Contents lists available at IJAHCI
International Journal of Advanced Human Computer Interaction
Journal Homepage: http://www.ijahci.com/
Volume 1, No. 1, 2023

Human–AI Co-Writing for Scientific Peer Review: An HCI Framework for Transparent, Calibrated Assistance

Madjid Siamandi

Department of Computer Engineering, Islamic Azad University, Mashhad, Iran

ARTICLE INFO

Received: 2023/08/29
Revised: 2023/10/02
Accepted: 2023/12/25

Keywords:
Human–Computer Interaction; AI Co-Writing; Peer Review; Explainable AI; Uncertainty Communication; Evidence Anchoring; Fairness; Transparency

ABSTRACT

Large language models (LLMs) increasingly assist reviewers in drafting, organizing, and justifying peer-review reports. While such tools promise faster, clearer feedback, they risk over-confident claims, ungrounded citations, and homogenized critique that can erode reviewer accountability and author trust. We study AI-assisted peer review as a human–computer interaction (HCI) problem centered on *co-writing* under constraints of fairness, transparency, and time. Through a multi-method investigation—(i) contextual inquiry with 24 active reviewers across computer science, industrial engineering, and HCI venues; (ii) two controlled experiments (N=96 reviews) comparing prompt patterns, evidence-linking, and uncertainty displays; and (iii) a six-week field deployment of a review-composition interface integrated with a manuscript viewer—we derive a design framework that aligns AI generation with reviewer judgment and venue policy. Interface patterns that (1) require *evidence anchoring* (inline links to exact manuscript spans), (2) enforce *claim typing* with uncertainty ranges, and (3) provide *counter-arguments on demand* improved rubric coverage (+22%), rationale specificity (+31%), and self-reported confidence calibration (+28%) while preserving reviewer voice. However, naive auto-summaries elevated superficiality and increased reliance on model phrasing. We contribute actionable guidelines, auditing checklists, and failure taxonomies for safe, transparent AI co-writing in scientific peer review.

1. Introduction

Peer review is a high-stakes writing task constrained by time, norms, and venue rubrics. Reviewers must read closely, assess novelty and rigor, and articulate actionable feedback. LLMs promise relief—summarizing sections, proposing probe questions, and restructuring comments. Yet, without careful interaction design, AI assistance can entrench cognitive shortcuts, propagate errors, and obscure accountability.

We position AI review assistance as *human–AI co-writing* that should: (i) preserve reviewer agency; (ii) surface uncertainty; (iii) anchor claims to manuscript evidence; and (iv) support deliberation, not only generation. We present a design framework and an interface that implements evidence-anchored drafting, claim typing, uncertainty badges, and counter-argument retrieval. We evaluate impacts on coverage, specificity, calibration, and perceived fairness, and provide governance artifacts for program chairs and venues.

2. Related Work

AI-assisted writing. Prior work shows mixed effects: AI can improve structure and clarity but risks factual drift and stylistic homogenization. HCI studies emphasize scaffolding, provenance, and iterative critique over one-shot generation.

Explainability and uncertainty. Communication of limits and confidence helps calibrate user trust; however, persuasive explanations can increase over-reliance. Structured uncertainty labels and access to underlying evidence mitigate this risk.

Scholarly review processes. Research on peer review highlights variability in rubric coverage, low inter-rater agreement, and the importance of specific, actionable feedback. Tools that align with venue rubrics and make rationales auditable improve transparency.

3. System Design

We implement a review co-writing interface layered over a manuscript viewer.

3.1. Core Principles

- **Evidence anchoring:** Every AI-assisted sentence must cite a manuscript span (figure, table, paragraph, or line range). Sentences lacking anchors are flagged.
- **Claim typing:** Review text is labeled as *Observation*, *Interpretation*, *Suggestion*, or *Question*. Each type has required fields (e.g., observations require anchor + quote).
- **Uncertainty badges:** Claims display discrete confidence bands (Low/Med/High) with tooltips for factors driving uncertainty (sample size, confounds, ambiguity).
- **Counter-arguments on demand:** A "steelman" toggle generates reasoned counter-views with anchors, encouraging balanced appraisal.
- **Rubric alignment:** A side panel mirrors venue criteria (originality, significance, methodology, clarity, ethics). Coverage meters indicate under-addressed areas.

3.2. Workflow

(1) Reviewer highlights manuscript spans; (2) selects a claim type; (3) optional AI suggestion appears *anchored* to the selected spans; (4) reviewer edits, sets uncertainty, and adds actionable suggestions; (5) a governance bar shows anchor completeness, citation hygiene, and policy checks (e.g., no identity inferences).

4. Methods

Contextual inquiry. We observed review sessions (think-aloud) with 24 reviewers—faculty, practitioners, and PhD students—across HCI, AI, and industrial engineering venues. We mapped friction points: locating evidence, maintaining rubric coverage, and writing actionable suggestions.

Controlled experiments. Two between-subject studies (N=96 reviews; manuscripts counterbalanced) compared interface variants: (A) with vs. without *evidence anchoring*; (B) with vs. without *claim typing + uncertainty badges*. Outcomes: rubric coverage (0–100), rationale specificity (1–5), edit distance from AI draft, perceived fairness (Likert), and calibration error (difference between self-rated confidence and blinded meta-reviewer ratings).

Field deployment. A six-week pilot with 34 active reviewers. We logged anchor density (anchors per 100 words), incomplete claims, rubric coverage, edit trajectories, and opt-out rates. Post-hoc interviews probed perceived accountability and effort.

5. Results

Evidence anchoring. Interfaces enforcing anchors increased rubric coverage by 22% (95% CI [14, 30]) and rationale specificity by 31% compared to free-form AI suggestions. Incomplete claims declined by 41%.

Claim typing + uncertainty. Structured typing reduced over-confident language; calibration error dropped by 28% with no increase in total time. Reviewers reported clearer separation between observations and interpretations, improving author trust.

Counter-arguments. Steelman toggles reduced polarity in recommendation rationales and improved fairness ratings, but excessive use increased time by a median 4.7 minutes; reviewers learned to deploy it selectively for borderline decisions.

Reviewer voice. Edit distance from AI drafts remained high (median 0.63), indicating preserved authorship. Homogenization concerns were lower when anchors and claim typing were present.

Failure modes. Naive auto-summaries encouraged superficial comments detached from evidence; reviewers over-relied on default phrasing. Anchoring and typing mitigated this by forcing contact with the manuscript.

6. Discussion

Our findings support an HCI stance: the utility of AI co-writing depends on scaffolds that bind generated language to evidence and make uncertainty legible. Anchors act as *friction* that improves rigor; claim typing externalizes the reviewer's internal structure; uncertainty badges shift from

persuasive tone to calibrated stance. Together, these patterns increased coverage and specificity without suppressing voice.

However, assistance can still nudge toward brevity and genericity. To counter this, interfaces should prioritize *extraction-first* (quote then paraphrase), require *actionable* suggestions (owner, locus, next step), and keep provenance visible. Venues should adopt governance artifacts (anchor completeness thresholds, audit trails, policy prompts).

7. Design Guidelines

1. **Make evidence the default.** Require a manuscript anchor for AI-assisted sentences; highlight unanchored text.
2. **Type every claim.** Use a small ontology (Observation/Interpretation/Suggestion/Question) with required fields.
3. **Expose uncertainty.** Pair discrete badges with tooltips listing drivers (data ambiguity, confounds, reviewer familiarity).
4. **Balance with counter-arguments.** Provide on-demand "steelman" generation; log use to prevent over-reliance.
5. **Align to rubrics.** Show live coverage meters mapped to venue criteria; prompt when a section is under-addressed.
6. **Preserve voice and accountability.** Track edit distance and authorship; surface provenance for program chairs.
7. **Constrain auto-summaries.** Prefer extract-then-paraphrase; cap unanchored summary length.

8. Limitations

Our studies use CS/HCI/IE manuscripts and reviewers; generalization to biomedicine or humanities may differ. We evaluated short-horizon outcomes (coverage, specificity, calibration), not long-term venue-level impacts (acceptance bias, diversity). The field pilot lacked double-blind program-committee feedback loops; future work should examine committee-level dynamics, disagreement resolution, and meta-review quality. Finally, our uncertainty badges rely on reviewer-provided factors; partial automation may misestimate context.

9. Ethical Considerations

AI co-writing can inadvertently amplify bias (e.g., penalizing non-native writing style) or leak confidential content if prompts include manuscript text. We recommend least-privilege data flows, no storage of manuscript content beyond the review session, red-team prompts for identity inferences, and audit trails for AI-assisted passages. Reviewers must retain full authorship responsibility; venues should disclose AI-assistance policies, require evidence anchoring, and support appeals when reviewers' unanchored claims influence decisions.

10. Conclusion

AI can help reviewers write clearer, more complete, and more transparent reports—when interfaces bind generation to evidence, structure claims, and normalize uncertainty. Our framework and evaluation show that evidence anchoring, claim typing, and counter-argument support improve rubric coverage and calibration while preserving reviewer voice. We offer concrete guidelines and governance checks for venues seeking safe, auditable adoption of AI co-writing in peer review.

Acknowledgments

We thank participating reviewers and program chairs for their time and feedback. Any opinions are our own and do not represent venue policy.

References

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*, 2016, pp. 1135–1144.

[2] Z. C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10), 2018.

[3] S. Tonekaboni, S. Joshi, M. McCradden, A. Goldenberg. What clinicians want: Contextualizing explainable ML for decision support. In *MLHC*, PMLR 106:359–380, 2019.

[4] A. B. Arrieta et al. Explainable AI: Concepts, taxonomies, opportunities and challenges. *Information Fusion*, 58:82–115, 2020.

[5] F. Poursabzi-Sangdeh, D. Goldstein, J. Hofman, et al. Manipulating and measuring model interpretability. In *CHI '21*, Article 273, 2021.

[6] M. Mitchell et al. Model cards for model reporting. In *FAT**, 2019.

[7] T. Gebru et al. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.