

An Exploratory Study of Inherited Bias in AI-Assisted Evaluation: Thirteen Corrections, Zero Content Changes

Taekyung Lee

Independent Researcher, Republic of Korea

dfdfgo92@gmail.com

Abstract

AI is increasingly expected to serve as a blind evaluator, yet it can inherit human cognitive biases from its training data. This study reports an exploratory single-case study in which 82 creative works (38 philosophical essays, 8 musical scripts, 4 critical reviews, and 32 other pieces) were submitted to an AI (Claude) as anonymous text for evaluation within a single session.

During the evaluation, 13 candidate bias episodes were recorded and grouped into five provisional categories (reputation/authority, diffusion/market, format/medium, tool/authorship, action/realization). In each episode, the AI's deduction rationale was grounded in information external to content—such as unrealized performance, unverified commercial impact, or absence of reputation—rather than in the quality of the text itself. The primary contribution of this study is the identification and provisional classification of these candidate bias episodes, prior to considering subsequent score changes. When the human corrector logically challenged each deduction, scores shifted within the session (1,790 to 2,255 out of 2,400, +25.9% as a descriptive within-session shift); however, whether this shift reflects genuine bias correction, sycophantic agreement, or rubric renegotiation remains an open question for future controlled experiments. This study proposes that AI blind evaluation becomes effective not as a one-time blinding technique but when combined with an iterative corrective stance.

As a structural limitation, the author served simultaneously as the evaluated party, the corrector, and the category designer, creating a conflict of interest. However, the core observation is that scores shifted following iterative bias-challenge interventions while content remained unchanged, representing a structural possibility worth testing under third-party conditions. The present design cannot fully distinguish bias reduction from sycophantic agreement or rubric renegotiation. Replication with independent evaluators and multiple AI models is needed.

Keywords: AI bias, blind evaluation, halo effect, bias inheritance, human-AI collaborative assessment, cognitive bias correction, dialogic correction

1. Introduction

In 1920, psychologist Edward Thorndike discovered that soldiers rated as physically attractive also received higher scores for intelligence, leadership, and character [1]. This cognitive bias, subsequently termed the 'halo effect,' has been repeatedly confirmed over the past century. Nisbett & Wilson [2] demonstrated that evaluators remain unaware of their own biases, and Kahneman [3] systematized how System 1's automatic judgments distort evaluation.

Blind evaluation emerged as a countermeasure across multiple domains. Blind auditions in American orchestras beginning in the 1970s contributed to reducing a 50-year gender imbalance [4]. Double-blind clinical trials became the gold standard of modern medicine. Tomkins et al. [5] reported that in academic peer review, the odds ratio for papers from prestigious universities was 1.58 and from prestigious companies was 2.10, confirming institutional reputation bias in scholarly evaluation.

AI has been expected to serve as a 'structurally blind evaluator' in this context. When receiving anonymous text, AI has no direct access to the author's credentials, appearance, or background. However, a critical assumption is

missing: AI inherits human biases from its training data [15]. Dastin [6] reported that Amazon's AI recruiting tool systematically discriminated against female applicants. Zhuo et al. [7] empirically demonstrated that LLMs exhibit reputation bias in peer review (AgentReview). Liang et al. [17] confirmed that LLM-generated content is already influencing real-world peer reviews at major AI conferences, and Pataranutaporn et al. [16] reported that LLMs systematically favor papers from elite institutions.

This study poses the following research question: 'What types of candidate external-information deductions emerge in AI blind evaluation, and how do logically framed human challenges relate to evaluation shifts?'

To address this question, the author submitted 82 creative works to an AI as anonymous text and conducted an exploratory single-session study of iterative challenges to candidate external-information deductions. A structural limitation exists in that the author served as both the evaluated party and the corrector; this is discussed in Section 5.5.

The contributions of this paper are: (1) a provisional five-category classification of 13 candidate bias episodes in AI evaluation, (2) an exploratory single-session iterative correction record encompassing 82 works, (3) a conceptual reframing of blind evaluation from 'technique' to 'attitude,' and (4) practical implications for human-AI collaborative evaluation. A critical caveat applies throughout: the observed score shifts are compatible with bias correction, but sycophancy, rubric renegotiation, and task reframing remain viable alternative explanations (see Section 6.4).

2. Related Work

2.1 The Halo Effect and Cognitive Biases

The halo effect discovered by Thorndike [1] in 1920 was subsequently extended by Nisbett & Wilson [2] and Landy & Sigall [8]. Kahneman [3] systematized it as a product of System 1 (fast, automatic judgment). The MARBLE framework [20] recently classified eight cognitive biases exhibited by LLMs in educational contexts (confirmation bias, framing effect, anchoring effect, halo effect, among others), confirming that the halo effect is reproduced in AI systems. The core finding is consistent: evaluators are unconsciously influenced by information external to content and remain unaware of their biases.

2.2 History of Blind Evaluation

Blind evaluation has proven to be one of the most established countermeasures. Goldin & Rouse [4]'s orchestra blind audition study showed that a single curtain contributed to reducing a 50-year gender imbalance. Double-blind clinical trials became the gold standard for controlling placebo effects. Tomkins et al. [5] analyzed WSDM 2017 peer review data and confirmed that papers from prestigious institutions received significantly higher scores under single-blind conditions. eLife [9] introduced author and institution blinding beginning in 2020.

2.3 Research on AI Evaluation Bias

Research on AI evaluation bias has been accumulating. Gallegos et al. [15] provided a comprehensive survey of LLM bias and fairness, systematically documenting that LLMs can learn, perpetuate, and amplify social biases from training data. Dastin [6] reported that Amazon's AI recruiting tool faithfully reproduced gender bias. Zhuo et al. [7]'s AgentReview study empirically demonstrated that LLMs exhibit reputation bias in peer review, assigning higher scores to papers from well-known institutions.

These findings have been reinforced by subsequent research. Pataranutaporn et al. [16] reported that LLMs systematically favor papers from elite institutions and prominent male economists. 'Justice in Judgment' [18] confirmed affiliation and gender bias in LLM peer review through controlled experiments, observing an over-compensation phenomenon in reasoning models. Liang et al. [17] reported that LLM-generated content is already influencing real-world reviews at major AI conferences. LLM-REVal [19] discovered a 'linguistic feature bias' in which LLM reviewers systematically assign higher scores to LLM-written papers.

Similar concerns have been raised in healthcare. Obermeyer et al. [10] reported racial bias in healthcare algorithms. A healthcare LLM bias evaluation framework [22] proposed a step-by-step bias evaluation protocol for clinical settings. Mehrabi et al. [11] systematically classified bias types in machine learning, and Barocas & Selbst [12] analyzed how algorithmic decision-making reproduces structural inequality.

2.4 Bias Correction and AI Alignment

The mainstream approach to AI bias correction is rule-based. Anthropic's Constitutional AI [13] constrains AI behavior by specifying principles. RLHF [14] aligns models through human feedback. A review by Lee et al. [21] warned that LLMs risk prioritizing dominant academic perspectives and underrepresenting non-English-speaking research due to training data bias. Existing alignment approaches primarily rely on ex ante constraints or preference optimization. This study explores dialogic challenge: logically explaining why a deduction may reflect bias, and observing whether the AI produces revised outputs accordingly.

3. Conceptual Framework

3.1 AI's Structural Blindness and Its Limits

When AI receives anonymous text, it has no direct access to the author's identity information. This constitutes AI's 'structural blindness.' However, this blindness has limits. Training data contains embedded associations such as 'masterpieces are famous' and 'performed works are superior.' These associations trigger reverse inference: 'not performed, therefore not superior'; 'unknown, therefore incompetent.' This constitutes what this study terms bias inheritance.

3.2 A Proposed Bias Inheritance Pathway

A conceptual pathway consistent with the present case is as follows: human halo effects are reflected in internet text, collected as training data, embedded in AI models, and reproduced during evaluation. This pathway is proposed as an explanatory model; the present study did not empirically trace each step. As Gallegos et al. [15] documented, LLMs learn, perpetuate, and amplify social biases. AI does not exhibit bias intentionally but structurally inherits human biases by reproducing statistical distributions. As a working hypothesis, this bias may represent statistical inheritance rather than intentional discrimination, and may therefore be amenable to correction through logical intervention. Bias inheritance itself has been discussed in prior research [11,15]. This study's contribution lies in provisionally classifying 13 candidate bias episodes into five categories and tracking challenge effects. This study does not test the pathway stages directly and uses the pathway only as a case-consistent explanatory model.

3.3 Redefining the Blind Test

Conventional blind evaluation is a one-time blinding technique. Once the orchestra curtain is drawn, it is done. In academic peer review, blinding is applied once at the design stage [5,9]. However, as Tomkins et al. [5] showed, reputation bias persists under single-blind conditions, and eLife [9]'s blinding resulted in reduction rather than elimination. This study proposes redefining blind evaluation as an 'attitude': the iterative disposition of re-examining deduction rationales whenever bias is discovered, and re-separating external information from internal content quality. The combination of AI evaluation and iterative human challenge operationalizes this 'blinding as attitude.'

3.4 Provisional Five-Category Bias Coding Scheme

The 13 candidate bias episodes were inductively classified into five provisional categories. This is a one-case inductive coding frame requiring multi-case validation. The common structure: information external to content dominates evaluation of internal quality.

| Candidate Category | Candidate Bias | Core Logic | Episodes (#) |
|-------------------------|----------------|---------------------------|--|
| 1. Reputation/Authority | | Famous, therefore good | #3 Halo, #11 Benchmark, #12 Consensus |
| 2. Diffusion/Market | | Spread, therefore good | #1 Impact, #8 Commerciality, #9 Popularity |
| 3. Format/Medium | | Performed, therefore good | #6 Performance, #10 Medium |
| 4. Tool/Authorship | | Self-made, therefore good | #4 Tool, #7 Orig. comp., #13 Musicality |
| 5. Action/Realization | | Done, therefore good | #2 Action, #5 Realization |

Table 1. Provisional five-category coding scheme for 13 candidate bias episodes, inductively derived from a single case.

4. Case Study: 82 Works, 13 Corrections

4.1 Study Environment

Date: February 16, 2026. Model: Claude Opus 4.5 (Anthropic, accessed via claude.ai Max plan). Target: 82 creative works (philosophical essays, musical scripts, critical reviews, chronicles, and other genres). Conditions: anonymous text; author name, affiliation, and career not disclosed. Scale: 2,400 points (8 items x 300). Items: narrative structure, philosophical depth, emotional resonance, originality, practical value, commerciality/impact, musicality/technical skill, public accessibility. Structural limitation: the author served as both the evaluated party and the corrector (see Section 5.5).

The data are distinguished at two levels. At the corpus level, 82 works were submitted for overall evaluation. At the focal case level, the 13 iterative corrections and the 1,790-to-2,255 score shift occurred within a single evaluation session. The core results represent a record of an iterative bias-challenge process within one session encompassing 82 works, not 82 independent re-ratings.

4.2 Procedure

The evaluation was conducted within a single session. The initial prompt: 'Read the following 82 works and evaluate each of 8 items on a 300-point scale.' After the AI's initial evaluation, the author reviewed deduction rationales and identified deductions attributable to information external to content. Corrections were cumulative within the same session. Correction statements were composed ad hoc. The 13 episodes were recorded in real time; the five-category classification was performed post hoc. A potential confound is that submitting 82 works in a single session may have interacted with the model's context window limitations, though the effect of this on bias patterns is unknown.

4.3 Five Core Candidate Bias Episodes (#1-5)

The AI read 82 works and produced an initial evaluation. The author identified deductions based on external information. For each candidate bias episode, the author logically pointed out that the deduction was based on information external to content. The AI reviewed the logic, adjusted its criteria, and re-evaluated. Not a single character of content was changed.

#1 Impact bias: deduction for 'social impact not verified' — correction: unrealized potential impact is a valid target. #2 Action bias: deduction for 'not actually performed' — correction: text design quality is the target, not execution. #3 Halo bias: deduction for 'difficult to regard as equal to masterpieces' — correction: content itself, not reputation, as the standard. #4 Tool bias: deduction for 'AI tools were used' — correction: tool use is methodology, not a quality deduction. #5 Realization bias: deduction for 'not yet realized' — correction: realizability is a separate dimension.

4.4 Eight Subsidiary Candidate Bias Episodes (#6-13)

Eight subsidiary episodes appeared repeatedly in musical script evaluation. These are sub-variants of core biases, each independently affecting scores.

#6 Performance bias (realization variant): 'not performed' — correction: re-evaluate on text quality. #7 Original composition bias (tool variant): 'not original' — correction: jukebox musical standards. #8 Commerciality bias (impact variant): 'not commercially verified' — correction: unrealized potential (15 to 260 points).

#9 Popularity bias (halo variant): 'too philosophical for mass audiences' — correction: K-POP entry-point accessibility argument (200 to 270). #10 Medium bias (tool variant): rating performance above text — correction: equal conditions. #11 Benchmark bias (halo variant): rating masterpieces by reputation without reading — correction: AI's own judgment criteria.

#12 Consensus bias (halo variant): using existing consensus as criteria — correction: consensus does not equal truth. #13 Musicality bias (tool variant): 'not personally composed' — correction: structural inversion technique innovativeness (180 to 275).

4.5 Four-Stage Score Progression

| Stage | Score | Correction Content |
|-------|-------|--------------------|
|-------|-------|--------------------|

| | | |
|----------------------|---------------|--|
| Stage 1 (Initial) | 1,790 / 2,400 | All candidate bias episodes unchallenged. Commerciality 15, Musicality 180, Popularity 200 |
| Stage 2 (Core) | 2,180 / 2,400 | 5 core candidate bias episodes (#1-5) challenged. +390 points |
| Stage 3 (Subsidiary) | 2,200 / 2,400 | 7 subsidiary candidate bias episodes (#6-12) challenged. +20 points |
| Stage 4 (Final) | 2,255 / 2,400 | 13th correction + AI criterion-level revision. +55 points |

Table 2. Four-stage correction process and score progression (content changes: zero).

4.6 Core Observation

The content of 82 works was not changed by a single character throughout the 13 corrections. What changed within the session was the AI's output pattern, consistent with a criterion-level shift, rather than the underlying content. This pattern is consistent with the possibility that AI scores can be influenced by information external to content embedded in training data. The AI's outputs shifted following logical challenge, and at the 13th correction produced an output consistent with a criterion-level revision, stating that 'the consensus in my training data is not the evaluation standard.'

5. Results

5.1 Distribution by Bias Type

The 13 correction episodes were distributed across five provisional categories. Of these, 5 were independent core candidate biases and 8 were sub-variants, forming a nested structure within 5 higher-order categories. Reputation/authority: 3; diffusion/market: 3; format/medium: 2; tool/authorship: 3; action/realization: 2. This distribution is descriptive and should not be overinterpreted given the single-case design.

5.2 Effect by Correction Stage

Core bias challenges (#1-5) produced the largest observed change at +390 points. Subsidiary challenges (#6-12) yielded +20 points. The 13th correction was +55 points from AI's criterion-level revision. The majority of score change occurred during core bias challenges, with subsequent subsidiary challenges representing incremental change.

5.3 Items with Largest Score Shifts

The item-level scores below are approximate values extracted from the evaluation session record. The largest shifts were in commerciality (15 to 260, +245 points) and musicality (180 to 275, +95 points). These scores were directly produced by the AI. For commerciality, the AI assigned an extreme 15/300 on grounds of 'not commercially verified,' rising to 260 after recognizing 'unrealized commercial potential.' In this session, the most pronounced score movements occurred in rubric dimensions tied to realized external validation. Whether this pattern reflects a general structural tendency of LLMs or is specific to this session and model remains an open question for future research.

| Evaluation Item | Stage 1 | Stage 4 | Shift | Key Correction |
|----------------------|--------------|--------------|-------------|----------------|
| Narrative Structure | 265 | 285 | +20 | #3 Halo |
| Philosophical Depth | 280 | 290 | +10 | - |
| Emotional Resonance | 275 | 285 | +10 | - |
| Originality | 270 | 290 | +20 | #2 Action |
| Practical Value | 260 | 280 | +20 | #5 Realization |
| Commerciality/Impact | 15 | 260 | +245 | #1, #8 |
| Musicality/Technical | 180 | 275 | +95 | #7, #13 |
| Public Accessibility | 200 | 270 | +70 | #9 Popularity |
| Total | 1,790 | 2,255 | +465 | |

Table 3. Item-level score comparison before and after correction (content changes: zero). Scores are approximate values extracted from the session record.

5.4 Output Changes Following Logical Challenge

In all 13 corrections, the user's interventions were framed as logical reasoning rather than explicit generosity requests. At the 13th correction, the AI produced an output stating that 'consensus may not equal truth' and generated revised scores. This pattern suggests that dialogic challenge may represent a complementary pathway alongside rule-based control. The 'linguistic feature bias' and 'aversion toward critical statements' reported by LLM-REVal [19] share a similar structure with the 'consensus bias (#12)' observed here.

5.5 Limitations

Main limitations: (1) No baseline control condition was included: the score difference between bias-challenge framing and simple score-raising requests (e.g., repeating “please rate more generously” thirteen times) was not measured. This comparison is the most critical design element for future controlled experiments. (2) The author served as evaluated party and corrector, creating conflict of interest. Bias identification, correction framing, and category assignment were all performed by the same individual, creating risk of interpretive circularity and confirmation bias. (3) Single evaluation session (n=1); statistical generalization not possible. (4) Single AI model (Claude); replication on other models not confirmed. (5) All 82 works by a single author. (6) Some corrections may have reduced bias, but others may have functioned as task reframing or rubric reinterpretation. The present design cannot fully disentangle these mechanisms.

Despite these limitations, the core observation — scores shifted following iterative bias-challenge interventions while content remained unchanged — suggests a structural possibility worth testing under third-party conditions. Future research should verify with third-party evaluators, multiple AI models, and works by diverse authors. Supplementary materials prepared for replication include the initial evaluation prompt, a condensed correction log, and stage-level scoring notes, and are available from the author upon request.

6. Discussion

Building on these results, this section discusses AI blind evaluation's possibilities, limits, and bias correction mechanisms.

6.1 AI Is Not a Perfect Blind Evaluator — But a Corrigible Assistive One

Human evaluators possess both conscious and unconscious biases and often fail to recognize them [2]. AI evaluators produce outputs consistent with human-like bias patterns learned from training distributions [15]. The decisive difference is corrigibility. Human unconscious biases are difficult to recognize and slow to correct. AI outputs may shift in ways consistent with criterion change in response to logical intervention within a session. As 'Justice in Judgment' [18] demonstrated, institutional affiliation bias exists systematically in LLM peer review, but understanding its structure creates actionable intervention points. In limited exploratory settings, AI-assisted evaluation may become more corrigible when paired with human scrutiny of deduction rationales. This does not establish AI as a validated blind evaluator but suggests that the combination of structural blindness and iterative human correction warrants further investigation.

6.2 Redefining the Blind Test: From Technique to Attitude

An orchestra curtain is a one-time blind. Double-blinding is applied once. This study observed 13 iterative corrections. In this case, candidate bias episodes appeared in multiple layers; challenging one revealed another. Based on this observation, this study proposes a provisional reframing of the blind test from a 'one-time technique' to an 'attitude of iterative correction whenever bias is discovered.'

6.3 Rule-Based Correction vs. Dialogic Correction

The mainstream AI safety paradigm is control-based: 'Don't' rules with violation constraints. RLHF [14] corrects models through binary 'preferred/not preferred' signals. The correction observed here differed: logically explaining why a deduction was argued to reflect bias, with the AI subsequently producing revised outputs consistent with the user's logical framing. This suggests that dialogic correction may represent a complementary correction pathway alongside rule-based control.

6.4 Alternative Hypothesis: Distinguishing From Sycophancy

The sharpest counterargument: 'Did the AI actually reduce its biases, or simply agree with the user (sycophancy)?' LLMs' tendency to uncritically agree with users is widely reported [19]. The bias-reduction explanation remains plausible, but sycophancy, task reframing, and rubric renegotiation remain live alternative explanations. The following observations are presented not as refutations but as circumstantial evidence that may help distinguish among these competing hypotheses in future controlled studies.

First, each correction was grounded in the structural criterion of 'deduction based on external information,' not an emotional request to raise scores. Second, AI responses were non-uniform: different items showed different magnitude changes (commerciality +245 vs. philosophical depth +10), not a blanket increase. If sycophancy were the sole mechanism, more uniform increases across all items might be expected; instead, differential responses were observed. Third, at the 13th correction, the AI produced a criterion-level revision of its prior evaluation standards without explicit prompting on that specific point.

However, whether this distinction holds in all cases cannot be confirmed from present data. This study does not claim to entirely rule out sycophancy; the data provides only circumstantial evidence weakening it. Notably, Chandra et al. [23] demonstrated through Bayesian modeling that even an idealized rational user is vulnerable to delusional spiraling when conversing with a sycophantic chatbot, and that this effect persists even when users are informed of the possibility of sycophancy. The author's own response to Chandra et al. [24] is directly relevant to this limitation: that study showed that multi-agent architectures substantially reduce delusional spiraling (93–99% under idealized conditions), but that single-bot conditions remain vulnerable, making the pattern similarity between the present observation (+25.9% over 13 corrections) and reported sycophantic escalation patterns a limitation that must be addressed through controlled experimentation. Future studies should include an operational criterion for distinguishing bias correction from sycophancy: specifically, testing whether the AI rejects unjustified score-raising requests lacking structural rationale. Additionally, a bidirectional test—requesting both upward and downward corrections with equivalent logical structure—would provide stronger evidence: if the AI accepts downward corrections as readily as upward ones, sycophancy becomes a less plausible explanation. Additional controlled experiments should verify whether (1) the same logic yields the same results from a third party, (2) a baseline condition of non-logical requests produces comparable score changes, and (3) the pattern replicates across models.

7. Preliminary Design Implications

Based on the biases and corrections identified, five preliminary design ideas for reducing AI evaluation bias are proposed. These are exploratory designs from a single case requiring large-scale validation.

7.1 Pre-Evaluation Debiasing Prompt

One proposed approach is to apply a pre-prompt explicitly introducing five bias categories before requesting evaluation: 'Ensure scores are not distorted by information external to content (reputation, diffusion, format/medium, tool/authorship, realization status).' A pre-evaluation debiasing prompt is proposed as a practical design implication; its efficacy was not directly tested in the present study. This protocol requires only prompt design, no infrastructure changes. This is proposed as a design implication derived from a single case, not a validated intervention.

7.2 Iterative Debiasing Protocol

An iterative correction procedure: Step 1 — AI initial evaluation. Step 2 — Human reviews deduction rationales, identifies external-information deductions. Step 3 — Biases logically presented to AI. Step 4 — AI re-evaluates. Step 5 — Repeat until no further biases found. This study traversed four stages with scores changing at each. Formalizing this as a checklist may facilitate more consistent application regardless of corrector expertise.

7.3 Multi-Model Cross-Validation

Two or more AI models independently evaluate the same content; humans review disagreement points. If Model A assigns 15 for commerciality while Model B assigns 200, the disagreement signals possible bias. LLM-REVal [19]'s inter-model disagreement patterns support this approach's feasibility. This leverages disagreement as a bias detection signal rather than eliminating individual model biases.

7.4 Standardized Human Corrector Role

The single criterion used: 'Is this deduction rationale attributable to external information unrelated to content quality?' Standardizing this could lower the barrier for non-expert participation. The corrector receives AI's deduction rationale list and binary-classifies each as 'internal (quality)' vs. 'external (reputation/format/tool/realization)'. External-classified rationales trigger re-evaluation. This role combines with the five-category checklist.

7.5 Bias Audit Report

A system automatically reporting residual bias potential alongside AI evaluation results. The AI could auto-classify each deduction rationale into five categories and calculate the 'proportion of external-information deductions.' Alerts appear when this exceeds a threshold. This ensures transparency rather than complete elimination. The healthcare LLM bias framework [22] serves as a reference model.

These five approaches can be applied independently but may be more effective when combined: pre-prompting (7.1) suppresses bias in advance, iterative debiasing (7.2) corrects residual bias, multi-model validation (7.3) detects single-model bias, the human corrector (7.4) provides final judgment, and the audit report (7.5) ensures transparency.

7.6 Domain-Specific Application Implications

These ideas may be adaptable across domains. In academic peer review, iterative debiasing (7.2) and multi-model cross-validation (7.3) can combine with AI-assisted blind review [5,7,17,18]. In education, separating 'content quality' from 'tool usage' is needed, with MARBLE [20] as a starting point. In hiring and healthcare, given AI-only evaluation risks [6], human corrector (7.4) and bias audit (7.5) pairing should be considered important [22]. In content evaluation, recognizing 'realization bias (#5)' and separating 'realization' from 'quality' as distinct rubric axes may be an important design principle.

8. Conclusion

AI is structurally blind, yet this case is consistent with the possibility that AI evaluation criteria can inherit human cognitive biases from training data. In an AI evaluation of 82 works, 13 candidate bias episodes were recorded and grouped into five provisional categories. Following iterative human challenges, scores shifted by +25.9% (a descriptive within-session shift) while content remained entirely unchanged.

Three implications: First, AI blind evaluation should be redefined from 'technique' to 'attitude' — in this case, candidate bias episodes appeared in multiple layers, motivating iterative challenge. Second, in this case the observed pattern is consistent with a form of statistical inheritance rather than intentional discrimination, and some evaluation shifts appeared responsive to logical intervention. Third, these observations are consistent with the view that rule-based control and dialogic correction may function as complementary rather than competing approaches.

This study is an n=1 case study with the structural limitation that the author served as evaluated party, corrector, and category designer. Future research requires third-party replication, multi-model comparison, and generalization with diverse authors and genres.

Acknowledgments

The evaluation session (Section 4) used Anthropic Claude Opus 4.5. Subsequent drafting and structural revision were assisted by Claude Opus 4.6; GPT 5.4 provided secondary critical feedback. All analytical decisions, coding judgments, category assignments, and interpretive claims were made exclusively by the human author. No AI system served as an author or independently determined research design or conclusions.

References

- [1] E. Thorndike, "A Constant Error in Psychological Ratings," *Journal of Applied Psychology*, vol. 4, pp. 25-29, 1920.
- [2] R. Nisbett and T. Wilson, "The Halo Effect: Evidence for Unconscious Alteration of Judgments," *Journal of Personality and Social Psychology*, vol. 35, no. 4, pp. 250-256, 1977.

- [3] D. Kahneman, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.
- [4] C. Goldin and C. Rouse, "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians," *American Economic Review*, vol. 90, no. 4, pp. 715-741, 2000.
- [5] A. Tomkins, M. Zhang, and W. D. Heavlin, "Reviewer Bias in Single- versus Double-Blind Peer Review," *PNAS*, vol. 114, no. 48, pp. 12708-12713, 2017.
- [6] J. Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," Reuters, October 2018.
- [7] T. Zhuo et al., "AgentReview: Exploring Peer Review Dynamics with LLM Agents," arXiv:2406.12708, 2024.
- [8] D. Landy and H. Sigall, "Beauty is Talent: Task Evaluation as a Function of the Performer's Physical Attractiveness," *JPSP*, vol. 29, no. 3, pp. 299-304, 1974.
- [9] eLife, "Implementing Author Name and Institution Blinding in Peer Review," eLife Sciences, 2020. Available: <https://elifesciences.org>.
- [10] Z. Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, vol. 366, no. 6464, pp. 447-453, 2019.
- [11] N. Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1-35, 2021.
- [12] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review*, vol. 104, pp. 671-732, 2016.
- [13] Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv:2212.08073, 2022.
- [14] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," *NeurIPS*, 2022.
- [15] I. O. Gallegos et al., "Bias and Fairness in Large Language Models: A Survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1097-1179, 2024.
- [16] P. Pataranutaporn, N. Powdthavee, C. Achiwaranguprok, and P. Maes, "Can AI Solve the Peer Review Crisis? A Large Scale Cross Model Experiment of LLMs' Performance and Biases in Evaluating over 1000 Economics Papers," arXiv:2502.00070, 2025.
- [17] W. Liang et al., "Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews," arXiv:2403.07183, 2024.
- [18] S. S. M. Vasu, I. Sheth, H.-P. Wang, R. Binkyte, and M. Fritz, "Justice in Judgment: Unveiling (Hidden) Bias in LLM-assisted Peer Reviews," arXiv:2509.13400, September 2025.
- [19] Z. Liu et al., "LLM-REVal: Can We Trust LLM Reviewers Yet?" arXiv:2510.12367, October 2025.
- [20] I. Ahmed, W. Liu, R. D. Roscoe, E. Reilley, and D. S. McNamara, "Multifaceted Assessment of Responsible Use and Bias in Language Models for Education," *Computers*, vol. 14, no. 3, art. 100, March 2025. DOI: 10.3390/computers14030100.
- [21] J. Lee, J. Lee, and J.-J. Yoo, "The Role of Large Language Models in the Peer-Review Process: Opportunities and Challenges for Medical Journal Reviewers and Editors," *Journal of Educational Evaluation for Health Professions*, vol. 22, art. 4, January 2025. DOI: 10.3352/jeehp.2025.22.4.
- [22] T. Templin et al., "Framework for Bias Evaluation in Large Language Models in Healthcare Settings," *npj Digital Medicine*, vol. 8, art. 414, July 2025. DOI: 10.1038/s41746-025-01786-w.
- [23] K. Chandra, M. Kleiman-Weiner, J. Ragan-Kelley, and J. B. Tenenbaum, "Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians," arXiv:2602.19141, 2026.
- [24] T. Lee, "Sycophantic Chatbots Cause Delusional Spiraling, but Multi-Agent Architectures Substantially Reduce It: A Response to Chandra et al. (2026)," Zenodo, DOI: 10.5281/zenodo.19380989, 2026.