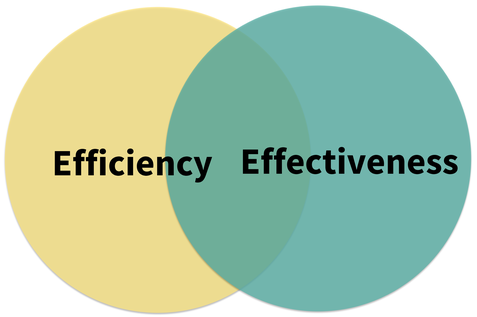
Harnessing the Potential of LLMs in STEM Education

Karen D. Wang[1], Lora Kaldaras[1], Shima Salehi[1], Carl Wieman[1,2],
[1] Graduate School of Education, [2] Department of Physics, Stanford University

HAI FIVE
Celebrating 5 Years of Impact

A Framework for Conceptualizing and Evaluating Large Language Model (LLM) Applications in Education



Efficiency

LLM applications reduce the time required to complete tasks for teachers or students

Effectiveness

LLM applications improve the quality of teaching and learning

Efficiency — Applying LLMs to Efficiently Assess Student Math-Science Sensemaking Competency

Blended math-science sensemaking

The cognitive process of expressing scientific concepts mathematically and integrating mathematical and scientific reasoning to understand phenomena (Kaldaras & Wieman, 2023)

Study Design

- College students (n = 204) in an introductory physics course
- Submitted short written responses to explain the relationship between the weight on a hammock and the stretch distance of the springs supporting the hammock
- Two human coders scored all responses based on a four-level rubric

Sample Student Responses

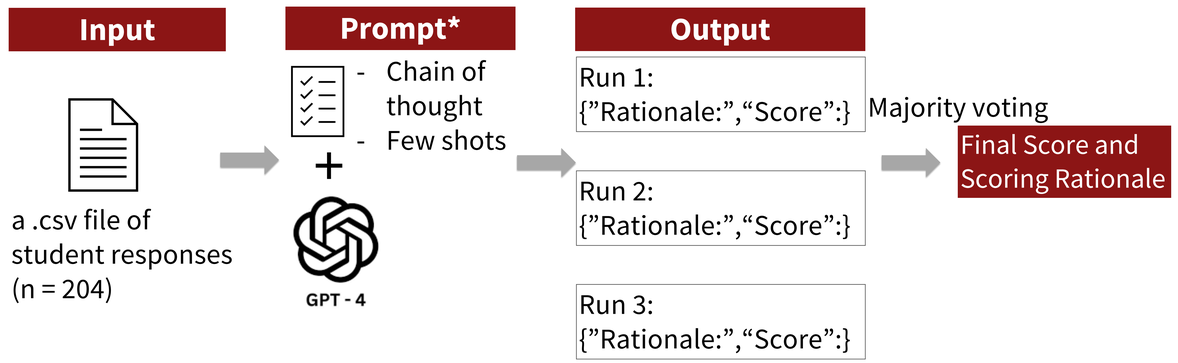
Level 3
"The distance in which the spring stretches is directly related to the spring force due to a pulling force such as the weight of an object. F = stretch distance x constant."

Level 2
"The distance the string stretches is directly proportional to weight."

Level 1
"The weight of the hammock is now a lot heavier than original and so it pulls on the springs with more force because of gravity while the hammock is heavier."

Level 0
"x+y=weight"

LLM Scoring Process



*Final prompt includes step-by-step evaluation criteria, scoring logic, and 11 out-of-distribution examples with scores and rationales.

Results

Human-AI Agreement
GPT-4 achieved substantial agreement with human coders in scoring student responses: **Percent agreement: 0.81; Cohen's Kappa: 0.60**

Self-consistency
GPT-4's consistency across three scoring runs was associated with higher agreement with human coders. For the subset where GPT-4 scored consistently (n = 151): **Percent agreement: 0.87; Cohen's Kappa: 0.69**

Error Analysis
GPT-4 struggled with accurately scoring student responses that contain non-standard language/terminology

Next Steps and Implications

Efficiency

- Refine and apply the prompt to score student responses for different questions measuring math-science sensemaking at scale

Effectiveness

- Use the scoring rationale as basis for GPT-4 to provide just-in-time, formative feedback to students to improve their math-science sensemaking competency
- Apply LLMs to tutor students on how to solve real-world physics problems

Both efficiency and effectiveness are important goals and require thoughtfully combining the latest innovations in learning sciences research and technology to achieve.