Thomas, D.R. et al. (2025). Improving Open-Response Assessment with LearnLM.
In: Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S. (eds) Artificial
Intelligence in Education. AIED 2025. Lecture Notes in Computer Science(), vol
15881. Springer, Cham. https://doi.org/10.1007/978-3-031-98462-4_28

Improving Open-Response Assessment with LearnLM

Danielle R. Thomas[0000−0001−8196−3252] [1], Conrad Borchers[0000−0003−3437−8979] [1], Shambhavi Bhushan[0009−0004−3695−2334] [1], Sanjit Kakarla[0009−0007−6508−8647] [1], Alex Houk[0009−0001−5933−6970] [1], Ralph Abboud[0000−0002−2332−0504] [2], Shivang Gupta[0000−0002−5713−3782] [1], Erin Gatz[0000−0002−6880−5740] [1], and Kenneth R. Koedinger[0000−0002−5850−4768] [1]

[1] Carnegie Mellon University
{drthomas,cborchers,shivangg,koedinger}@cmu.edu
{shambhab,sanjitk,ahouk,egatz}@andrew.cmu.edu
[2] Learning Engineering Virtual Institute
rabboud@levimath.org

Abstract. Automated grading of learners' open responses remains challenging due to the complexity of language and the subjective nature of human evaluation. Recent advances in generative AI, particularly large language models (LLMs), offer new possibilities to improve assessment. Off-the-shelf LLMs, such as GPT-4, have been applied to this task, as well as dedicated education-oriented models, such as LearnLM. However, little is known about their effectiveness compared to general-purpose models. In this study, we evaluate GPT-4o, GPT-4-turbo, and Gemini-Pro and compare their performance to LearnLM to determine their effectiveness in assessing learning, specifically the professional development of adult tutors. We find that LearnLM outperforms other models on tasks requiring tutor learners to predict the most appropriate response to students. We hypothesize that this is due to the model's fine-tuning on tutor-student interaction data and suggest that LearnLM may be particularly useful in scenario-based tutor training. To further improve automated assessment methods, we challenge the concept of human "ground truth" by proposing alternative validation methods. Specifically, we introduce a predictive validity method by relating open-response scores with corresponding multiple-choice scores that demonstrate statistically significant and moderate correlations, particularly with LearnLM. Our novel method demonstrates predictive validity but should be combined with additional measures to ensure a more comprehensive assessment. This study contributes an open source dataset, human annotation rubrics, and LLM prompts, to improve future assessment applications of LLMs.

Keywords: Generative AI, LLMs, LearnLM, Assessment

1 Introduction and Related Work

Automated grading of learners' textual responses has long been a challenging task due to the complexity of language, context, and the subjective nature of human judgment [6, 10, 17]. Recent breakthroughs in generative AI, particularly large language models (LLMs), have improved automated assessment. These models, such as GPT-4o [1], GPT-4-turbo, and Gemini-Pro [18], offer scalable solutions to evaluate textual responses. Recently, LearnLM [11], a new family of Gemini models fine-tuned for learning, has been developed to assist with education-related tasks. Fine-tuned for pedagogical instruction from both synthetic and human-written datasets, LearnLM shows promise with results preferred by expert raters over GPT-4o and Gemini-Pro models across several learning scenarios, such as a tutor helping a student solve a math problem [19]. Despite promising preliminary results, the LearnLM's ability to assess pedagogical interaction more broadly, such as equity-related tasks and adult tutor learning, as opposed to student learning and content-related tutoring, is underexplored.

Further Validation of LearnLM. Recent advances in generative AI have significantly influenced the development of specialized models aimed at improving educational outcomes. Introduced by Google in 2024, LearnLM–specifically `learnlm-1.5-pro-experimental`–is a pedagogically fine-tuned extension of the Gemini 1.5 Pro model, designed to emulate behaviors associated with effective human tutors [2]. Its core contribution lies in its ability to follow nuanced pedagogical instructions, enabling targeted responses aligned with specific educational objectives [19, 11]. LearnLM has been evaluated through multistage human assessments, where pedagogy experts simulated learner interactions with various AI systems. These evaluations consistently demonstrated that LearnLM outperformed contemporaneous models like GPT-4o, Claude 3.5, and Gemini 1.5 Pro in areas such as adapting to learners' needs, managing cognitive load, and promoting active learning [2, 19]. Experts preferred LearnLM across varying scenarios (e.g., a shy student trying to practice addition or a disinterested student wanting a quick response) with mean performance strength 31% higher than GPT-4o and 13% more compared to Gemini 1.5 Pro [19]. Despite promising results, the LearnLM team [19] acknowledges the limitations of their evaluation and invites further extrinsic validation, which this work aims to address.

Revisiting Human Coding as Ground Truth. The reliability of human grading as the "gold standard" for evaluating learners' responses has limitations: inconsistency, subjective interpretation, and bias [3, 8, 15]. Studies reveal significant variability in human grading practices, often due to poor inter-rater reliability and inconsistent self-grading among human judges [8, 15]. Individual biases remain even when rubrics are present. The emerging trend of using LLM-as-a-judge [26] has introduced a novel framework for evaluating performance in open-ended responses without relying on predefined human annotations. However, this approach introduces its own bias problems [9, 13, 14], sparking the urgency to develop alternative validation frameworks that reduce dependence on subjective human judgment. To address this limitation, we draw from predictive validity, which is the degree to which a test or assessment accurately forecasts or correlates with future performance on a related measure [23]. In this context, we use multiple-choice questions (MCQs) that assess the same learning objectives as open-response questions. Predictive validity evaluates whether performance on MCQs can reliably predict outcomes on open-response questions or other measures of learner understanding. MCQs are known to be effective and more efficient than open response tasks when instructional time is limited [21, 7, 16], in addition to being more objective than human scoring of open responses. This approach ensures MCQs serve as effective proxies for more complex assessments.

In this study, we address these gaps by evaluating the performance of several LLM models and comparing their absolute performance to LearnLM to determine LearnLM's ability to generalize to tangential tasks beyond its fine-tuning. We also propose an alternative method to humans as "ground truth." We hypothesize that the fine-tuned model LearnLM will be better at assessing textual responses in a way that is aligned with tutor learning due to its fine-tuning and aim to demonstrate a more objective and efficient method to evaluate its performance. This work also contributes an open source dataset and LLM prompts. We ask: **RQ1:** How does LearnLM perform compared to GPT-4o, GPT-4 turbo, and Gemini Pro on adult tutor learning and equity-related tasks? **RQ2:** How does the alternative validation method of using predictive validity by mapping to the corresponding MCQs (instead of human-scored open response questions as traditional ground truth) compare in evaluating LLM performance?

2 Methods

Participants, Setting, and Lesson Design. The open-source dataset was derived from 81 college students employed as paid tutors for a remote tutoring organization. The lesson was delivered through an online tutoring platform, PLUS.[3] Tutor participants accessed the lesson through the PLUS tutoring app in their own time and submitted their constructed responses through open-text boxes. The focal lesson for this present work, titled *Helping Students Manage Inequities*, was developed by a university research team specializing in learning sciences. The lesson trains human tutors to support students facing various inequities, such as a student not having access to the internet at home to do their homework or a student not being able to hear the teacher during class instruction because they sit in the back of the classroom [20]. All data were anonymously logged, in line with the approved IRB protocol and consisted of responses to a pretest and posttest tutoring scenario where participants were asked to *predict* the best response within an open-response question (i.e., *"What would you say to the student..."*), followed by an MCQ (i.e., *"Which of the following tutor responses..."*). Then, participants were asked to *explain* their rationale in an open-response question (i.e., *"Why did you choose to..."*) and a MCQ (i.e., *"Which of the following explains your rationale..."*). All participants' responses were binary coded as incorrect or correct (0/1) by two human researchers. The interrater reliability for the *predict* open responses was 89% agreement ($\kappa = 0.75$) and 87% agreement ($\kappa = 0.75$) for *explain* open responses.

Comparing LearnLM Performance and Predictive Validity. To evaluate the performance of various models in scoring learners' responses, we used a

[3] https://www.tutors.plus/

quantitative comparison based on key metrics: accuracy, AUC, and F1 score. The four models used and their specific model strings were GPT-4o (`gpt-4o-2024-11-20`), GPT-4-turbo (`gpt-4-turbo-2024-04-09`), Gemini 1.5 Pro (`gemini-1.5-pro-002`), and the experimental Gemini model LearnLM (`learnlm-1.5-pro-experimental`). Each model assessed both open-response question types: *predicting* the best response and *explaining* the rationale behind that response. Few-shot prompting was applied uniformly across all models, with temperature settings adjusted for select models to observe variability in performance [5]. In this context, temperature is a parameter that controls the randomness of the model's responses—lower values make outputs more deterministic, while higher values increase variability in text generation. Few-shot prompting has been found to outperform zero-shot methods in nuanced lesson scenarios [21, 22]. Chain-of-thought prompting was used that required the models to provide rationale for their scores [24]. The dataset, lesson, and LLM system prompts are available via GitHub.[4]

To evaluate the predictive validity of the LLM open response scores on MCQ scores, we calculated the correlation of participants' MCQ scores (0-2 pts) and their LLM scores on open-ended responses (0-2 pts). We also performed correlational analysis of the higher performing LLMs with MCQ scores to determine predictive validity on better performing models.

3 Results and Discussion

RQ1: Model Performance Comparisons with LearnLM. Table 1 summarizes model performance in scoring learners' responses in *predicting* the best response and *explaining* their rationale, with accuracy, AUC, and F1 score as evaluation metrics. Confidence intervals were obtained via bootstrapping. The better performing models are shown in bold. GPT-4o demonstrates the highest overall performance, with accuracies of 0.84 and 0.85 for *predict* and *explain* open responses, respectively. GPT-4-turbo lags behind with significantly lower accuracy for assessing *predict* questions (0.62), but improves substantially for *explain* question (0.78). Gemini-pro-1.5 shows consistent performance, with accuracy and F1 scores improving as temperature increases, peaking at temp=1. LearnLM performs well, second to GPT-4o, achieving accuracy in the low-80s across both tasks, with stable AUC and F1 scores. Temperature adjustments for LearnLM show minimal impact on performance compared to Gemini-pro.

Fig. 1 illustrates the pretest and posttest scores, two MCQs and two open-response questions, for a total of four points each at temperature 0. In line with previous work, human scores were generally higher for the pretest and posttest compared to all LLMs, demonstrating leniency among human graders [4]. Although LearnLM performed well, particularly on assessing learners' open-response tasks involving prediction of the best approach, this is not visually apparent due to its comparatively lower performance on evaluating learners' open-response tasks involving explanation of their chosen approach.

[4] https://github.com/conradborchers/learnLM-open-response

Table 1: Performance against human scores with 95% confidence intervals. All temperatures at 0, unless otherwise specified. Better performing models in bold.

Model	Predicting the best approach			Explaining their rationale		
	Accuracy	AUC	F1	Accuracy	AUC	F1
GPT-4o	**0.84** [0.79, 0.89]	**0.87** [0.82, 0.92]	**0.88** [0.83, 0.92]	**0.85** [0.79, 0.90]	**0.86** [0.81, 0.91]	**0.85** [0.79, 0.90]
GPT-4-turbo	0.63 [0.57, 0.7]	0.73 [0.69, 0.78]	0.65 [0.57, 0.72]	0.78 [0.72, 0.83]	0.79 [0.74, 0.85]	0.75 [0.66, 0.82]
Gemini-Pro-1.5	0.75 [0.68, 0.82]	0.74 [0.67, 0.81]	0.82 [0.76, 0.87]	0.84 [0.78, 0.89]	0.84 [0.79, 0.89]	0.85 [0.79, 0.90]
Gemini-Pro-1.5, T=1	0.75 [0.68, 0.81]	0.74 [0.66, 0.81]	0.81 [0.75, 0.86]	**0.87** [0.82, 0.92]	**0.87** [0.82, 0.92]	**0.88** [0.82, 0.93]
Gemini-Pro-1.5, T=2	0.73 [0.67, 0.79]	0.72 [0.65, 0.79]	0.80 [0.74, 0.85]	0.86 [0.81, 0.91]	0.87 [0.81, 0.92]	0.87 [0.81, 0.92]
LearnLM	**0.82** [0.76, 0.87]	**0.82** [0.76, 0.88]	**0.86** [0.81, 0.91]	0.83 [0.77, 0.88]	0.84 [0.79, 0.89]	0.82 [0.76, 0.88]
LearnLM, T=1	0.79 [0.73, 0.85]	0.78 [0.71, 0.85]	0.85 [0.79, 0.89]	0.82 [0.77, 0.88]	0.84 [0.79, 0.88]	0.82 [0.75, 0.87]
LearnLM, T=2	0.81 [0.75, 0.87]	0.79 [0.72, 0.86]	0.86 [0.81, 0.91]	0.81 [0.75, 0.86]	0.82 [0.77, 0.87]	0.80 [0.72, 0.86]

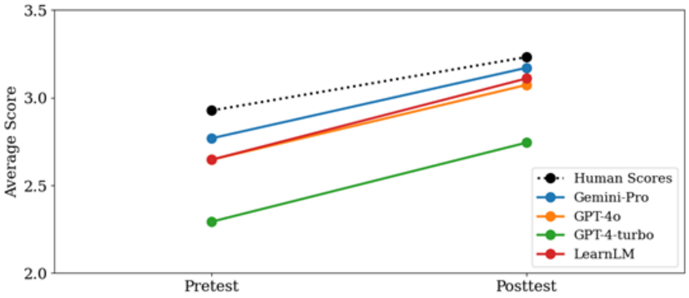

Fig. 1: Comparison of average human and LLM scores at pre/posttest (T=0).

Referring to Table 1, LearnLM performs considerably better than other models, aside from GPT-4o, on assessing responses where tutors *predict the best approach*. For instance, the learner response of what to say to a student who is experiencing an educational inequity is as follows: *“I would discuss how Alexis could present her problem to the teacher”* is an incorrect response as determined by human-expert graders as it does not actively state *what* the tutor will say to the student. Similarly, LearnLM (few-shot and zero-shot prompting at temperatures of 0, 1, and 2) effectively scored the response as incorrect. However, GPT-4o, GPT-4-turbo, and Gemini-Pro-1.5 all ineffectively scored the response as correct. In contrast, the response, *“I am sorry you are going through this. Is there any way we can present this to your teacher in a way she provides an alternative?”* was scored correct by human experts and the LearnLM models, with the remaining models scoring it incorrectly, giving the response a 0. Several more incidences occurred where human truth and LearnLM agreed in opposition to the remaining models. We hypothesize that LearnLM's strong performance in approach prediction stems from its more curated instruction fine-tuning, which aligns strongly with following pre-set pedagogical practices. In contrast, we conjecture that this same fine-tuning also explains LearnLM's less competitive performance on explaining tutor rationale, as LearnLM, compared to its initial generally fine-tuned version, has explicitly been trained to avoid making explicit pedagogical choices, making it less adept at justifying approaches compared to following them. There were 26 learner open responses where LearnLM (few shot, T=0) and human truth were in agreement and GPT-4o (few shot, T=0) was in disagreement, with the majority related to questions where learners had to *predict* the best response. More specifically, there were seven incidences among the 352 open responses in which human truth and LearnLM were in agreement and the other LLMs (few-shot, T=0) were in opposition. All responses, except one, were prediction tasks and scored as correct by humans and LearnLM.

RQ2: Alternative Validation Approaches to Human as Ground Truth. The analysis revealed a significant, positive correlation between MCQ scores and human-graded open-response scores, $r(86) = 0.42, p < .001$. While this correlation is statistically significant, it is not particularly large. Potential alternative contributors to the moderate size of this correlation are (1) the few test items, just two each, used to produce the MCQ and open-response scores and (2) correlation strength is limited by the imperfect reliability and ambiguity of human grading of open responses. To compare this correlation to the validity of LLM assessment, we computed the correlation between MCQ scores and LLM scores. We correlated MCQ scores with GPT-4o-scored open responses, yielding $r(86) = 0.41, p < .001$; and MCQ scores with LearnLM-scored open responses, yielding $r(86) = 0.48, p < .001$. This descriptive evidence of a difference did not reach statistical significance, z = -1.04, p = .297 but merits further investigation.

We find that LLM scores have significant predictive validity and *this validity can be determined without open-ended grading by humans.* Notably, the LearnLM correlation of 0.48 is 0.06 higher than the human-scored correlation of 0.42. In other words, with this tighter comparison, we find that the predictive validity of the LLM scoring is comparable to that of human scoring. This result supports predictive validity as a complementary method for evaluating model performance, though it should be combined with additional measures to ensure a more comprehensive assessment. Other alternative approaches to using human scores as "ground truth" include: using the average score among several adversarial models (e.g., LearnLM vs. GPT-4o) to establish reliability rather than comparing human judgments; and applying LLM self-consistency measures by comparing evaluations across different prompts to check robustness.

4 Implications, Future Work, and Conclusion

This work presents two main implications for the use of LLMs in education: (1) insight into the boundaries of tasks for which LearnLM is best suited, and

(2) the introduction of predictive validity as a complementary method to using human scoring as ground truth for evaluating open responses. Future research will integrate these implications by using LearnLM to score open responses with much larger sample sizes and across a variety of tutor training lessons. In addition, future work will involve the use of ensemble methods such as bagging and boosting to leverage strengths of different LLMs [25, 12]. Our findings reveal that LearnLM can match or even outperform other models on tasks that require tutors to interact with students in situational contexts, especially when they require the generation of specific tutor dialogue moves. Similarly, LearnLM can match or even outperform human scorers using predictive validity methods showing potential for scaling the intricate task of open-response grading.

Acknowledgments

This work was made possible with the support of the Learning Engineering Virtual Institute. The opinions, findings and conclusions expressed in this material are those of the authors.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. AI, G.: Learnlm | gemini api | google ai for developers. https://ai.google.dev/gemini-api/docs/learnlm, accessed: 2025-02-18
3. Andrade, H.G.: Teaching with rubrics: The good, the bad, and the ugly. College teaching **53**(1), 27–31 (2005)
4. Awidi, I.T.: Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (ai) tool. Computers and Education: Artificial Intelligence **6**, 100226 (2024)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. International journal of artificial intelligence in education **25**, 60–117 (2015)
7. Butler, A.C.: Multiple-choice testing in education: Are the best practices for assessment also good for learning? Journal of Applied Research in Memory and Cognition **7**(3), 323–331 (2018)
8. Chen, G.H., Chen, S., Liu, Z., Jiang, F., Wang, B.: Humans or llms as the judge? a study on judgement biases. arXiv preprint arXiv:2402.10669 (2024)
9. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. Computational Linguistics pp. 1–79 (2024)
10. Haller, S., Aldea, A., Seifert, C., Strisciuglio, N.: Survey on automated short answer grading with deep learning: from word embeddings to transformers. arXiv preprint arXiv:2204.03503 (2022)

11. Jurenka, I., Kunesch, M., McKee, K.R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S.M., Hermann, K., Kasenberg, D., Bhoopchand, A., et al.: Towards responsible development of generative ai for education: An evaluation-driven approach. arXiv preprint arXiv:2407.12687 (2024)
12. Lee, G.G., Latif, E., Wu, X., Liu, N., Zhai, X.: Applying large language models and chain-of-thought for automatic scoring. Computers and Education: Artificial Intelligence **6**, 100213 (2024)
13. Lee, J., Hicke, Y., Yu, R., Brooks, C., Kizilcec, R.F.: The life cycle of large language models in education: A framework for understanding sources of bias. British Journal of Educational Technology **55**(5), 1982–2002 (2024)
14. Liang, P.P., Wu, C., Morency, L.P., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: International Conference on Machine Learning. pp. 6565–6576. PMLR (2021)
15. Messer, M., Brown, N.C., Kölling, M., Shi, M.: How consistent are humans when grading programming assignments? arXiv preprint arXiv:2409.12967 (2024)
16. Moore, S., Bier, N., Stamper, J.: Assessing educational quality: Comparative analysis of crowdsourced, expert, and ai-driven rubric applications. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 12, pp. 115–125 (2024)
17. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review **55**(3), 2495–2527 (2022)
18. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
19. Team, L., Modi, A., Veerubhotla, A.S., Rysbek, A., Huber, A., Wiltshire, B., Veprek, B., Gillick, D., Kasenberg, D., Ahmed, D., et al.: Learnlm: Improving gemini for learning. arXiv preprint arXiv:2412.16429 (2024)
20. Thomas, D., Yang, X., Gupta, S., Adeniran, A., Mclaughlin, E., Koedinger, K.: When the tutor becomes the student: Design and evaluation of efficient scenario-based lessons for tutors. In: LAK23: 13th International Learning Analytics and Knowledge Conference. pp. 250–261 (2023)
21. Thomas, D.R., Borchers, C., Kakarla, S., Lin, J., Bhushan, S., Guo, B., Gatz, E., Koedinger, K.R.: Do tutors learn from equity training and can generative ai assess it? In: Proceedings of the 15th International Learning Analytics and Knowledge Conference. pp. 505–515 (2025)
22. Thomas, D.R., Borchers, C., Kakarla, S., Lin, J., Bhushan, S., Guo, B., Gatz, E., Koedinger, K.R.: Does multiple choice have a future in the age of generative ai? a posttest-only rct. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference. pp. 494–504 (2025)
23. Trochim, W.M., Donnelly, J.P., Arora, K.: Research methods: The essential knowledge base (2016)
24. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
25. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. Journal of Research in Science Teaching **59**(10), 1765–1794 (2022)
26. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems **36**, 46595–46623 (2023)