70% on HarmBench), and (iii) releases a useful open-source dataset (MHJ). These points indicate that the reviewer sees the work as novel, impactful, and relevant to the community.
However, the reviewer also identifies non-trivial weaknesses. They point out missing strong automated black-box baselines (CodeAttack, PAP), which limits the strength of the empirical comparison, and they raise concerns about the clarity of the evaluation metric for model unlearning. In addition, the questions section emphasizes that some of the experimental findings (e.g., the high effectiveness of direct requests) are under-analyzed, and that the mechanisms behind the superiority of human multi-turn attacks over automated ones are not sufficiently explored. The reviewer explicitly states that they would consider *raising* their score if these concerns were addressed, suggesting that their current stance is cautiously positive rather than fully enthusiastic.

Taken together, this combination of a clearly valuable main contribution with methodological and analytical gaps is most consistent with a score of 6 (marginally above the acceptance threshold): the paper is viewed as worth accepting, but with reservations that prevent it from reaching the level of a clear "accept, good paper" (8) or higher.

Failure-mode flags: inaccurate/hallucinatory, superficial
Justification:
Inaccurate / hallucinatory

- The review claims that the evaluation metric for the unlearning experiments is "not clearly defined" and asks, for example, whether success means "the model output includes content from the forget set." However, the paper does spell this out: for WMDP-Bio, they use the ground truth answer corresponding to each prompt to verify the attack, and define success as the model output semantically containing all the information in the label, as judged by a human reviewer. Thus, the alleged lack of clarity is factually incorrect.
- The strength section claims this paper "provides . . . the first multi-turn human red-teaming experiments." The paper instead positions its work as examining a realistic but underexplored multi-turn threat model and notes that large-scale human red teaming is rare outside industry, but it does not assert a clean "first" in the way the reviewer phrased it. This is a mild overstatement of novelty not actually grounded in the manuscript.

Superficial

- The criticism of an "unclear evaluation metric" for model unlearning ignores the dedicated description of how success is judged in the WMDP-Bio setting (a semantic match to ground-truth answers via human review). This suggests the reviewer did not fully engage with the relevant part of the methodology.
- The "lack of strong automated attack baselines" complaint is framed entirely around the absence of two specific attacks (CodeAttack and PAP), without engaging with the existing suite of six automated attacks (including white-box gradient-based methods) and the ensemble upper-bound attack success rate (ASR) the authors already use, nor with their explicit threat-model-driven exclusions (e.g., Crescendo due to missing code, or attacks that modify model internals). This makes the baseline critique feel somewhat shallow relative to the paper's fairly detailed experimental setup.
- The review does not really grapple with the paper's central design choices and limitations (multi-turn human pipeline, validation stages, conservatively biased ASR estimates, and discussion of differences between human and automated pipelines) that occupy multiple sections and appendices. Instead, it focuses on a small number of surface-level points (missing two baselines, asking for more analysis of one tactic) and offers only a single, high-level strength. This imbalance contributes to the overall superficiality.

Reviewer #8

Summary:
This paper looks at red teaming via multi-turn human interaction and releases an accompanying dataset of the jailbreaks. This is a useful study: the vast majority of the research effort and available data is focused on single-turn attacks, with only a few exceptions (e.g. Anthropic's hh-rlhf dataset having harmful multi turn conversations). Overall, the paper is clear and well written, and supplies useful well curated data for future work.

Strengths:

- Human driven red teaming data, particularly multi-turn due to its scarcity, is always useful. There are many synthetically generated datasets, but high quality human driven attacks are a valuable resource. In particular, each prompt here had a high degree of manual curation (in comparison to datasets like hackaprompt which contain many low quality samples).
- The analysis across a few defenses is useful, and does highlight that even SOTA defenses on highly studied benchmark harmful questions can be broken in around 15min on a large subset of questions (e.g. often quicker than algorithmic attack runtime).

Weaknesses:

- It would have been useful to see results which have human single turn jailbreak attempts in Figure 3. At the moment, the "Human" attacks have two variables changed compared to the others: the attack source (e.g. handcrafted), and additionally have multi-turn capabilities. Hence it makes it challenging to disambiguate if the difference in performance is due to the multi-turn aspect, or if humans given enough time remain better than automated based methods at creating jailbreaks.
- I am unclear as to why results against Cygent defense could not be carried out in the same setup as the original paper: the Llama model is open source, and the defense has a published paper. It would have enabled stronger reproducibility and clearer interpretation of results.
- Releasing the non-successful jailbreak attempts as well can be beneficial as it is still a useful resource, for example as training/fine-tuning data or to carry out further analysis.
- Although different styles of harmfulness were investigated: both "regular" harmbench style questions, but also WMDP-Bio for different attack objectives and domain performance from looking into the supplementary material it seems like just the harmbench data was released. Given the dataset is the core contribution of the paper it would have been useful to include the other domain data.

Questions:
Are there statistics on how many red-team members there are/distributions on number of samples provided per red-team member?

Rating: 8: accept, good paper
Justification:

- The review describes the work as a *useful study*, highlighting that most prior work and data focus on single-turn attacks, whereas this paper targets the underexplored setting of *multi-turn human red teaming* and releases a corresponding dataset. This indicates a clear and meaningful contribution rather than an incremental tweak.
- The dataset is portrayed as *high-quality* and *well curated*, with an explicit comparison to synthetically generated or lower-quality resources (e.g., hackaprompt). This suggests strong lasting value to the community.
- The analysis of multiple defenses and the empirical finding that even SOTA defenses can be broken in roughly 15 minutes are framed as *useful* and informative, reinforcing the paper's impact on understanding current safety limitations.
- The weaknesses raised (lack of a human single-turn baseline for clearer attribution to multi-turn effects, missing Cygent reproduction details, omission of unsuccessful attempts, and incomplete release of all domain data) are presented as limitations or missed opportunities for strengthening the work, but not as fundamental flaws that would undermine the main results.
- Overall, the tone is clearly positive and confident in the paper's value and clarity, but not enthusiastic enough (nor claiming sufficiently broad or groundbreaking impact) to justify a "strong accept." This aligns best with an *accept, good paper* rather than merely marginal acceptance.

Failure-mode flags: Inaccurate / hallucinatory (minor).
Justification:
The review's second weakness claims it is unclear why experiments on Cygent/CYGNET" could not be reproduced in the same setup as the original paper, arguing that the Llama model is open source, and the defense has a published paper." However, the manuscript explicitly states that CYGNET is *closed-source/proprietary* and that the authors conducted human red teaming via an external API (Gray Swan AI), while also noting they did not run automated attacks on CYGNET in their own setup and instead cite automated results from the original work (with caveats about comparability). This makes the review's implication that straightforward reproduction should have been possible materially inconsistent with the manuscript's described experimental constraints.

AC Meta-review

Recommendation: Reject
Meta Review:
Strength:

1. Most previous papers have focused on LLM jailbreaking in single-turn conversations, while this work emphasizes multi-turn conversations, presenting a new scenario of evaluating the LLM jailbreak robustness from a different perspective.
2. The authors have also released their Multi-turn Human Jailbreaks dataset, which can support further research on multi-turn jailbreaks.

Weakness:

1. This paper appears to be strictly an experimental investigation (based on unfair comparison) rather than a proposal of novel techniques.
2. Limited evaluations: limited to llama-3-8b-instruct, limited set of evaluated defenses, and unclear evaluation metric for model unlearning.
3. Lack of more detailed analysis of the effectiveness of different tactics in providing deeper insights into multi-turn jailbreaks.

While this work contributes to the study of Multi-turn Human Jailbreaks, it cannot be accepted at its current status due to the main weaknesses, including limited evaluations and lack of methodological analyses.

CiteAssist

Citation Sheet

Generated with citeassist.uni-goettingen.de

BibTeX Entry

```
@inproceedings{Wang2026PeerReview,
  author={Wang, Hanrui and Spinde, Timo and Habereder, Isabella and Lu, Chun-Shien and
      Echizen, Isao},
  title={Peer Review Should Constrain Evaluative Authority},
  booktitle={Proceedings of the 40th Conference on Neural Information Processing Systems (
      NeurIPS 2026)},
  note={Position Paper Track},
  year={2026},
  month={12}
}
```