

Reliability of Generative AI Based Reflection Chatbot Using the 5Rs Model

Akiko KAI^{a*}, Takuto WADA^b & Katsuaki SUZUKI^c

^a*Aoyama Gakuin University, Japan*

^b*Towers Quest Inc., Japan*

^c*Musashino University, Japan*

*a-kai@aim.aoyama.ac.jp

Abstract: This study examines the reliability of a generative AI based reflection chatbot using the 5Rs model with dialogue logs from student employees. Results show moderate agreement but strong consistency in relative ordering. Discrepancies reflect interpretive differences, suggesting that reflection-level estimates function as internal signals for structuring dialogue and that low-cost models are sufficient in this purpose.

Keywords: Generative AI, Reflection Support, 5Rs Model, Inter-rater Reliability

1. Introduction

Reflection plays a key role in transforming experience into learning, yet providing timely qualitative feedback remains challenging in resource-limited settings. It is conceptualized as purposeful mental processing (Moon, 1999; 2004) and structured through the 5Rs model for reflection (1: Reporting, 2: Responding, 3: Relating, 4: Reasoning, and 5: Reconstructing; Bain *et al.*, 2002), describing progression from descriptive reporting to deeper reflective thinking (Ryan & Ryan, 2013). Such progression requires dialogic support, but sustaining it is difficult in practice. Recent work has explored chatbot systems supporting structured reflection through guided dialogue (e.g., Ammersdörfer & Inkermann, 2025), highlighting the potential of generative AI. However, a key challenge lies in balancing estimation reliability and operational sustainability, given limited evidence on low-cost model reliability.

This study examines the reliability of AI-based reflection-level estimation in a workplace-based setting and addresses the following research question: To what extent are reflection ratings consistent across different generative AI models?

2. Method

2.1 Research Design and Context

This study examines the consistency of AI-based reflection-level estimates across AI models. The analysis is based on dialogue logs from 25 student employees working in a university learning commons. Due to rotating shifts, communication was conducted asynchronously via Slack (<https://slack.com>), where reflection activities were integrated into routine work through a chatbot system. The data were collected through regular work rather than an experimental intervention, supporting the ecological validity of the study.

2.2 Reflection Support System

The reflection support system used in this study was developed by adapting an existing chatbot framework (Kai & Wada, 2025) and deployed as a Slack-based application. The system operates within students' regular communication environment, enabling reflection to be embedded in everyday work practices. In addition, supervisors can configure the structure and content of prompts for AI, allowing the dialogue to reflect specific pedagogical

intentions. A session is initiated when a student clicks a designated button within Slack, triggering a sequence of five core questions.

Within the chatbot, reflection-level is estimated for each student response and used solely as internal logic for generating follow-up prompts. These estimates are not disclosed to users or administrators and were not used for the analytical evaluation in this study. Based on this internal estimation, the system performs the following pedagogical steps to scaffold progressively deeper reflection:

1. 5Rs-based Estimation: Internal estimation of reflection depth level.
2. Validation: Empathetic responses to sustain engagement.
3. Scaffolding: Targeted questions to prompt deeper reflection.
4. Integration: Synthesis of feedback into a single, cohesive prompt.

Table 1 shows an example of a chatbot–student reflection session through this process. The dialogue progresses from reporting a task (Turns 1–2) to an affective response (3–4), reflection on process and improvement (5–6), and future-oriented application (7–9). This pattern suggests that the chatbot’s questioning naturally guides the student toward progressively deeper reflection, consistent with the 5Rs framework.

Table 1. *Example of a Chatbot–Student Reflection Session (Excerpt)*

	Message
1	Chatbot: What was something memorable in your work this week? It can be something small. Please tell me freely.
2	Student: I filmed the 27th video today. I just finished.
3	Chatbot: Great job on your 27th video. Your project is moving forward steadily. How do you feel now after finishing the filming?
4	Student: I’m happy I can move on to editing.
5	Chatbot: You feel happy that you can move to the editing stage. From the 27 videos so far, did you try anything new or notice anything this time?
6	Student: Before, scripting, filming, and editing were done by different people, so the process was not smooth. This time, I filmed while keeping editing in mind.
7	Chatbot: Great improvement. How will you use this idea next week?
8	Student: I want to share what I noticed with the whole team.
9	Chatbot: Sharing your ideas with the team will help everyone improve. How will you do this next week? Start small and try it. Let me know how it goes.

Note. Original dialogue in Japanese; translated into English.

2.3 Data and Analysis

Data were collected over a six-week period (December 2025–January 2026) from student employees engaged in weekly chatbot-based reflection as part of their regular duties. The dataset consists of dialogue logs, including message content and session identifiers. All personally identifiable information was removed prior to analysis.

A total of 85 reflection sessions were analyzed, with each session treated as a unit of analysis. Three generative AI models—ChatGPT 5.2 (Auto), Gemini 3 (Fast), and Claude (Haiku 4.5)—were used to estimate reflection-level to each session, independent of the chatbot system. All models were given identical prompts and applied a five-level reflection scheme aligned with the 5Rs framework. These levels represent ordered stages of reflection rather than equal-interval scores; however, they were treated as approximately interval-scaled for reliability analysis, following common practice in educational research.

Inter-rater reliability was assessed using ICC for agreement and consistency, along with Cronbach’s alpha and Spearman’s rank correlation. Model-generated rationales were also analyzed qualitatively to examine interpretive differences.

3. Results (Consistency Among AI Models)

Agreement among the three AI models was in the moderate to substantial range ($ICC(2,1) = .64$; $ICC(3,1) = .66$), with high internal consistency ($\alpha = .85$). Gemini produced the highest mean reflection-level estimate ($M = 4.00$, $SD = 0.93$), followed by ChatGPT ($M = 3.80$, $SD = 1.10$) and Haiku ($M = 3.62$, $SD = 1.12$) (Figure 1). Despite these differences, a strong rank-order correlation was observed between Haiku and ChatGPT ($r_s = .73$; Figure 2), indicating broadly consistent relative ordering across models.

Representative cases with large discrepancies (difference ≥ 2 in reflection-level estimates) were selected for qualitative comparison of model rationales (Table 2). Analysis of these cases indicated that disagreement often arose from differences in interpretive criteria, particularly in how reflective elements were interpreted. Some models emphasized explicit articulation (e.g., the presence of concrete plans), whereas others inferred higher-level reflection from implicit cues such as emerging insights or contextual understanding.

A qualitative review with reference to human judgments (first and third authors) was conducted to interpret these discrepancies. In Talk #273, the authors estimated it as level 5, identifying a complete cycle including reconstructing, whereas Haiku estimated it as level 2, emphasizing the absence of a concrete plan. Gemini, in contrast, interpreted the same response as higher-level reconstruction. In Talk #41, ChatGPT and Haiku assigned level 1 based on factual reporting, whereas the authors assigned level 2, recognizing an emerging problem identification. Gemini, however, estimated it a substantially higher level (4), suggesting a tendency to infer reflective depth from limited cues.

Across cases, Haiku tended to estimate lower reflection-level than the authors, but its relative ordering of cases was broadly consistent with human judgments. These patterns suggest that differences in interpretive criteria—particularly in recognizing implicit versus explicitly articulated elements of reflection—can lead to divergent estimations. It should be noted that judgments by the authors are presented as an interpretive reference rather than formal validation.

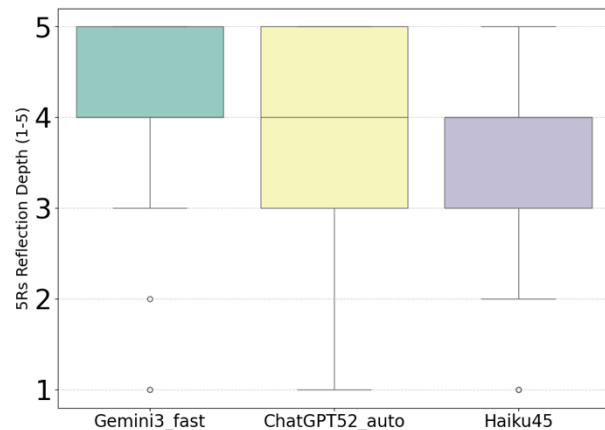


Figure 1. Level Distribution by Model

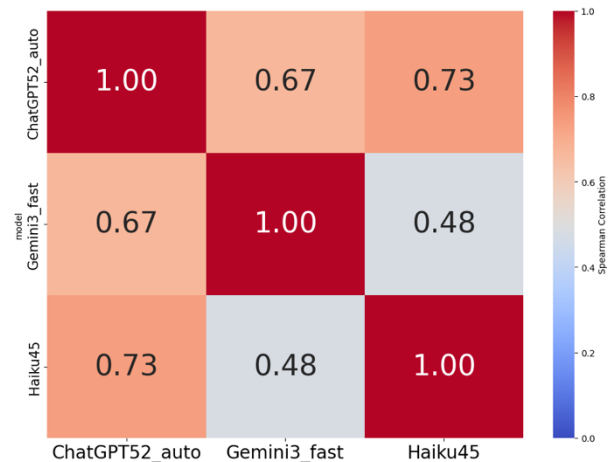


Figure 2. Correlation Between Models

Table 2. Comparison of Rationales for Reflection-Level Estimates in Discrepant Cases

Talk ID	C	G	Ha	Hu	Key Interpretations
41	1	4	1	2	factual reporting (C/Ha); implicit problem identification (G)
102	3	4	1	3	descriptive response (Ha); situational relating (C); problem-focused reasoning (G)
273	4	5	2	5	absence of a concrete plan (Ha); reasoning with solution (C); explicit action planning (G)

Note. C = ChatGPT, G = Gemini, Ha = Haiku, Hu = Human. Differences ≥ 2 levels.

4. Discussion and Conclusion

This study provides preliminary evidence that AI based reflection-level estimation is more useful for supporting reflective dialogue than as a measurement tool. Although consistency across models was moderate, strong agreement in relative ordering was observed despite variation in absolute estimates.

These discrepancies appear to reflect differences in interpretive criteria—particularly in how implicit versus explicit expressions of reflection are treated—rather than random measurement error. From a design perspective, this supports the use of reflection-level estimates as internal, non-disclosed signals to structure dialogue. Despite variation in absolute estimates, their relative consistency allows them to function as scaffolding for guiding reflective progression without imposing evaluative pressure on student employees.

Taken together, these findings suggest that low-cost models are sufficient to support reflection-oriented interaction in this context, because reflection-level estimation functions not as a precise evaluative output but as a resource for structuring dialogue in formative contexts. Accordingly, marginal gains in model accuracy may offer limited pedagogical benefit. Instead, effective reflection support depends on interaction design—for example, through structured questioning aligned with the 5Rs framework.

From a broader perspective, these findings point to a broader implication for instructional design: the effectiveness of AI-supported learning environments depends less on maximizing model accuracy and more on aligning system behavior with pedagogical goals. This highlights the importance of an instructional design perspective, in which methods are selected based on learning objectives and theoretical grounding rather than defaulting to more advanced technologies. In this sense, the value of generative AI lies not in achieving higher accuracy, but in enabling pedagogically grounded design at scale.

These implications should be interpreted with caution, as this study is based on a limited number of sessions in a specific context and does not examine whether repeated use leads to deeper reflection over time. Further validation is required to assess generalizability, longer-term use, and its impact on the reflective development of student employees.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 21K13646 and 25K06527.

References

- Ammersdörfer, T., & Inkermann, D. (2025). Supporting structured reflection in engineering design by chatbots: Potentials and concept for a reflection chatbot. *Proceedings of the Design Society*, 5, 2951–2960.
- Bain, J. D., Ballantyne, R., Mills, C., & Lester, N. C. (2002). Professional reflection: Deteriorating or developing?. Queensland University of Technology.
- Kai, A., & Wada, T. (2025). Development and Evaluation of a Generative AI-Integrated Chatbot System Tailored for Personalized Learner Support (in Japanese). *JSiSE Research Report*, 39(7), 9–16.
- Moon, J. A. (1999). Reflection in learning and professional development: Theory and practice. Kogan Page.
- Moon, J. A. (2004). A handbook of reflective and experiential learning: Theory and practice. Routledge.
- Ryan, M., & Ryan, M. (2013). Assessing a reflection scale for the 5Rs framework of reflective practice. *Assessment & Evaluation in Higher Education*, 38(2), 244-257.