

Exploring Trust in Human-LLM Feedback Systems: Observation of Student Behaviour in Software Engineering Education

Yiwen Liao
University of New South Wales
Sydney, NSW, Australia
yiwen.liao@student.unsw.edu.au

Yuchao Jiang
University of New South Wales
Sydney, NSW, Australia
yuchao.jiang@unsw.edu.au

Madhushi Bandara
University of Technology Sydney
Sydney, NSW, Australia
Madhushi.Bandara@uts.edu.au

Yuekang Li
University of New South Wales
Sydney, NSW, Australia
yuekang.li@unsw.edu.au

Iromie Samarasekara
University of New South Wales
Sydney, NSW, Australia
i.samarasekara@unsw.edu.au

Zixiu Guo
University of New South Wales
Sydney, NSW, Australia
z.guo@unsw.edu.au

Drishtant Leuva
University of Technology Sydney
Sydney, NSW, Australia
drishtant.a.leuva@student.uts.edu.au

George Joukhadar
University of New South Wales
Sydney, NSW, Australia
g.joukhadar@unsw.edu.au

Abstract

AI systems can deliver scalable feedback in large courses and often achieve high consistency with human grading. Yet, students continue to view human feedback as more credible, actionable, and trustworthy, motivating growing interest in hybrid approaches that combine AI and human input. Despite advances in this area, three key challenges remain: (1) understanding when and why students escalate from AI to human support, (2) identifying how design interventions such as transparency, explainability, and endorsement affect trust and uptake, and (3) evaluating whether hybrid AI+human feedback models improve learning outcomes compared to AI-only or human-only systems. This study uses controlled experiments and structured interviews to investigate these questions. The results aim to guide the responsible design of hybrid feedback systems that align scalability with pedagogical credibility.

ACM Reference Format:

Yiwen Liao, Yuchao Jiang, Madhushi Bandara, Yuekang Li, Iromie Samarasekara, Zixiu Guo, Drishtant Leuva, and George Joukhadar. 2026. Exploring Trust in Human-LLM Feedback Systems: Observation of Student Behaviour in Software Engineering Education. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.2 (SIGCSE TS 2026)*, February 18–21, 2026, St. Louis, MO, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3770761.3777342>

1 Introduction and Motivation

Feedback is central to student learning [6], yet timely and personalized responses remain difficult to provide at scale. Recent advances in large language models (LLMs) enable automated feedback that can improve equity and consistency [2], but students still view human feedback as more credible, actionable, and fair [1]. Trust in

AI also declines once its source is revealed [4], motivating interest in hybrid human-LLM systems that combine scalability with pedagogical credibility.

Early automated feedback tools addressed surface-level issues [5] but often lacked the contextual depth needed for trustworthy guidance [3].

Specifically, despite advances in AI feedback research, three key gaps remain for building reliable hybrid human-LLM systems. First, little is known about *when and why* students escalate from AI to human feedback, as most work compares overall perceptions rather than specific decision points (RQ1). Second, while interventions such as explainability, endorsement, and transparency are often proposed to build trust, few studies directly measure their effects on trust and uptake (RQ2). Third, although hybrid AI+human models are widely recommended, there is limited empirical comparison with AI-only or human-only systems to show when and why hybrid approaches are more effective (RQ3).

This study addresses these gaps by observing student behaviour in a hybrid feedback environment in software engineering education, examining escalation decisions, the influence of design features on trust and uptake, and how hybrid workflows may support learning. The findings aim to guide the design of scalable yet credible human-LLM feedback systems.

2 Methods

2.1 Prototype System

We developed a hybrid feedback platform in which students first receive LLM-generated feedback on a short programming or writing explanation task. At this stage, they may choose to *escalate* to a human tutor for additional feedback. This design enables analysis of both behavioural decision points (when escalation occurs) and attitudinal responses (how trust and credibility are formed).

2.2 Participants and Setting

The study will involve 50 undergraduate software engineering students in a supervised laboratory environment. Each session will



This work is licensed under a Creative Commons Attribution 4.0 International License. *SIGCSE TS 2026, St. Louis, MO, USA*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2255-4/2026/02
<https://doi.org/10.1145/3770761.3777342>

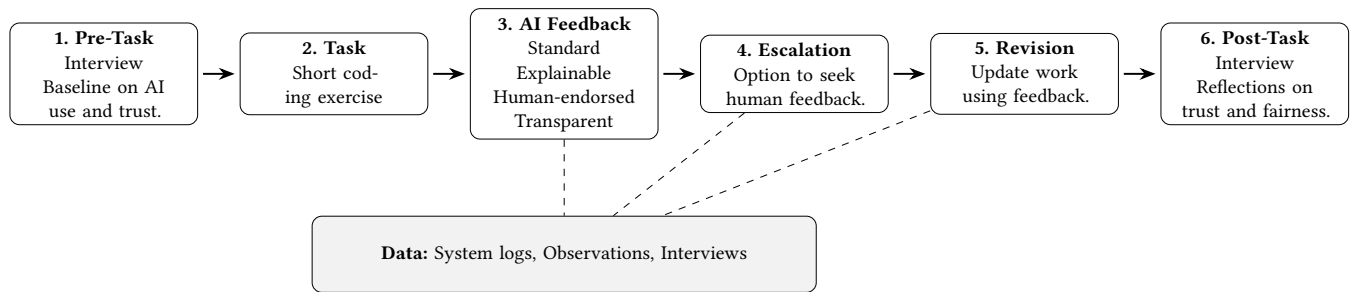


Figure 1: Experiment flow and data sources.

last approximately 30 minutes and follow standardized instructions. A researcher will be present to observe behaviour and record field notes.

2.3 Procedure

As shown in Figure 1, each session followed a structured sequence capturing both behavioural and perceptual aspects of hybrid feedback use. Participants began with a short pre-task interview, then completed a time-limited programming task under controlled conditions.

After submission, they received LLM feedback under one of four conditions—standard, explainable [3], human-endorsed [2], or transparent versus blind source [3, 4]. These conditions vary in whether the feedback includes a rationale, a human-endorsement message, or a disclosure that it is AI-generated, allowing us to isolate the effects of explainability, endorsement, and transparency. Participants could then choose to escalate to a human tutor, providing a behavioural indicator of reliance and trust [1]. Each session ended with a brief post-task interview on usefulness, fairness, and confidence.

2.4 Data Collection

Data were collected from system logs, researcher observations, and structured interviews to capture engagement and trust formation. System logs recorded timestamps, feedback views, and revisions [2, 5]. Researcher observations noted escalation decisions and visible cues such as hesitation, frustration, or confidence, offering contextual insight into interaction patterns [6]. Pre- and post-task interviews elicited reflections on accuracy, fairness, and usefulness of AI versus human feedback. This triangulated approach aligns with emerging recommendations that effective evaluation of AI feedback systems should integrate both behavioural and perceptual evidence, combining objective interaction data with subjective experience measures [1, 3].

3 Planned Analysis and Expected Results

We will analyze qualitative data from system logs, researcher observations, and interviews to examine engagement and trust formation within the hybrid feedback system. Key areas of analysis include:

- **Escalation patterns:** Interpreting escalation as a behavioural signal of trust, confidence, and perceived sufficiency of AI feedback, examining when students escalate and why across conditions.

- **Impact of design interventions:** Evaluating whether features such as explainability, human endorsement, or transparency increase trust and uptake of AI feedback.
- **Learning behaviours:** Assessing how feedback types affect revision quality, learner confidence, and fairness, with quality rated using a brief rubric and confidence captured through a short rating scale and interview responses.

Findings are expected to guide the design of hybrid feedback systems that balance AI scalability with human credibility, supporting responsible integration of LLMs in education.

4 Conclusion and Contributions

Our study introduces a controlled experimental design to examine how transparency, explainability, and human endorsement shape trust, escalation behaviour, and learning outcomes in hybrid human-LLM feedback systems. It lays the groundwork for evidence-based design of trustworthy AI feedback tools.

Contributions:

- **Framework:** A replicable experiment design combining system logs, observations, and interviews to capture behavioural and perceptual aspects of hybrid feedback use.
- **Theoretical groundwork:** A critical review highlighting the unresolved gap in empirical evidence on the influence of design interventions on trust and escalation in hybrid AI-human feedback systems.
- **Design Implications:** Empirically informed guidelines for integrating AI feedback in ways that balance scalability with human credibility in educational contexts.

References

- [1] Erkan Er, Gökhan Akçapınar, Alper Bayazıt, Omid Noroozi, and Seyyed Kazem Banihashem. 2025. Assessing student perceptions and use of instructor versus AI-generated feedback. *British Journal of Educational Technology* 56 (2025), 1074–1091. doi:10.1111/bjet.13558
- [2] Alwyn Vwen Yen Lee. 2023. Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation. *Studies in Educational Evaluation* 77 (2023), 101250. doi:10.1016/j.stueduc.2023.101250
- [3] E. D. Lindsay, M. Zhang, A. Johri, and J. Bjerva. 2025. The Responsible Development of Automated Student Feedback with Generative AI. <https://arxiv.org/abs/2308.15334>
- [4] Tanya Nazaretsky, Paola Mejia-Domenzain, Vinitra Swamy, Jibril Frej, and Tanja Käser. 2024. AI or Human? Evaluating Student Feedback Perceptions in Higher Education. In *Proceedings of EC-TEL 2024 (Lecture Notes in Computer Science, Vol. 15159)*. doi:10.1007/978-3-031-72315-5_20
- [5] H. Shi and V. Aryadoust. 2024. A systematic review of AI-based automated written feedback research. *ReCALL* 36, 2 (2024), 187–209. doi:10.1017/S0958344023000265
- [6] Valerie Shute. 2008. Focus on Formative Feedback. *Review of Educational Research* 78 (03 2008), 153–189. doi:10.3102/0034654307313795