



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

Karen D. Wang, Jialin Li, Carl Wieman, Leonora Kaldaras

PEAF Workshop

Jun 28, 2026

-
1. What to Assess: **Blended Math-Science Sensemaking**
 2. How to Assess: **Adapting an LLM to Score Students' Short Responses Using a Learning Progression Framework**
 3. Results i: **LLM Scoring Performance**
 4. Results ii: **LLM Feedback Performance**

Agenda

Blended Math-Science Sensemaking

Definition

The **integration** of **mathematical** and **scientific thinking** to interpret **quantitative relationships** in scientific phenomena



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.

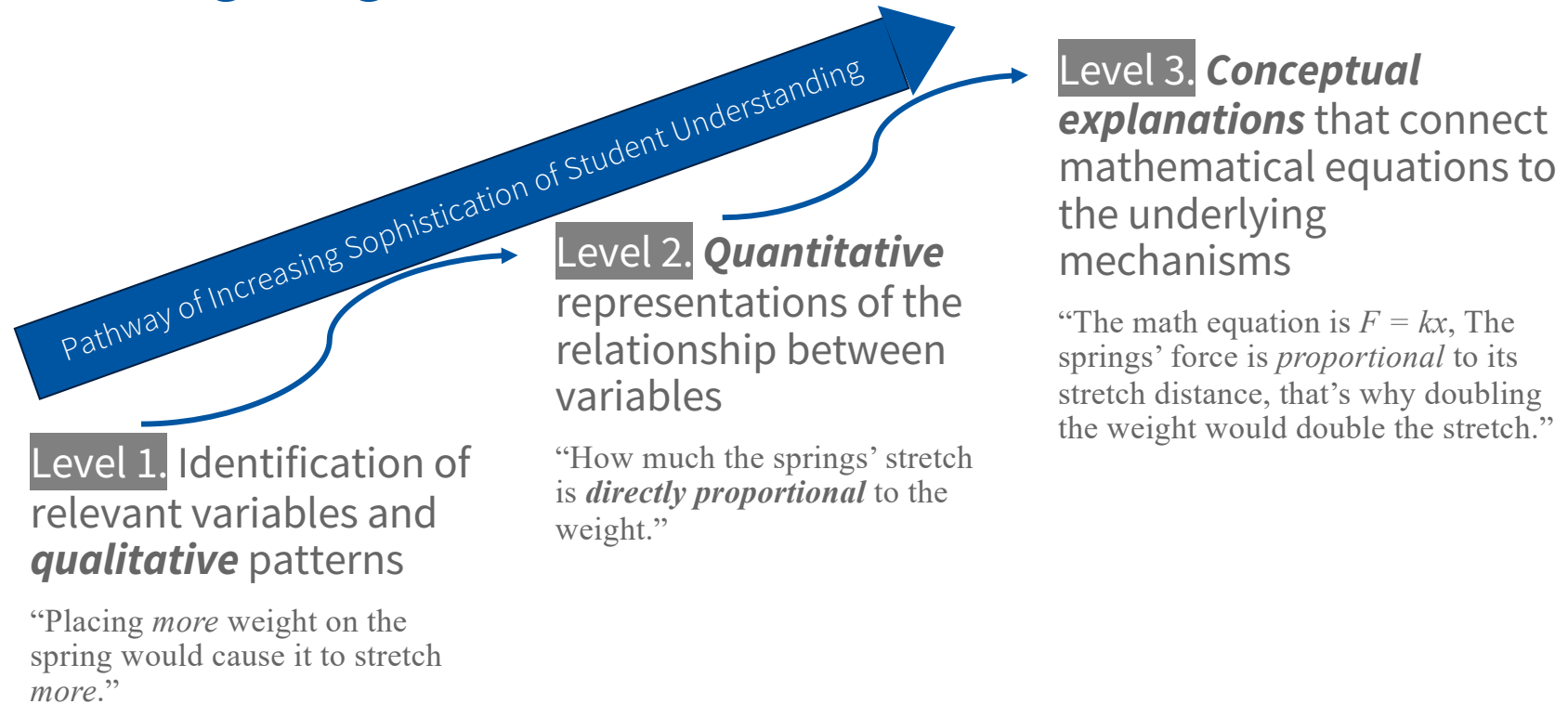
L. Kaldaras, C. Wieman, Cognitive framework for blended mathematical sensemaking in science, International Journal of STEM Education 10 (2023) 18.

An Item Assessing Math-Science Sensemaking

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.**

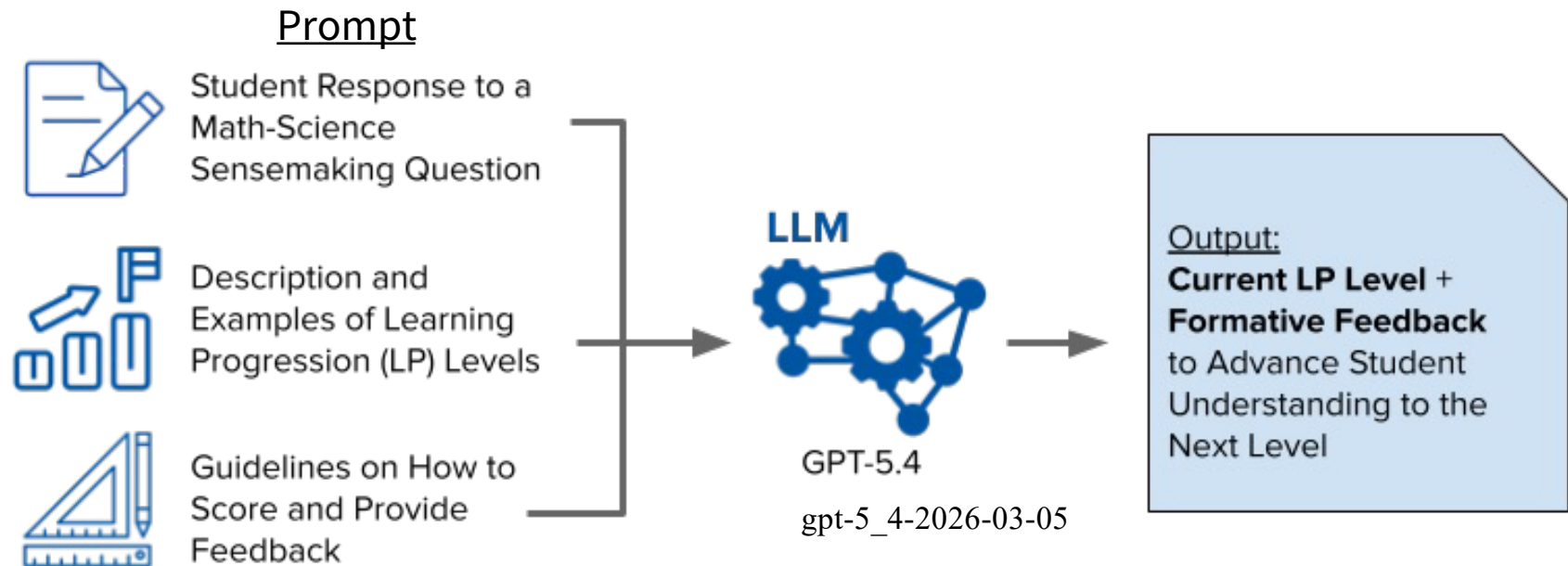


Learning Progression Framework



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

Scoring Process



LLM Performance

- Scored 191 Student Responses; Scoring Accuracy: 86%; Kappa: 0.79
- Main Scoring Challenge: Students' conceptually accurate yet non-standard language

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. <i>[Underscoring: did not give credit for non-standard language]</i>
"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. <i>[Overscoring: gave credit for implied rather than stated understanding]</i>

LLM Performance

- LLM Provided Feedback to a Stratified Sample of 39 Student Responses
- 37/39 LLM-generated Feedback Satisfied All Five Criteria

Feedback Evaluation Criteria

5. Clarity in Addressing Student Mistakes

4. Specificity to Student Language

3. Learning Progression-aligned Scaffolding

2. Supportive Tone

1. Content Accuracy