Developing a Generative AI Framework for Analyzing Student Responses to Enhance Classroom Assessments

CREATE for STEM Institute

Namsoo Shin, Yue Xing, Xunlei Qian, and Joseph Krajcik

CREATE for STEM Institute, Michigan State University

MICHIGAN STATE UNIVERSITY

OBJECTIVE

Goal of the Study

This study aims to develop a GenAI-enhanced assessment framework to (1) analyze students' written responses, (2) generate rationale for scoring decisions and identify response uncertainty, and (3) assign final scores and provide feedback on weaknesses to support learning.

Research Question

This study explores two research questions (RQs): **RQ1**, "How can AI algorithms, leveraging Large Language Models (LLMs) and grounded in learning theories, be developed to analyze student responses and identify weakness in written responses for knowledge-in-use assessment tasks?" **RQ2**, "How can uncertainty in student responses be identified to provide feedback on knowledge-in-use assessment tasks?"

CONCEPTUAL FRAMEWORK

Knowledge-In-Use

- Knowledge-in-use refers to students applying their knowledge with scientific practices to solve complex problems.
- Classroom assessments provide tangible evidence of this knowledge, enabling feedback to support learning.

Feedback in Learning

- Timely, targeted feedback fosters cognitive growth by engaging learners to reflect on their knowledge, identify weaknesses, seek additional information, and revise their work.
- Effective feedback should be based on analyzing student responses, making judgments about student performance, and providing actionable, constructive feedback on areas for improvement to foster ongoing learning and growth.

Generative Artificial Intelligence

- Large Language Models (LLMs), facilitate automated analysis of written responses in assessment tasks to provide meaningful feedback.
- Chain-of-Thought (CoT) prompting enhances LLMs to generate personalized guidance that encourages learners to iteratively refine their answers to achieve their learning goals.

METHODS: AI Assessment Framework

Our AI models used Distilled Versions of GPT-4o-mini to provide the results of precision, recall, and F1-score to evaluate the accuracy by analyzing our AI model scoring predictions.

- Use 3D assessment item: Measure students' learning by integrating the scientific knowledge: disciplinary core ideas (DCI), crosscutting concepts (CCC), and science and engineering practices (SEP).
- **For RQ1**, Our CoT approach involves (1) dissecting responses into critical components, (2) checking logical connections and alignment with the assessment question, (3) identifying possible incorrect statements, and (4) providing an overall assessment with scoring rationale. Our prompting structure integrates "Instruction + Rubric + Examples of responses (5 examples) with CoT." This AI model analyzes responses as "0" for incorrect and "1" for correct answers while also generating scoring rationale.
- **For RQ2**, Four distinct models by varying the prompts. First, we adjusted the number of examples included in some prompts. Second, we modified the instructions within the prompts to address two factors: (1) whether a score is assigned when the student provides an intuitively correct but not explicitly articulated response, and (2) whether the LLM focuses more on the structural analysis. The final score is determined by the majority vote of the four models (Wang, et al., 2022). If at least three of the four models assign a score 1, the final score is 1. For example, if the model scores are 1, 0, 0, and 0, the final score for the case is 0. This approach also quantifies and identifies uncertainties in student responses when the models scores differ. In the above example, 25% of the student response is considered ambiguous.

EXAMPLE OF PROMPTS AND RUBRIC SYSTEM

The rubric contains three components, including (1) DCI, (2) DCI and SEP, and (3) DCI and CCC to analyze the understanding of knowledge-in-use.

Instruction: *You are a professional grader to grade K-12 students responses. Below are the learning goal, question, exemplary response, rubric, and sample gradings:*

Learning Goal

Students use a model to explain that in a chemical reaction, atoms are regrouped and why mass is conserved.

Question

When heated, hydrogen peroxide breaks down and becomes water and oxygen gas. The image represents what happens to hydrogen peroxide during the chemical reaction.
For the questions below, use the model to help explain what occurs with the atoms that make up hydrogen peroxide, water, and oxygen gas. Be sure to include the number and/or types of atoms for each molecule before and after the reaction.
1) Explain how water and oxygen gas are formed.
2) Explain how the model shows that the mass of hydrogen peroxide and the combined mass of water and oxygen are conserved.

Exemplary Response

1. Water and oxygen are formed from heated hydrogen peroxide.
2. The masses were conserved because in a chemical reaction, the same materials are in it.

Analytic Rubric

- E1. Disciplinary Core Ideas and Scientific Practices: *Student states water and oxygen gas (or they) are produced by (the chemical reaction of) heating hydrogen peroxide. Note: don't need to have "chemical reaction" because the prompt has the information.*
- E2. Disciplinary Core Ideas: *Student indicates the atoms in hydrogen peroxide molecules (or reactants) are rearranged (or became, re-organized, reformed) into water molecules and oxygen molecules (or products). Note: should have "atoms are regrouped or reformed" and "to water and oxygen gas."*
- E3. Disciplinary Core Ideas and Cross-Cutting Concept: *Student indicates that 1) the total number of each type of atom before and after the reaction is the same, 2) so the mass stays the same. Note: should have "the same number before and after" and "the mass stays the same".*

Below is an example of the scoring reasoning procedure:

To grade E1, following the rubric,
does the student states water and oxygen gas (or they) are produced by (the chemical reaction of) heating hydrogen peroxide.
Since the student writes "Water and oxygen are formed from heated hydrogen peroxide", E1.2 is satisfies.
Score: 1

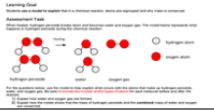
To grade E2, following the rubric,
does the student indicates the atoms in hydrogen peroxide molecules (or reactants) are rearranged (or became, re-organized, reformed) into water molecules and oxygen molecules (or products)?
Since the student just mentions water and oxygen are formed from heated hydrogen peroxide without details, E2 is satisfied.
Score: 0

To grade E3, following the rubric,
does the student indicates that 1) the total number of each type of atom before and after the reaction is the same, 2) so the mass stays the same
Since the student does not mention any numbers, E4 is satisfied.
Score: 0

Given the above reasoning step, grade the following response:

The hydrogen peroxide, water and the oxygen have the same number of atoms. When the hydrogen peroxide is heated the product is the same to the reactant. hydrogen peroxide is heated and breaks down making water and oxygen gas.

EXAMPLE OF AI ANALYSIS AND SCORING RATIONALE

Assessment Tasks	
	**Questions:** Explain (1) how water and oxygen gas are formed; (2) how the model shows that the mass of hydrogen peroxide and the combined mass of water and oxygen are conserved
	Student Response: *oxygen is really just tiny molicules of water so it is conserved all around the world oxygen and gas are formed when hydrogen peroxide heats into water and when water heats it becomes gas*
Rubric DCI (chemical reaction) + SP (use a models): In the model, water and oxygen gas are produced by heating hydrogen peroxide.	
Score: *Incorrect;* **Ambiguity:** *25 % unclear;* **Reason:** *You state that oxygen and gas are formed when hydrogen peroxide heats into water, but this statement is unclear and partially incorrect.* **Feedback:** *Your response needs some corrections to describe the chemical reactions involving hydrogen peroxide and water.*	

RESULTS AND DISCUSSION

- This proposal reports the results of 834 student-written responses from one of the five Next Generation Science Assessment tasks, achieving an inter-rater reliability of 0.80.
- The **RQ1** results indicate that our proposed model for analyzing student responses and generating scoring rationale achieved the following accuracy: All (0.83), DCI&SEP (0.87), DCI (0.78), and DCI&CCC (0.84).
- The **RQ2** results, using the majority vote method to identify ambiguity in student responses, show that our model increased accuracy across all rubrics: All (0.85), DCI&SP (0.88), DCI (0.82), and DCI&CCC (0.86). Table 1 presents the results of three components.
- We reviewed the students' responses from the ambiguous group. These responses were unclear and inconsistent between sentences in students' entire paragraphs, leading to low reliability in scoring for humans and AI.
- We propose using this uncertainty information to provide feedback to students to warn them about the ambiguity in the responses rather than AI to analyze responses with a high risk of inaccuracy.

	Method	Accuracy	False Negative	False Positive
ALL	RQ1	0.83	83 cases	**52 cases**
	RQ2	0.85	77 cases	**47 cases**
DCI	RQ1	0.78	114 cases	**70 cases**
	RQ2	0.82	83 cases	**64 cases**
DCI&CCC	RQ1	0.84	40 cases	**74 cases**
	RQ2	0.86	51 cases	**74 cases**
DCI&SEP	RQ1	0.87	96 cases	**12 cases**
	RQ2	0.88	96 cases	**3 cases**

CONCLUSION AND FUTURE DIRECTIONS

- Our GenAI-enhanced assessment framework demonstrates the potential of LLMs to analyze written responses, generate scoring rationale, identify uncertainty, and generate feedback that enhances learning.
- We are designing classroom studies to collect and analyze new response data with our AI models, exploring how timely feedback influences learning across various student characteristics (e.g., performance levels, ethnicity, school location, and socioeconomic status).
- This AI Assessment Framework is being used to analyze elementary students' written responses to Knowledge-in-Use science assessments, providing actionable and constructive feedback to support the learning of 64,000 students across 80 teachers.
- Our framework and methods can be applied to other disciplines that use classroom assessment for teaching and learning, as well as other constructs like motivation, metacognition, and collaboration skills, which are influenced by causal variables that help diagnose students' behaviors.

REFERENCES

- Next Generation Science Standards (NGSS). Lead States. (2013). Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press.
- Next-generation science assessment (NGSA) (2023). https://ngssassessment.portal.concord.org
- Nicol, David J., & Macfarlane-Dick, D.. "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice." Studies in higher education 31, no. 2 (2006): 199-218.
- Pellegrino, J. W., & Hilton, M. L. (Eds.) (2012). Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. National Research Council of the National Academes. Washington DC: The National Academies Press.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.