frontiers | Frontiers in Education

TYPE Brief Research Report
PUBLISHED 10 July 2025
DOI 10.3389/feduc.2025.1601789

Check for updates

OPEN ACCESS

EDITED BY
Davide Girardelli,
University of Gothenburg, Sweden

REVIEWED BY
Isaiah T. Awidi,
University of Southern Queensland, Australia
Kyung-Mi O,
Dongduk Women's University,
Republic of Korea

*CORRESPONDENCE
Lhea Reinhold
✉ lhea.reinhold@fau.de

RECEIVED 28 March 2025
ACCEPTED 09 June 2025
PUBLISHED 10 July 2025

CITATION
Reinhold L, Händel M and Naujoks-Schober N (2025) AI-teacher agreement in evaluating learning diaries. *Front. Educ.* 10:1601789. doi: 10.3389/feduc.2025.1601789

COPYRIGHT
© 2025 Reinhold, Händel and Naujoks-Schober. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

AI-teacher agreement in evaluating learning diaries

Lhea Reinhold[1]*, Marion Händel[2] and Nick Naujoks-Schober[2]

[1]Faculty of Humanities, Social Sciences, and Theology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberg, Germany, [2]Faculty of Media, Ansbach University of Applied Sciences, Ansbach, Germany

Learning diaries are reflective tools, often used as formative assessments in adult education with the aim to promote cognitive and metacognitive learning strategies. As grading of and feedback on learning diaries is effortful for teachers, artificial intelligence (AI) may assist teachers in evaluating learning diaries. A prerequisite is that AI's ratings show high accordance with the teachers' ratings. AI accuracy, measured via absolute accuracy and bias, is the focus of the current study with $N = 540$ learning diary entries focusing on learning strategies, seven teachers, and ChatGPT-4o. Findings revealed that AI evaluations align closely with teacher assessments, indicated by high overall accuracy and low bias. Interestingly, the accuracy varied based on the types of learning strategies assessed in the diaries. Additionally, individual teacher assessments influenced the alignment between human and AI evaluations, suggesting that teachers applied their profession-specific expertise to the assessment process while AI produced somewhat generic evaluations. Overall, the study results indicate that AI can enhance the efficiency of formative assessments while providing timely feedback to learners.

KEYWORDS

learning diary, AI assessment, AI accuracy, AI-teacher agreement, adult education

1 Introduction

The future of learning and assessment is expected to change significantly, with formative assessments gaining importance (Bürgermeister and Saalbach, 2018). Open assessment formats such as learning diaries aim to support learning and foster reflection on professional practices (Alt et al., 2022; Chang et al., 2016; Campbell et al., 1999; Trif and Popescu, 2013). However, their evaluation remains time-consuming. Integrating artificial intelligence (AI) into the evaluation process can provide meaningful support for teachers in the assessment process (Järvelä et al., 2025; Molenaar, 2022), thereby enabling immediate feedback to learners (Mao et al., 2024).

Against this background, the present study investigates the potential of AI as a teaching assistant in formative assessment settings. More specifically, it examines the accuracy and bias of AI-supported assessment in the context of reflective learning diaries in adult education. Drawing on a multiple-rater design, the study explores the extent to which (a) the type of learning strategy assessed and (b) the individual teacher influences the level of agreement between human and AI-based evaluations.

2 Theoretical concepts

2.1 Learning diaries

In an open learning diary, learners write about their own learning in a reflective way with the aim to deepen their knowledge and to apply learning strategies (Chang et al., 2016; Naujoks and Händel, 2020). By structuring open learning diaries in alignment with different types of learning strategies, namely cognitive and metacognitive strategies (Glogger et al., 2012; Nückles et al., 2009; Wilkens, 2020), learners receive targeted support in tracking their progress and engaging in systematic reflection (Schmitz and Wiese, 2006; Wallin and Adawi, 2018).

2.1.1 Cognitive learning strategies

Organization and elaboration strategies are cognitive learning strategies with the goal to help learners process, memorize, and retrieve information. Organizational strategies are used to structure knowledge (e.g., via summarizing) with the aim to transform learning content into a readily comprehensible and retrievable form that allows for the systematic integration of information into existing memory structures (Winne, 2001). Elaboration strategies inherent more complex cognitive processes and focus on linking new information with prior knowledge or allow for a transfer of information learned to new contexts (Marton and Säljö, 1976).

2.1.2 Metacognitive learning strategies

Metacognitive strategies help learners to monitor current understanding or problems regarding their learning process (e.g., identifying knowledge gaps) and to regulate future learning. Overall, especially metacognitive strategies play an important role in the success of the self-regulated learning process with positive effects on academic performance as indicated by meta-analyses (Anthonysamy et al., 2020; Broadbent and Poon, 2015).

In line with models of self-regulated learning, which encompass both cognitive and metacognitive processes, empirical research on learning diaries suggests that both types of strategies influence each other and therefore require the intertwined use of all strategies (Nückles et al., 2009; Roelle et al., 2017).

2.1.3 Guiding question for each learning strategy

Following Wilkens (2020), learning diaries can be structured along the three core learning strategies discussed above—organization, elaboration, and metacognition. To support learners in applying these strategies reflectively, Wilkens (2020) proposes the use of guiding prompts:

1. Organization: How can I summarize the central content of the topic?
2. In-depth elaboration: What connections can I make to my prior knowledge?
3. Transfer-supporting elaboration: Where and how can I apply the presented theories or models in practice?
4. Metacognition: What will I do next to clarify remaining questions or to deepen my understanding?

However, the open format of learning diaries makes evaluation as well as feedback time-consuming. Hence, the current study investigates how AI can support teachers in the assessment process of learning diaries.

2.2 AI in assessment

Integrating AI into assessment practices is highly transformative. While traditional assessment systems are often perceived as burdensome, discrete, uniform, inauthentic, and antiquated (Mao et al., 2024; Swiecki et al., 2022), AI-based assessment practices could offer a "paradigm shift" (Agostini et al., 2024, p. 3) especially with regard to evaluating performance. For example, AI can automatically grade tasks and provide immediate feedback to learners, thus enabling adaptive assessment experiences and simultaneously reducing the workload for educators—especially in formative assessment settings.

2.2.1 AI in grading

Trained AI tools showed reliable and valid outcomes when automatically grading closed and open short-answer questions in several domains like programming education (Grivokostopoulou et al., 2017; Messer et al., 2024) or written and oral tasks in language assessment contexts (Huang et al., 2023; Kumar and Boulanger, 2020). In contrast, generative AI models not specifically trained for educational assessment purposes, such as ChatGPT, so far led to ambivalent results of agreement between human and AI assessments when grading open answer questions and essays (Alers et al., 2024; Kooli and Yusuf, 2024; Lundgren, 2024). Nevertheless, their broad accessibility for non-specialist users and the lack of development-related implementation costs render them a potentially attractive solution for educational settings.

2.2.2 AI and learning diaries

As learning and assessment are increasingly "shifting from product-focused to process-focused assessment" (Corbin et al., 2025, p. 7), learning diaries may gain importance due to their ability to capture learners' metacognitive and reflective processes. Initial research has also explored AI-assisted grading of reflective essays in higher education (Awidi, 2024), yet the extent to which generative AI aligns with human ratings in evaluating structured and open learning diaries remains unexplored.

2.3 Research questions

This study investigates learning diary assessments performed by teachers and AI, namely the ChatGPT-4o model by OpenAI. The aim of the current study is to analyze the agreement between teachers and ChatGPT-4o by examining four separately assessed learning strategy categories of a learning diary in adult education. The first research question focuses on the differences between the four learning strategy categories, which are based on different information processing levels:

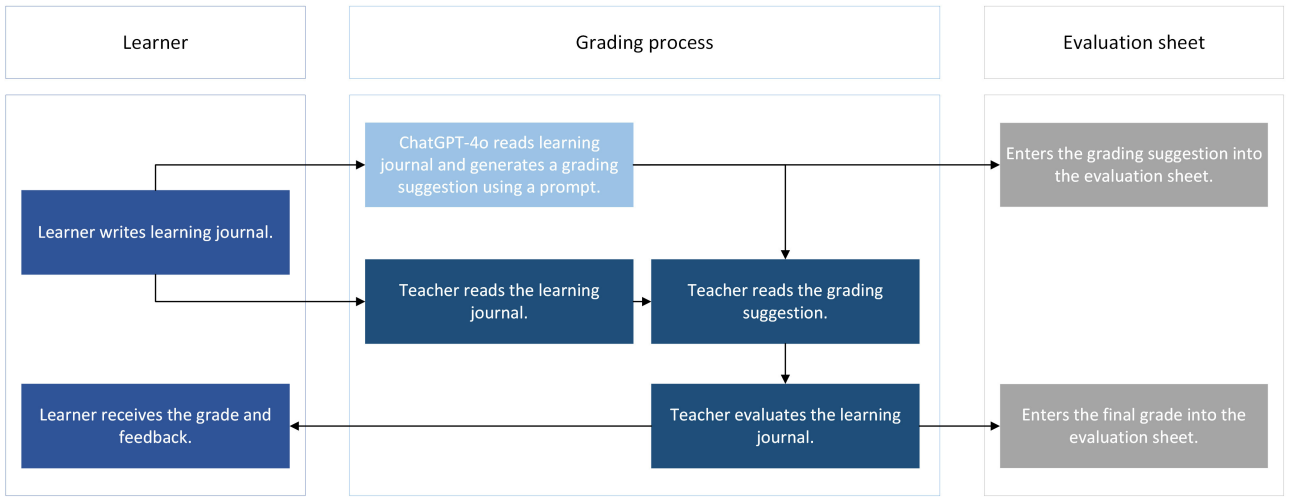

FIGURE 1
The assessment process at velpTEC GmbH.

RQ1: To what extent does the AI-teacher agreement vary across and between the learning strategy categories assessed in the learning diaries?

Additionally, teacher-specific patterns are studied:

RQ2: To what extent does the AI-teacher agreement vary by teacher across the four learning strategy categories?

3 Methodology

3.1 Sample

Between June and September 2024, 135 learners[1] of a publicly funded continuing education program in Germany submitted each four digital learning diaries as graded formative assessments.[2] Each diary was independently assessed by one of seven teachers (human raters)[3] and ChatGPT-4o. The study was conducted at velpTEC GmbH, an adult education institute, and was reviewed and approved by the institute's and Friedrich-Alexander Universität Erlangen-Nürnberg data protection officers.[4]

3.2 Object of investigation

The study used a digital, open, and structured learning diary comprising four learning strategy categories (Wilkens, 2020). Learners responded to four categories—organization, in-depth elaboration, transfer-supporting elaboration, and metacognition—in free-text form (Wilkens, 2020).

3.3 Study design and procedure

Overall, learners could achieve a maximum of eight points across the four categories at each measurement occasion. Teachers[5] as well as AI were trained to evaluate criteria-based per learning strategy category.

Both, the teachers and the AI assessed the learning diaries (see Figure 1). A prompt created by the educational institute supported the ChatGPT-4o model. Only the content of the learning diary was transmitted to the AI, and no additional personal data was provided for teachers or AI. During the evaluation process, teachers

1 Learners submitted between two and eleven learning diary entries depending on the duration of their continuing education program ($n = 573$). For analytical purposes, we selected learners who had submitted exactly four entries ($n = 135$), as this was the most common submission frequency and provided a standardized basis for formative assessment.

2 No additional personal data of the learners were collected as part of the study. Since the assessment was implemented within the regular workflow of continuing education programs at velpTEC GmbH, which prepare participants for evolving professional roles in areas such as information technology, the study relied exclusively on existing instructional processes. In line with the applied nature of the setting and to minimize data collection, no further demographic information was gathered.

3 No additional personal data of the teachers were collected as part of the study.

4 Additionally, all stakeholders were informed and instructed that participation in the study would not lead to any form of discrimination or employment-related disadvantages. Accordingly, the teachers' data were anonymized.

5 In preparation for the study, all participating teachers were introduced to the learning diary as an assessment format and the associated evaluation criteria in April 2024 (Wilkens, 2020). Sample solutions were provided and discussed in weekly workshops to support their understanding, following the concept of informed judgment (Südkamp et al., 2012). In addition, individual sessions were held in which an expert reviewed example diary entries with each rater and answered specific questions regarding the assessment format and evaluation standards. In May 2024, teachers corrected learning diaries without AI being present. This extensive preparation was necessary, as the learning diary was not commonly used in the funded continuing education sector, and none of the seven raters had prior experience with this format.

TABLE 1 Descriptive statistics *M (SD)* for overall absolute accuracy and bias of ChatGPT-4o regarding learners' performance (in relation to the human raters and aggregated over all learning strategy categories).

Human rater	*N*	Overall absolute accuracy	Overall bias
Overall	135	93.26 (7.31)	1.09 (7.71)
Rater 1	25	94.75 (4.58)	−2.25 (5.30)
Rater 2	19	91.45 (6.15)	5.26 (7.44)
Rater 3	17	92.64 (8.04)	1.10 (8.90)
Rater 4	13	93.57 (7.27)	−3.13 (7.86)
Rater 5	19	91.61 (10.26)	3.13 (8.20)
Rater 6	23	97.69 (5.43)	−0.14 (4.76)
Rater 7	19	89.80 (7.13)	3.95 (9.13)

received AI-generated suggestions including a score and short rationale (Swiecki et al., 2022). Teachers then recorded their final human evaluation per learning strategy category. This workflow is illustrated in detail in Figure 1.

3.4 Measurement of learners' performance

The assessment of the learning diary entries followed an analytic scoring model (Jönsson et al., 2021) with 0 to 2 points assigned per learning strategy category. A score of 2 indicated a high-quality response with well-reasoned and differentiated engagement; a score of 1 reflected a surface-level but sufficient response; and 0 points were given when the category was not addressed or the response lacked meaningful engagement.

3.5 Data analyses

3.5.1 AI-teacher agreement

For the evaluation of the AI-teacher agreement, the teacher ratings served as the reference. Therefore, we adapted the established measures absolute accuracy as well as a bias by Schraw (2009) to AI-assessment. While the original scores consider differences of learners' actual and self-judged performance, accuracy and bias in our study are based on differences between the human and AI ratings using the same formulas.

Absolute accuracy represents the summed absolute difference between the teacher rating and the ChatGPT-4o rating over all four measurement occasions. By transforming this difference into a percentage score, the maximum value of 100 indicates a perfect match between the human and AI ratings. Bias was calculated as the difference between ChatGPT-4o ratings and human ratings (range from −100 to 100). It reveals the extent to which ChatGPT-4o overestimated (positive value) or underestimated (negative value) the learners' performance relative to the human rating that served as a relative reference point for this comparison. A value of zero signifies balanced judgments. Absolute accuracy and bias were calculated for each learning strategy category as well as for overall AI-teacher agreement across the categories.

3.5.2 Statistical analyses

A multivariate mixed design based on the ANOVA procedure with the four learning strategy categories as repeated measures (within-subjects factor), assigned teacher (between-subjects factor), and absolute accuracy and bias (dependent variables) was conducted to test for AI-teacher agreement. The influence of the learning strategy categories on AI-teacher agreement was analyzed by the within-subjects effects (RQ1). Additionally, the interaction effect between the learning strategy category and assigned teacher on the AI-teacher agreement was tested to answer RQ2.

4 Results

4.1 AI-teacher agreement

All accuracy measures indicated an overall high AI-teacher agreement. The descriptive values in Table 1 show, that average overall absolute accuracy was close to the possible maximum of 100%, indicating a high agreement between teachers and ChatGPT-4o regarding learners' overall performance. Bias indicated a slight overestimation by ChatGPT-4o of learners' performance in relation to the human ratings when summarizing all four learning categories.

4.2 Category-specific differences in AI-teacher agreement (RQ1)

Both accuracy measures of AI-teacher agreement (see Table 2) showed significant differences between the four learning strategy categories, indicating that AI can evaluate learners' performance more accurately for some strategies than others (see Tables 3, 4). The results of the multivariate mixed-design revealed a significant main effect for the learning strategy category, $F_{(6, 768)} = 6.08$, $p < 0.001$, $\eta_p^2 = 0.05$. According to Cohen (1988), this effect size was small to medium. The differences were small for absolute accuracy and medium for bias. The Greenhouse-Geisser adjustment was used to correct for violations of sphericity. *Post-hoc* tests showed significant differences between the learning strategy categories (see different letters in Figures 2A, B). The learning strategy category of metacognition yielded a bias closest to zero (see Table 2). Together with the high absolute accuracy of this category, this points to a high AI-teacher agreement without underestimating or overestimating the category of metacognition by the AI.

4.3 Rater independence of category-specific differences in AI-teacher agreement (RQ2)

In addition to differences due to the learning strategy category, the results of the multivariate mixed design revealed a significant main interaction effect between the learning strategy category and teacher, $F_{(36, 768)} = 2.15$, $p < 0.001$, $\eta_p^2 = 0.09$. This interaction effect became significant for both absolute accuracy and bias (see Tables 3, 4). All effect sizes were of medium size and indicated that the agreement differences between learning strategy categories were

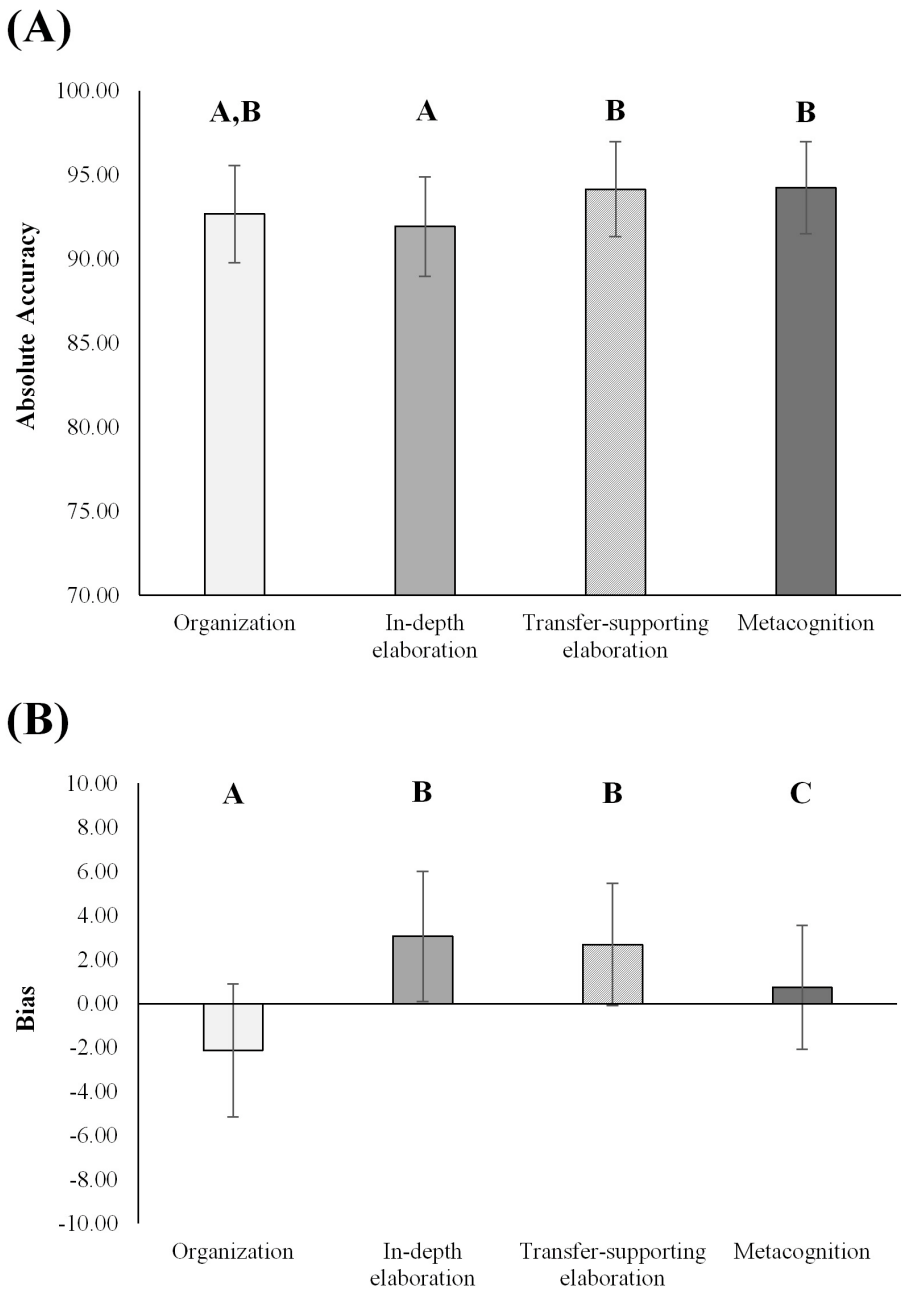

FIGURE 2
Pairwise comparisons of category specific absolute accuracy **(A)** and bias **(B)** between ChatGPT-4o and human raters. Different letters point to significant differences between learning strategy categories. Error bars show standard errors. Positive (negative) values indicate an overestimation (underestimation) of learners' performance by ChatGPT-4o in relation to the human ratings.

dependent from the teacher the learning diaries were assigned to. Differences between raters regarding absolute accuracy and bias are displayed in Figures 3A, B.

5 Discussion

5.1 Category-specific differences in AI-teacher agreement (RQ1)

This study examined the feasibility of using a publicly available AI model, guided by a customized prompt, to support teachers in assessing learning strategy categories recorded in learning diaries for adult education, with particular emphasis on two distinct accuracy measures. The AI-teacher agreement between ChatGPT-4o and human raters varied significantly across learning strategy categories. While absolute accuracy was consistently high across all four categories—indicating overall adequate AI performance—substantial differences emerged in bias: ChatGPT-4o tended to overestimate learner performance in in-depth elaboration and transfer-supporting elaboration, whereas metacognition yielded only minimal bias.

The differences in bias values might be explained as follows: AI overestimated the quality of learners' diary entries

TABLE 2 Descriptive statistics *M (SD)* for absolute accuracy and bias regarding the four learning strategy categories.

Agreement	Organization	In-depth elaboration	Transfer-supporting elaboration	Metacognition
Absolute accuracy	92.69 (10.73)	91.94 (11.48)	94.17 (9.99)	94.26 (9.00)
Bias	-2.13 (12.46)	3.06 (11.48)	2.69 (9.46)	0.74 (9.99)

TABLE 3 Results of the multivariate mixed design for differences in absolute accuracy between the within-subjects factor learning strategy categories, the between-subjects factor teacher, and their interaction.

Effects	*F*	*df*	*p*	η_p^2
Learning strategy category	2.88	(2.69, 343.73)	0.042	0.02
Teacher	142.55	(6, 128)	0.011	0.12
Learning strategy category x teacher	2.48	(16.11, 343.73)	0.001	0.10

TABLE 4 Results of the multivariate mixed design for differences in bias between the within-subjects factor learning strategy categories, the between-subjects factor teacher, and their interaction.

Effects	*F*	*df*	*p*	η_p^2
Learning strategy category	9.40	(2.82, 361.46)	<0.001	0.07
Teacher	180.96	(6, 128)	0.004	0.14
Learning strategy category x teacher	1.99	(16.94, 361.46)	0.011	0.09

in the categories in-depth elaboration and transfer-supporting elaboration. Teachers, with their expertise in profession-specific applications, may be better trained to provide nuanced assessments of practical examples and transfer—both of which are critical in these categories. It is possible that ChatGPT-4o lacks the contextual understanding necessary to accurately evaluate these aspects, resulting in a systematic overestimation of learner performance. Lundgren (2024), for example, also observed that AI tends to apply assessment criteria more "optimistically" and "politely" than human raters. As a result, essential regulations in learning may not be feedbacked, since the AI may fail to detect certain learner deficits, leaving them unaddressed. By contrast, in the category metacognition, AI demonstrated the highest absolute accuracy and the lowest bias. Within semi-automated assessment settings, this category thus appears to be a promising candidate for partial automation (Molenaar, 2022). This is further supported by the fact that metacognition is conceptually distinct from other learning strategy categories, allowing for clearer boundaries in automated evaluation and potentially reducing teacher workload. While these results are encouraging, the potential for automation of metacognitive assessment must be interpreted with caution. Although AI achieved high accuracy and low bias in this category, it remains unclear whether the AI captures the qualitative patterns that teachers attend to in metacognitive reflection.

In terms of full automation, learners can be provided with transparent access to the AI-based evaluation and can use this to make potential learning improvements and therefore support the self-directed learning process. This allows learners to critically evaluate their learning progress, identify strengths and weaknesses, and refine their learning strategies accordingly. This aligns with findings from Roe and Perkins (2024), who highlight that AI has the potential to enhance self-directed learning and to foster autonomy by offering on-demand, personalized assistance. Future research could explore how AI-assisted evaluation can be optimized to further strengthen self-directed learning in adult education. In addition to that, longitudinal studies comparing groups with immediate AI-generated feedback to those receiving delayed feedback from human teachers could help determine which type of feedback more effectively fosters sustained learning development (Henderson et al., 2025).

5.2 Rater independence of category-specific differences in AI-teacher agreement (RQ2)

The results of the multivariate mixed design indicated a significant interaction effect between the learning strategy categories and the teachers. This finding suggests that inter-rater agreement differences between ChatGPT-4o and teachers depend on the specific teacher. Within a heterogeneous pool of educators, ChatGPT-4o's grading may be interpreted not as a deviation from a universal "true" score, but as a distinct assessment style. Thus, ChatGPT-4o may serve as an additional assessor complementing the human grading team rather than replacing it (Gentile et al., 2023; Lameras and Arnab, 2022). However, it should be noted that each teacher assessed a different set of learning diary entries, which means that no direct inter-rater agreement among teachers could be calculated. This lack of overlapping assessments represents a potential limitation of the study. Accordingly, future research should include overlapping ratings to clarify whether differences in AI agreement are due to rater effects or variations in learner texts. An additional limitation concerns the potential influence of AI-generated suggestions on teacher judgments. Since teachers had access to ChatGPT-4o's scoring and rationale before recording their final decisions, anchoring effects cannot be ruled out. This may have affected the independence of human evaluations and should be considered when interpreting the AI-teacher agreement.

6 Conclusion

The study examined the agreement between teachers and ChatGPT-4o in assessing learning diaries in adult education.

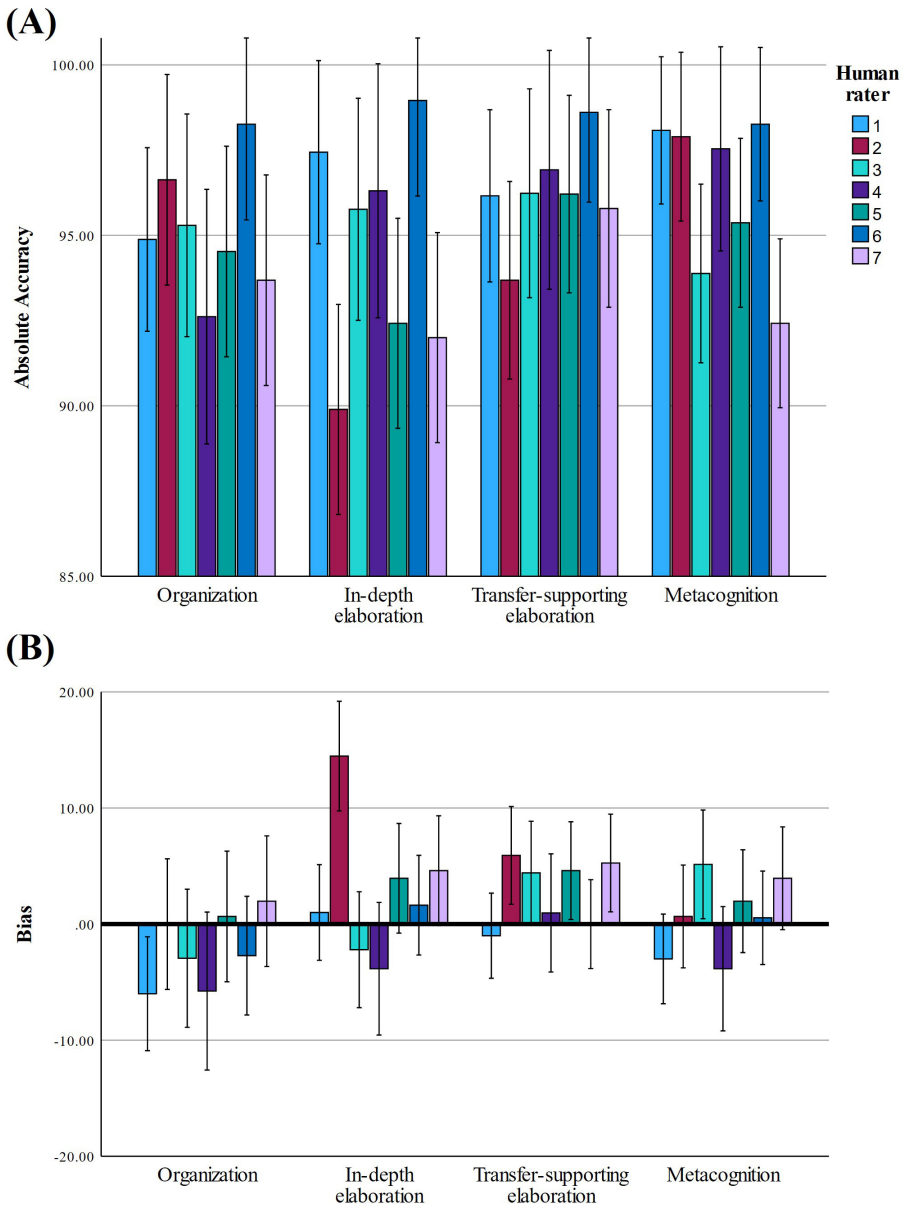

FIGURE 3
Differences in absolute accuracy **(A)** and bias **(B)** of ChatGPT-4o between human raters across the learning strategy categories. Error bars show 95% confidence intervals.

Results showed that AI-teacher agreement varied depending on both the type of learning strategy and the individual teacher. As a result, human expertise remains essential—particularly in evaluating elaboration strategies that require contextual and domain-specific judgment. At the same time, the high overall accuracy and low bias suggest that ChatGPT-4o may serve as an additional assessor complementing the human grading team rather than replacing it. From this perspective, the integration of ChatGPT-4o can be seen as a step toward hybrid intelligence, in which human expertise and AI-generated assessments complement each other (Dellermann et al., 2019). In such systems, the goal is not to replace human judgment, but to enhance it through collaborative evaluation processes. Practically, this may provide educators with immediate secondary input and foster more differentiated and reflective assessment practices. Future research should investigate how hybrid intelligence systems affect teacher judgment and the quality of formative assessments in diverse educational contexts.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The study involving humans was approved by the Data Protection Officer of FAU Erlangen-Nürnberg. It was conducted in accordance with local legislation and institutional requirements. All participants provided their written informed consent to participate in the study.

Author contributions

LR: Project administration, Conceptualization, Methodology, Writing – review & editing, Data curation, Investigation, Writing – original draft. MH: Writing – review & editing, Conceptualization, Methodology, Supervision. NN-S: Data curation, Writing – original draft, Conceptualization, Supervision, Writing – review & editing, Formal analysis, Methodology.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript for translation purposes.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agostini, D., Lazareva, A., and Picasso, F. (2024). Advancements in technology-enhanced assessment in tertiary education. *Australas. J. Educ. Technol.* 40. doi: 10.14742/ajet.10122

Alers, H., Malinowska, A., Meghoe, G., and Apfel, E. (2024). "Using ChatGPT-4 to grade open question exams," in *Advances in Information and Communication, Lecture Notes in Networks and Systems,* ed. K. Arai (Berlin; Cham: Springer), 1–9.

Alt, D., Raichel, N., and Naamati-Schneider, L. (2022). Higher education students' reflective journal writing and lifelong learning skills: insights from an exploratory sequential study. *Front. Psychol.* 12:707168. doi: 10.3389/fpsyg.2021.707168

Anthonysamy, L., Koo, A.-C., and Hew, S.-H. (2020). Self-regulated learning strategies and non-academic outcomes in higher education blended learning environments: a one decade review. *Educ. Inf. Technol.* 25, 3677–3704. doi: 10.1007/s10639-020-10134-2

Awidi, I. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative Artificial Intelligence (AI) tool. *Comput. Educ. Artif. Intell.* 6:100226. doi: 10.1016/j.caeai.2024.100226

Broadbent, J., and Poon, W. L. (2015). Self-regulated learning strategies and academic achievement in online higher education learning environments: a systematic review. *Internet High. Educ.* 27, 1–13. doi: 10.1016/j.iheduc.2015.04.007

Bürgermeister, A., and Saalbach, H. (2018). Formative assessment: an approach to foster individual learning. *Psychol. Erzieh. Unterr.* 65:194. doi: 10.2378/peu2018.art11d

Campbell, C., Parboosingh, J., Gondocz, T., Babitskaya, G., and Pham, B. (1999). A study of the factors that influence physicians' commitments to change their practices using learning diaries. *Acad. Med.* 74, 34–36. doi: 10.1097/00001888-199910000-00033

Chang, C.-C., Liang, C., Shu, K.-M., Tseng, K.-H., and Lin, C.-Y. (2016). Does using e-portfolios for reflective writing enhance high school students' self-regulated learning? *Technol. Pedagogy Educ.* 25, 317–336. doi: 10.1080/1475939X.2015.1042907

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Routledge.

Corbin, T., Dawson, P., and Liu, D. (2025). Talk is cheap: why structural assessment changes are needed for a time of GenAI. *Assess. Eval. High. Educ.* 1–11. doi: 10.1080/02602938.2025.2503964

Dellermann, D., Ebel, P., Söllner, M., and Leimeister, J. M. (2019). Hybrid intelligence. *Bus. Inf. Syst. Eng.* 61, 637–643. doi: 10.1007/s12599-019-00595-2

Gentile, M., Città, G., Perna, S., and Allegra, M. (2023). Do we still need teachers? *Navigating the paradigm shift of the teacher's role in the AI era. Front. Educ.* 8:1161777. doi: 10.3389/feduc.2023.1161777

Glogger, I., Schwonke, R., Holzäpfel, L., Nückles, M., and Renkl, A. (2012). Learning strategies assessed by journal writing: prediction of learning outcomes by quantity, quality, and combinations of learning strategies. *J. Educ. Psychol.* 104, 452–468. doi: 10.1037/a0026683

Grivokostopoulou, F., Perikos, I., and Hatzilygeroudis, I. (2017). An educational system for learning search algorithms and automatically assessing student performance. *Int. J. Artif. Intell. Educ.* 27, 207–240. doi: 10.1007/s40593-016-0116-x

Henderson, M., Bearman, M., Chung, J., Fawns, T., Shum, S. B., Matthews, K. E., et al. (2025). Comparing generative AI and teacher feedback: student perceptions of usefulness and trustworthiness. *Assess. Eval. High. Educ.* 1–16. doi: 10.1080/02602938.2025.2502582

Huang, X., Zou, D., Cheng, G., Chen, X., and Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educ. Technol. Soc.* 26, 112–131. doi: 10.30191/ETS.202301_26(1).0009

Järvelä, S., Zhao, G., Nguyen, A., and Chen, H. (2025). Hybrid intelligence: human–AI coevolution and learning. *Br. J. Educ. Technol.* 56, 455–468. doi: 10.1111/bjet.13560

Jönsson, A., Balan, A., and Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assess. Educ. Princ. Policy Pract.* 28, 212–227. doi: 10.1080/0969594X.2021.1884041

Kooli, C., and Yusuf, N. (2024). Transforming educational assessment: insights into the use of ChatGPT and large language models in grading. *Int. J. Human Comput. Interact.* 41, 3388–3399. doi: 10.1080/10447318.2024.2338330

Kumar, V., and Boulanger, D. (2020). Explainable automated essay scoring: deep learning really has pedagogical value. *Front. Educ.* 5:572367. doi: 10.3389/feduc.2020.572367

Lameras, P., and Arnab, S. (2022). Power to the teachers: an exploratory review on artificial intelligence in education. *Information* 13:14. doi: 10.3390/info13010014

Lundgren, M. (2024). Large language models in student assessment: comparing ChatGPT and human graders. arXiv [Preprint]. arXiv.2406.15610. doi: 10.2139/ssrn.4874359

Mao, J., Chen, B., and Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends* 68, 58–66. doi: 10.1007/s11528-023-00911-4

Marton, F., and Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process*. *Br. J. Educ. Psychol.* 46, 4–11. doi: 10.1111/j.2044-8279.1976.tb02980.x

Messer, M., Brown, N. C. C., Kölling, M., and Shi, M. (2024). Automated grading and feedback tools for programming education: a systematic review. *ACM Trans. Comput. Educ.* 24:10. doi: 10.1145/3636515

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *Eur. J. Educ.* 57, 632–645. doi: 10.1111/ejed.12527

Naujoks, N., and Händel, M. (2020). Cram for the exam? Distinct trajectories of cognitive learning strategy use during the term. *Unterrichtswissenschaft* 48, 221–241. doi: 10.1007/s42010-019-00062-7

Nückles, M., Hübner, S., and Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learn. Instr.* 19, 259–271. doi: 10.1016/j.learninstruc.2008.05.002

Roe, J., and Perkins, M. (2024). *Generative AI in self-directed* learning: a scoping review. arXiv [Preprint]. arXiv.2411.07677. doi: 10.48550/arXiv.2411.07677

Roelle, J., Nowitzki, C., and Berthold, K. (2017). Do cognitive and metacognitive processes set the stage for each other? *Learn. Instr.* 50, 54–64. doi: 10.1016/j.learninstruc.2016.11.009

Schmitz, B., and Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: time-series analyses of diary data. *Contemp. Educ. Psychol.* 31, 64–96. doi: 10.1016/j.cedpsych.2005.02.002

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacogn. Learn.* 4, 33–45. doi: 10.1007/s11409-008-9031-3

Südkamp, A., Kaiser, J., and Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *J. Educ. Psychol.* 104, 743–762. doi: 10.1037/a0027627

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., et al. (2022). Assessment in the age of artificial intelligence. *Comput. Educ. Artif. Intell.* 3:100075. doi: 10.1016/j.caeai.2022.100075

Trif, L., and Popescu, T. (2013). The reflective diary, an effective professional training instrument for future teachers. *Procedia Soc. Behav. Sci.* 93, 1070–1074. doi: 10.1016/j.sbspro.2013.09.332

Wallin, P., and Adawi, T. (2018). The reflective diary as a method for the formative assessment of self-regulated learning. *Eur. J. Eng. Educ.* 43, 507–521. doi: 10.1080/03043797.2017.1290585

Wilkens, R. (2020). Assessment without examination: competence-oriented, semesteraccompanying performance measurement of students. *Hochschullehre* 6, 499–503. doi: 10.3278/HSL2038W

Winne, P. H. (2001). "Self-regulated learning viewed from models of information processing," in *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives*, eds. J. B. Zimmerman and D. H. Schunk (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 153–189.