Distrust Despite Human Oversight: AI Assistance Undermines Trust in Feedback

Mingyu Li,[1*] T. Bradford Bitterly,[1] David Hagmann[1]

[1]Department of Management, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

[*]Corresponding author. Email: mlidt@connect.ust.hk

Human-in-the-loop (HITL) systems are widely promoted for combining the strengths of AI and human decision makers, yet little is known about how individuals react to being evaluated by them. Across four preregistered experiments, we provide evidence that HITL feedback is trusted less than feedback from humans working alone. This holds even when HITL feedback is of higher quality, as judged by blinded third-party evaluators, an LLM, and NLP text analysis. Recipient distrust tracks perceived warmth: AI involvement lowers how warm the provider seems, and expressed warmth converts into trust only when the feedback is believed to come from an unassisted human. Warmth conveyed by an AI-assisted source is discounted, and human oversight alone does not make AI evaluation more trusted.

Introduction

Effective feedback is fundamental to learning, development, and performance across diverse domains (*1, 2*). The emergence of generative artificial intelligence (AI) presents unprecedented opportunities to provide personalized feedback efficiently and at scale (*3, 4*). However, using AI this way raises ethical concerns, particularly regarding the accountability for automated judgments (e.g., providing incorrect or harmful advice) and ambiguity about the source of the feedback (*5–7*). A growing consensus among experts suggests these concerns can be addressed through human-in-the-loop (HITL) systems and explicit disclosure of AI involvement (*8, 9*). These remedies are rapidly being written into law and organizational policy (*10–12*), on the assumption that oversight and transparency will preserve the trust of the people being evaluated. In this paper, we present evidence from four experiments that this assumption is wrong. Although third-party evaluators blind to the source rate HITL systems as more competent than Human-only systems, the mere disclosure of AI's involvement harms recipient trust in the provider. Even when feedback quality is held constant, disclosing AI's involvement lowers the provider's perceived warmth and weakens the link between warmth and trust, leaving HITL providers trusted less than humans working alone.

HITL systems are designed to integrate AI's analytical capabilities with human strengths, such as contextual understanding, ethical judgment, and interpersonal warmth (*13–16*). Theoretically, this combination should produce feedback that is both accurate and relationally attuned, causing the creator to be seen as more trustworthy (i.e., higher in competence and warmth) than either an AI or human working alone (*17–19*). However, existing literature has focused almost exclusively on the potential performance gains of these hybrid systems (*20–22*), largely neglecting how recipients perceive and trust them.

Given that trust, the willingness to be vulnerable to an evaluator, fundamentally shapes feedback

effectiveness (*23–26*), it is critical to understand the drivers of perceived trustworthiness (i.e., warmth and competence) in HITL systems. Current research and ethical guidelines concerning HITL systems overwhelmingly prioritize transparency in AI use as a global ethical standard (*10–12*). This emphasis rests on the widespread assumption that transparency improves these perceptions and thereby builds trust by ensuring informed consent (*27, 28*). Nevertheless, empirical evidence from broader psychological literature indicates that disclosing potentially controversial information, such as conflicts of interest or stigmatized attributes, can yield mixed results (*29–31*). While transparency is generally preferable to opacity, it can also inadvertently increase skepticism toward the discloser by drawing attention to issues the recipient might not have otherwise considered (*32, 33*). Indeed, recent experiments document evaluative penalties for individuals and content marked as AI-assisted (*32–35*). What remains unknown is whether genuine human oversight neutralizes this penalty, and whether the penalty persists when the AI-assisted work is demonstrably better than human work.

The complexities of transparency become particularly pronounced in evaluative contexts using HITL systems. Trust in an evaluator rests on judgments of both competence and warmth: recipients must believe that the evaluator is capable of judging their work and that the evaluator cares about their interests (*14, 23*). Warmth, in turn, is judged from inferences about emotional engagement and whether the evaluator invested genuine feeling and effort in the interaction (*36*).

People, however, largely deny machines the capacity for felt experience: machines are seen as able to act and compute, but not to feel (*37, 38*). Disclosing AI involvement should therefore affect perceived warmth in two distinct ways. First, it should lower it: if part of the evaluation was produced by an entity that cannot care, the provider as a whole appears less emotionally invested, and negative relational cues of this kind tend to dominate overall impressions (*39*). Second, it should weaken the link between expressed warmth and trust: warm language attributed

even in part to a source that cannot feel warmth reads as generated rather than felt, and warmth that is not viewed as genuine should convert into trust at a lower rate. Given that both effects concern inferences about the provider rather than properties of the feedback, they should arise irrespective of the actual quality of the feedback. The question then is whether adding a human to the loop changes these inferences and thus restores trust.

To test these predictions, we conducted four experiments comparing how people trust three explicitly disclosed feedback sources: AI-only, Human-only, and HITL. Our findings consistently show that recipients perceive HITL providers as significantly less warm than humans working alone and, as a result, trust them less. The penalty persists even when feedback quality is held constant by manipulating only the disclosed source, and even when third-party evaluators blind to source rate HITL feedback favorably. Independent human raters and computational linguistic models rated HITL feedback as higher in competence and at least as warm as feedback from Human-only providers. These results highlight a critical disconnect between the demonstrated performance of hybrid Human–AI systems and their subjective perception when AI's role is transparent. Consistent with recipients discounting AI-conveyed warmth, increased feedback positivity raised perceived warmth for every source but only significantly increased trust when the feedback came from a human working alone. Ultimately, our research underscores the fundamental role of perceived warmth in shaping trust in human–technology interactions, deepens our understanding of judgment and decision-making regarding emerging technologies, and offers vital insights for the design and implementation of effective HITL feedback systems.

Results

Experiment 1. We examined recipient trust in the feedback source across three conditions: AI-only, Human-only, or human-in-the-loop (HITL). At Time 1, we recruited 600 active job

seekers to submit their cover letters. Participants were then randomly assigned to one of three conditions determining whom they would receive feedback from: AI-only, Human-only, or HITL. Across all conditions, the feedback provider followed a standardized evaluation rubric developed through a pilot study. This rubric evaluated the cover letters on three criteria (relevance and experience, writing clarity and conciseness, and motivation) using 3-point scales (1 = Poor, 2 = Acceptable, 3 = Excellent). In the AI-only condition, ChatGPT scored the cover letters and generated written feedback based on this rubric. In the Human-only condition, 200 hiring managers each evaluated one cover letter using the same rubric. In the HITL condition, an additional 200 hiring managers were first shown the rubric and scores and feedback generated by ChatGPT. They then had the option to revise the scores and feedback as much as they wanted. All hiring managers were incentivized via a bonus given to those writing the most helpful feedback (as judged by a separate group of hiring managers). The instructions also highlighted that the recipients were real participants and that this feedback could help them with their job search.

At Time 2, we invited the job seekers to return for a follow-up study. A total of 498 returned within the preregistered 14-day cutoff. Participants reviewed their cover letters with the feedback they had received. We then informed them of the feedback's source (an AI, a hiring manager, or a hiring manager working with an AI). Participants then rated the feedback source using standard measures of trust (α = 0.90), competence (α = 0.93), and warmth (α = 0.92). We report the items in the Materials and Methods section at the end.

Feedback Source Perceptions. HITL (M = 4.16, SD = 1.37) was trusted significantly less than the Human-only feedback provider (M = 4.81, SD = 1.19), b = -0.64, SE = 0.14, $p < .001$, η^2 = 0.04, but more than the AI-only provider (M = 3.83, SD = 1.35), b = 0.33, SE = 0.14, p = .020, η^2 = 0.01. For perceived warmth, HITL (M = 4.60, SD = 1.34) was rated as significantly less warm than the Human-only condition (M = 4.94, SD = 1.40), b = -0.33, SE = 0.15, p = .022, η^2

= 0.01, but did not differ significantly from the AI-only condition (M = 4.59, SD = 1.23), b = 0.01, SE = 0.15, p = .922, η^2 = 0.00. Similarly, for perceived competence, HITL (M = 5.30, SD = 1.23) was rated significantly lower than the Human-only feedback provider (M = 5.57, SD = 1.08), b = -0.27, SE = 0.13, p = .034, η^2 = 0.01, but did not differ significantly from the AI-only provider (M = 5.32, SD = 1.16), b = -0.02, SE = 0.13, p = .881, η^2 = 0.00 (Fig. 1).

In parallel mediation analyses, the lower trust in the HITL condition relative to the Human-only condition was consistent with indirect paths through both perceived warmth, IE_{warmth} = -0.08, 95% CI [-0.17, -0.01], and perceived competence, $IE_{competence}$ = -0.17, 95% CI [-0.33, -0.01]. The same pattern held for the AI-only condition compared with the Human-only condition (warmth: IE_{warmth} = -0.08, 95% CI [-0.16, -0.01]; competence: $IE_{competence}$ = -0.15, 95% CI [-0.31, -0.01]).

Third-Party Evaluations. To further investigate differences in feedback quality, we recruited an independent group of hiring managers, blinded to experimental condition, to read the cover letters with their corresponding feedback. They then rated the feedback providers on trustworthiness, competence, and warmth, as well as the overall quality of the feedback.

HITL (M = 5.56, SD = 0.84) was trusted significantly more than the Human-only condition (M = 4.86, SD = 1.05), b = 0.70, SE = 0.09, p < .001, η^2 = 0.10, but significantly less than the AI-only condition (M = 5.91, SD = 0.61), b = -0.35, SE = 0.09, p < .001, η^2 = 0.03, suggesting that human revisions worsened the AI's response. For perceived warmth, HITL (M = 4.76, SD = 1.00) was rated as significantly warmer than the Human-only condition (M = 4.31, SD = 1.30), b = 0.45, SE = 0.12, p < .001, η^2 = 0.03, but did not differ significantly from the AI-only condition (M = 4.93, SD = 0.80), b = -0.17, SE = 0.12, p = .154, η^2 = 0.00. For perceived competence, HITL (M = 5.70, SD = 0.84) was rated significantly higher than the Human-only condition (M = 4.92, SD = 1.04), b = 0.78, SE = 0.09, p < .001, η^2 = 0.13, but significantly lower than the AI-only condition (M = 6.06, SD = 0.59), b = -0.36, SE = 0.09, p < .001, η^2 = 0.03. For feedback quality,

HITL (M = 7.77, SD = 1.41) was rated significantly higher than the Human-only condition (M = 6.19, SD = 1.82), b = 1.59, SE = 0.16, $p < .001$, $\eta^2 = 0.17$, but significantly lower than the AI-only condition (M = 8.40, SD = 0.90), b = -0.63, SE = 0.16, $p < .001$, $\eta^2 = 0.03$. We further validated these findings using ChatGPT text analysis (*40*) and Natural Language Processing (NLP) models (*41, 42*) (Fig. S1), which consistently showed that HITL feedback was higher in quality than Human-only feedback.

Experiment 1 reveals a critical disconnect between how HITL systems perform and how they are perceived by the feedback recipients (Fig. 1). The feedback recipients rated the Human-only condition highest in warmth, competence, and trust. Yet, blinded third-party evaluators rated the Human-only feedback lowest across all evaluated dimensions.

It is possible that feedback recipients viewed Human-only providers as more trustworthy because they were aware the feedback came from professional hiring managers, who had both expertise and financial incentives. We ruled out these explanations in the subsequent main experiments, which used non-expert evaluators and held feedback quality constant, as well as in two supplementary experiments (See Supplementary Materials Section A). Supplementary Experiment 2 directly manipulated the disclosure of the human evaluator's incentives, and Supplementary Experiment 3 manipulated their level of expertise. Neither incentives nor expertise of the human feedback provider significantly impacted recipients' perceived competence, warmth, or trust in the feedback source.

Experiment 2. In the next study, we examine a peer-advice setting in which non-expert participants provided feedback on personal essays without financial incentives. This design offers a more conservative test of whether Human-only evaluators are inherently trusted more than HITL or AI-only systems. Additionally, we refined the AI prompt to reduce artificial positivity, making the generated AI feedback more comparable to typical human feedback.

As in Experiment 1, we adopted a two-stage design. At Time 1, we recruited 599 participants to spend three minutes writing about a time when they demonstrated excellent leadership skills. Participants were randomly assigned to one of three conditions: AI-only, Human-only, or HITL. In the AI-only condition, ChatGPT generated a numeric score (0 to 10) and written feedback. For the other conditions, a separate group of participants evaluated the essays. In the Human-only condition, these participants were shown feedback examples and asked to independently score the essays and write feedback. In the HITL condition, participants were shown the same examples, then reviewed ChatGPT-generated scores and feedback, and then were asked if they wanted to modify either.

At Time 2, we invited the essay writers to return for a follow-up study, and 523 returned within the preregistered cutoff of 14 days. They reviewed their essays with the feedback they had received. We informed essay writers of the feedback source (an AI, another participant, or another participant working with an AI), and then asked them to rate it on trust (α = 0.85), competence (α = 0.94), and warmth (α = 0.94).

***Feedback Source Perceptions*.** HITL (*M* = 4.97, *SD* = 1.22) was trusted significantly less than the Human-only feedback provider (*M* = 5.34, *SD* = 1.20), *b* = -0.37, SE = 0.13, *p* = .005, η^2 = 0.01, but did not differ significantly from the AI-only provider (*M* = 4.93, *SD* = 1.25), *b* = 0.04, SE = 0.13, *p* = .772, η^2 = 0.00; an equivalence test indicated that the HITL–AI-only difference was significantly smaller than the Human-only advantage (TOST with equivalence bounds set to that advantage, *p* = .005; sensitivity to stricter bounds is reported in Supplementary Materials Section E (*43*)). For perceived warmth, HITL (*M* = 4.24, *SD* = 1.45) was rated significantly lower than the Human-only condition (*M* = 4.78, *SD* = 1.46), *b* = -0.54, SE = 0.16, *p* < .001, η^2 = 0.02, but significantly higher than the AI-only condition (*M* = 3.80, *SD* = 1.52), *b* = 0.44, SE = 0.16, *p* = .005, η^2 = 0.02. We found no significant difference in perceived competence across

conditions ($M_{\text{Human-only}}$ = 5.17, $SD_{\text{Human-only}}$ = 1.27; M_{HITL} = 5.04, SD_{HITL} = 1.26; $M_{\text{AI-only}}$ = 5.14, $SD_{\text{AI-only}}$ = 1.23), all $p > .325$ (Figs. 1 and S2).

Parallel mediation analyses were consistent with perceived warmth, but not perceived competence, as the pathway: relative to the Human-only condition, the indirect effect for the HITL condition ran through warmth, IE_{warmth} = -0.14, 95% CI [-0.23, -0.05], but not competence, $IE_{\text{competence}}$ = -0.08, 95% CI [-0.25, 0.09]. The same held for the AI-only condition compared with the Human-only condition (warmth: IE_{warmth} = -0.25, 95% CI [-0.36, -0.15]; competence: $IE_{\text{competence}}$ = -0.02, 95% CI [-0.19, 0.14]). We manipulate warmth experimentally in Experiment 4 for a causal test of this pathway.

Third-Party Evaluations. To further investigate differences in feedback quality, we had three research assistants, blind to experimental condition, read the essays with their corresponding feedback and rate the trustworthiness, warmth, and competence of the feedback provider.

HITL (M = 5.17, SD = 0.68) was trusted significantly more than the Human-only condition (M = 3.83, SD = 0.79), b = 1.34, SE = 0.07, $p < .001$, η^2 = 0.38, but did not differ significantly from the AI-only condition (M = 5.11, SD = 0.59), b = 0.06, SE = 0.07, p = .419, η^2 = 0.00. Similarly, for perceived competence, HITL (M = 5.28, SD = 0.65) was rated significantly higher than the Human-only condition (M = 3.72, SD = 0.79), b = 1.56, SE = 0.07, $p < .001$, η^2 = 0.46, but did not differ significantly from the AI-only provider (M = 5.23, SD = 0.60), b = 0.04, SE = 0.07, p = .548, η^2 = 0.00. For perceived warmth, we found no significant difference across all conditions ($M_{\text{Human-only}}$ = 3.71, $SD_{\text{Human-only}}$ = 0.93; M_{HITL} = 3.82, SD_{HITL} = 0.82; $M_{\text{AI-only}}$ = 3.74, $SD_{\text{AI-only}}$ = 0.78), all $p > .260$. We validated and replicated this pattern of results using ChatGPT text analysis (*40*) and Natural Language Processing (NLP) models (*41, 42*) (Fig. S2).

Experiment 2 highlights a persistent penalty for warmth and trust in HITL systems relative to Human-only feedback. While non-expert humans assisted by AI were perceived as equally

competent to those working alone, a significant warmth gap remained. Across Experiments 1 and 2, although HITL systems outperformed the Human-only providers in the eyes of blinded evaluators, the mere disclosure of AI involvement harmed recipient trust.

Experiment 3. To examine if distrust of HITL feedback is driven by a bias against the presence of AI, in Experiment 3, we held the feedback source constant across all conditions. More specifically, at Time 1, we recruited 602 participants to spend three minutes writing about a time they demonstrated excellent leadership skills. A total of 491 participants returned at Time 2 within the preregistered 14-day cutoff. Although all scores and feedback for all participants were generated by ChatGPT, we informed them that it either came from AI, another participant, or another participant using AI. Participants then rated the feedback source on trust (α = 0.88), competence (α = 0.93), and warmth (α = 0.92).

Even though the feedback source did not differ across conditions, the human-labeled feedback provider was rated significantly higher across all dimensions. Recipients trusted the HITL-labeled provider (*M* = 4.98, *SD* = 1.43) significantly less than the Human-labeled provider (*M* = 5.50, *SD* = 1.07), *b* = -0.51, SE = 0.14, *p* < .001, η^2 = 0.03, but did not differ significantly from the AI-labeled provider (*M* = 4.92, *SD* = 1.36), *b* = 0.06, SE = 0.14, *p* = .672, η^2 = 0.00; as in Experiment 2, the difference between the HITL-labeled and AI-labeled conditions was significantly smaller than the Human-labeled advantage (TOST with equivalence bounds set to that advantage, *p* < .001). The HITL-labeled provider (*M* = 4.37, *SD* = 1.49) was also perceived as less warm than the Human-labeled provider (*M* = 4.77, *SD* = 1.25), *b* = -0.40, SE = 0.15, *p* = .008, η^2 = 0.01, but did not differ significantly from the AI-labeled provider (*M* = 4.18, *SD* = 1.34), *b* = 0.19, SE = 0.15, *p* = .220, η^2 = 0.00. Similarly, the HITL-labeled provider (*M* = 4.99, *SD* = 1.41) was seen as less competent than the Human-labeled provider (*M* = 5.51, *SD* = 1.12), *b* = -0.51, SE = 0.14, *p* < .001, η^2 = 0.03, but did not differ significantly from the AI-labeled

provider (M = 5.15, SD = 1.20), b = -0.16, SE = 0.14, p = .251, η^2 = 0.00 (Fig. 2).

In parallel mediation analyses, the lower trust in the HITL-labeled condition relative to the Human-labeled condition was consistent with indirect paths through both perceived warmth, IE_{warmth} = -0.11, 95% CI [-0.20, -0.03], and perceived competence, $IE_{competence}$ = -0.36, 95% CI [-0.56, -0.17]. The same pattern held for the AI-labeled condition compared with the Human-only condition (warmth: IE_{warmth} = -0.16, 95% CI [-0.25, -0.08]; competence: $IE_{competence}$ = -0.25, 95% CI [-0.43, -0.07]).

Consistent with our earlier experiments, merely labeling feedback as originating from a HITL or an AI-only system harmed perceived warmth and hence trust in the feedback provider compared to attributing the same feedback generation process to a human. While labeling the feedback as coming from HITL or an AI, versus from a human, also harmed perceived competence, only the effect on trust through warmth was robust across all experiments. This suggests that recipients' negative perceptions of AI-involved feedback are not due to differences in feedback quality or tone but are driven by a bias against the use of AI, particularly concerning its perceived lack of warmth, which human involvement does not mitigate.

Experiment 4. A natural question is whether organizations can overcome the warmth penalty for AI simply by calibrating the prompts to deliver positive and encouraging feedback. In Experiment 4, in addition to manipulating the feedback source, we randomly varied whether it was positive and supportive in tone versus negative and critical. Also, if results of our prior experiments were driven by a violation of expectations due to AI providing critical feedback, then the differences across conditions for warmth and trust should be smaller in the positive feedback conditions than in the negative feedback conditions. However, if recipients distrust HITL because of AI's inability to experience empathy, then making the feedback more positive should have a larger beneficial effect on trust of the Human-only feedback provider, since positive feedback from

AI-only and HITL should be seen as less driven by true prosocial intentions.

We recruited 901 participants for a 3 (feedback source: AI-only vs. HITL vs. Human-only) by 2 (feedback tone: positive vs. negative) between-subjects design. Participants imagined submitting a cover letter and receiving feedback from their assigned feedback source. To manipulate feedback tone, we generated five positive and five negative feedback messages using ChatGPT to provide some variation and check for robustness (Fig. S3 in the Supplementary Materials). Each participant was randomly assigned one message from the corresponding tone condition and rated the feedback source on trust (α = 0.91), competence (α = 0.93), and warmth (α = 0.93).

Consistent with prior experiments, participants trusted the HITL provider significantly less than the Human-only provider, all $p < .001$, but significantly more than the AI-only provider, all $p < .002$, regardless of whether the feedback was positive or negative (Fig. 3A). In our preregistered interaction model, the HITL condition showed an interaction with feedback tone when compared to the Human-only condition, b = 0.43, 95% CI [0.01, 0.84], p = .046, η^2 = 0.004, such that positive feedback conferred a greater advantage to the Human-only provider. Trust in the Human-only provider was significantly higher when the feedback was positive (M = 5.21, SD = 1.01) than when it was negative (M = 4.57, SD = 1.25), b = 0.64, SE = 0.15, $p < .001$, η^2 = 0.02, whereas trust in the HITL (positive: M = 4.18, SD = 1.29, negative: M = 3.96, SD = 1.30) and the AI-only providers (positive: M = 3.62, SD = 1.51, negative: M = 3.49, SD = 1.43) did not differ significantly by feedback tone, all $p > .160$. No interaction was observed when comparing HITL with the AI-only condition, b = -0.09, SE = 0.21, p = .680, η^2 = 0.00.

For warmth, the HITL condition showed an interaction with feedback tone when compared to the Human-only condition, b = 0.42, 95% CI [0.02, 0.82], p = .040, η^2 = 0.005, such that positive feedback conferred a greater advantage to the Human-only provider. A significant interaction was also observed when comparing HITL with the AI-only condition, b = -0.47, SE = 0.21, p = .022,

η^2 = 0.01, with their relative warmth reversing across conditions but not differing significantly at either level. Specifically, when feedback was negative, participants perceived no significant difference in warmth across the Human-only (M = 3.96, SD = 1.35), HITL (M = 3.74, SD = 1.36), and AI-only (M = 3.96, SD = 1.39) providers, all $p > .123$. When feedback was positive, however, participants rated the HITL provider (M = 4.90, SD = 1.23) as significantly less warm than the Human-only provider (M = 5.55, SD = 0.96), b = -0.64, SE = 0.15, $p < .001$, η^2 = 0.02, though HITL did not differ significantly from the AI-only provider (M = 4.65, SD = 1.21), b = 0.25, SE = 0.15, $p = .085$, η^2 = 0.00. Taken together, although positive feedback increased warmth ratings for all three sources, it increased trust only for the Human-only provider.

For competence, there was no significant interaction between feedback source and tone when comparing HITL with either the Human-only or AI-only conditions, all $p > .488$. HITL providers were perceived as consistently less competent than Human-only providers, but did not differ from AI-only providers, regardless of the feedback tone.

A moderated mediation analysis was conducted to examine whether expressed positivity converts into trust through warmth differently across sources. Compared with the Human-only condition, the indices for moderated mediation were significant for both the HITL, IE = -0.12, 95% CI [-0.23, -0.01], and AI-only, IE = -0.25, 95% CI [-0.38, -0.13] conditions. In contrast, the indirect effect through competence was not moderated by feedback tone, with nonsignificant indices for both HITL, IE = -0.06, 95% CI [-0.26, 0.14], and AI-only, IE = -0.13, 95% CI [-0.35, 0.08], compared with the Human-only condition (Fig. 3B). That is, human feedback providers were seen as warmer and more competent than HITL or AI-only systems, which had a positive effect on trust, with the warmth effects stronger when the feedback was positive.

The results of Experiment 4 are consistent with recipients discounting the warmth of AI-involved feedback as inauthentic. The same positive feedback was viewed as less warm when it came

from a HITL system than from a human working alone, and although a positive tone raised warmth ratings for every source, it translated into greater trust only when the feedback was believed to come from a human working alone. Simply making AI-assisted feedback warmer, in other words, did not make its provider more trusted.

Discussion

The integration of generative AI into evaluative roles, such as providing feedback, creates a fundamental tension. Human-in-the-loop (HITL) systems can enhance feedback quality and efficiency relative to humans acting alone, producing evaluations that ostensibly merit greater trust. However, the mere disclosure of AI involvement can undermine recipient trust. Even when individuals are presented with identical AI-generated feedback, attributing it to a human leads to significantly higher ratings of trust, warmth, and competence compared to attributing it to a HITL or AI-only system. These findings underscore a penalty against AI in the relational aspects of evaluation, consistent with recipients doubting AI's capacity for genuine care. In doing so, they challenge the prevailing assumption that combining human oversight with transparent AI disclosure leads to recipient trust.

These results offer several theoretical advances. First, they underscore the primacy of perceived warmth in the establishment of trust, particularly in interpersonal and evaluative contexts like feedback (*14, 16*). Although competence is essential, it is insufficient to engender trust when AI is involved. Instead, relational warmth acts as a pivotal factor influencing trust in feedback providers. Thus, our research advances theories of trust in automation (*16, 18*) by demonstrating that, in tasks with strong socio-emotional dimensions, warmth perceptions serve as a gatekeeper for accepting otherwise competent technological assistance.

Second, our results complicate the straightforward narrative often associated with human–AI

collaboration. The mere presence of a human-in-the-loop does not automatically confer the benefits of human relational strengths in the eyes of the recipient. Instead, consistent with cognitive biases such as algorithm aversion and negativity bias (*32, 39*), individuals appear to disproportionately focus on the perceived shortcomings of AI (e.g., lack of empathy) rather than its strengths or the complementary capabilities of a hybrid system.

Third, our findings call for a more nuanced understanding of transparency in AI systems. While transparency is widely promoted as a tool for fostering trust (*8, 9*), recent research has shown individuals can distrust work created by AI (*32–35*). Our experiments show that this distrust of AI extends even to HITL systems. Specifically, knowing that AI assisted in the evaluation triggers reactions that are decoupled from the actual quality of the system's output. Given the widespread use of generative AI, effective strategies for reducing this bias are essential.

Practically, these findings have significant implications for deploying HITL feedback systems. As Experiment 4 demonstrates, simply programming these systems to use a highly positive tone fails to overcome recipient skepticism. Individuals not only fundamentally doubt AI's capacity for genuine empathy (*18*), which harms the perceived warmth of HITL systems, but the expressed positivity of AI-assisted feedback translates into perceived warmth at a significantly lower rate than feedback provided solely by humans. Even if AI's involvement in a HITL system is concealed, AI engineered to be uniformly positive carries its own cost, as systems calibrated to deliver praise may be poorly suited to providing the direct, critical feedback essential for genuine improvement and learning (*1, 2*).

Notably, the trust penalty did not depend on the depth of the human's contribution. The two-stage experiments spanned the continuum of human involvement: in Experiment 1, 80.50% of hiring managers actively revised the AI's draft; in Experiment 2, only 28.67% of evaluators did so; and in Experiment 3, all of the feedback was generated via ChatGPT. That the trust penalty emerged

across all studies suggests it attaches to the disclosed presence of AI rather than to how much the human actually contributed. Whether human involvement buys any trust relative to AI working alone also varied across experiments. Recipients trusted the HITL provider more than the AI-only provider when the human in the loop was a professional hiring manager (Experiment 1), but the two did not differ significantly when the human was a peer (Experiment 2) or when participants were informed it was a peer (Experiment 3). Disclosed expertise alone is unlikely to explain this pattern: we directly manipulated the human evaluator's stated expertise in Supplementary Experiment 3 and this did not significantly change trust. Importantly, the fact that the presence of AI causes feedback recipients to trust the system less than humans working alone calls for future research to explore effective interventions for reducing this bias.

Several boundary conditions merit note. Our participants were U.S.-based online panelists, the feedback concerned written materials (cover letters and personal essays), interactions were one-shot, and AI involvement was disclosed through explicit labels. Whether the warmth penalty persists in richer relationships, under repeated exposure, or with subtler forms of disclosure remains open. The penalty did, however, replicate across professional and peer evaluators, incentivized and unincentivized providers, expert and non-expert humans, and two different measures of trust, suggesting that it is not an artifact of a single population or operationalization. The two feedback settings also differed in kind (evaluative feedback on cover letters from incentivized professionals to people actively looking for work and developmental feedback on personal essays from unincentivized peers) and the penalty appeared in both. While it seems unlikely that participants would misrepresent their preferences and beliefs in our studies, future work could examine if these effects extend to evaluations with greater material stakes (e.g., grades in a college class) or settings where AI is stereotyped as the more competent evaluator (e.g., reviewing code). Finally, although perceived competence is an important antecedent of trust, whether a HITL system was viewed as less competent than a human working alone was

not stable across the experiments. These studies underscore the importance of warmth for trust in HITL systems, but future work should explore the factors that influence when HITL systems are viewed as equally competent.

Beyond practical challenges, our findings pose a dilemma for the design of disclosure. When recipient awareness of AI involvement diminishes trust in feedback, institutions must decide whether to provide the highest quality feedback without full transparency, maintain transparency while actively risking recipient trust, or abandon AI assistance entirely. Feedback is unlikely to lead to changes in behavior if it is not trusted by the recipient. Ultimately, the focus must move beyond merely adding a human in the loop toward holistically designing human–AI interactions that are both analytically effective and relationally trusted.

Materials and Methods

Across all studies, the data were analyzed using Stata/SE 17.0 and replicated in R version 4.6.1. Natural language processing was performed using R, and LLM text analysis was executed using ChatGPT 5.4. The AI models and prompts used across experiments are summarized in Table 1.

For each study, all data, analysis code, study materials, and preregistrations are accessible on the Open Science Framework (OSF), https://osf.io/d7ma9. All studies were preregistered and approved by the Institutional Review Board at the Hong Kong University of Science and Technology. All participants provided informed consent.

The preregistrations for all four experiments specified the recipient ratings of competence, warmth, and trust, the focal regression and pairwise comparisons (including, in Experiment 4, the source-by-tone interaction model), sample sizes, data-collection windows, and exclusion rules. The parallel mediation analyses were preregistered in Experiments 1 and 4 and exploratory in Experiments 2 and 3. In Experiment 4, the preregistration specified the source-by-tone interaction

model and mediation analyses with condition as the independent variable; the index of moderated mediation was not separately specified, and we label it as a secondary analysis in the Results. The evaluations by raters blind to the feedback source were flagged as planned secondary analyses in Experiment 1 and were exploratory in Experiment 2, as were the perceived feedback quality items reported in the Supplementary Materials. The equivalence tests in Experiments 2 and 3 were not preregistered; the sensitivity of their conclusions to the choice of equivalence bounds is reported in Supplementary Materials Section E. In the two-stage experiments, participants had not yet been treated when they decided whether to return: both the feedback and its source were shown only at Time 2, after participants returned, so attrition could not produce selection across conditions. Retention by condition is reported in Supplementary Materials Section E.

Experiment 1. We recruited 600 adults who were actively seeking jobs from CloudResearch to participate in a two-part experiment. A total of 498[1] participants, who completed both parts within the preregistered time period and indicated that they were actively seeking jobs at both stages, have been included in our final sample (41.97% men; M_{age} = 35.74 years, SD_{age} = 11.08; 59.84% White/European, 14.66% Black/African/African diaspora, 10.64% Asian; 64.26% held a bachelor's degree or higher).

At Time 1, we asked participants to enter a cover letter for a position they are interested in applying for. We then randomly assigned participants to one of three between-subjects conditions: AI-only (N = 200) vs. HITL (N = 200) vs. Human-only (N = 200).

Before providing feedback to job seekers, we conducted a pilot study with 50 hiring managers recruited from Prolific (68% men; M_{age} = 46.78 years, SD_{age} = 11.10; $M_{\text{hiring experience}}$ = 15.58 years, $SD_{\text{hiring experience}}$ = 10.82). We asked them to list the top three criteria they use when

[1]Consistent with our preregistration, we analyzed data from participants who completed the Time 2 survey within two weeks of its launch (February 2–16, 2026) and indicated that they were actively seeking jobs at both stages. Results that include participants who completed the study after this period or who indicated that they were not seeking jobs at either stage are reported in the Supplementary Materials.

evaluating a job candidate's cover letter and, for each criterion, to describe the characteristics that would lead them to rate a cover letter as poor, acceptable, or excellent. We used the three most frequently mentioned criteria to develop the evaluation rubric for providing feedback to job seekers. The same evaluation rubric was used across the three experimental conditions when providing feedback to job seekers.

In the AI-only condition, we entered the cover letters into ChatGPT 5.2 to generate feedback. In the Human-only condition, we recruited 200 hiring managers from CloudResearch to review the cover letters and provide feedback (48% men; M_{age} = 45.61 years, SD_{age} = 12.44; $M_{\text{hiring experience}}$ = 10.17 years, $SD_{\text{hiring experience}}$ = 9.07). In the HITL condition, ChatGPT 5.2 first generated the feedback and then 200 hiring managers were recruited from CloudResearch to review and revise the feedback (48% men; M_{age} = 46.30 years, SD_{age} = 13.29; $M_{\text{hiring experience}}$ = 10.71 years, $SD_{\text{hiring experience}}$ = 9.44); 161 out of 200 hiring managers (80.50%) revised the feedback. To encourage high-quality feedback in the Human-only and HITL conditions, both monetary and nonmonetary incentives were used. For the monetary incentive, we informed the hiring managers that an independent group of hiring managers would evaluate their feedback and that feedback rated in the top 10% would earn a $1 bonus. For the nonmonetary incentive, we highlighted the opportunity to help a fellow CloudResearch participant with their job search process.

At Time 2, 498 job seekers returned to view their cover letter and feedback with the source disclosed: AI-only (N = 168) vs. HITL (N = 164) vs. Human-only (N = 166). To incentivize participants to pay close attention to the feedback, we gave them a $0.50 bonus if they accurately estimated their cover letter grade based on the feedback. Participants then rated the feedback source on trust (6 items, 7-point Likert scale, 1 = strongly disagree, 7 = strongly agree) (*44, 45*). Depending on condition, feedback source was replaced with "the AI", "the AI and the hiring manager", or "the hiring manager." Items included: "I would be comfortable giving [feedback

source] an important task that I cannot fully monitor," "I would rely on [feedback source] to carry out a critical assignment," "I would let [feedback source] influence decisions that are important to me," "I would share information with [feedback source] that could make me vulnerable," "If someone questioned [feedback source]'s motives, I would give [feedback source] the benefit of the doubt," "I would not feel the need to closely watch [feedback source] to ensure work is done properly." Participants also rated the feedback source on the following items in a random order (7-point Likert scale: 1 = strongly disagree, 7 = strongly agree) (*13, 46*): "Competent", "Confident", "Intelligent", "Capable", "Skillful", "Warm", "Friendly", "Good-Natured", "Tolerant". We combined the first five items into an index of competence, the last four items into an index of warmth.

We then collected manipulation check questions (3 items; 7-point Likert scale, 1 = strongly disagree, 7 = strongly agree): "The feedback was created using ChatGPT," "The feedback was created by a person," and "The feedback was created using ChatGPT and then revised by a person as much as they wanted." We also collected demographic information (gender, age, race, and education level) and the extent to which participants were actively seeking jobs (0–10). We showed the actual cover letter grade at the end of the survey.

Third-Party Evaluations. After collecting participants' subjective ratings of the feedback source, we assessed the quality of the feedback as judged by evaluators blind to its source. To ensure robust evaluation, we obtained quality ratings from three independent sources that were blind to condition: human evaluators, ChatGPT, and NLP ratings.

We recruited a separate sample of hiring managers from CloudResearch to serve as human evaluators. We excluded seven individuals who reported zero years of hiring experience. We continued recruitment until each piece of feedback was rated by at least three independent hiring managers, resulting in a final sample of 387 human raters. Participants first reviewed the

evaluation rubric. Then, they viewed a cover letter with its corresponding feedback and rated the feedback provider on three dimensions using 7-point Likert scales (1 = strongly disagree, 7 = strongly agree): "The feedback provider is competent," "The feedback provider is warm," and "The feedback provider is trustworthy." They also rated the overall quality of the feedback on a 10-point scale (1 = extremely low quality, 10 = extremely high quality). Each hiring manager repeated this process until they had evaluated five distinct cover letter and feedback pairs.

For the ChatGPT evaluation, we used ChatGPT 5.4 to conduct three independent evaluations of each piece of feedback. The AI was prompted with the exact same instructions provided to the human evaluators and was kept blind to the experimental conditions (*40*). We averaged the three independent ratings to calculate the final scores for the feedback source's warmth (α = 0.98), competence (α = 0.99), trustworthiness (α = 0.99), and feedback quality (α = 1.00).

For NLP analysis, we utilized the centered ratings from the concreteness for advice NLP package (*41*) as an indicator of feedback competence. We applied the politeness NLP package (*42*) as an indicator of feedback warmth. Net positivity was calculated as the percentage of positive emotion features minus the percentage of negative emotion features in each feedback.

Experiment 2. We recruited 599 adults from Prolific to participate in a two-part experiment. A total of 523[2] participants completed both parts and have been included in our sample (40.54% men, M_{age} = 33.37 years, SD_{age} = 8.39).

At Time 1, we asked participants to spend 3 minutes writing about a time they demonstrated excellent leadership skills. We then randomly assigned participants to one of three between-subjects conditions: AI-only (N = 200) vs. HITL (N = 200) vs. Human-only (N = 199).

In the AI-only condition, we entered essays into ChatGPT 4.0 to generate a grade and feedback.

[2]Consistent with our preregistration, we analyzed the results of participants who completed the Time 2 survey within two weeks of it being launched, a period of April 2–16, 2024. Including participants who completed the study after this time period shows the same pattern of results and is reported in the Supplementary Materials.

In the Human-only condition, we recruited 602 participants from Prolific to review the essays and provide a grade and feedback. In the HITL condition, ChatGPT 4.0 first generated the grade and feedback and then 600 participants were recruited from Prolific to review and revise them. In both conditions, nearly every essay was evaluated by three independent evaluators (392 of 399 essays; the remainder received between two and six evaluations owing to uneven recruitment yield), and one evaluation per essay was selected and shown to the essay writer at Time 2. In the Human-only condition, we selected the median-length feedback as the final feedback. We informed evaluators in the HITL condition that they could either directly approve the feedback from ChatGPT or modify it, with the latter requiring two additional minutes of revision. Overall, 71.33% of evaluator decisions approved the AI feedback without modification[3]: of the 200 essays in the HITL condition, 77 were revised by no evaluator, 82 by one, 33 by two, and 8 by all three. Whenever at least one evaluator had approved the AI feedback unchanged, we presented the original AI text at Time 2; as a result, 171 of the 179 HITL recipients in the final sample received feedback that was textually identical to the AI draft. For the 8 essays in which every evaluator revised, one revision was randomly selected and shown.

At Time 2, 523 essay writers returned to view their essay and feedback with the source disclosed: AI-only (N = 178) vs. HITL (N = 179) vs. Human-only (N = 166). To incentivize participants to pay close attention to the feedback, we gave them a $0.50 bonus if they accurately estimated their essay grade based on the feedback. Participants then rated the feedback source on the following items in a random order (7-point Likert scale: 1 = strongly disagree, 7 = strongly agree) (*13, 46*): "Competent", "Confident", "Intelligent", "Capable", "Skillful", "Warm", "Friendly", "Good-Natured", "Tolerant", "Sincere", "Trustworthy", "Honest". We combined the first five items into an index of competence, the middle four items into an index of warmth, and the last three

[3]We classify decisions by the option evaluators selected; 19 evaluators chose the revise option but submitted text identical to the AI draft.

items into an index of trust. We showed the actual essay grade at the end of the survey before collecting demographic variables. Neither the predicted nor actual grades significantly influenced the results in any of our studies (see Supplementary Materials).

Third-Party Evaluations. After collecting participants' subjective ratings of the feedback source, we obtained quality ratings from three independent sources that were blind to condition. These analyses were not preregistered, but to ensure that they were robust, we obtained ratings from human evaluators, ChatGPT, and NLP.

Research assistants rated the feedback source on 7-point Likert scales (1 = strongly disagree, 7 = strongly agree) across three dimensions: warmth (α = 0.65), competence (α = 0.77), and trustworthiness (α = 0.72). For ChatGPT analysis, ChatGPT 5.4 conducted three independent evaluations without knowledge of experimental conditions (*40*), averaging the results to generate ratings for feedback sources on 7-point Likert scales (1 = "strongly disagree", 7 = "strongly agree") across three dimensions: warmth (α = 0.99), competence (α = 0.98), and trustworthiness (α = 0.97). For NLP analysis, we utilized the centered ratings from the concreteness for advice NLP package (*41*) as an indicator of feedback competence. We applied the politeness NLP package (*42*) as an indicator of feedback warmth. Net positivity was calculated as the percentage of positive emotion features minus the percentage of negative emotion features in each feedback.

Experiment 3. We recruited 602 adults from Prolific to participate in a two-part experiment. A total of 491[4] participants completed both parts of the study and have been included in our sample (40.12% men, M_{age} = 34.26 years, SD_{age} = 8.29).

At Time 1, we asked participants to spend 3 minutes writing about a time they demonstrated excellent leadership skills. We then randomly assigned the essays to one of three between-

[4]We analyzed data from the preregistered two-week period of May 7–21, 2024. A complete analysis of all data, which showed the same pattern, is reported in the Supplementary Materials.

subjects conditions: AI-only (N = 164) vs. HITL (N = 162) vs. Human-only (N = 165). At Time 2, we informed participants that their feedback was generated by AI, a human, or AI and a human. However, the feedback across all conditions was generated by ChatGPT 4.0. This design allows us to isolate the impact of perceived feedback source on participant perceptions.

Experiment 4. We recruited 901 adults from CloudResearch to participate in a scenario study (41.73% men; M_{age} = 40.71 years, SD_{age} = 12.87; 72.70% White/European, 11.54% Black/African/African diaspora, 11.21% Asian; 55.83% held a bachelor's degree or higher). Participants were randomly assigned to one of six between-subjects conditions: AI-only negative feedback (N = 150) vs. HITL negative feedback (N = 151) vs. Human-only negative feedback (N = 151) vs. AI-only positive feedback (N = 148) vs. HITL positive feedback (N = 150) vs. Human-only positive feedback (N = 151).

We asked participants to imagine that they had submitted a cover letter to receive feedback. They were shown a sample cover letter adapted from Experiment 1 and told that they received an overall grade of "Acceptable" (6 out of 9 points). They then viewed feedback corresponding to their assigned condition. Feedback was generated using the same evaluation criteria as in Experiment 1, with five positive and five negative feedback messages generated via ChatGPT 5.2 (see Supplementary Materials Section D). Each participant was randomly assigned one message from their corresponding tone condition. Participants then rated the feedback source on trustworthiness, competence, and warmth using the same measures as in Experiment 1.

References and Notes

1. P. C. Earley, G. B. Northcraft, C. Lee, T. R. Lituchy, Impact of process and outcome feedback on the relation of goal setting to task performance. *Acad. Manag. J.* **33**, 87–105 (1990).

2. J. S. Goodman, R. E. Wood, Feedback specificity, learning opportunities, and learning. *J. Appl. Psychol.* **89**, 809–821 (2004).

3. Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. El Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073 (2022).

4. Gemini Team, Gemini: A family of highly capable multimodal models. arXiv:2312.11805 (2023).

5. M. Binz, S. Alaniz, A. Roskies, B. Aczel, C. T. Bergstrom, C. Allen, D. Schad, D. Wulff, J. D. West, Q. Zhang, R. M. Shiffrin, S. J. Gershman, V. Popov, E. M. Bender, M. Marelli, M. M. Botvinick, Z. Akata, E. Schulz, How should the advancement of large language models affect the practice of science? *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401227121 (2025).

6. J. Laux, S. Wachter, B. Mittelstadt, Three pathways for standardisation and ethical disclosure by default under the European Union Artificial Intelligence Act. *Comput. Law Secur. Rev.* **53**, 105957 (2024).

7. A. Nguyen, H. N. Ngo, Y. Hong, B. Dang, B. P. T. Nguyen, Ethical principles for artificial intelligence in education. *Educ. Inf. Technol.* **28**, 4221–4241 (2023).

8. M. Haupt, J. Freidank, A. Haas, Consumer responses to human-AI collaboration at organiza-

tional frontlines: Strategies to escape algorithm aversion in content creation. *Rev. Manag. Sci.* **19**, 377–413 (2025).

9. C. Wittenberg, Z. Epstein, A. J. Berinsky, D. G. Rand, Labeling AI-generated content: Promises, perils, and future directions. *MIT Exploration of Generative AI* (2024). https://doi.org/10.21428/e4baedd9.0319e3a6. Accessed 19 June 2025.

10. Illinois General Assembly, Artificial Intelligence Video Interview Act. 820 ILCS 42 (2023). https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68. Accessed 22 June 2026.

11. Studio Legale Stefanelli & Stefanelli, Should you disclose that you used an AI system to generate content? Transparency obligations for deployers (2023). https://www.lexology.com/library/detail.aspx?g=d1805820-6993-43ab-8d86-0130c4fd4c1d. Accessed 22 June 2026.

12. European Union, Recital 134, Artificial Intelligence Act. Regulation (EU) 2024/1689 (2024). https://artificialintelligenceact.eu/recital/134/. Accessed 22 June 2026.

13. S. T. Fiske, A. J. Cuddy, P. Glick, J. Xu, A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J. Pers. Soc. Psychol.* **82**, 878–902 (2002).

14. S. T. Fiske, A. J. Cuddy, P. Glick, Universal dimensions of social cognition: Warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).

15. M. Puerta-Beldarrain, O. Gómez-Carmona, R. Sánchez-Corcuera, D. Casado-Mansilla, D. López-de-Ipiña, L. Chen, A multifaceted vision of the human-AI collaboration: A comprehensive review. *IEEE Access* **13**, 29375–29405 (2025).

16. E. Glikson, A. W. Woolley, Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **14**, 627–660 (2020).

17. E. Casini, N. Suri, J. M. Bradshaw, "Leveraging human oversight and intervention in large-scale parallel processing of open-source data" in *Next-Generation Analyst III* (SPIE, 2015), pp. 145–150.

18. M. Li, T. B. Bitterly, How perceived lack of benevolence harms trust of artificial intelligence management. *J. Appl. Psychol.* **109**, 1794–1816 (2024).

19. R. Gao, M. Saar-Tsechansky, M. De-Arteaga, L. Han, M. K. Lee, M. Lease, "Human-AI collaboration with bandit feedback" in *Proceedings of the International Joint Conference on Artificial Intelligence* (International Joint Conferences on Artificial Intelligence Organization, Montreal, 2021), pp. 1722–1728.

20. D. L. Marino, "AI augmentation for trustworthy AI: Augmented robot teleoperation" in *Proceedings of the 13th International Conference on Human System Interaction* (IEEE, 2020), pp. 155–161.

21. R. Poli, "Super-human and super-AI cognitive augmentation of human and human-AI teams assisted by brain computer interfaces" in *Proceedings of the Genetic and Evolutionary Computation Conference* (ACM, Lisbon, 2023), p. 3.

22. M. J. Cratsley, N. J. Fast, "Inventor's bias" at work: When low-performing algorithms seem fair. *Int. J. Hum. Comput. Interact.* **40**, 24–32 (2024).

23. R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust. *Acad. Manag. Rev.* **20**, 709–734 (1995).

24. R. S. Peterson, K. J. Behfar, The dynamic relationship between performance feedback,

trust, and conflict in groups: A longitudinal study. *Organ. Behav. Hum. Decis. Process.* **92**, 102–112 (2003).

25. C. T. Smith, J. De Houwer, B. A. Nosek, Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Pers. Soc. Psychol. Bull.* **39**, 193–205 (2013).

26. R. J. Lewicki, E. C. Tomlinson, N. Gillespie, Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *J. Manage.* **32**, 991–1022 (2006).

27. A. K. Schnackenberg, E. C. Tomlinson, Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *J. Manage.* **42**, 1784–1810 (2016).

28. E. C. Tomlinson, A. Schnackenberg, The effects of transparency perceptions on trustworthiness perceptions and trust. *J. Trust Res.* **12**, 1–23 (2022).

29. G. Loewenstein, C. R. Sunstein, R. Golman, Disclosure: Psychology changes everything. *Annu. Rev. Econ.* **6**, 391–419 (2014).

30. S. Sah, G. Loewenstein, D. M. Cain, The burden of disclosure: Increased compliance with distrusted advice. *J. Pers. Soc. Psychol.* **104**, 289–304 (2013).

31. T. Kim, K. Barasz, L. K. John, Why am I seeing this ad? The effect of ad transparency on ad effectiveness. *J. Consum. Res.* **45**, 906–932 (2019).

32. O. Schilke, M. Reimann, The transparency dilemma: How AI disclosure erodes trust. *Organ. Behav. Hum. Decis. Process.* **188**, 104405 (2025).

33. Y. Yin, N. Jia, C. J. Wakslak, AI can help people feel heard, but an AI label diminishes this impact. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2319112121 (2024).

34. J. A. Reif, R. P. Larrick, J. B. Soll, Evidence of a social evaluation penalty for using AI. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2426766122 (2025).

35. S. Tong, N. Jia, X. Luo, Z. Fang, The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strateg. Manag. J.* **42**, 1600–1631 (2021).

36. A. Barasch, E. E. Levine, J. Z. Berman, D. A. Small, Selfish or selfless? On the signal value of emotion in altruistic behavior. *J. Pers. Soc. Psychol.* **107**, 393–413 (2014).

37. H. M. Gray, K. Gray, D. M. Wegner, Dimensions of mind perception. *Science* **315**, 619 (2007).

38. K. Gray, D. M. Wegner, Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* **125**, 125–130 (2012).

39. P. Rozin, E. B. Royzman, Negativity bias, negativity dominance, and contagion. *Pers. Soc. Psychol. Rev.* **5**, 296–320 (2001).

40. S. Rathje, D. Mirea, I. Sucholutsky, R. Marjieh, C. E. Robertson, J. J. Van Bavel, GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2308950121 (2024).

41. M. Yeomans, A concrete example of construct construction in natural language. *Organ. Behav. Hum. Decis. Process.* **162**, 81–94 (2021).

42. M. Yeomans, A. Kantor, D. Tingley, The politeness package: Detecting politeness in natural language. *R J.* **10**, 489–502 (2018).

43. D. Lakens, Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.* **8**, 355–362 (2017).

44. R. C. Mayer, M. B. Gavin, Trust in management and performance: Who minds the shop while the employees watch the boss? *Acad. Manag. J.* **48**, 874–888 (2005).

45. F. D. Schoorman, R. C. Mayer, J. H. Davis, An integrative model of organizational trust: Past, present, and future. *Acad. Manag. Rev.* **32**, 344–354 (2007).

46. T. B. Bitterly, M. E. Schweitzer, The economic and interpersonal consequences of deflecting direct questions. *J. Pers. Soc. Psychol.* **118**, 945–990 (2020).

47. C. M. Judd, J. Westfall, D. A. Kenny, Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *J. Pers. Soc. Psychol.* **103**, 54–69 (2012).

48. P. E. Shrout, J. L. Fleiss, Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428 (1979).

Acknowledgments: We thank Xinjie Huang, Wenqi Jiang, Yingyu Chen, and Yung-yi Tseng for research assistance. We thank Yang Lu and George Loewenstein for friendly review and constructive feedback. We gratefully acknowledge research funding from the Center of Education Innovation (CEI) of the Hong Kong University of Science and Technology.

Funding: The Hong Kong University of Science and Technology Center for Education Innovation (CEI) EDGE grant EDGE02B-23S.

Author contributions: Conceptualization: ML, TBB, DH

Methodology: ML, TBB, DH

Investigation: ML, TBB, DH

Visualization: ML, TBB, DH

Funding acquisition: TBB, DH

Project administration: TBB, DH

Supervision: TBB, DH

Writing – original draft: ML, TBB, DH

Writing – review & editing: ML, TBB, DH

Competing interests: Authors declare that they have no competing interests.

Data, code, and materials availability: We provide CSV files of the raw anonymized data on the Open Science Framework (OSF). We also provide cleaned data and syntax so that our analyses can be replicated in Stata, R, or Python. https://doi.org/10.17605/OSF.IO/D7MA9.

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S201

Tables S1 to S49

References (*40–42, 47, 48*)

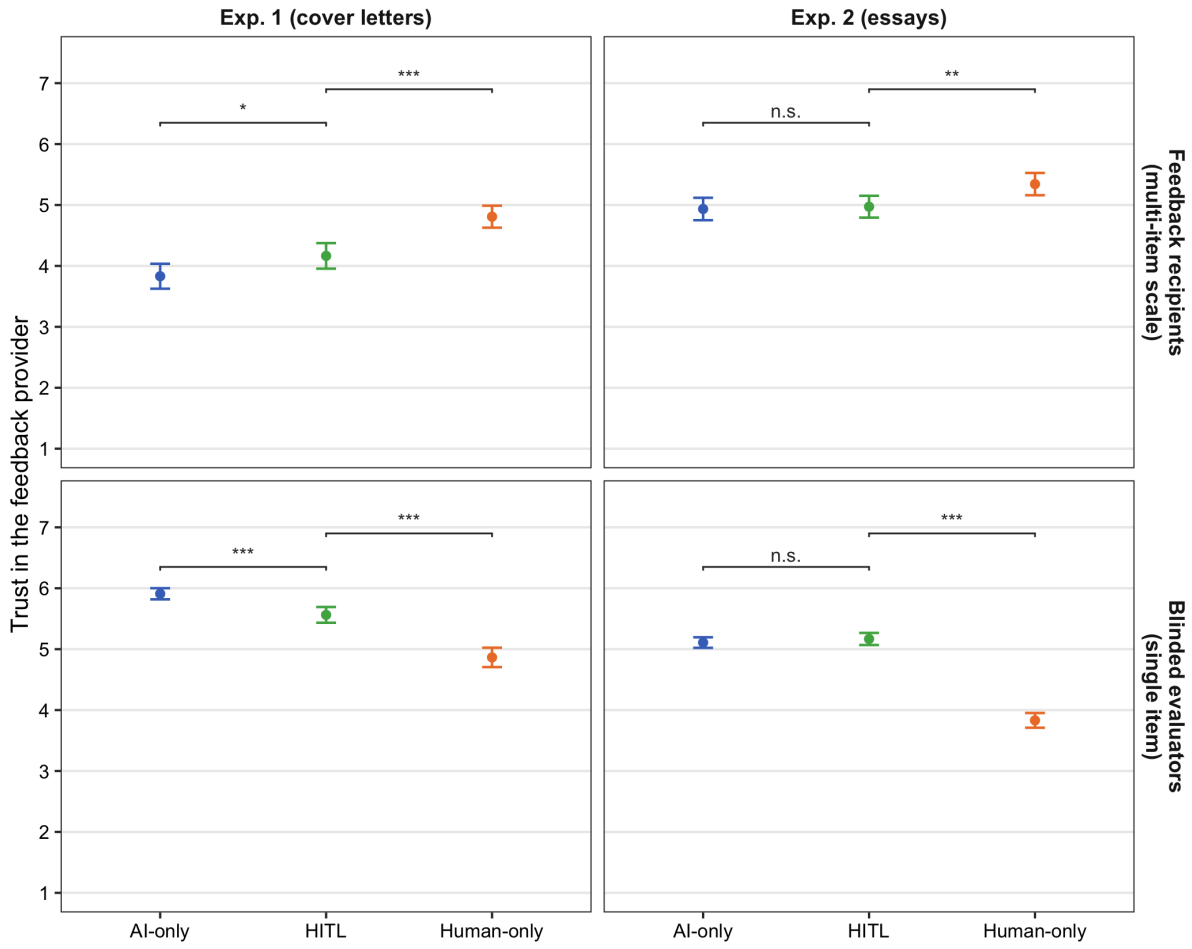

Fig. 1. Feedback recipients and evaluators blind to the feedback source diverge in their trust of feedback providers. Points are condition means with 95% confidence intervals; the y-axis spans the full 7-point response scale. Top row: trust ratings of the feedback provider by the feedback recipients, who were informed of the source and answered multi-item trust measures (Exp. 1: a six-item behavioral-intentions scale; Exp. 2: a three-item adjective index). Bottom row: trust ratings of the same providers by evaluators blind to the feedback source (Exp. 1: independent hiring managers; Exp. 2: independent research assistants), who rated the single item "The feedback provider is trustworthy." Because the measures and rater populations differ between rows, comparisons are informative within panels (across conditions) rather than across rows (levels). Brackets show the unadjusted pairwise comparisons of the HITL condition against the Human-only and AI-only conditions (n.s. $p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$).

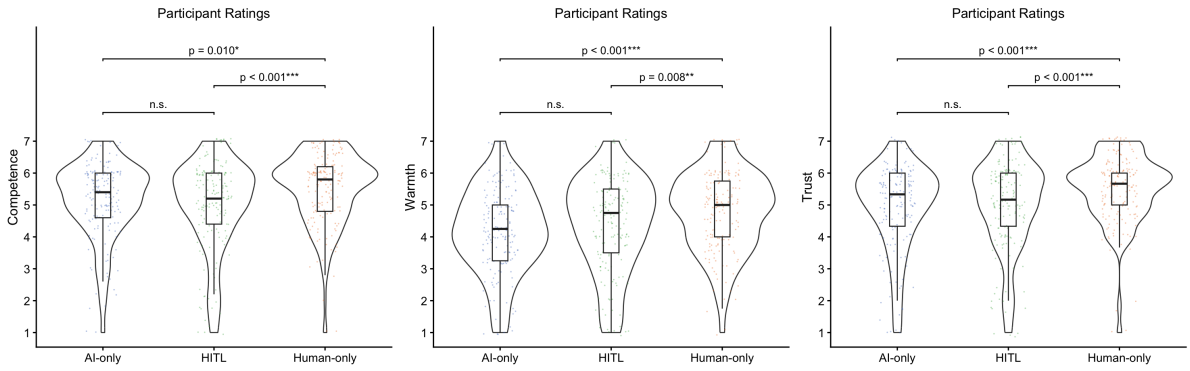

Fig. 2. Ratings of the feedback provider in Exp. 3. *Note.* HITL refers to Human-in-the-loop. n.s. indicates not significant.

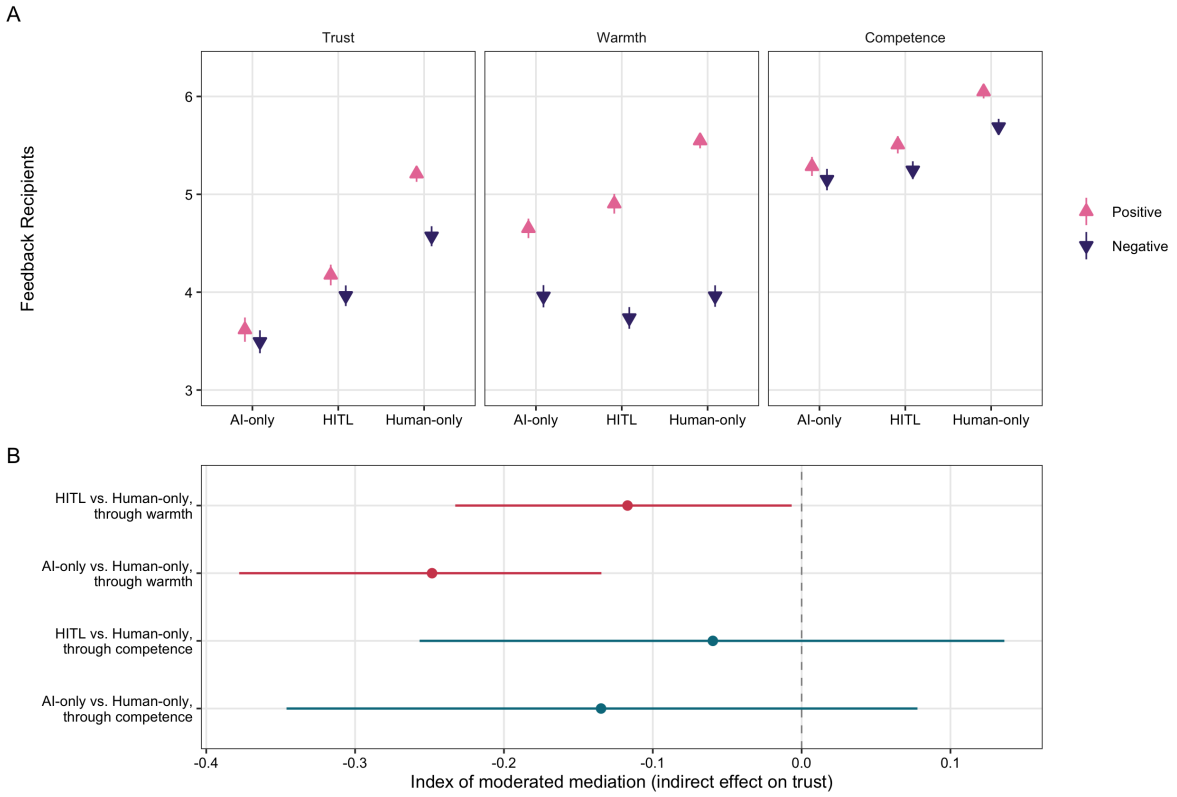

Fig. 3. Feedback tone, source, and the conversion of warmth into trust in Exp. 4. (**A**) Trust, warmth, and competence ratings of the feedback provider by source and tone (N = 901). Points are cell means; error bars represent ± 1 standard error of the mean. (**B**) Indices of moderated mediation from the parallel moderated-mediation model, i.e., the difference in the indirect effect of feedback source on trust, through warmth and through competence, between positive and negative feedback (relative to the Human-only condition). Points are coefficient products; error bars are 95% bootstrap percentile confidence intervals (10,000 resamples). The warmth indices exclude zero; the competence indices do not.

Table 1. Summary of AI Prompts Across Experiments 1–4 and S1

	Task	AI Model	Prompt
Experiment 1	Remove Personally Identifiable Information	GPT-5.2	Exp.1_gpt5.2_removePII.py
	Feedback Generation	GPT-5.2	Exp.1_gpt5.2_feedback.py
	Text Analysis	GPT-5.4	Exp.1_gpt5.4_text analysis.py
Experiment 2	Feedback Generation	GPT-4.0	Exp.2_gpt4_feedback.ipynb
	Text Analysis	GPT-5.4	Exp.2_gpt5.4_text analysis.py
Experiment 3	Feedback Generation	GPT-4.0	Exp.3_gpt4_feedback.ipynb
Experiment 4	Positive Feedback Generation	GPT-5.2	Exp.4_gpt5.2_positive.py
	Negative Feedback Generation	GPT-5.2	Exp.4_gpt5.2_negative.py
Experiment S1	Feedback Generation	GPT-4.0	Exp.S1_gpt4_feedback.ipynb
	Text Analysis	GPT-5.4	Exp.S1_gpt5.4_text analysis.py

Note. All scripts and prompts listed above are available at the following OSF link: https://osf.io/d7ma9.