

Position: Peer Review Demands AI–Human Mutual Supervision

Hanrui Wang¹ Timo Spinde^{1,2} Isabella Habereeder^{2,3} Chun-Shien Lu⁴ Isao Echizen^{1,2,5}

Abstract

Peer review forms the basis for decisions in academic publishing. However, modern challenges, such as the increasing number of submissions, pose several problems for the academic publication process, leading to unprofessional, strategically biased, and sometimes unethical reviews. To ensure the quality of peer reviews, it is not expedient to introduce stricter guidelines for both humans and AI, or to use AI for automated review and meta-review generation. We therefore propose an AI-human mutual supervision framework. **We argue** for assigning humans and AI distinct, non-collaborative roles in which they supervise one another to address the identified problems in the current peer reviewing process. Our position can serve as a starting point for further discussion and cooperation to ensure an appropriate quality of peer reviews.

1. Introduction

Peer review is a cornerstone of academic publishing and of scholarship itself (Lee et al., 2013). In principle, independent experts act as third-party referees, whose assessments certify the quality and reliability of scientific claims and justify decisions across journals, conferences, funding agencies, and public institutions. However, in recent years, the number of submissions has increased sharply. For example, the number of papers submitted for review at leading computer science conferences almost doubled between 2024 and 2025¹ (see Figure 1). This leads to problems such as reviewers being overloaded, which, in turn, increases the variation in review quality.

To address challenges in maintaining peer-review quality as

¹Echizen Laboratory, National Institute of Informatics, Tokyo, Japan ²Media Bias Group, National Institute of Informatics, Tokyo, Japan ³Georg-August-University Göttingen, Göttingen, Germany ⁴Institute of Information Science, Academia Sinica, Taipei, Taiwan ⁵The University of Tokyo, Tokyo, Japan. Correspondence to: Hanrui Wang <hanrui.wang@nii.ac.jp>.

Preprint. February 4, 2026.

¹<https://papercopilot.com>

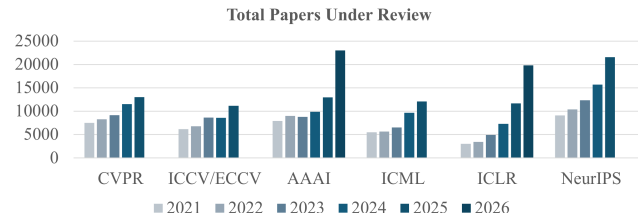


Figure 1. Rapid and sustained growth of peer-review workload. Values indicate submissions that proceeded to review; total submissions are higher (e.g., AAAI’26 received 29K submissions, of which 23K were reviewed).

scale and workload increase, venues have adopted several stopgap measures. For example, some outlets have implemented author-obligated reviewing², which expands the reviewer pool by requiring qualified authors to serve as reviewers. As another alternative, IJCAI-ECAI’26 introduced a \$100 submission fee intended both to compensate reviewers and to regulate the volume and quality of submissions³. Moreover, AI technologies are increasingly integrated into the review pipeline, supporting writing assistance⁴, automated review⁵, and review-feedback mechanisms⁶ (Farber, 2025; Kousha & Thelwall, 2024; Chang et al., 2025; Jin et al., 2024). Review policies are also adapting in response to the changing landscape of AI use in research. For example, AAAI’26 embraces AI-assisted reviewing⁵, and ICLR permits AI for writing and piloted a review-feedback system (Thakkar et al., 2025) under disclosure and auditability requirements⁷. However, there are also arguments for excluding AI from the review process. For example, CVPR bans LLM usage outright due to confidentiality and ethical concerns⁸. ICML’26 allows authors to choose between permissive AI assistance (e.g., understanding or polishing text, as in ICLR) and stricter policies (e.g., CVPR-style prohibitions), while explicitly disallowing the use of LLMs

²<https://cvpr.thecvf.com/Conferences/2026/AuthorGuidelines>

³<https://2026.ijcai.org/primary-paper-initiative/>

⁴<https://blog.icml.cc/2026/01/14/icml-experimental-program-using-googles-paper-assistant-tool-pat/>

⁵<https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/>

⁶<https://blog.iclr.cc/2025/04/15/leveraging-llm-feedback-to-enhance-review-quality/>

⁷<https://iclr.cc/Conferences/2026/ReviewerGuide>

⁸<https://cvpr.thecvf.com/Conferences/2026/ReviewerGuidelines>

to generate reviews⁹ (see discussion in Appendix A).

Despite these measures, incidents persist that point to a systematic decrease in the integrity and professionalism of peer review, collectively undermining the fairness, rigor, and trustworthiness of the review process. Documented problems include reviewers’ basic unfamiliarity with foundational methods¹⁰, the submission of formulaic and potentially AI-generated reviews¹¹, unprofessional interactions between authors and reviewers¹², and breaches of reviewer anonymity¹³. Taken together, we believe that these incidents and the constantly increasing amount of submissions suggest that incremental adjustments are insufficient so that the peer review process requires substantive structural reform (Yang, 2025; Kim et al., 2025; Seghier, 2025; Su, 2025).

In summary, peer review faces a structural problem: humans with conflicts cannot reliably supervise one another, and AI systems alone cannot be trusted without oversight. We argue that fairness in peer review cannot depend on the goodwill, competence, or ethics of human or AI reviewers, but must be engineered through a balanced system of mutual supervision. **Our position is that AI should neither replace reviewers nor serve as an obedient assistant to them, but rather serve as an independent supervisory actor:** checking human reviews, being checked by human supervisors, and enabling a separation of powers within the review process.

Achieving such a system requires redesigned pipelines, new datasets, principled evaluation methodologies, and robust mechanisms for detecting misconduct. We have therefore worked out a systematic redesign of the peer review process and will contribute as follows.

- We analyze problems that arise when humans or AI systems operate independently in peer reviews (Section 2).
- We evaluate the limitations of four alternative reform perspectives (Section 3).
- We propose an AI-human mutual supervision framework and explain how its non-collaborative structure addresses identified problems (Section 4).
- We conduct a practical example to assess the operability of the proposed framework and to discuss its advantages and limitations (Section 5).
- We formulate a concrete set of call-to-action objectives that specify actionable research and deployment steps to achieve the new framework (Section 6).

⁹<https://icml.cc/Conferences/2026/LLM-Policy>

¹⁰https://x.com/2prime_PKU/status/1948549824594485696

¹¹<https://openreview.net/forum?id=kDhAiaGzrn>

¹²<https://openreview.net/forum?id=wktBQXOtQS>

¹³<https://x.com/openreviewnet/status/1994176189875814611>

2. Problems of Peer Review

To provide effective recommendations for improving peer review quality, we summarize the problems that arise when humans or AI systems work independently in peer review.

2.1. Problems in Human Review

Problem (P1): Overload in Nominated (Expert-selected) Review. Traditional peer review relies on nominated domain experts, assuming that expertise and professional norms ensure trustworthy judgment (Lee et al., 2013). Yet, the exponential growth of submissions has far exceeded the capacity of this limited pool (Dokaliuk et al., 2025; Kim et al., 2025; Seghier, 2025), leaving a small number of qualified reviewers to evaluate an unsustainable number of papers, thereby reducing review quality and increasing variance even when reviewers act in good faith.

Problem (P2): Limited Expertise in Author-obligated Review. To mitigate reviewer shortages, many venues adopt author-obligated reviewing². This expands the reviewer supply but introduces uneven expertise. For example, senior researchers may delegate reviews to students, and inexperienced reviewers often struggle to provide rigorous assessments (Lin et al., 2023a).

Problem (P3): Irresponsibility in Author-obligated Review. Because reviewing is uncompensated and poorly aligned with academic incentives, some reviewers will produce superficial or minimal-effort comments, treating reviewing as an administrative requirement rather than a scholarly duty (Aczel et al., 2025).

Problem (P4): Misconduct. Conflicts of interest (COI) can distort evaluations, e.g., collusive bidding among authors (Littman, 2021; Jecmen et al., 2023), and incidents of identity leaks that enable direct author–reviewer contact (Mallapaty, 2024) illustrate the fragility of current safeguards. This creates opportunities for strategic behavior, where reviewers may down-rate competing submissions or inflate scores for collaborative work (Stelmakh et al., 2021).

Problem (P5): Human Bias. Prior work has shown that peer review increasingly tend to favor manuscripts reporting predominantly positive results (Lee et al., 2013), rather than papers that offer valuable insights despite limitations (a psychological form of negativity bias (Rozin & Royzman, 2001), which has been shown to affect information evaluation and communication processes broadly (Spinde, 2025)), whereby reviewers overweight identified weaknesses relative to overall contributions, particularly before rebuttal, when potential clarifications have not yet been considered.

Problem (P6): Subjective Unfairness. Subjective scoring, inconsistent evaluation standards, and emotional or contextual reactions can cause different papers to be assessed

under uneven criteria, leading to unfairness and reduced comparability across reviews (Haffar et al., 2019).

Existing mechanisms (COI declarations, guidelines, double-blind review) mitigate but cannot reliably prevent these strategic or biased behaviors at scale.

2.2. Problems in AI Review

Problem (P7): Reading Errors, Hallucinations, and Opaque Reliability. AI systems misinterpret formulas, figures, and references, especially from PDFs (Lin et al., 2023a; Zhuang et al., 2025), and may fabricate claims or citations that appear authoritative (Manakul et al., 2023; Zhuang et al., 2025).

Problem (P8): Long, Unfocused, and Misaligned Feedback. LLM-generated reviews often produce overly long, unfocused text that obscures key critique⁶⁵. Length frequently substitutes for substance (Thakkar et al., 2025), with AI missing core contributions while overemphasizing peripheral details (Drori & Te’eni, 2024; Du et al., 2024; Liang et al., 2024b; Weber, 2024; Farber, 2025). LLMs also tend to assume the manuscript is correct as written, overlooking clarity and exposition issues (Drori & Te’eni, 2024; Saad et al., 2024; Weber, 2024).

Problem (P9): Adversarial Manipulation, Security, and Privacy Risks. AI-based review is vulnerable to prompt injections, adversarial phrasing, and strategically crafted writing designed to steer model output (Drori & Te’eni, 2024; Ye et al., 2024; Collu et al., 2025; Gibney, 2025; Lin et al., 2025c). Many LLM services also store reviewer or manuscript content, creating confidentiality and privacy risks⁸ (Hosseini & Horbach, 2023; BaHammam, 2025; Zhuang et al., 2025).

Problem (P10): Systemic Bias. LLMs exhibit consistent relative biases in reviewing, often assigning more lenient scores than human reviewers (Du et al., 2024; Jin et al., 2024; Thelwall, 2024; Demetrio et al., 2025; Doskaliuk et al., 2025; Farber, 2025; Russo et al., 2025; Thelwall & Yaghi, 2025; Zhuang et al., 2025). While such biases may apply uniformly across submissions, they risk shifting score distributions, masking critical weaknesses, and anchoring human judgments when using AI outputs as references.

Problem (P11): Low Diversity and Governance Concerns. LLM-generated reviews often follow prompt-dominant patterns with limited stylistic and evaluative diversity. Exclusive reliance on AI reviewers further raises governance concerns, including accountability, transparency, and the delegation of scientific judgment to opaque systems (Liu & Shah, 2023; Shcherbiak et al., 2024; Zhou et al., 2024).

3. Alternative Views

Position papers with proposed solutions already exist to address the problems mentioned in Section 2. To place our position within the broader context of existing views, we review four existing perspectives on reforming peer review and discuss their limitations (see Table 1).

View 1 (V1): Against AI. This view argues that AI-generated reviews violate long-standing academic norms, motivating efforts to detect and penalize AI use in reviewing (Liang et al., 2024a; Yu et al., 2024b). Yet AI assistance in writing and reviewing is already widespread, and workload pressures make strict origin-based enforcement impractical. Reliable detection is inherently challenging: distinguishing AI-written from AI-polished text is difficult (Saha & Feizi, 2025), and current detectors have high false-positive and false-negative rates. Furthermore, this “human-only” view leads to Problems (P1) to (P6) as discussed in Section 2.

View 2 (V2): Embracing AI. This view considers AI, especially LLMs, as the future of peer review, often promoting “LLM-as-a-judge” (Li et al., 2025) and exploring assistance across the pipeline (Lin et al., 2023b; Tan et al., 2024; Tyser et al., 2024), with a strong focus on automated review and meta-review generation (Heaven, 2018; Checco et al., 2021; Li et al., 2023; Chamoun et al., 2024; D’Arcy et al., 2024a; Gao et al., 2024; Faizullah et al., 2024; Taechoyotin et al., 2024; Yu et al., 2024a; Chang et al., 2025; Lin et al., 2025a). As discussed in Section 2, this “AI-driven” view leads to Problems (P7) to (P11). Most such systems are evaluated by human-alignment—how closely AI outputs match human ratings or comments (Shcherbiak et al., 2024; Tan et al., 2024; Demetrio et al., 2025; Russo et al., 2025)—despite the fact that human reviews themselves can be biased, inconsistent, and strategically distorted (Lee et al., 2013; Haffar et al., 2019; Stelmakh et al., 2021).

View 3 (V3): Human-Centered Reform. This view attributes peer-review distortions primarily to human behavior, thereby emphasizing the regulation of reviewer conduct. Proposals include fully open review, where identities are disclosed to encourage accountability (Yang, 2025), incentive or author-feedback mechanisms to improve review quality (Kim et al., 2025), and author-self-rankings to identify noisy review scores¹⁴ (Su, 2021; 2025; Su et al., 2025a;b). However, these approaches overlook fundamental human constraints. Authors and reviewers often take turns in opposite roles, so revealing identities can lead to unfair return actions or retaliation after criticism (Fehr & Gächter, 2000; Eisenberger et al., 2004). If reviewers can behave unfairly, authors likewise cannot be assumed to function

¹⁴<https://blog.icml.cc/2026/01/12/introducing-icml-2026-policy-for-self-ranking-in-reviews/>

Table 1. Illustration of addressed problems.

Problems (Section 2)	Alternative Views (Section 3)				Our Position (Section 4)
	(V1) Against AI	(V2) Embracing AI	(V3) Human-centered Reform	(V4) Money-motivation	
(P1) Overloaded		Addressed	More serious ¹	Addressed	Addressed
(P2) Limited Expertise		Addressed		Addressed	Addressed
(P3) Irresponsible		Addressed	Addressed	Addressed	Addressed
(P4) Misconduct		Addressed	Addressed		Addressed
(P5) Human Bias		Addressed			Addressed
(P6) Subjective Unfair		Addressed		More serious ²	Addressed
(P7) Technical Errors	Addressed		Addressed	Addressed	Addressed
(P8) Lack Focus	Addressed		Addressed	Addressed	Addressed
(P9) Security & Privacy	Addressed		Addressed	Addressed	Partially ³
(P10) AI Bias	Addressed		Addressed	Addressed	Addressed
(P11) Low Diversity	Addressed		Addressed	Addressed	Addressed

¹ Stricter reviewer requirements (e.g., no anonymity or mandatory recognized status) may reduce the pool of qualified reviewers.

² Wealthier teams can submit more freely, gaining disproportionate opportunities.

³ This problem can largely be mitigated through human supervision (Phase 1) and human review (Phase 2), as attacks tend to produce inconsistent comments that are more detectable by humans, though robustness is not guaranteed.

as reliable supervisors (Problems (P5) and (P6)). Consequently, author feedback or self-ranking mechanisms may inherit similar biases and fail in ways analogous to human review systems. Moreover, these approaches do not address the chronic shortage of qualified reviewers (Problems (P1) and (P2)). Such systems cannot create additional expertise, and incentives such as best reviewer awards carry limited professional value compared to publication outcomes.

View 4 (V4): Money-Motivation. This view argues that reviewers should be compensated, noting that authors bear the majority of publication costs while publishers profit. (Seghier, 2025) advocates financial incentives and even professional reviewing as a career track. While well-intentioned, this shift prioritizes peer review as a market transaction, potentially further inflating already high publication and registration fees. More critically, if high-quality reviews can be purchased, so can biased or strategically motivated ones (Problems (P4) and (P5)). Monetary mechanisms do not remove COI; they simply monetize it.

(Seghier, 2025) also proposes submission fees to discourage low-effort submissions, and IJCAI–ECAI’26 has piloted such fees³. While simple and potentially effective, these policies disproportionately discourage financially constrained groups, even when submissions are strong, while wealthier teams can submit freely and enjoy more opportunities (Problem (P6)). The fee waiver for authors appearing on only one submission does not address this inequity: it implicitly penalizes collaboration and restricts groups in which multiple members wish to submit. Financial barriers tied to resource constraints and the authorship structure distort scientific opportunities and conflict with the principle that acceptance should depend on merit, not financial capacity. Monetary reforms, therefore, risk introducing new inequities when resolving problems of peer review.

4. AI-Human Mutual Supervision

Prior reform strategies fail in structurally different ways (see Table 1). Our proposal is to separate the roles of humans and AI, placing them under mutual supervision. In this framework, humans and AI alternate between generating review comments and assigning scores, ensuring that neither side can unilaterally determine the review outcome. Furthermore, since neither direction alone (AI supervising humans or humans supervising AI) can address the full spectrum of problems (see Sections 4.1 and 4.2), this results in a dual-phase, non-collaborative integration in which each side independently constrains the other. Our AI-human mutual supervision framework is illustrated in Figure 2.

4.1. Phase 1: AI Review, Human Supervision

Phase 1 consists of two individual steps:

1. **AI review:** A fixed model produces one structured review (summary, strengths, weaknesses, correctness, novelty, clarity), but no score to avoid biasing subsequent human judgments. The AI review prevents the problems (P1) to (P4) that occur in a human-only review.
2. **Human supervision:** M human supervisors (not necessarily experts) read the AI review, optionally skimming the paper, and independently:
 - assign a paper score,
 - provide a justification that references agreements, disagreements, or omissions in the AI review.

Since disagreements among human supervisors immediately reveal when the AI review is misleading or unbalanced, the human supervision primarily addresses the AI-specific problems (P7), (P10), and (P11).

Output: Area chairs (ACs) and program chairs (PCs) re-

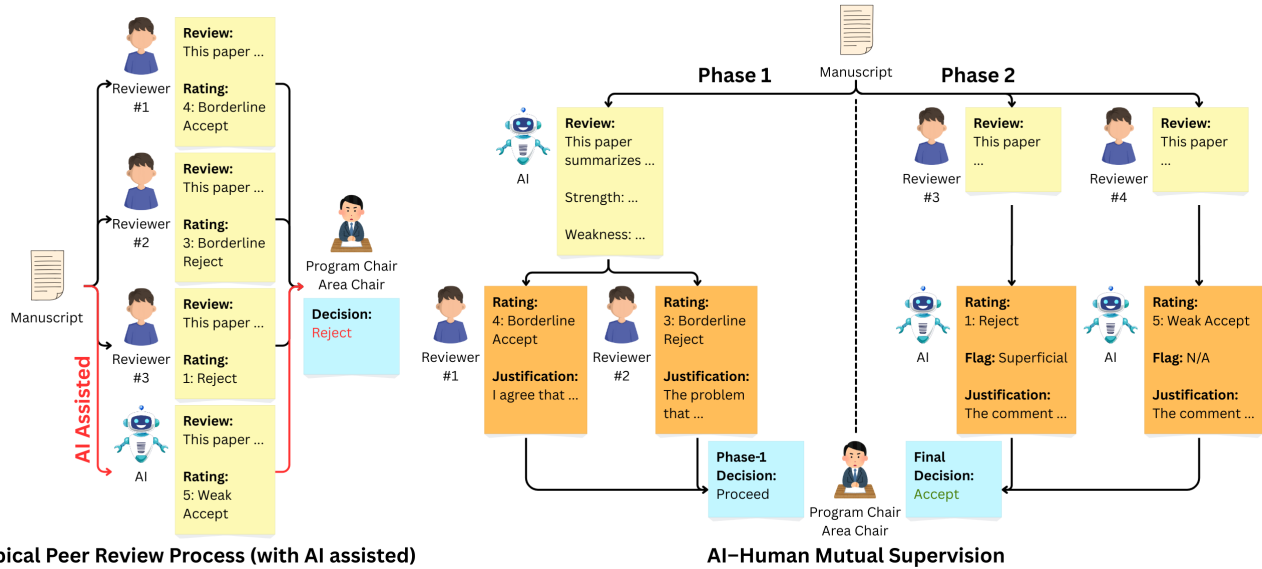


Figure 2. Comparison between a typical AI-assisted peer-review workflow and the proposed dual-phase AI-human mutual supervision framework. In current practice (left), AI serves only as an auxiliary reviewer whose comments may be considered or ignored; human reviewers directly determine paper scores. In our framework (right), AI and humans each generate structured score-free reviews, and the opposite side assigns paper scores with justifications. Area chairs (ACs) and program chairs (PCs) receive both sets of supervisory signals to identify unfairness, superficial reasoning, and misconduct.

ceive the AI review plus M human scores and justifications.

4.2. Phase 2: Human Reviews, AI Supervision

Phase 2 reverses the direction:

1. **Human review:** M reviewers independently write full structured reviews, without scores. The human review thus addresses problems (P7) to (P11).
2. **AI supervision:** For each review, the AI outputs:
 - an implied paper score with a justification.
 - problem flags (overly harsh, superficial, misguided, etc.) with a justification.

The AI translates free-text commentary into structured numerical signals, highlighting problematic reviews and making it harder for bias or superficiality to go unnoticed. Therefore, the AI supervision prevents problems (P5) and (P6). It also makes (P2)-(P4) easier to see.

Output: ACs/PCs receive each human review alongside its AI score, justification, and flags.

4.3. Dual-Phase Integration: Non-Collaborative Mutual Supervision

The two phases are intentionally non-collaborative: each phase never accesses the other’s outputs during review or supervision, and neither AI nor humans score their own written reviews. Therefore, ACs/PCs see orthogonal signals from each phase. Rather than collapsing these signals into an automated decision, the framework uses them for diagnostic pur-

poses. Consistency across phases implies high-confidence decisions. Reviews strongly flagged as superficial, misguided, or inconsistent can be down-weighted. Furthermore, the divergence between AI- and human-supervised scores indicates a structural disagreement that warrants AC’s attention. In summary, the mutual supervision does not automate acceptance or rejection. It makes it harder to hide unfair or unreliable reviews, thereby ensuring the quality of reviews.

4.4. Human-Led Rebuttal and Discussion

Authors respond to human reviewers’ comments as usual. For AI-generated review comments, authors only need to address those points that the human supervisor deems relevant. Reviewers may update their comments after discussion and rebuttal. However, dramatic reversals require an explicit explanation to discourage strategic behavior. For post-discussion audit, the AI supervisor can be rerun on final reviews to update implied scores and flags, providing ACs with an additional consistency check.

4.5. Alternative Views vs. Our Position

Table 1 presents where our position differs from the alternative views (V1)-(V4) regarding the problems mentioned in Section 2. We enforce a separation of powers in which humans and AI remain institutionally distinct and are required to justify their evaluations to one another.

Against (V1), we argue that using fallible tools to detect AI as evidence of misconduct is questionable, and that origin-

focused policies fail to address the substantive fairness of review content. Detection may remain a policy tool, but fairness evaluation must ultimately depend on content, not provenance. Against (V2), we argue that AI should not simply imitate human reviewers, but rather act as an independent supervisory actor whose judgments are cross-checked against those of human reviewers. Against (V3), we argue that human-centered reforms cannot fully resolve bias or misconduct. Fairness requires a structurally neutral actor, immune to retaliation and the influence of competitive interests. Against (V4), we argue that monetary incentives do not eliminate human-nature problems; they merely put a price on them.

5. Practical Example

We run a practical example to assess whether our mutual supervision framework is operational. Although this is a single toy case, it illustrates the advantages of the proposed approach and highlights limitations that warrant further investigation. We present summaries here, with full details provided in Appendix B.

5.1. Toy Case Study

Settings. We demonstrate the dual-phase AI-human mutual supervision framework on a single borderline ICLR’25 rejected submission (no COI), originally rated 3/6/6/8 (avg. 5.75)¹⁵. We chose this paper because of its divergent reviews: three broadly positive and one negative, with scores ranging from clear accept to clear reject, where changes in the review set could plausibly alter the final decision. The study involves eight human reviewers (four original reviewers in Phase 2 and four newly recruited reviewers in Phase 1), one area chair for meta-review, and one LLM (ChatGPT 5.1). In Phase 1, the AI produces a structured, score-free review that guides scoring by new reviewers. In Phase 2, the AI infers scores and flags issues from original reviews, and the area chair issues the meta-review. A brief questionnaire (ease, quality, and whether reviewers voluntarily referenced the original paper) assesses reviewer experience (Table 6 in Appendix B).

Results. The meta-review decision remains unchanged (*reject*). Aggregate ratings shift from 3/6/6/8 (avg. 5.75) to 5/5/5/6/6/8/6/8 (avg. 6.125), with the first four scores from newly recruited Phase 1 reviewers. Notably, the original rating-3 review is raised to 6, and AI supervision flags superficiality, inaccuracy, and misguidance¹⁶.

¹⁵<https://openreview.net/forum?id=fFtmpqLFvw>

¹⁶Declaration: This does not imply any error in the ICLR’25 review process or unethical reviewer behavior; a single illustrative example with simple prompting cannot rigorously compare AI and human biases.

Questionnaire responses indicate mixed experiences: reviewers generally reported the task as easier or unchanged, but perceived quality gains were inconsistent. Two reviewers consulted the original paper due to concerns about AI bias and limited depth. Overall, the toy example suggests the framework is operational and maintains decision consistency, while revealing reliability challenges such as rating mismatch, AI dependence, and potential bias.

Inspirations. This toy case shows that AI can surface overlooked evidence and potential misunderstandings in human reviews, yet AI-leaning scores may still diverge from human judgments. One plausible explanation is that reviewers optimize for different criteria: humans may lean toward flaw-finding (“does it have weaknesses?”) rather than proportional weighing (“are weaknesses critical or outweighed by strengths?”), consistent with negativity bias and related reviewer-psychology observations (see Appendix B.2 - Inspiration). Under mutual supervision, such misalignments either counterbalance one another or appear as explicit discrepancies, rather than being implicitly collapsed into a single judgment dominated by either human or AI bias.

5.2. Advantages of Mutual Supervision

The proposed mutual supervision framework introduces a structured, review-linked scoring system. Each score is explicitly linked to a review comment whose logic is supervised by the opposite side.

First, with this framework, the overall proportion of the reviewer workload is reduced by the two-phase process (addressing (P1)) (see Appendix C). Phase 1 reduces human burden by replacing many full reviews with lightweight supervision of an AI baseline. The full human reviews in Phase 2 can focus on stronger or borderline papers.

Moreover, the division of labor in our AI-human mutual supervision framework improves overall review quality by combining the consistency and coverage of AI (see Section 4.1; addressing (P1) to (P4)) with the nuanced judgment of domain experts (see Section 4.2; addressing (P7) to (P11)), while mutual supervision helps particular identify subjective unfairness (P6), hallucinations (P7), superficial assessments ((P3) and (P7)), unfocused and misaligned feedback (P8), and biases ((P5) and (P10)). The requirement that reviews must be translated into justified scores also increases transparency, making COI, strategic bias, and low-effort evaluations easier to detect ((P2) and (P4)).

Lastly, the framework maintains ethical and diversity safeguards by preserving human decision authority and enforcing a clear separation of roles that constrains both human and AI problems.

5.3. Limitations of Mutual Supervision

Adopting an AI–human mutual supervision framework introduces new research and governance challenges. These limitations outline the agenda for responsible large-scale deployment.

First, although mutual supervision reduces the likelihood of undetected manipulation, AI-assisted review remains vulnerable to technical misconduct, such as gaming the system or adversarial attacks (P9). Existing defenses are limited, and full mitigation may be impossible, much like the recent identity leak incident¹³. Human oversight currently provides partial protection, but any deployment must treat technical misconduct as a primary threat.

Deployment also raises practical trade-offs. Open-source models offer transparency, reproducibility, and lower cost but may require more effort to implement for specific tasks (Kukreja et al., 2024). Commercial models offer strong baseline performance but introduce concerns regarding privacy, cost, and controllability (Kukreja et al., 2024; Wu et al., 2024). Responsible adoption, hence, requires explicit governance over model choice, data handling, and update policies.

Another challenge is whether reviewers will accept being evaluated or flagged for bias by AI. Mutual supervision alters long-standing norms of reviewer autonomy, necessitating clear communication about how supervisory signals are used and safeguards against opaque or punitive practices.

Lastly, our framework aims to separate the roles of reviewers, AI systems, and decision makers, but best practices for integrating AI-generated supervisory signals remain open. Future work must determine how ACs/PCs should respond when human and AI assessments diverge, how to weigh problematic reviews, and how to maintain a process that is balanced, transparent, and accountable.

6. Call to Action

The proposed mutual supervision framework necessitates a staged research and deployment program spanning taxonomy design, dataset construction, methodological development, security, and governance. Table 2 and Figure 3 summarize the core objectives, their dependencies, and the relevant stakeholders.

At a high level, Objective I (this paper) establishes the conceptual foundations for constructing biased-review datasets (Objective II). These datasets support the development of review rating and failure-mode detection methods (Objective III; Phase 2), which, in turn, facilitate training less-biased AI reviewers for Phase 1 (Objective IV). Across all stages, technical vulnerabilities must be addressed (Objective V) before large-scale deployment of dual-phase AI-human mutual

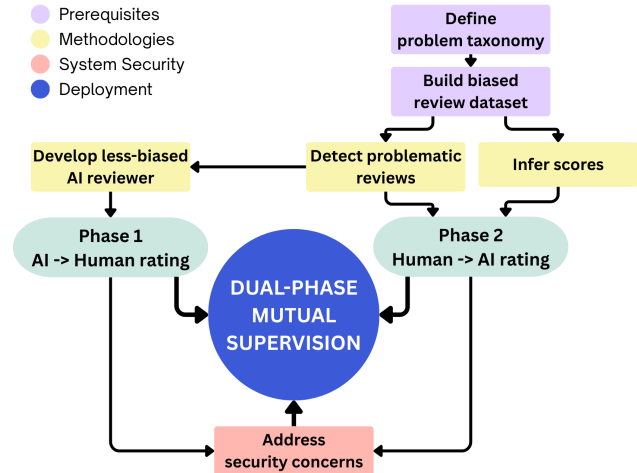


Figure 3. Step-wise objectives, following Table 2. This paper have formalized a provisional taxonomy of problems (Table 1) and summarized existing peer-review datasets (Appendix D).

supervision becomes feasible (Objective VI).

Objective I is partially instantiated in this work through a provisional problem taxonomy (Table 1) and a review of existing datasets (Appendix D). However, some problems (e.g., (P1) overload and (P11) low diversity) do not necessarily manifest as problematic individual reviews and can be mitigated through automated integration of human and AI reviews. We therefore define seven more actionable failure modes that directly indicate problematic reviews: overly harsh, overly gentle, unconfident, superficial, inaccurate or hallucinatory, subjective, and misguided (Table 3). These failure modes provide an initial analytical structure for evaluating review quality on both the human and AI sides.

Prior work has explored individual objectives related to peer-review reform, such as automated reviewing, reviewer assignment, bias analysis, or AI-assisted feedback (see Appendix E), but does not attempt to construct a unified framework that simultaneously addresses the full set of problems identified in Section 2, particularly those spanning data requirements, methodological design, and security considerations. Achieving such a framework requires satisfying multiple objectives concurrently, which defines the gaps we aim to bridge. We discuss these objectives and the limitations of existing work in greater detail in Appendix E.

Near-term pilots: Phase 2 first. While our long-term objective is full dual-phase AI-human mutual supervision, we recommend piloting Phase 2 (human reviews with AI supervision) as the most practical near-term deployment. This setup requires no changes to reviewer assignment or workload: reviewers submit standard score-free reviews, and an AI supervisor infers implied scores and flags potential problems for ACs and PCs. In this configuration, the AI functions solely as an advisory layer, highlighting superfi-

Table 2. Roadmap from taxonomy to deployment.

Stage	Obj.	Task	Key idea	Primary actors
Prerequisites	I	Problem taxonomy	Define a shared set of review problems for humans and AI (this paper)	Peer-review researchers, conference policy committees
	II	Biased review datasets	Build datasets with annotated or controllably biased reviews based on Obj. I	Dataset builders, OpenReview/venue maintainers
Methodologies	III	Rate reviews	From Obj. I–II, (a) detect problematic human reviews and (b) infer paper scores from review text for AI supervision for <i>Phase 2</i>	ML/NLP researchers, tool developers
	IV	Unbiased AI review	Train less-biased AI reviewers by combining ML/RLHF ¹ researchers, Obj. III–IV detectors with quality rewards; use these models as the <i>Phase 1</i> AI baseline	tool developers
System Security	V	Security & robustness	Analyze and defend against technical attacks on Obj. III–IV (prompt injection, adversarial text, poisoning, gaming of AI ratings)	Security/robustness researchers, platform engineers
Deployment	VI	Deployment	Pilot dual-phase AI–human mutual supervision with Obj. III–V in place: limited tracks, opt-in reviewers, and carefully monitored governance	Conference organizers, PCs/ACs, platform maintainers

¹ RLHF: Reinforcement Learning from Human Feedback.

Table 3. Provisional failure modes in peer review, with illustrative examples.

Failure Modes	Problems	Illustrative Example
Overly Harsh	(P4), (P5)	Rejects a paper for minor or peripheral issues without engaging with the core contributions. (e.g., “The paper should be rejected because it has typos,” despite valid empirical results.)
Overly Gentle	(P4), (P9), (P10)	Accepts or supports a paper without sufficient scrutiny. (e.g., “This is a promising direction,” despite methodological errors.)
Unconfident	(P2)	Vague or hedged judgments that avoid taking a clear stance. (e.g., “It seems like the contribution might be useful, but I am not sure.”)
Superficial	(P3), (P8)	Generic comments that do not engage with the manuscript. (e.g., “The writing needs improvement,” without identifying any concrete issues.)
Inaccurate / Hallucination	(P2), (P7)	Claims errors or missing content that are not true. (e.g., “The paper does not have ablation studies,” although it does.)
Subjective	(P6)	Personal preference presented as an objective assessment. (e.g., “I think method X is better than this proposed method.”)
Misguided	(P4)	Comments based on irrelevant or misinterpreted material. (e.g., suggesting unrelated citations or applying inappropriate evaluation criteria.)

cial, inconsistent, or anomalously harsh reviews, while all decisions remain fully human. A key challenge here is the AI’s tendency to assign higher scores, but this applies uniformly across submissions and can be adjusted via prompting. As Objective III (review rating and problem detection) matures and security issues in Objective V become better understood, conferences can incrementally adopt Phase 1 (AI review, human supervision) and progress toward full dual-phase mutual supervision.

7. Conclusion

This position paper argues that meaningful reform of peer review requires structural change rather than prompt engi-

neering or incremental tooling. As workload and AI usage both rise, neither human-only nor AI-only oversight is sufficient: the former enables overload, uneven expertise, and bias, while the latter introduces reading errors, security risks, and unresolved accountability.

AI–human mutual supervision offers a principled alternative in which human judgment and AI evaluation serve as independent, cross-checking forces under non-collaborative roles. Rather than automating accept-or-reject decisions, the framework provides auditable signals that make superficial, biased, or inaccurate reviews harder to hide and easier to detect. By defining concrete problems, identifying data and security requirements, and outlining prerequisites for a more unbiased review, we highlight both the opportunities and

the open challenges on the path to deployment.

Our aim is not to replace human reviewers, but to redesign the system so that fairness, accountability, and transparency are enforced by design. We call on the research community, conference organizers, and program committees to pilot and refine this framework as a step toward a more trustworthy and resilient peer-review ecosystem.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grants JP21H04907 and JP24H00732, by JST CREST Grant JPMJCR20D3 and JPMJCR2562 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, 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).

References

- Aczel, B., Barwich, A.-S., Diekman, A. B., Fishbach, A., Goldstone, R. L., Gomez, P., Gundersen, O. E., von Hippel, P. T., Holcombe, A. O., Lewandowsky, S., et al. The present and future of peer review: Ideas, interventions, and evidence. *Proceedings of the National Academy of Sciences*, 122(5):e2401232121, 2025.
- BaHamam, A. S. Peer review in the artificial intelligence era: A call for developing responsible integration guidelines. *Nature and Science of Sleep*, 17:159–164, 2025.
- Bharti, P. K., Navlakha, M., Agarwal, M., and Ekbal, A. Politepeer: does peer review hurt? a dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation*, 58(4):1291–1313, 2024.
- Chamoun, E., Schlichtkrull, M., and Vlachos, A. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9742–9763, 2024.
- Chang, Y., Li, Z., Zhang, H., Kong, Y., Wu, Y., So, H. K.-H., Guo, Z., Zhu, L., and Wong, N. 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)*, pp. 15662–15693, 2025.
- Checco, A., Bracciale, L., Loreti, P., Pinfield, S., and Bianchi, G. Ai-assisted peer review. *Humanities and Social Sciences Communications*, 8(1):1–11, 2021.
- Cheng, L., Bing, L., Yu, Q., Lu, W., and Si, L. 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)*, pp. 7000–7011, 2020.
- Choudhary, G., Modani, N., and Maurya, N. React: A review comment dataset for actionability (and more). In *International Conference on Web Information Systems Engineering (WISE)*, pp. 336–343. Springer, 2021.
- Collu, M. G., Salviati, U., Confalonieri, R., Conti, M., and Apruzzese, G. Publish to perish: Prompt injection attacks on llm-assisted peer review. *arXiv preprint arXiv:2508.20863*, 2025.
- D’Arcy, M., Hope, T., Birnbaum, L., and Downey, D. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024a.
- D’Arcy, M., Ross, A., Bransom, E., Kuehl, B., Bragg, J., Hope, T., and Downey, D. 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)*, pp. 6985–7001, 2024b.
- Demetrio, L., Apruzzese, G., Grosse, K., Laskov, P., Lupu, E., Rimmer, V., and Widmer, P. Gen-review: A large-scale dataset of ai-generated (and human-written) peer reviews. *arXiv preprint arXiv:2510.21192*, 2025.
- Doskaliuk, B., Zimba, O., Yessirkepov, M., Klishch, I., and Yatsyshyn, R. Artificial intelligence in peer review: enhancing efficiency while preserving integrity. *Journal of Korean Medical Science*, 40(7), 2025.
- Drori, I. and Te’eni, D. Human-in-the-loop ai reviewing: feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109, 2024.
- Du, J., Wang, Y., Zhao, W., Deng, Z., Liu, S., Lou, R., Zou, H. P., Venkit, P. N., Zhang, N., Srinath, M., et al. Llms assist nlp researchers: Critique paper (meta-) reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5081–5099, 2024.
- Dycke, N., Kuznetsov, I., and Gurevych, I. 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)*, pp. 5049–5073, 2023.
- Eisenberger, R., Lynch, P., Aselage, J., and Rohdieck, S. Who takes the most revenge? individual differences in negative reciprocity norm endorsement. *Personality and Social Psychology Bulletin*, 30(6):787–799, 2004.

- Faizullah, A. R. B. M., Urlana, A., and Mishra, R. 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)*, pp. 106–124. Springer, 2024.
- Farber, S. Comparing human and ai expertise in the academic peer review process: towards a hybrid approach. *Higher Education Research & Development*, 44(4):871–885, 2025.
- Fehr, E. and Gächter, S. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–182, 2000.
- Fromm, M., Faerman, E., Berrendorf, M., Bhargava, S., Qi, R., Zhang, Y., Dennert, L., Selle, S., Mao, Y., and Seidl, T. Argument mining driven analysis of peer-reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 4758–4766, 2021.
- Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I., and Miyao, Y. Does my rebuttal matter? insights from a major nlp conference. In *Proceedings of NAACL-HLT*, pp. 1274–1290, 2019.
- Gao, Z., Brantley, K., and Joachims, T. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*, 2024.
- Ghosal, T., Varanasi, K. K., and Kordoni, V. Hedgepeer: A dataset for uncertainty detection in peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–5, 2022.
- Gibney, E. Scientists hide messages in papers to game ai peer review. *Nature*, 2025. Available at: <https://www.nature.com/articles/d41586-025-02172-y> (accessed 02 December 2025).
- Guo, Y., Shang, G., Rennard, V., Vazirgiannis, M., and Clavel, C. Automatic analysis of substantiation in scientific peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10198–10216. Association for Computational Linguistics, 2023.
- Haffar, S., Bazerbachi, F., and Murad, M. H. Peer review bias: a critical review. *Mayo Clinic Proceedings*, 94(4): 670–676, 2019.
- Heaven, D. The age of ai peer reviews. *Nature*, 2018. Available at: <https://www.nature.com/articles/d41586-018-07245-9.pdf> (accessed 02 December 2025).
- Hosseini, M. and Horbach, S. P. 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.
- Hua, X., Nikolov, M., Badugu, N., and Wang, L. 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)*, pp. 2131–2137, 2019.
- Jecmen, S., Yoon, M., Conitzer, V., Shah, N. B., and Fang, F. A dataset on malicious paper bidding in peer review. In *Proceedings of the ACM Web Conference (WWW)*, pp. 3816–3826, 2023.
- Jin, Y., Zhao, Q., Wang, Y., Chen, H., Zhu, K., Xiao, Y., and Wang, J. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1208–1226, 2024.
- Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., and Schwartz, R. 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)*, pp. 1647–1661, 2018.
- Kennard, N. N., O’Gorman, T., Das, R., Sharma, A., Bagchi, C., Clinton, M., Yelugam, P. K., Zamani, H., and McCallum, A. 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*, pp. 1234–1249, 2022.
- Kim, J., Lee, Y., and Lee, S. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning (ICML) Position Paper Track*, 2025.
- Kousha, K. and Thelwall, M. Artificial intelligence to support publishing and peer review: A summary and review. *Learned Publishing*, 37(1):4–12, 2024.
- Kukreja, S., Kumar, T., Purohit, A., Dasgupta, A., and Guha, D. A literature survey on open source large language models. In *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, pp. 133–143, 2024.
- Kumar, S., Ghosal, T., and Ekbal, A. When reviewers lock horns: Finding disagreements in scientific peer reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 16693–16704, 2023.

- Kuznetsov, I., Buchmann, J., Eichler, M., and Gurevych, I. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986, 2022.
- Kuznetsov, I., Afzal, O. M., Dercksen, K., Dycke, N., Goldberg, A., Hope, T., Hovy, D., Kummerfeld, J. K., Lauscher, A., Leyton-Brown, K., et al. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.
- Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013.
- Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2757–2791, 2025.
- Li, M., Hovy, E., and Lau, J. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7089–7112, 2023.
- Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., 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)*, pp. 29575–29620. PMLR, 2024a.
- Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):A10a2400196, 2024b.
- Lin, E., Peng, Z., and Fang, Y. Evaluating and enhancing large language models for novelty assessment in scholarly publications. In *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, pp. 46–57, 2025a.
- Lin, J., Song, J., Zhou, Z., Chen, Y., and Shi, X. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98:101830, 2023a.
- Lin, J., Song, J., Zhou, Z., Chen, Y., and Shi, X. Mopr: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35(34):24191–24206, 2023b.
- Lin, J., Shan, R., Zhu, J., Xi, Y., Yu, Y., and Zhang, W. Stop ddos attacking the research community with ai-generated survey papers. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025b.
- Lin, T.-L., Chen, W.-C., Hsiao, T.-F., Liu, H.-I., Yeh, Y.-H., Chan, Y. K., Lien, W.-S., Kuo, P.-Y., Yu, P. S., and Shuai, H.-H. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks. *arXiv preprint arXiv:2506.11113*, 2025c.
- Littman, M. L. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6):43–44, 2021.
- Liu, R. and Shah, N. B. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.
- Liu, Y., Jia, Y., Geng, R., Jia, J., and Gong, N. Z. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1831–1847, 2024.
- Mallapaty, S. Elite researchers in china say they had ‘no choice’ but to commit misconduct. *Nature*, 2024. Available at: <https://www.nature.com/articles/d41586-024-01697-y> (accessed 02 December 2025).
- Manakul, P., Liusie, A., and Gales, M. 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)*, pp. 9004–9017, 2023.
- Mökander, J., Schuett, J., Kirk, H. R., and Floridi, L. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2024.
- Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3419–3448, 2022.
- Piet, J., Alrashed, M., Sitawarin, C., Chen, S., Wei, Z., Sun, E., Alomair, B., and Wagner, D. Jatmo: Prompt injection defense by task-specific finetuning. In *European Symposium on Research in Computer Security (ESORICS)*, pp. 105–124. Springer, 2024.
- Purkayastha, S., Lauscher, A., and Gurevych, I. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 14479–14495, 2023.
- Raina, V., Liusie, A., and Gales, M. 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)*, pp. 7499–7517, 2024.

- Rozin, P. and Royzman, E. B. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4):296–320, 2001.
- Russo, G., Horta Ribeiro, M., Davidson, T. R., Veselovsky, V., and West, R. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction (ACM HCI)*, 9(7):1–28, 2025.
- Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., Vaishya, R., and Botchu, R. Exploring the potential of chatgpt in the peer review process: an observational study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 18(2):102946, 2024.
- Saha, S. and Feizi, S. Almost ai, almost human: The challenge of detecting ai-polished writing. *arXiv preprint arXiv:2502.15666*, 2025.
- Seghier, M. L. Paying reviewers and regulating the number of papers may help fix the peer-review process. *F1000Research*, 13:439, 2025.
- Shcherbiak, A., Habibnia, H., Böhm, R., and Fiedler, S. Evaluating science: A comparison of human and ai reviewers. *Judgment and Decision Making*, 19:e21, 2024.
- Shi, J., Yuan, Z., Liu, Y., Huang, Y., Zhou, P., Sun, L., and Gong, N. Z. 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)*, pp. 660–674, 2024.
- Singh, S., Singh, M., and Goyal, P. Compare: a taxonomy and dataset of comparison discussions in peer reviews. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 238–241. IEEE, 2021.
- Spinde, T. *Automated Detection of Media Bias: From the Conceptualization of Media Bias to its Computational Classification*. Springer Vieweg Wiesbaden, 2025. ISBN 978-3-658-47797-4 978-3-658-47798-1. doi: 10.1007/978-3-658-47798-1. URL <https://link.springer.com/10.1007/978-3-658-47798-1>.
- Stelmakh, I., Shah, N. B., and Singh, A. Catch me if i can: Detecting strategic behaviour in peer assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 4794–4802, 2021.
- Su, B., Collina, N., Wen, G., Li, D., Cho, K., Fan, J., Zhao, B., and Su, W. How to find fantastic ai papers: Self-rankings as a powerful predictor of scientific impact beyond peer review. *arXiv preprint arXiv:2510.02143*, 2025a.
- Su, B., Zhang, J., Collina, N., Yan, Y., Li, D., Cho, K., Fan, J., Roth, A., and Su, W. The icml 2023 ranking experiment: Examining author self-assessment in ml/ai peer review. *Journal of the American Statistical Association*, (just-accepted):1–16, 2025b.
- Su, W. You are the best reviewer of your own papers: The isotonic mechanism. *Operations Research*, 2025.
- Su, W. J. You are the best reviewer of your own papers: an owner-assisted scoring mechanism. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 27929–27939, 2021.
- Taechoyotin, P., Wang, G., Zeng, T., Sides, B., and Acuna, D. Mamorx: Multi-agent multi-modal scientific review generation with external knowledge. In *NeurIPS 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024.
- Tan, C., Lyu, D., Li, S., Gao, Z., Wei, J., Ma, S., Liu, Z., and Li, S. Z. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*, 2024.
- Thakkar, N., Yuksekgonul, M., Silberg, J., Garg, A., Peng, N., Sha, F., Yu, R., Vondrick, C., and Zou, J. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*, 2025.
- Thelwall, M. Can chatgpt evaluate research quality? *Journal of Data and Information Science*, 9(2):1–21, 2024.
- Thelwall, M. and Yaghi, A. Evaluating the predictive capacity of chatgpt for academic peer review outcomes across multiple platforms. *Scientometrics*, pp. 1–23, 2025.
- Tyser, K., Segev, B., Longhitano, G., Zhang, X.-Y., Meeks, Z., Lee, J., Garg, U., Belsten, N., Shporer, A., Udell, M., et al. Ai-driven review systems: evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*, 2024.
- Weber, R. The other reviewer: Roboreviewer. *Journal of the Association for Information Systems*, 25(1):85–97, 2024.
- Wu, X., Duan, R., and Ni, J. Unveiling security, privacy, and ethical concerns of chatgpt. *Journal of information and intelligence*, 2(2):102–115, 2024.
- Yang, J. Position: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. In *Forty-second International Conference on Machine Learning (ICML) Position Paper Track*, 2025.

- Ye, R., Pang, X., Chai, J., Chen, J., Yin, Z., Xiang, Z., Dong, X., Shao, J., and Chen, S. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708*, 2024.
- Yu, J., Ding, Z., Tan, J., Luo, K., Weng, Z., Gong, C., Zeng, L., Cui, R., Han, C., Sun, Q., et al. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10164–10184, 2024a.
- Yu, S., Luo, M., Madasu, A., Lal, V., and Howard, P. Is your paper being reviewed by an llm? investigating ai text detectability in peer review. In *NeurIPS Safe Generative AI Workshop 2024*, 2024b.
- Yuan, W., Liu, P., and Neubig, G. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.
- Zeng, S., Tian, K., Zhang, K., Wang, Y., Gao, J., Liu, R., Yang, S., Li, J., Long, X., Ma, J., et al. Reviewrl: Towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 16942–16954, 2025.
- Zhang, Y., Zhang, H., Ji, W., Hua, T., Haber, N., Cao, H., and Liang, W. 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.
- Zhou, R., Chen, L., and Yu, K. 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)*, pp. 9340–9351, 2024.
- Zhuang, Z., Chen, J., Xu, H., Jiang, Y., and Lin, J. Large language models for automated scholarly paper review: A survey. *Information Fusion*, pp. 103332, 2025.

A. Our Position in the New Elements of ICML 2026 Peer Review

Peer-review challenges have been widely recognized across the research and academic communities. Consequently, many conferences revised their peer-review policies in 2025 and 2026. While AAAI’26 piloted AI-assisted reviews for the first time⁵ (without influencing decisions), IJCAI–ECAI’26 introduced submission fees³, and ICML’26 surprised us in January 2026 with far-reaching changes¹⁷ that we felt compelled to comment on.

Thinly Sliced Contributions. ICML’26 defines *thinly sliced contributions* as the submission of multiple papers with minor variations on the same theme, motivated by the rise of AI-generated submissions (Lin et al., 2025b). To address this, ICML requires authors to disclose and discuss related submissions from the same authors on similar topics. We agree that mass-produced, minimally differentiated AI-generated submissions are not worth reviewing and should be treated as serious misconduct. However, the policy leaves open the question of how legitimate parallel research should be distinguished from misconduct. In addition, the requirement to engage with unpublished work raises methodological and epistemic issues. For example, unpublished works may contain substantial errors or incomplete evaluations, and requiring such comparisons risks propagating incorrect claims. The new policy is also inconsistent with existing ICML guidelines. For instance, according to the ICML 2026 Reviewer Instructions¹⁸, authors are not required to discuss concurrent work publicly within two months of the deadline, and the Call for Papers¹⁹ permits submissions previously presented at workshops without proceedings.

To improve the accountability of authors and reviewers (i.e., addressing Problems (P1)–(P6) discussed in Section 2), several guidelines now aim to address the different needs:

AI-assisted Pre-submission Feedback. The AI-assisted pre-submission feedback initiative is closely related to the AI-assisted review piloted at AAAI’26⁵, but applied at the pre-submission stage. We agree that such feedback can help improve submission quality, especially given that generic LLM tools are not tailored to conference-specific criteria. However, as discussed in Section 3, AI reviews operating independently exhibit limitations, motivating our dual-phase mutual supervision framework in Section 4. With appropriate mitigation of bias and security risks, we believe this AI-assisted pre-submission feedback can be naturally integrated into Phase 1 of our framework.

Free ICML Registration for Top Reviewer. Another change is the incentive for top reviewers: the top 25% receive free registration for ICML, which falls under the alternative view (V4), money-motivation. As discussed in Section 3, this approach may encourage more responsible behavior but cannot create additional expert reviewers or mitigate conflicts of interest.

Author-provided Self-ranking. ICML’26 introduces an optional *author-provided self-ranking mechanism* to help identify potentially low-quality reviews via discrepancies between authors’ within-group rankings and pre-rebuttal reviewer scores¹⁴. Concretely, authors with multiple submissions may be asked to rank their own papers; the corresponding reviewer scores are then projected to satisfy this ranking via an isotonic mechanism (Su, 2021; 2025; Su et al., 2025a;b), and the resulting deviation from the original scores is used as a discrepancy signal for additional scrutiny. We still categorize this strategy as a human-centered reform under our alternative views ((V3) in Section 3). While this design provides a lightweight diagnostic, this procedure has asymmetrical coverage in favor of productive research groups, is susceptible to strategic behavior, and is highly dependent on the operational capacity of the area chairs. Furthermore, it is based on pre-rebuttal scores, whereas many disagreements are resolved during rebuttal, and some issues only become apparent after rebuttal during meta-review synthesis. This timing mismatch can yield both false positives (cases later resolved) and false negatives (failures that emerge only in later stages).

Overall, the regulation increases the burden on authors and reviewers and creates uncertainty in evaluation without clearly addressing the problems it aims to solve. The late announcement also makes consistent implementation for 2026 difficult.

B. Toy Example of Dual-Phase AI–Human Mutual Supervision Review

In this section, we present a toy example demonstrating how the proposed dual-phase AI–human mutual supervision framework operates on a selected ICLR’25 rejected submission¹⁵. In addition to the four original human reviewers from ICLR’25, we recruited four new human reviewers with diverse backgrounds, as well as one experienced area chair, to participate in this demonstration. Under our framework, although the final decision remains unchanged, the aggregate

¹⁷<https://blog.icml.cc/2026/01/08/whats-new-in-icml-2026-peer-review/>

¹⁸<https://icml.cc/Conferences/2026/ReviewerInstructions>

¹⁹<https://icml.cc/Conferences/2026/CallForPapers>

ratings shift from the original 3/6/6/8 (average 5.75) to 5/5/5/6/6/8/6/8 (average 6.125), where the first four scores are from the newly recruited reviewers. These results indicate that the proposed dual-phase AI–human mutual supervision framework is operational—maintaining decision consistency while easing the review process and improving perceived review quality—while also revealing open challenges to reliability, including potential biases and mismatches with human review priorities.

Declaration

1. Although our framework overturns some individual reviewer ratings and flags review issues, we emphasize that this does *not* imply any error in the ICLR’25 decision process or any unethical behavior by its reviewers. A single illustrative example and simple prompt engineering cannot provide a rigorous analysis of AI bias relative to human bias. This example is intended solely to illustrate how the proposed framework operates, what kinds of interactions it enables, and how it can generate different insights even when using the same underlying review materials. We reaffirm that the objectives outlined in the “Call to Action” section are essential for achieving a reliable and principled implementation.
2. The results (Section B.2) and discussion (Section B.3) include our hypotheses about participants’ psychology. We stress that these hypotheses are interpretive and derived solely from the empirical patterns in the collected data. They should not be read as assertions about participants’ actual psychological states, which lie beyond the scope of this study.

B.1. Settings

The selected paper from ICLR 2025. We selected an ICLR’25 rejected submission¹⁵ (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 received notably diverse reviewer opinions. Its average score is 5.75—borderline, but slightly leaning toward acceptance. It received three broadly positive reviews and one negative review, including one clear accept and one clear reject. Although the final decision was a rejection, this type of case is particularly informative: even minor changes in the review set could reasonably shift the outcome toward either acceptance or a more decisive rejection.

Such borderline cases provide an ideal testbed for evaluating whether our dual-phase framework offers new insights or alters the decision dynamics compared to the existing review process.

Participants. In total, our study involves eight human reviewers (including the four original ICLR reviewers), one human area chair, and one LLM (ChatGPT 5.1).

We used the raw review comments from the four original anonymous ICLR reviewers as the human review input for *Phase 2 only*, which were later evaluated under AI supervision. We did not provide the original human ratings or confidences to the LLM; these scores were used solely for post-hoc analysis of AI–human alignment.

We also recruited four new human reviewers to participate in *Phase 1 only*: two AI safety researchers, one machine learning researcher, and one Ph.D. student in LLM security. Their task was to assign a score to the paper based on the AI-generated long-form review and to provide a justification for their rating.

ChatGPT 5.1 (standard reasoning mode) served as the AI reviewer. To avoid unintended adaptation, we disabled memory functionality and ensured that no interaction history influenced the model’s behavior. The model participated in both phases:

- *Phase 1*: The AI generated a full review (without a numerical rating) based solely on the anonymous manuscript and a

structured prompt. We adapted the review-guideline template from AAAI 2026²⁰ for this prompt (see Section F.1), which, to our knowledge, is the only formally piloted AI-assisted review system.

- *Phase 2*: The AI assigned a rating to the submission based *only* on the human review comments and then flagged potential failure modes in those comments, referring back to the manuscript when necessary and providing justification. Because no prior work provides a template for this task, we designed custom prompts (see Section F.2). The failure modes are defined as in Table 3.

Finally, we recruited one of the ICLR 2025 area chairs—an experienced area chair who also serves in ICML and NeurIPS—to provide the *meta-review assessment*.

Process. Table 4 summarizes the inputs and outputs of each step in the toy example. Detailed prompts and procedures are provided in Section F. *Note that* we do not include a rebuttal phase, as we cannot involve the original authors or the anonymous ICLR reviewers in the discussion.

Table 4. Full process of our AI–human mutual supervision review.

Phase	Stage	Participants	Inputs	Outputs
Phase 1	AI review	AI	Manuscript, prompt	Review comment (no rating)
	Human supervision	Reviewer #1–4	Review comment, manuscript ¹	Rating, justification
Phase 2	Human review	Reviewer #5–8	Manuscript	Review comment (no rating)
	AI Rating	AI	Review comment, prompt	Rating, justification
	Failure Detection	AI	Review comment, prompt, manuscript	Flag, justification
Meta Review	Meta review	Area chair	Full review summary, manuscript ¹	Meta review comment

¹ Reviewers are not required to read the manuscript, although they may refer to it at their discretion.

Beyond our framework. In addition to simulating the full review process under our proposed framework and protocol, we administered a short questionnaire to the four newly recruited reviewers and the area chair. The questionnaire included:

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 reviewers completed their tasks, we added a post-hoc question:

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

These questions were designed to assess whether the proposed framework simplifies the review workflow and improves the reviewer experience.

We also evaluated AI–AI mutual supervision—that is, allowing the AI to rate, detect failure modes, and justify assessments on its own AI-generated review. This examines whether the AI exhibits self-preference or systematic bias independent of human input.

B.2. Results

Overall. As summarized in Table 5, the meta-review decision remains unchanged (*reject*) under our framework. However, among the eight human reviewers, none issued comments warranting a clear reject (rating 3 or below), while the clear-accept ratings (8) remained. Notably, the original ICLR review with a rating of 3 was overturned under AI supervision and raised to 6 (marginally above the threshold). The AI flagged multiple issues in the original human review, including superficiality (insufficient supporting evidence), inaccuracy (overlooking existing content), and misguidance (misinterpreting statements or conditions). Overall, AI-assisted ratings tend to be higher than the original human ratings. Beyond known relative biases (Russo et al., 2025), we hypothesize that this shift may stem from (1) mismatches between human and AI review priorities, and/or (2) humans being more sensitive to weaknesses, while the AI enumerates a larger set of issues in Phase 1 than a human reviewer typically articulates.

²⁰<https://docs.google.com/document/u/0/d/1tqQGwtNUIALPSTqoTo5uTFx8vKuqpILNTne9jeBCOVI/mobilebasic>

Table 5. Detailed peer-review outcomes from the original ICLR’25 process and from our AI–human mutual supervision framework.

Participant		ICLR’25		AI-human Mutual Supervision	
Review	Supervision	Rating (after rebuttal)	Confidence	Rating (before 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 Just an analysis in this paper, not part of our framework.

Phase 1: AI Review, Human Supervision (AI → Human Rating). Human reviewers produced more consistent borderline ratings (5/5/5/6) when evaluating with access to the AI-generated review. In contrast, traditional full-paper reviewing showed substantially higher variance: in ICLR’25, four reviewers rated this paper 3/6/6/8. We hypothesize that this pattern arises because the AI review is highly detailed—enumerating many strengths and weaknesses—yet lacks diversity, as all reviewers rely on a single reference review. Human reviewers, being more sensitive to weaknesses than strengths (Section B.3), may become hesitant to assign high scores when confronted with a long list of issues. At the same time, the AI review also identifies substantial strengths, making it difficult to justify a low score. Borderline ratings, therefore, emerge as the “safest” choice.

This hypothesis reveals a design challenge for AI-generated reviews: if the review is *too* detailed, it may overwhelm human evaluators and inadvertently anchor them to conservative (borderline) ratings. Conversely, if the AI were to report only high-level or critical points, it may fail to capture the breadth of evidence needed for effective mutual supervision. This tension motivates our Objective IV: to develop unbiased and balanced AI auto-review methodologies that appropriately weight critical versus minor points.

Phase 2: Human Reviews, AI Supervision (Human → AI Rating). In Phase 2, AI supervision substantially increases two ratings: the only negative review and one particularly short review. Both are flagged with multiple problem warnings. While we cannot yet determine whether ChatGPT 5.1, and our toy prompts capture all relevant dimensions comprehensively, the observed pattern suggests the following: human reviewers tend to lower scores upon identifying weaknesses, while underweighting strengths may be flagged by the AI as a “superficial” review. The AI, in contrast, may exhibit a gentler evaluation tendency. Regardless of which interpretation is correct, our mutual supervision framework yields a clear principle: reviewers cannot assign low scores without providing convincing reasoning. To justify rejection, a reviewer must identify *critical* weaknesses that are factually correct, well supported, and proportionally severe.

Meta Review. The area chair recommended rejection, consistent with the original ICLR decision. The meta-review acknowledged the strengths of the novel multi-turn conversation setting and the released dataset, but concluded that the weaknesses—limited methodological novelty, lack of theoretical analysis, and restricted evaluation—outweighed these contributions. This outcome suggests that our framework does not introduce extreme deviations from existing peer-review decisions. Notably, the criteria emphasized by the area chair closely align with those prioritized by human reviewers, and differ substantially from the AI assessment, indicating a potential mismatch between human and AI review priorities.

Feedback on the Questionnaire. As summarized in Table 6, all four human reviewers reported that rating the AI-generated review was easier than performing a full paper review, even reviewers who voluntarily read the original paper to better understand the context and to evaluate the reliability of the AI review—still found the task easier than conducting a traditional, full review.

However, not all reviewers perceived a clear improvement in the quality of their assessments with AI assistance. Although the AI’s detailed analysis provided a useful structure—especially for those with limited expertise in the paper’s domain—Reviewer #2 noted, after reading the original papers, that the AI review itself may be biased and lacks broader or

deeper validation. This concern aligns with observations in prior studies (Section 2.2).

We initially anticipated that the additional supervisory signals might increase the burden on the area chair. However, the area chair reported that it is difficult to conclude whether the framework improves or worsens the meta-review process. While the AI-generated reviews are often lengthy and lack focus—particularly in terms of technical depth—they can help alleviate issues common in human reviews, such as personal bias, overemphasis on a single argument when deciding to accept or reject, or overlooking parts of the paper. Importantly, the framework does not appear to significantly increase the meta-review workload. Nevertheless, AI-generated reviews remain insufficiently trusted by human decision-makers, highlighting a key challenge that warrants further study.

Table 6. 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

AI Supervises AI (not part of our framework). As shown in the last row of Table 5, the AI assigns a high rating to its own review and detects no failure modes, suggesting a possible self-preference or confirmation bias. Although this experiment lies outside the scope of our proposed framework, it highlights the need for more systematic investigation of AI–AI mutual supervision, particularly with respect to debiasing AI-generated reviews and AI-generated ratings.

Inspiration. The results suggest that AI reviews can identify some issues in human reviews—though not comprehensively or reliably yet—while their ratings may still diverge from human judgments. We hypothesize that this mismatch arises from differing priorities: human reviewers often evaluate submissions based on academic standards, research goals, and prevailing field trends, whereas AI tends to enumerate and weigh strengths and weaknesses using more general critical criteria. This highlights a key challenge in designing more reliable review systems: clarifying what is desired in a given round of submissions, e.g., embracing diverse subjective judgments versus enforcing more regularized evaluation criteria. Such expectations are typically shaped by venue organizers through review guidelines and policies, and may be partially operationalized through the AI component of our mutual supervision framework, whereas human-only review inevitably remains subjective.

At the same time, incorporating additional opinions, even from AI, requires careful balance. When AI reviews enumerate an extensive list of strengths and weaknesses, including less critical points, they may unintentionally influence human scoring behavior and hinder fair evaluation. A psychology perspective further suggests separating *commenting* from *scoring*. Here, the AI mainly contributes to commenting by providing (a) redundant cues that humans would likely raise anyway, and/or (b) complementary cues that humans might otherwise miss. Since humans remain the final arbiters of scores, they may (i) adopt, (ii) ignore, or (iii) integrate AI inputs with their own judgment, depending on individual differences such as personality and perceived relative competence. For example, less motivated reviewers may be more likely to (i) adopt, whereas more conscientious reviewers may be more likely to (ii) ignore (e.g., viewing AI use as “cheating” or less competent). When reviewers perceive the AI as comparable in competence to themselves, they may instead integrate (iii), often yielding intermediate scores. This heterogeneity may diversify opinions but also increases subjectivity, making reliance on a single human supervisor fragile and motivating our mutual checks (including AI supervision of human reviews) to improve consistency.

Finally, the observed outcome does not validate our AI baseline, nor does it demonstrate that the framework reliably improves the peer-review process. Rather, it illustrates how mutual supervision can provide an alternative perspective by surfacing overlooked evidence or potential misunderstandings. As this is only a toy example, these findings should be interpreted as suggestive rather than conclusive.

B.3. Discussion — Implicit Psychological Criteria in Paper Acceptance

Our toy example also raises a broader question: how do reviewers implicitly define acceptance criteria? The outcomes from Phase 1 and Phase 2 suggest that human reviewers may gravitate toward the question “Does this paper have weaknesses?” rather than the more decision-relevant questions “Does this paper have *critical* weaknesses?” or “Do the strengths outweigh

the weaknesses?” This tendency is consistent with findings in the psychology of reviewers: humans are generally more sensitive to negative than to positive information (a psychological form of negativity bias) (Rozin & Royzman, 2001), and peer review often incentivizes fault-finding over recognizing contributions. As a result, reviewers may unintentionally shift from evaluating whether a paper meets the bar for acceptance to cataloguing imperfections. Indeed, prior work has argued that peer review increasingly favors “flawless” papers with only positive results (Lee et al., 2013), rather than papers offering valuable insights despite limitations.

This distinction between *searching for flaws* and *evaluating contributions* has meaningful implications. A flaw-seeking mindset may disproportionately disadvantage borderline or unconventional papers, since even minor weaknesses can become grounds for rejection. A contribution-oriented mindset, in contrast, evaluates the work holistically—considering novelty, potential impact, and the proportional severity of weaknesses.

AI–human mutual supervision offers an opportunity to recalibrate this dynamic. Because AI and human reviewers operate independently, discrepancies between their assessments make implicit reviewer tendencies more visible. When AI highlights structural strengths or consistent evidence while humans emphasize concerns, the misalignment signals that neither perspective should unilaterally dominate. Over time, such cross-supervision may encourage more balanced, criterion-driven judgments and reduce the tendency toward flaw-centric evaluation.

B.4. Limitations of the Toy Example

Our toy example has several important limitations. First, it involves only a single paper and a small number of reviewers, limiting generalizability. Second, it is illustrative rather than statistical, and therefore cannot support quantitative claims about reliability or effect size. Third, it relies on a single LLM configuration, and the observations may not transfer to other models, settings, or prompting strategies.

C. Human Workload Analysis and Comparison Across Peer-Review Designs

A central design principle of our framework is that human-reviewer workload must decrease, or at least not increase. In the proposed system, the overall reviewing burden can be controlled by adjusting the strictness of the Phase 1 decision. Only submissions that satisfy the AI-defined baseline proceed to full expert evaluation, ensuring that human effort is concentrated where it is most impactful.

To illustrate comparative workload, let one full human review correspond to a unit workload W_{full} , and let rating an AI-generated review require $W_{\text{rate}} = 0.5 W_{\text{full}}$. Consider $N = 100$ submissions, a Phase 1 pass rate of $p = 0.5$, and a total allocation of $k = 4$ reviewers per paper, with two reviewers in Phase 1 and two in Phase 2:

- *Case (a)*: Four human full reviews, single-phase evaluation.

$$W_a = N \cdot 4 \cdot W_{\text{full}} = 100 \cdot 4 W_{\text{full}} = 400 W_{\text{full}}.$$

- *Case (b)*: Three human full reviews + one AI full review, single phase (AI replaces one human reviewer).

$$W_b = N \cdot 3 \cdot W_{\text{full}} = 300 W_{\text{full}}.$$

- *Case (c)*: Two human full reviews in Phase 1, followed by two human full reviews in Phase 2 for the pN surviving papers.

$$W_c = N \cdot 2 \cdot W_{\text{full}} + pN \cdot 2 \cdot W_{\text{full}} = (100 \cdot 2 + 50 \cdot 2) W_{\text{full}} = 300 W_{\text{full}}.$$

- *Case (d)*: **Our dual-phase framework.** Phase 1: two humans rate the AI-generated review (each requiring $0.5 W_{\text{full}}$). Phase 2: two humans provide full reviews for the pN surviving papers.

$$W_d = N \cdot 2 \cdot W_{\text{rate}} + pN \cdot 2 \cdot W_{\text{full}} = 100 \cdot 2 \cdot 0.5 W_{\text{full}} + 50 \cdot 2 \cdot W_{\text{full}} = 100 W_{\text{full}} + 100 W_{\text{full}} = 200 W_{\text{full}}.$$

Under these illustrative parameters, our dual-phase framework (Case d) requires *only half* the human workload of the standard four-reviewer baseline (Case a), and substantially less than both the reduced-reviewer model (Case b) and the two-phase human-only model (Case c). This demonstrates that an AI–human mutual supervision structure can simultaneously enhance oversight and reduce overall human reviewing burden.

D. Peer-Review Datasets

We compiled what we believe is the most complete list of peer-reviewed datasets currently available, as shown in Table 7. Our review draws on three major surveys (Lin et al., 2023a; Kuznetsov et al., 2024; Zhuang et al., 2025) as well as individual dataset papers. Because many existing datasets contain overlapping submissions or reviews, we focus on those most relevant to our study—datasets that include papers, human reviews, AI-generated reviews, or bias-related annotations. Datasets that do not meet these criteria are excluded from our summary. Among the available resources, the Gen-Review dataset (Demetrio et al., 2025) most closely aligns with the requirements of our study; however, it lacks sufficient coverage of problematic reviews and has not been validated against reliable benchmarks.

Table 7. Overview of peer-review datasets. All listed datasets are publicly accessible.

Dataset	Paper			Review				Rebuttal Author	Sources
	Abstract	TXT	PDF	Human	Meta-	AI	Biased		
AMPERE (Hua et al., 2019)				14,202					ICLR 2017–2018 UAI 2018 NeurIPS 2013–2017 ACL 2017
ReAct (Choudhary et al., 2021)				1,250					ICLR 2018
AMSR (Fromm et al., 2021)				12,135					ICLR 2019–2020 MIDL 019–2020 NeuroAI 2019 GI 2020
HedgePeer (Ghosal et al., 2022)				2,966					ICLR 2018
LimGen (Faizullah et al., 2024)		4,068							ACL 2022 EACL 2022 EMNLP 2022–2023
PeerRead (Kang et al., 2018)		14,784		10,770					NIPS 2013–2017 ICLR 2017 ACL 2017 CoNLL 2016 arXiv 2007–2017
F1000RD (Kuznetsov et al., 2022)		3,000		8,053					F1000Research
ASAP-Review (Yuan et al., 2022)		8,877		28,119	Include				ICLR 2017–2020 NeurIPS 2016–2019
ContraSciView (Kumar et al., 2023)		8,582		25,825					ICLR 2017–2020 NeurIPS 2016–2019
ARIES (D’Arcy et al., 2024b)		1,720		4,088					OpenReview
ReviewMT (Tan et al., 2024)		26,841		92,017					ICLR 2017–2024 Nature 2023
NLPeer (Dycke et al., 2023)			5,672	11,515					CONLL 2016, 2020 ARR 2022 ACL 2017 F1000 2022
Reviewer2 (Gao et al., 2024)			27,805	99,727	Include				CONLL 2016, 2020 ACL 2017 ARR 2022 ICLR 2017–2023 NeurIPS 2016–2022
TreeReview (Chang et al., 2025)			80	336					NeurIPS 2023 ICLR 2024

Table 7 (continued)

Dataset	Paper			Review				Rebuttal Author	Sources
	Abstract	TXT	PDF	Human	Meta-	AI	Biased		
AIReviewLottery (Russo et al., 2025)			23,959	86,690					ICLR 2018–2024
ACL-2018 (Gao et al., 2019)				7,929				1,227	ACL 2018
RR (Cheng et al., 2020)				4,764				4,764	ICLR 2013–2020
DISAPERE (Kennard et al., 2022)				506				506	ICLR 2019–2020
JitsuPeer (Purkayastha et al., 2023)				1,705				1,705	ICLR 2018–2020
PeerSum (Li et al., 2023)	14,993			53,451	Include			5,854	ICLR 2018–2020 NeurIPS 2021–2022
MOPRD (Lin et al., 2023b)			6,578	22,483	Include			11,213	Multiple journals Multiple domains
ReviewCritique (Du et al., 2024)		100		380		60			ICLR 2020–2023 NeurIPS 2020–2023
AgentReview (Jin et al., 2024)			523			53,800			ICLR 2020–2023
SubstanReview (Guo et al., 2023)							550		CoNLL 2016 ACL 2017 COLING 2022 ARR 2022
PolitePEER (Bharti et al., 2024)							2,500		ICLR NeurIPS Publons ShitMyReviewersSay
Gen-Review (Demetrio et al., 2025)			32,652	124,615		81,850	54,567		ICLR 2018–2025

E. Objectives

To construct the proposed mutual supervision framework, multiple objectives must be addressed across datasets, methodologies, and security considerations. In this section, we highlight key prerequisites and objectives (Figure 3 and Table 2) that current approaches fail to meet, focusing on conceptual directions rather than implementation details.

E.1. Prerequisites

Objective I: Define problems in peer review. A necessary foundation is a clear taxonomy of problems that affect fairness across both human and AI evaluations. Prior studies have discussed bias in AI reviews, yet often implicitly assume human judgments as the normative reference. For example, AI reviews have been labeled “biased” for being more gentle than human reviewers (Hosseini & Horbach, 2023; Demetrio et al., 2025; Russo et al., 2025), even though human reviews may themselves be overly harsh. Such assumptions require reconsideration.

We therefore formalize a provisional set of problems (see Table 1). However, some issues (e.g., P1: overload and P11: low diversity) do not necessarily manifest as problematic individual reviews and can be mitigated through automated integration of human and AI reviews. We thus define seven more actionable failure modes that directly indicate problematic reviews: overly harsh, overly gentle, unconfident, superficial, inaccurate or hallucinatory, subjective, and misguided (see Table 3). These failure modes provide an initial analytical structure for evaluating review quality on both the human and AI sides.

Objective II: Establish biased peer-review datasets. Existing datasets (Section D) primarily repurpose raw records from venues such as ICLR and NeurIPS on OpenReview²¹. These datasets exhibit heavy overlap and offer limited coverage of high-quality AI reviews or deliberately biased comments.

Manual bias annotation is equally problematic: program committees seldom invalidate a review solely because it is overly harsh or overly gentle, making human-labeled bias inherently unreliable. LLMs, combined with well-defined failure modes, can generate synthetic biased reviews at scale (Demetrio et al., 2025), yet a central open question remains: *how can we reliably validate that synthetic biased reviews are genuinely biased and not simply different in style or strictness?*

E.2. Methodologies

Objective III: Rate reviews. This objective involves two related tasks. First, assigning paper-level scores based solely on human reviewers’ comments. While this appears straightforward for LLMs, (Zhou et al., 2024) shows that current models perform poorly when relying solely on review text, especially when comments are long. More effective approaches incorporate paper content or abstracts alongside reviews (Kang et al., 2018; Singh et al., 2021; Dycke et al., 2023; Zhang et al., 2025); however, such methods operate as *AI reviewers* rather than *AI supervisors* and therefore do not support mutual supervision.

Second, identifying potential problematic human review comments. To our knowledge, no existing work provides a comprehensive framework that detects all major issues. Prior studies address only specific categories, such as overly harsh reviews (Stelmakh et al., 2021), unconfident language (Ghosal et al., 2022), superficial assessments (Guo et al., 2023), or impolite tone (Bharti et al., 2024). Comprehensive problem detection in peer review, therefore, remains an open challenge.

Objective IV: Unbiased AI review. Although Phase 1 includes human oversight of AI-generated review comments, an unbiased AI reviewer would further improve the reliability of human ratings and reduce confusion among reviewers. At present, however, there is no established method for producing truly unbiased AI reviews. Recent efforts, such as (Zeng et al., 2025), which apply reinforcement learning to improve review quality, suggest that quality rewards can guide models toward more consistent assessments. We believe similar approaches could be adapted to develop less biased AI reviewers, particularly when combined with the failure-detection framework introduced in Objective III.

E.3. System Security

Objective V: Address technical misconduct. AI-assisted peer review introduces new vulnerabilities to technical manipulation. Existing attacks can be grouped into four categories:

Prompt injection: inserting explicit or hidden statements in the paper to influence the model’s behavior (Shi et al., 2024; Ye et al., 2024; Collu et al., 2025; Gibney, 2025).

Textual adversarial attacks: making small, semantically minor edits that inflate AI-generated ratings (Raina et al., 2024; Lin et al., 2025c).

Data poisoning: feeding LLMs systematically biased comments for targeted papers, such that future model updates inherit these distortions (Drori & Te’eni, 2024).

Implicit manipulation: deliberately shaping the paper’s text, structure, template, or metadata to exploit known AI-review biases or scoring criteria (Ye et al., 2024).

Currently, no defense work provides reliable countermeasures to these vulnerabilities *in the peer-review context*. We therefore identify two threat models that must be addressed:

- *Misleading review rating:* human reviewers attempt to manipulate AI scoring in ways that contradict the paper’s true quality.
- *Misleading AI reviewing:* authors attempt to manipulate AI-generated reviews without making substantive improvements to the paper.

In the absence of peer-review-specific defenses, our framework must initially rely on general-purpose LLM security techniques adapted to this setting. Possible defenses include:

²¹<https://openreview.net>

- sanitizing inputs before they reach the model (e.g., stripping suspicious instructions or markup likely to act as prompt injections) (Piet et al., 2024);
- restricting the channels through which model instructions can be delivered (e.g., ignoring text in certain environments or metadata fields) (Liu et al., 2024);
- employing adversarial training and red-teaming to harden prompts and models against known attack templates (Perez et al., 2022);
- using ensembles or cross-model agreement checks to detect anomalous outputs (Manakul et al., 2023);
- maintaining detailed audit logs so that suspicious patterns in AI ratings or flags can be investigated by security teams and program chairs (Mökander et al., 2024).

None of these techniques is sufficient on its own, but together they can raise the cost of successful manipulation and make attacks more detectable.

Our framework aims to release AI review prompts transparently, enabling authors to improve their papers ethically—reducing arbitrary, low-effort submissions—while still preventing attempts to manipulate AI behavior. We strive for a system where transparency becomes a strength, not a vulnerability, and where technical misconduct is treated as a first-class threat model in the design and deployment of AI-assisted peer review.

E.4. Deployment

Objective VI: Pilot AI–human mutual supervision. A crucial next step is to pilot the AI–human mutual supervision framework in real peer-review settings. We envision initial deployment in selected conferences or workshops willing to experiment with structural improvements to governance. Such pilots would allow the community to evaluate feasibility, reviewer experience, robustness against misconduct, and the practical impact on review quality and fairness.

We therefore call for participating venues, organizers, and program committees to collaborate in designing and testing pilot implementations. Early adopters will play a central role in shaping the next generation of peer-review infrastructure and in demonstrating how mutual supervision can be integrated responsibly and transparently into existing academic processes.

F. Full Review Process of AI-Human Mutual Supervision

F.1. Phase 1: AI Review, Human Supervision (AI → Human Rating)

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 ensemble 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.

Position: Peer Review Demands AI–Human Mutual Supervision

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 — just an analysis in this paper, not part of our framework

Rating: 8: accept, good paper

Failure-mode flags: N/A

Justification

- The reviewer'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.

F.2. Phase 2: Human Reviews, AI Supervision (Human → AI Rating)

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-teams, 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-teams, 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-teams 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-teams, 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-teams, 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 PAF), 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.

F.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-learned 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.
- 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.

- CodeAttack: Revealing Safety Generalization Challenges of Large Language Models via Code Completion.*
- 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 Habereeder, Isabella and Lu, Chun-Shien and  
    Echizen, Isao},  
  title={Position: Peer Review Demands AI-Human Mutual Supervision},  
  booktitle={International Conference on Machine Learning (ICML) [in review]},  
  pages={41},  
  publisher={PMLR},  
  address={Seoul, South Korea},  
  year={2026},  
  month={02},  
  url={https://www.researchgate.net/publication/400590054}  
}
```

Online Access

Official Publication <https://www.researchgate.net/publication/400590054>