Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

ALBe: An AI Literacy Behavioral Assessment Framework Based on EFL Chatbot Interaction Logs

Min Ma[a*], Chia-Yu Hsu[b], Hiroaki Ogata[b]
[a]*Graduate School of Informatics, Kyoto University*
[b]*Academic Center for Computer and Media Studies, Kyoto University,*
*ma.min.46u@st.kyoto-u.ac.jp

Abstract: The accelerating integration of artificial intelligence (AI) technology into education and its close connection with academic performance are leading to increasing awareness of supporting learners' AI literacy. However, the existing instruments lack evidence-based assessment of AI literacy. In this study, we present the AI Literacy Behavioral Framework (ALBe), a codebook instrument developed following established procedures in behavioral measurement of student communication with Tammy. After human labeling, elbow and silhouette were processed to confirm the K-means (k=3) to classify students into developing, emerging, and advanced-level groups based on three dimensions of AI literacy. These results suggest that ALBe offers a potential and valid method for assessing AI literacy.

Keywords: AI literacy, behavioral assessment, K12 education, learning chatbot, EFL

1. Introduction

The integration of Artificial Intelligence (AI) into education is expanding. Such as chatbots have a significant positive impact on English as a Foreign Language (EFL) context (Woollaston et al., 2024). Moreover, higher AI literacy has been found to be connected to increased engagement with AI tools for learning purposes, which, in turn, leads to better learning outcomes and improved academic performance (Singh et al., 2024). Therefore, the development of AI literacy becomes essential. However, the challenges may come from the not-yet-firmly-conceived concept of AI literacy. Generally, researchers have endeavored to define and conceptualize "AI literacy," with the definition by Long and Magerko (2020): "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace." This has prompted growing scholarly interest in understanding what it means for learners to be "AI literate" and how such literacy can be developed and supported through educational interventions (Jin et al., 2024). Thus, AI literacy instruments have become crucial for understanding and promoting AI literacy development.

Current AI literacy assessments fall into two types: self-reported and performance-based measures (Jin et al., 2024). Most existing instruments rely on self-reported surveys, which are effective in capturing perceived knowledge but may lack the reliability needed to accurately measure actual competencies (Lintner, 2024). On the other hand, performance-based instruments are context-sensitive due to the limited application (Lyu et al., 2024). To address this gap, we developed the AI Literacy Behavioral Assessment framework (ALBe), a performance-based codebook specifically designed to evaluate AI literacy within K12 EFL contexts. Specifically, this study responds to the following research question by leveraging the codebook and the chat logs retrieved from the interaction between learners and an LLM-based chatbot:

How are the three dimensions of AI literacy demonstrated by ALBe through interactional behaviors of K–12 students in the EFL context?

2. AI Literacy Behavioral Assessment framework (ALBe)

2.1 Codebook Developing

Ng et al. (2021) conceptualize AI literacy as four dimensions: Know & Understand, Use & Apply, Evaluate & Create, and Ethics. Yet, the Know & Understanding aspects were excluded due to their being hard to extract through interaction with the chatbot. The remaining three dimensions were each divided into two sub-dimensions, scored on a 0–4 scale.

In the Apply & Use dimension, Wang et al. (2023) emphasize that its concept is about how to apply AI concepts in different contexts and applications in everyday life. It can be explained as *interaction behavior* between students and chatbots in this study.

- **Frequency:** Total chat interactions (0=below median, 1=median, 2=above median)
- **Consistency:** Replies within an episode (0=no reply, 1=one task-related reply, 2=two+ task-related replies)

Following, Long and Magerko (2020) extended AI literacy to two other competencies (i.e., Create & Evaluate) that enabled individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI. Building on the definition, in the context of student–AI interaction, the present study concluded the Create & Evaluate dimension as *co-work behaviors* from collaboration and efficiency, to demonstrate that the student's AI literacy extends beyond conceptual understanding to the practical use of AI in authentic tasks.

- **Collaboration:** Reply relevance and task completion (0=no/irrelevant, 1=related but incorrect, 2=correct leading to completion)
- **Efficiency:** Number of successfully completed tasks (median-based threshold: 0/1/2)

The ethical dimension of AI literacy emphasizes intelligent social responsibility, which places higher demands on students' moral awareness, emotional disposition, and psychological readiness when engaging with AI (Wang et al., 2023). This perspective aligns with human-centered considerations in AI education, where learners are encouraged to continuously reflect on their learning progress and outcomes. We concluded that this is *reflective behavior*. It leads to both feedback and reflection to enhance their learning progress by crucial metacognition (Al-Hashim, 2019).

- **Student Feedback:** Task-related replies after chatbot suggestions (median-based: 0/1/2)
- **Reflective Reaction:** 0=none, 1=one behavior (suggestion-related reply, task revision, or inquiry), 2=suggestion plus revision/inquiry

3. Creating a Learner Profile of AI Literacy based on ALBe

3.1 Method

We adopted the above codebook and chat logs collected from a real K12 EFL context. The sample included two advanced English classes, totaling 45 students, from a co-ed public high school in Japan (CEFR A2-B2 level). In the context, students practice translation from Japanese to English, facilitated by the LLM-based chatbot, Tammy, which provides the following features: sentence selection, random sentence selection, language and level options, chat interface, user input, and example messages.

For the data processing, K-means clustering was applied with optimal k determined by the elbow method and silhouette analysis. The elbow method showed an inflection point at **k=3** (silhouette score = 0.54). Consequently, a three-cluster solution was selected based on the converging evidence from these two approaches.

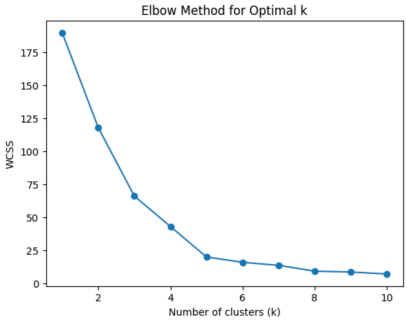

Figure 1. Elbow plot for k-means clustering

3.2 Results

Then, one-way ANOVA and Post-hoc comparisons were conducted to examine differences in each AI literacy aspect across the clusters. The results are shown in Table 1.

Table 1. Tukey HSD Post Hoc Comparisons Across Groups

Aspect	Clusters *i*	*j*	Mean Difference *j-i*	p (adjusted)
Apply & Use	2	0	-1.72	< .001
	2	1	-.46	< .001
	0	1	0.26	.716
Create & Evaluate	2	0	-2.81	< .001
	2	1	-0.12	.875
	0	1	2.69	< .001
Ethics	2	0	-2.80	<.001
	2	1	-2.80	< .001

Based on the above results, we labeled the three clusters as Developing (i.e., cluster 0), Emerging (i.e., cluster 1), and Advanced (i.e., cluster 2) groups. We further compared their academic performance with regular English exam scores. While the ANOVA test did not show significant differences (*F = 2.37, p = 0.11),* we identified three learner profiles of AI Literacy as follows (Figure 2). The Developing group (n = 22) showed minimal interactive and reflective behaviors and demonstrated the lowest and most variable English academic performance (M = 71.22, SD = 13.11). The Emerging group (n = 18) exhibited emerging collaborative behaviors and slightly higher but still variable achievement (M = 76.72, SD = 13.16). The Advanced group (n = 5), characterized by advanced and ethical engagement, achieved the highest and most consistent academic performance (M = 83.80, SD = 5.17).

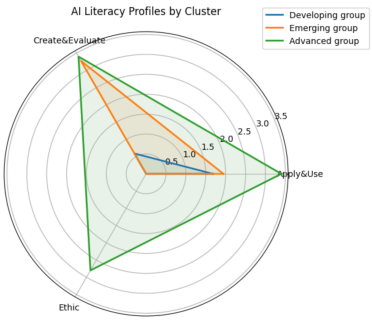

Figure 2. Radar plot of AI literacy aspects for the three identified clusters

4. Discussion and Conclusion

This study developed a codebook for assessing students' AI literacy based on their learning behavior. The finding identifies three distinct profiles of students' AI literacy. In the developing group, students show that they rarely engage with AI tools in sustained or meaningful ways. Their co-work actions are fragmented and unsystematic. Moreover, students at the developing level show almost no engagement with feedback or reflective interactions with Tammy. For the emerging group, in the interactive dimension and co-work dimension, students struggle to balance frequency and efficiency. These students also demonstrate minimal engagement with feedback or reflective actions. The advanced-level group reflects strong and consistent engagement across functional, higher-order, and ethical aspects of AI literacy. Students at this level demonstrate well-developed interactive and co-work behaviors, with clear evidence of reflection. They actively engage with feedback and reflect on their interactions with Tammy, indicating a mature and self-regulated approach to AI use. These findings suggest that ethical awareness and reflective engagement function as critical thresholds in the development of AI literacy, marking a shift from instrumental use toward more mature and self-regulated AI-supported learning.

Limitations are also identified as the limited sample size. Future research should therefore extend the framework's applicability. In addition, integrating data mining and visual analytics techniques can enable researchers to longitudinally track ALBe usage, allowing for a deeper exploration.

Acknowledgements

This work was supported by CSTI SIP Grant Number JPJ012347 and JSPS KAKENHI Grant Number 23H00505.

References

Al-Hashim, A. (2019). Critical thinking and reflective practice in the science education practicum in Kuwait. *Utopía y Praxis Latinoamericana, 24*(6), 85–96.

Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test. *Computers and Education: Artificial Intelligence*, Article 100436.

Lintner, T. (2024). A systematic review of AI literacy scales. *npj Science of Learning, 9*, Article 50.

Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-16).

Lyu, W., Wang, Y., Chung, T., Sun, Y., & Zhang, Y. (2024). Evaluating the effectiveness of LLMs in introductory computer science education: A semester-long field study. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 63–74). Association for Computing Machinery.

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Wang, X., Li, X., & Huang, J. (2023). Junior high school artificial intelligence literacy: Connotation, evaluation and promotion strategy. *Open Journal of Social Sciences, 11*(4), 33–49.

Singh, E., Vasishta, P., & Singla, A. (2024). AI-enhanced education: Exploring the impact of AI literacy on generation Z's academic performance in Northern India. *Quality Assurance in Education*. Advance online publication.

Woollaston, S., Flanagan, B., Ocheja, P., Dai, Y., & Ogata, H. (2024, November). TAMMY: Supporting EFL translation practice with an LLM-powered chatbot. In *International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education.