Using a Learning Progression Framework to Guide LLM-Based Formative Assessment in STEM Education

Karen D. Wang[1,*], Jialin Li[2], Carl Wieman[2] and Leonora Kaldaras[3]

[1]San Jose State University, 1 Washington Square, San Jose, CA 95112, USA
[2]Stanford University, 450 Jane Stanford Way, Stanford, CA 94305, USA
[3]University of Houston, Houston, TX 77204, USA

Abstract

This study examines how a learning progression (LP) framework can adapt LLMs for formative assessment of open-ended math-science sensemaking responses. We developed LP-aligned prompts for scoring and feedback generation and evaluated LLM scoring performance on 191 student responses and feedback quality on a stratified subsample. GPT-5.4 achieved 85.9% agreement and a weighted kappa of 0.79 with human scores. Feedback evaluation indicated that 95% of generated feedback instances satisfied all five pedagogical quality criteria. Findings demonstrate the promise of grounding LLM-based formative assessment in LP frameworks.

Keywords

Learning progression, Formative feedback, Automated scoring, Large language models, STEM education

1. Introduction

Compared to multiple-choice items, constructed-response tasks offer a richer window into how students reason and make sense of complex ideas [1]. Yet the labor involved in evaluating open-ended responses and generating timely, individualized feedback makes formative assessment at scale a persistent challenge [2, 3]. Recent advances in large language models (LLMs) offer hope for addressing this challenge [4, 5]. LLMs can automate formative assessment, scoring student responses and generating individualized feedback [6, 7, 8]. However, this requires evidence that LLM-generated scores align closely with human judgment and that the feedback generated is pedagogically effective [9]. One central challenge in adapting LLMs for formative feedback is calibrating the level of support: providing sufficient scaffolding to move student thinking forward without over-scaffolding in ways that reveal answers or bypass productive struggle.

Learning progressions (LPs) offer a theory-driven approach for structuring both the scoring of student responses and the generation of feedback [10]. LPs describe ordered pathways of increasing sophistication in student understanding, providing a framework for locating a learner within a developmental trajectory and identifying the next step towards more sophisticated understanding [11]. By making intermediate levels of understanding explicit, LPs enable more precise diagnosis of student thinking and more targeted feedback. In the context of LLM-based systems, LPs can guide how models assign student responses to a corresponding LP level and generate feedback to advance their thinking toward the next level [12], thus grounding model behavior in learning theories and avoiding the model's default tendency to provide correct answers over scaffolding student thinking.

The present study examines how an LP framework can be used to adapt an LLM for formative assessment of constructed responses in a math–science sensemaking task. We focus on first scoring student responses by mapping them to levels within the LP, then generating feedback to advance student understanding to the next level. This work contributes to the design of effective LLM-based formative assessment and offers an approach to evaluating automated feedback that extends beyond accuracy to consider pedagogical effectiveness. This study is guided by the following research questions:

PEAF 2026, June 28, 2026, Seoul, South Korea
*Corresponding author.
karen.wang02@sjsu.edu (K. D. Wang); lijialin@stanford.edu (J. Li); cwieman@stanford.edu (C. Wieman); lkaldara@cougarnet.uh.edu (L. Kaldaras)

- **RQ1**: How well can an LLM score student open-ended responses to a math-science sensemaking question?
- **RQ2**: How well can an LLM generate LP-aligned formative feedback as measured by a set of pedagogical quality criteria?

2. Methods

2.1. Assessment of Math-Science Sensemaking

This study focuses on a constructed-response item designed to measure students' blended math–science sensemaking competency, defined as the integration of mathematical and scientific thinking to interpret quantitative relationships in scientific phenomena [13, 14]. The question presents a real-world scenario and asks students to propose and justify a mathematical relationship between the key variables. All scoring and feedback results reported in this study are derived from this item.

> A group of friends set up a hammock with two springs. They noticed that doubling the weight on the hammock caused the springs to stretch twice as far, and that a new set of springs stretched more but showed the same pattern. **Can you suggest a mathematical relationship between the variables in this problem that could explain why the new set of springs stretches more than the original springs? Justify your answer.**

We collected a dataset of 191 student responses from an introductory physics course at an R1 university in the US. Student responses ranged from 1 to 137 words with a median of 38 words. In our previous work, we have developed and validated the LP framework for math-science sensemaking with three increasingly sophisticated levels [14]:

- **Level 1**: Identification of relevant variables and qualitative patterns in a scientific phenomenon
- **Level 2**: Quantitative representations of the relationship between variables
- **Level 3**: Conceptual explanations that connect mathematical equations to underlying mechanisms

Using a rubric aligned with this LP framework, two authors (J.L. and L.K.) collaboratively scored all 191 student responses. Several rubric adaptations were made during the scoring process to accommodate conceptually accurate responses that used non-standard language or terminology. Disagreements were resolved through discussion, yielding a final inter-rater reliability of 0.95.

2.2. LLM Prompts for Scoring and Feedback Generation

To automate scoring and feedback generation, we implemented a Python script that called the GPT-5.4 model (`gpt-5_4-2026-03-05`) API and adapted the model for each task through instructional prompts. The scoring prompt followed established best prompt engineering practices, including clearly defined role and goal specification, chain-of-thought instructions directing the model to evaluate student responses step-by-step against each LP level [15, 16], and out-of-distribution illustrative examples for each level [17]. The prompt also specified a structured JSON output format containing a score and a brief justification for each scoring decision.

The feedback prompt extended the scoring prompt by adding level-specific feedback generation guidelines. These guidelines directed the model to: (1) acknowledge what the student demonstrated at their current LP level, (2) identify the specific element(s) needed to advance to the next level, (3) provide a guiding question or hint that scaffolds thinking toward the next level without revealing the answer, and (4) use an encouraging, growth-oriented tone. Crucially, the guidelines also instructed the model to mirror students' own language and avoid introducing terminology or variables not present in students' responses. Feedback was constrained to 2–3 sentences. Figure 1 illustrates the scoring and feedback generation workflow.

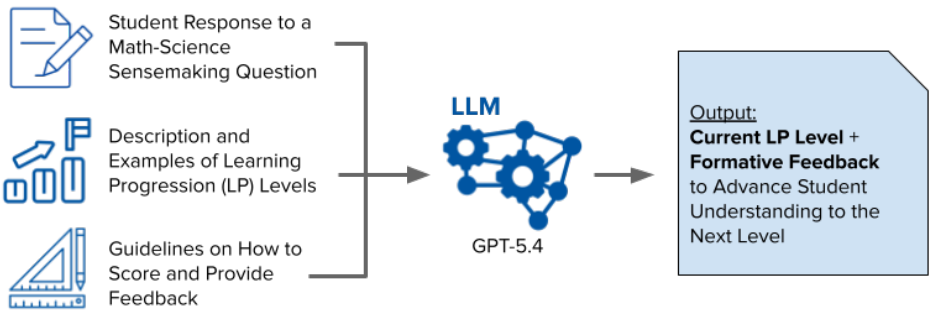

Figure 1: Overview of the LLM-based scoring and feedback generation workflow

2.3. Evaluation of LLM Scoring and Feedback

To address RQ1, we evaluated LLM scoring performance for all 191 student responses using two quantitative metrics: (1) raw agreement between LLM-assigned and human scores, and (2) weighted Cohen's kappa, which accounts for chance agreement and the ordinal nature of LP levels [18]. Cases where LLM and human scores diverged were reviewed to identify and categorize the underlying causes of LLM scoring inaccuracies. To address RQ2, we evaluated LLM-generated feedback on a stratified sample of 39 (20%) responses across five dimensions: (1) content accuracy, (2) supportive tone, (3) LP-aligned scaffolding, (4) specificity to student language, and (5) clarity in addressing student mistakes (if applicable). Each dimension was scored on a binary scale (0 or 1). Feedback evaluation was conducted by one author (J.L.) who was not directly involved in prompt design.

3. Results

3.1. LLM Scoring Performance and Error Analysis

GPT-5.4 achieved a raw agreement of 86% and a linear weighted Cohen's kappa of 0.79 with human scores, indicating substantial agreement. Among the 27 mis-scored cases, the model assigned a lower level than human raters in 16 cases and a higher level in 11. The majority of mis-scores were due to the model's difficulty in accurately evaluating non-standard language and terminology in student responses, a challenge human raters navigated more reliably and flexibly. Table 1 presents illustrative examples of mis-scored cases.

Table 1
Examples of LLM mis-scored cases

Student Response	Human Level	GPT-5.4 Level	GPT-5.4 Scoring Rationale
"Spring distance is directly proportional to weight, sW = D (W is weight, s is some scalar representing the stretch of the spring, D is distance stretched)."	3	2	Includes a quantitative description and identifies key variables and a coefficient term, but the equation is not valid because it does not clearly relate weight and stretch with the coefficient in a correct form. *[Underscoring: did not give credit for non-standard language]*
"There is no mathematical relationship that explains why the new springs stretched more. The springs must have been different themselves, the ratio of the springs stretching remain the same."	1	2	Gives a quantitative description by stating the ratio of stretching remains the same, implying a proportional pattern, but does not provide a valid equation. *[Overscoring: gave credit for implied rather than stated understanding]*

3.2. LLM Feedback Quality

LLM-generated feedback demonstrated strong performance across the five evaluation dimensions, with 95% (37/39) of feedback instances received full scores. Table 2 presents representative examples of LLM-generated feedback across LP levels. The two cases that did not receive full scores each revealed a distinct limitation. In the first, the feedback failed the LP-aligned scaffolding dimension by introducing the correct answer in a guiding question. In the second, the feedback failure was downstream of a scoring error: because the model underscored the student's response due to non-standard language, the feedback called on the student to demonstrate understanding they had already expressed.

Table 2
Examples of LLM-generated formative feedback across LP levels

Student Response	**GPT-5.4 Level**	**GPT-5.4 Feedback [Dimensions Met]**
"$y = x^x$: y indicates the spring stretch and x indicates the weight of the hammock. The stretch is dependent on the amount of weight enforced on the hammock which depends on how much force is put upon it."	1	You identified the two important parts of the situation: the spring stretch and the weight, and you explained that the stretch depends on how much force is put on the hammock. Take another look at your relationship and describe the pattern more quantitatively: when the weight changes, how does the stretch change in relation to it? [CA - 1, ST - 1, LPS - 1, SSL - 1, CSM - 1]
"When weight is added, the string stretches proportionally to the weight of the person in the hammock. The weaker the spring, the more it must stretch to accommodate the weight of the person as the tension force it exerts on the hammock is weaker."	2	You clearly connected the weight in the hammock to how much the spring stretches, and you also explained that a weaker spring stretches more. To move up, take another look at turning your idea that the stretch is "proportional to the weight" into a mathematical relationship that also includes the spring's strength. [CA - 1, ST - 1, LPS - 1, SSL - 1, CSM - N/A]
"The mathematical expression (Hooke's Law) can be expressed as F = kx. Where F is force, x is the displacement of the spring position and k is the stiffness of the spring. The stretching of the spring is directly proportional to the weight applied to it. If the weight is doubled, the displacement (x) is doubled as well."	3	You clearly stated Hooke's Law as F = kx and explained that the stretching is directly proportional to the weight applied, which shows strong math-science sensemaking. You do not need to revise your response; for extension, consider how changing the stiffness of the spring in your equation could help predict why a new set of springs stretches more than the original. [CA - 1, ST - 1, LPS - 1, SSL - 1, CSM - N/A]

CA: content accuracy; ST: supportive tone; LPS: Learning Progression-aligned scaffolding;
SSL: specificity to student language; CSM: clarity in addressing student mistakes (if applicable)

4. Discussion

This study demonstrates that LLMs can be adapted for formative assessment of higher-order competencies such as math-science sensemaking in STEM education. The results suggest that LLMs can locate student responses within a developmental progression, with interpretation of non-standard language being the main limitation. With explicit LP-aligned guidelines, the model also produced feedback that was largely accurate, responsive to student language, and appropriately scaffolded to support student progression towards the next LP level. However, both scoring accuracy and feedback quality required iterative prompt refinement to achieve acceptable performance on a single assessment item, and the stochastic nature of LLM outputs is a practical limitation. Our next steps include a student study using an attempt-feedback-revision design to assess the effectiveness of LP-aligned automated feedback.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude for proofreading and formatting assistance. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] R. Glaser, N. Chudowsky, J. W. Pellegrino, Knowing what students know: The science and design of educational assessment, National Academies Press, 2001.

[2] V. J. Shute, Focus on formative feedback, Review of educational research 78 (2008) 153–189.

[3] D. J. Nicol, D. Macfarlane-Dick, Formative assessment and self-regulated learning: A model and seven principles of good feedback practice, Studies in higher education 31 (2006) 199–218.

[4] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, A. R. Srinivasa, Automatic assessment of text-based responses in post-secondary education: A systematic review, Computers and Education: Artificial Intelligence 6 (2024) 100206.

[5] C. Narreddy, S. Joordens, S. Prompiengchai, Harnessing large language models for scalable and effective formative assessment in higher education: A review, Trends in Higher Education 4 (2025) 65.

[6] G.-G. Lee, E. Latif, X. Wu, N. Liu, X. Zhai, Applying large language models and chain-of-thought for automatic scoring, Computers and Education: Artificial Intelligence 6 (2024) 100213.

[7] C. Impey, M. Wenger, N. Garuda, S. Golchin, S. Stamer, Using large language models for automated grading of student writing about science, International Journal of Artificial Intelligence in Education (2025) 1–35.

[8] W. Xie, J. Niu, C. J. Xue, N. Guan, Grade like a human: Rethinking automated assessment with large language models, in: Proceedings of the International Conference on Research in Adaptive and Convergent Systems, 2025, pp. 1–8.

[9] L. Kaldaras, H. O. Akaeze, M. D. Reckase, Developing valid assessments in the era of generative artificial intelligence, in: Frontiers in education, volume 9, Frontiers Media SA, 2024, p. 1399377.

[10] R. Duschl, S. Maeng, A. Sezen, Learning progressions and teaching sequences: A review and analysis, Studies in Science Education 47 (2011) 123–182.

[11] L. Kaldaras, J. Krajcik, Development and validation of knowledge-in-use learning progressions, in: Handbook of research on science learning progressions, Routledge, 2024, pp. 70–87.

[12] L. Kaldaras, K. Haudek, J. Krajcik, Employing automatic analysis tools aligned to learning progressions to assess knowledge application and support learning in stem, International Journal of STEM Education 11 (2024) 57.

[13] F. Zhao, A. Schuchardt, Development of the sci-math sensemaking framework: Categorizing sensemaking of mathematical equations in science, International Journal of STEM Education 8 (2021) 10.

[14] L. Kaldaras, C. Wieman, Cognitive framework for blended mathematical sensemaking in science, International Journal of STEM Education 10 (2023) 18.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[16] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[18] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174.