# Providing Automated Feedback on Formative Science Assessments: Uses of Multimodal Large Language Models

Ha Nguyen
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
ha.nguyen@unc.edu

Saerok Park
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
saerok@unc.edu

## Abstract

Formative assessment in science education often involves multimodality and combines textual and visual representations. We evaluate the capacity of multimodal large language models (MLLMs), including Anthropic's Claude 3.5 Sonnet, Google's Gemini 1.5 Flash, and OpenAI's GPT-4o and GPT-4 Turbo, to score and provide feedback on multimodal science assessments. Overall, the MLLMs can accurately transcribe students' hand-written text. The best performing models (Claude and GPT4-o) show moderate to substantial agreement with human evaluators in assessing students' scientific reasoning. MLLMs provided with example responses, scores, and explanations (few-shot learning) generally perform better than those without examples (zero-shot learning). Thematic analysis reveals cases where the models misevaluate the depth in students' answers, add details not included in the input (i.e., hallucinate), or show incorrect numerical reasoning. Findings demonstrate the feasibility of and considerations for using MLLMs to provide in-time feedback for science assessments. Such feedback can help to revise students' understanding and inform teachers' instructional practices.

## CCS Concepts

• **Computing methodologies** → **Natural language generation**;
• **Applied computing** → **Education**.

## Keywords

multimodal large language model, science assessment, automated evaluation

**ACM Reference Format:**
Ha Nguyen and Saerok Park. 2025. Providing Automated Feedback on Formative Science Assessments: Uses of Multimodal Large Language Models. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3706468.3706480

## 1 Introduction

Formative assessment—the process of eliciting students' understanding through tasks, interpreting answers, and responding to advance understanding [37]—is a common practice to help teachers adjust instructional practices. Formative assessment in science education often involves multimodality—combining textual, audio, and visual representations [3, 35]. For example, an assessment may invite students to draw the interactions within an ecosystem and provide reasoning in writing. Analyses that integrate multimodal responses yield a more comprehensive picture of students' science understanding than relying on single modalities [13, 31, 32]. Formative assessment is the most effective when it is coupled with in-time feedback. Such feedback can come in the form of rubric scores, written evaluations, follow-up opportunities for student discussion, and adjustment of learning activities [10]. Feedback should be aligned with learning objectives and detailed enough to guide students' refinement of science knowledge [4]. However, delivering tailored feedback to every student is time-consuming for teachers. Researchers have thus developed automated scoring and feedback that showed high agreement with humans and helped students refine scientific explanations [20, 22, 40, 43, 44].

While promising, computer-based science assessments rarely employ automated feedback for open-ended science questions [17]. This is because traditionally, developing automated scores and feedback relies on large corpora of hand-labeled data that are costly to collect [20, 22]. Multimodal large language models (MLLMs) present opportunities to generate automated evaluation of multimodal science assessments. The technology can effectively interpret handwritten text and reason about visual inputs such as graphs [3, 14, 42]. In this work, we investigate the feasibility of using emergent, multimodal AI to develop automated feedback in science classrooms. While related terms (e.g., auto-feedback, auto-scoring, AI-powered evaluation) exist, we use *automated feedback* because the MLMMs not only assign scores, but also explanations that serve as feedback for students. We ask the following questions:

**RQ1: How effective are MLLMs in transcribing and evaluating responses to formative science assessments?**

**RQ2: What are areas of improvement to align the automated feedback with human feedback?**

To answer these questions, we used MLLMs to (1) transcribe images of students' responses to two science assessments ($n$ submissions = 82), and (2) score the responses based on the models' task interpretation and students' answers. The data came from two learning units in sixth grade. We experimented with several MLLMs, including Anthropic's Claude 3.5 Sonnet, Google's Gemini 1.5 Flash, and OpenAI's GPT-4o and GPT-4 Turbo. For RQ1, we evaluated the extent to which the models accurately transcribed students' hand-written text and agreed with human evaluations. For RQ2, we performed thematic analysis to identify systemic errors in the generated output. Our research has key contributions to learning analytics research and science education. We demonstrate the feasibility of using MLLMs to provide at-scale feedback for open-ended,



*LAK 2025, March 03–07, 2025, Dublin, Ireland*

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706480

multimodal assessments. We experiment with different prompt conditions, including *few-shot* (providing example scores and explanations; [6]) and *zero-shot learning* (not providing examples). We highlight areas where MLLMs' evaluations are not well-aligned with human feedback as opportunities for future research.

## 2 Background

### 2.1 Multimodal Formative Assessments in Science Education

The formative assessment process includes three practices: eliciting, interpreting, and responding [37]. *Eliciting* involves using activities such as whole-class and group discussions, worksheets, and exit tickets to gather quick insights about student learning [1]. *Interpreting* such evidence consists of analyzing students' thinking and identifying actionable insights [10]. Finally, *responding* comes in the form of feedback to students or adjustment to instruction [11].

Assessment design can integrate multimodality and invite students to convey ideas through different modes, including writing, talking, drawing, visualizing data, and modeling [1, 10, 12]. Multimodal formative assessments provide opportunities to engage students in *disciplinary literacy*, defined as the ability to participate in "currently valued forms of disciplinary knowledge" [28] (p. 4) leveraging language and visual representations [35, 38]. Further, analyzing responses across visual, written, and verbal modes presents a more comprehensive view of learning than relying on single modalities [13, 29, 30, 32]. Students reveal conceptual understanding and language in their written and oral presentations [38, 41]. They visualize abstract phenomena and complex, dynamic relationships through drawing and modeling [21]. To illustrate, Grapin & Llosa [13] found that fifth-graders, particularly English learners, conveyed different aspects of their scientific ideas in modeling tasks through distinct modalities (writing and drawing).

### 2.2 Automated Evaluation of Formative Assessments

In-time feedback following formative assessments can deepen students' science understanding. To allow students to refine their thinking, the feedback needs to be detailed and aligned with learning objectives [4]. Researchers have leveraged machine learning to develop automated feedback systems for formative assessments, by training the systems on human scoring and high-quality responses [22, 24, 43]. Automated feedback can deepen scientific reasoning [20, 44] and reveal important snapshots of student learning to help teachers adjust instruction [23]. For example, Liu et al. [24] trained c-rater-ML using human scores of written science assessments. The automated scores showed high agreement (Cohen's $\kappa$ range .62-.90) and high correlation with human scores (Pearson's $r$ range .66-.91). Researchers have also explored the potential of automated feedback for multimodal assessments [31–34]. Smith et al. [32] proposed a framework for automated assessment in a multimodal learning task about magnetism. Their framework involved a model employing convolutional neural network to assess writing and a model using topology to evaluate drawings.

Notably, prior efforts that leverage machine learning often rely on hand-labeled human scoring that are expensive to collect and not always generalizable to other contexts. Additionally, misspellings and linguistic diversity in students' responses may reduce the accuracy of the automated evaluations [9, 24]. Multimodal large language models that are trained on extensive data corpora have demonstrated enhanced capacity to comprehend human language and visual representations [3, 7]. They offer opportunities to develop automated feedback scalable to different assessment contexts.

### 2.3 Multimodal Large Language Models

Large language models (LLMs) are AI applications capable of taking textual input to generate human-like text [6]. Multimodal LLMs (MLLMs; also referred to as Large Vision-Language Models) extend the capacity of LLMs to incorporate visual and textual inputs [7]. MLLMs can respond to text prompts and engage in multi-turn exchanges combining text and images [42]. These models have shown promising results in vision-language understanding tasks, including identifying emotions in images, synthesizing multiple images, and reasoning with mathematical and scientific knowledge based on visual cues [25, 42, 45]. In education settings, researchers have developed MLLMs to comprehend complex scientific diagrams and generate automated captions and analyses [14].

Despite these promising findings, MLLM-generated output is not on par with human performance [25]. MLLMs may fail to parse complex image details and struggle in complicated, mathematical reasoning [8]. They may hallucinate and describe elements that the input does not contain [15]. Prompt engineering, or systematically adjusting instruction for LLMs and MLLMs, may address these challenges and improve model performance [26]. Specifying the instructional task and feedback *context and rubric* can enhance the quality of the generated responses [19]. Researchers have prompted the models to engage in *chain of thought* and "think step by step", to articulate the reasoning paths leading to the output [39]. Additionally, *few-shot learning* [6], or providing example submissions and feedback, might also improve the feedback quality [19, 36]. We build on these different strategies—specifying context, chain of thought, and few-shot learning—to construct the prompts to MLLMs to provide feedback on multimodal science assessments.

## 3 Methods

### 3.1 Data Sources

Our analyses included two science assessments in separate learning units in sixth grade. We selected these assessments because they involved multimodality (graphs and text) and required the MLLMs to reason about both modalities to evaluate students' responses. Figure 1 showed example responses for the assessments. Both datasets had participants' consent to use de-identified data in future research. The first dataset, **Mulch**, consisted of identical pre-tests and post-tests collected from a science curriculum to study plant restoration in Spring 2020. The learning activities took place in a public school in the Southwestern United States. Students experimented with how different mulch conditions (no mulch, bark, or leaf mulch) related to soil moisture and decomposition rates. Our analyses focused on one test question that leveraged multimodality ($n$=37 scanned responses from paper submissions). The question provided students with graphs depicting changes in soil moisture

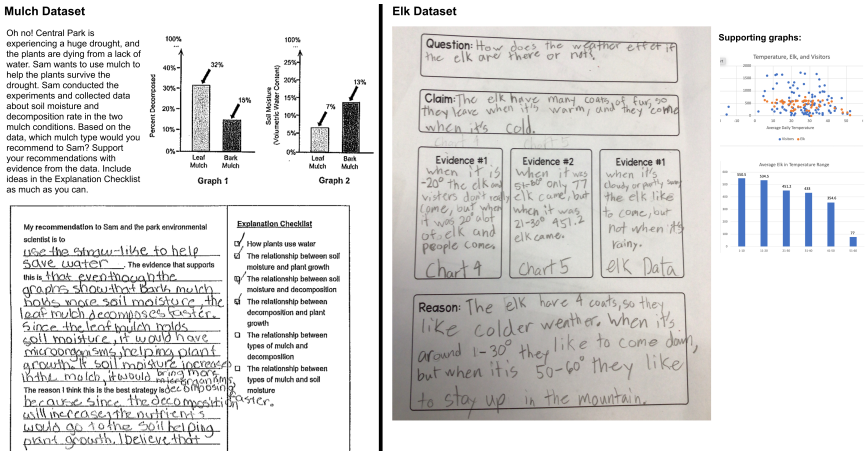


**Figure 1: Examples of the assessments: Mulch (left) and Elk dataset (right)**

and decomposition rates in the two mulch conditions (Figure 1). Students provided written recommendations to the state park about which mulch to use. Their responses were coded for three categories: *Element*, *Evidence*, and *Causal Coherence*. The rubric was developed with science teachers. Each category was scored on a scale of 1 to 3 for increasing levels of elaboration.

The second dataset, **Elk**, involved an in-class worksheet to examine the impact of human activities on local ecosystems ($n$=45 responses). The data were collected in Spring 2024 at a public charter school in the Western United States. Students were provided with several data sources, including articles and data linking outdoor recreation, temperature, and elk counts at a local ranch. They used Claim-Evidence-Reasoning [27] to articulate whether human activities and weather positively or negatively affected the elks, cite the evidence that supported their claims through graphing the data, and provide written explanations (Figure 1). Responses were scored using a Claim-Evidence-Reasoning rubric adapted from [2] and refined with teachers' feedback. The rubric scale was 0-2 for *Claim*, 0-3 for *Evidence*, and 0-2 for *Reasoning* based on increasing levels of elaboration and relevance to the science phenomenon. Human coders established substantial inter-rater agreement for each dataset based on 25% of the data. Cohen's $\kappa$ were .73, .92, and .88 for Element, Evidence, and Causal Coherence for the Mulch dataset and .82, .82, and 1 for Claim, Evidence, and Reasoning for the Elk dataset.

## 3.2 Procedures

We made API calls in August 2024 to several MLLMs that had shown state-of-the-art performance in multimodal reasoning tasks, including Claude 3.5 Sonnet, Gemini 1.5 Flash, GPT-4o and GPT-4 Turbo (max tokens=1024; temperature=1). For **RQ1**, each model was assessed for the (1) **transcription** of students' hand-written text and (2) **evaluation** of students' responses. For the first task, we reported the proportion of correctly transcribed words per student response. For the second task, we provided cleaned students' responses as input and compared the MLLM-generated scores and feedback with human scores. Because the rubric scores were ordinal, we reported weighted Cohen's $\kappa$ and Spearman's rank correlation coefficient ($\rho$). Weighted Cohen's $\kappa$ suggests the agreement strength: ≤0=poor, .01-.20-slight, .21-.40=fair, .41-.60=moderate, .61-.80=substantial, and .81-1=almost perfect [18]. Spearman's $\rho$ ranges between -1 and 1. Values closer to 1 (or -1) indicate stronger, positive (or negative) relationships and values closer to 0 denote weaker relationships.

We experimented with two prompt conditions: with and without examples (few-shot and zero-shot learning). Both conditions contained task descriptions and graphs, rubrics, and images of student responses. The few-shot condition also included example responses, scores, and explanations (Figure 2). LLMs with few-shot learning showed higher agreement with humans than zero-shot learning in generating scores and feedback [19, 36]. We generated three runs per MLLM and used majority voting to select the final scores. We observed some variation but substantial agreement (Fleiss' $\kappa$ range .67-.89) between runs. In total, we generated 888 evaluations (37 submissions*3 evaluations*4 models*2 prompt conditions) for the Mulch dataset and 1080 evaluations for the Elk dataset.

For **RQ2: areas to improve the feedback**, we performed thematic analysis [5] to examine cases where MLLMs disagreed with humans and models' explanations for giving specific scores. Two researchers familiar with the assessments and MLLMs' prompting performed open coding of the evaluations. The codes (e.g., accuracies/inaccuracies; underscoring) largely overlapped between coders. Coders then discussed emerging themes in two discussion rounds.

# 4 Results

## 4.1 RQ1: MLLMs' Ability to Provide Automated Feedback

*4.1.1 Transcription.* We first investigated the MLLMs' ability to interpret images of the assessments and demonstrate reasoning in answering them, as if they were students. The images included text and data visualizations (bar graphs and scatter plots; Figure 1). **All four MLLMs showed accurate and coherent reasoning in interpreting the text and graphs**. Consider the following response from GPT-4o. The response incorporated specific data points from the graphs.

**TASK & RUBRIC**

You are a science teacher who is evaluating students' responses based on the following rubric. Think **step by step** in your reasoning. + {rubric}

Example rubric:
**Elements**: Your output should follow the format: {Score: score of 0, 1, or 2; Reasoning: your reasoning}.
First, count the number of elements recognized by students. Then, return 0 if 0 or 1 factor is identified; return 1 if 2 to 4 factors are identified; and return 2 if 5 or more factors are identified. Examples of elements: human/social activities (fire, pollution, trash, weeding, fertilizing); biotic (insects, microorganisms, plant, mulch, seedling, native plant, weeds, animals); abiotic (gases like air, oxygen, CO2, water, energy/sun, weather, physical conditions); processes (competition, facilitation, invasion, water cycle, decomposition, photosynthesis).
**Evidence**: Your output should follow the format: {Score: score of 0, 1, or 2; Reasoning: your reasoning}.
Return a score of 0 if students are not using any evidence like common knowledge, data, information from the graphs, or scientific knowledge to support their claims. Return a score of 1 if students are using some forms of evidence and knowledge to support some of their claims. Return a score of 2 if the student refers to both graphs. They show a clear connection between claim and evidence.
**Depth and Causal Coherence**: Your output should follow the format: {Score: score of 0, 1, or 2; Reasoning: your reasoning}.
Return a score of 0 if the response does not contain any unobservable science ideas, and students are only making simple connections. Return a score of 1 if students are making connections between the observation and theoretical science ideas. Return a score of 2 if students show sophisticated connections between observable and unobservable science ideas, and there is a clear, coherent explanation of causal mechanisms.

**ASSESSMENT TASK**

Here is the task students are responding to + {task prompt}.

Here is the graphs included in the task + {images of graphs}.

**EXAMPLES (*only for few-shot learning*)**

Here are some examples:
Response 1: "The evidence that supports this is mulch".
Scores: 0 for Elements (mulch; 1 component); 0 for Evidence; 0 for Causal Coherence.
Reasoning: The student does not cite any evidence. The evidence is not linked to causal claims.

Response 2: "… moisture will be good for the plant during the drought. In graph 2, there is more soil moisture in the bark mulch".
Scores: 1 for Elements (moisture, plant, drought, bark mulch; 4 components); 1 for Evidence; 1 for Causal Coherence.
Reasoning: The student is citing one graph and explains the link between soil moisture and plant growth. The explanation can have more depth by adding connections to science knowledge or citing specific numbers from the graphs. [... Response 3 omitted to save space]

**STUDENT SUBMISSIONS**

Here is the student's submission + {image of student submission}.

**Figure 2: Prompt components to the MLLMs**

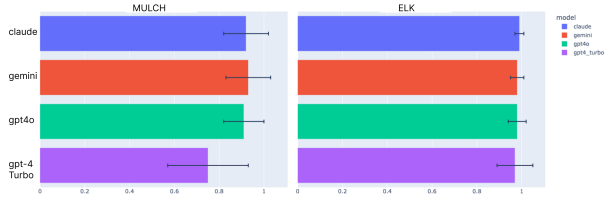


**Figure 3: Descriptive statistics of transcription accuracy (*M, SD*). Values closer to 1 indicate perfect transcription**

> My recommendation to Sam is to use leaf mulch. The evidence that supports this is that the data shows leaf mulch had a significantly higher decomposition rate at 32% compared to bark mulch at 15% (Graph 1), and it also maintained a higher soil moisture content with 7% for leaf mulch compared to 13% for bark mulch (Graph 2).[...]

Second, we explored how the MLLMs transcribed students' handwritten text (Figure 3). **The models accurately transcribed most text** for the Mulch dataset (range .75-.93) and almost all text for the Elk dataset (range .97-.99). The average proportions of accurately transcribed text for the Mulch dataset were: Claude = .92 (*SD* = .10), Gemini = .93 (*SD* = .10), GPT-4o = .91 (*SD*= .09), and GPT-4 Turbo = .75 (*SD* = .18). For Elk dataset: Claude = .99 (*SD* = .02), Gemini = .98 (*SD* = .03), GPT-4o = .98 (*SD* = .04), and GPT-4 Turbo = .97 (*SD* = .08). Responses with a large proportion of mistranscribed text tended to have students erasing and rewriting their answers, making the text harder to parse.

Analysis of variance (ANOVAs) showed significant differences between the MLLMs and the proportions of accurate transcription for the Mulch dataset, $F(3, 144) = 17.84$, $p < .001$. Post hoc comparisons using the Tukey HSD test indicated that the mean score for GPT-4 Turbo was significantly lower than the other models ($p < .001$). GPT-4 Turbo tended to claim a large proportion of text as illegible (e.g., "handwritten text obscured"). There was no significant difference between the MLLMs for the Elk dataset, $F(3, 176) = 1.29$, $p = .28$.

We observed five cases of hallucination (three for GPT-4 Turbo; 3.65%, two for Claude; 2.44%), where the MLLMs generated extended output that was not present in the visual input instead of stating that the text was illegible. For example, a Claude transcription reads "The evidence that supports this is bark mulch retains moisture the best. It keeps the soil damp and it helps decomposition. It survives the longest." This markedly differs from the actual text: "The evidence that supports this is plants water daily when the soil was dry the bark gives it water it needs to survive".

*4.1.2 Agreement with Human Feedback.* We calculated the inter-rater agreement (weighted Cohen's $\kappa$) and correlations between the MLLMs and human scores (Spearman's $\rho$; Table 1). We compared performance across MLLMs. For the Mulch dataset, **Claude 3.5 Sonnet and GPT-4o were generally the best performing**. They showed consistent, moderate to substantial inter-rater agreement across rubric categories in the few-shot condition; Claude 3.5: weighted Cohen's $\kappa$ for Element = .69, Evidence = .71, Causal Coherence = .73; GPT-4o: Element = .72, Evidence = .57, Causal Coherence = .84. For the Elk dataset, **Claude** was the best performing; weighted Cohen's $\kappa$ for Claim, Evidence, and Reasoning was .73, .74, .63 in the zero-shot condition and .73, .69, .68 in the few-shot condition, with significant correlations with human evaluations ($p < .001$). For each model, we generally observed **higher inter-rater**

**Table 1: Inter-rater Agreement with Human Scores: Weighted Cohen's $\kappa$ and Spearman's $\rho$**

| | Zero-shot | | | | | | | | Few-shot | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Claude | | Gemini | | GPT4o | | GPT4turbo | | Claude | | Gemini | | GPT4o | | GPT4turbo | |
| | $\kappa$ | $\rho$ | $\kappa$ | $\rho$ | $\kappa$ | $\rho$ | $\kappa$ | $\rho$ | $\kappa$ | $\rho$ | $\kappa$ | $\rho$ | $\kappa$ | $\rho$ | $\kappa$ | $\rho$ |
| Mulch dataset | | | | | | | | | | | | | | | | |
| Element | .51 | .63*** | .19 | .34* | .60 | .73*** | .16 | .28 | .69 | .78*** | .41 | .59*** | **.72** | **.78*** ** | .52 | .60*** |
| Evidence | **.75** | **.80*** ** | .34 | .60*** | .62 | .73*** | .39 | .64*** | .71 | .75*** | .40 | .63*** | .57 | .66*** | .46 | .55*** |
| Coherence | .48 | .56*** | .28 | .41* | .78 | .78*** | .63 | .68*** | .73 | .73*** | .38 | .48** | **.84** | **.84*** ** | .56 | .63*** |
| Elk dataset | | | | | | | | | | | | | | | | |
| Claim | **.73** | **.75*** ** | .30 | .38* | .43 | .46** | .44 | .47** | **.73** | **.73*** ** | .37 | .57*** | .58 | .59*** | .32 | .35* |
| Evidence | **.74** | **.78*** ** | .24 | .62*** | .52 | .64*** | .48 | .56*** | .69 | .78*** | .28 | .69*** | .67 | .77*** | .54 | .65*** |
| Reasoning | .63 | .65*** | .11 | .28 | .62 | .64*** | .17 | .23 | **.68** | **.71*** ** | .04 | .36* | .43 | .45** | .42 | .49*** |

*Notes*: ***$p$<.001, **$p$<.01, *$p$<.05. **Bold**=best performing for each rubric category.

**agreement with human coders in few-shot learning**, when the models were provided with example scores and explanations, compared to zero-shot learning. For instance, for the Mulch dataset, the weighted Cohen's $\kappa$ for Gemini was .19, .34, and .28 in the zero-shot condition and .41, .40, and .38 in the few-shot condition.

## 4.2 RQ2: Areas of Improvement for MLLMs' Feedback

To improve the feedback, we examined areas of disagreement between MLLMs' and human evaluations for each rubric category. An area of discrepancy was **how the models assessed the depth in student understanding**, compared to humans. In some cases, the MLLMs assigned higher scores than humans when evaluating Evidence. MLLMs interpreted the quotation marks ("") that students added as strong evidence, even when students did not specify or misspecified their sources. In other cases, MLLMs assigned lower scores for Evidence, Causal Coherence, and Reasoning than human evaluators. For instance, for the Elk dataset, 13 submissions that were given a 2 by human evaluators received a 1 from Gemini for Evidence (28.89% of submissions). Consider the student submission in Figure 4. The human evaluator gave the student a 3 for Evidence and 2 for Reasoning, because the student clearly identified evidence from each article and their reasoning was coherently linked to the claim and evidence. In comparison, Gemini and Claude claimed that the response lacked depth. Human evaluators had contextual knowledge of the lessons and students' science understanding. Their scores reflected considerations of the whole-class response patterns. Meanwhile, the MLLMs were given one student response at a time and might not possess this contextual knowledge.

We observed **incorrect reasoning by introducing details not included** in the tasks, rubrics, and student responses (hallucination) [25]. For instance, in evaluating the student's claim "Human activities impact both the elk and humans in a *negative* way", Gemini stated that "The claim is related to the problem (human activities and elk) but *doesn't specify* the relationship as positive or negative". This assessment was false, since the student had specified a "negative" relationship.

We also found **errors in the MLLMs' numerical reasoning**. Specifically, when the rubric required evaluators to count the ecosystems components in students' responses and assign scores based on the number of components (Mulch dataset; Element), the models listed the components successfully but assigned scores incorrectly. For example, GPT-4 Turbo gave the student a score of 1 (identifying 2-4 components) instead of 2 (>4 components), even though it correctly listed the five elements in student's response: "Score: 1; Elements: [bark mulch, leaf mulch, soil moisture, plants, water]. The student *identifies 5 elements. This falls into the range of 2-4 factors* identified, so the score is 1."

## 5 Discussion

Our research illuminates opportunities of MLMMs in learning analytics and science education. Researchers have called for developing new methods to analyze multimodal data in formative assessments [3, 31, 32, 34]. Our work demonstrates the feasibility of using MLLMs to transcribe and provide automated feedback on science assessments that teachers frequently administer. The best performing models (Claude; GPT-4o) showed moderate to substantial agreement with human evaluators. Similar to work with LLMs [19, 36], we found that MLLMs' performance generally improved when provided with example submissions and scores. Leveraging MLLMs is promising for the generalizability to different assessment contexts, without relying on extensive hand-labeled data or large corpora of example responses.

While these results are promising, we observe cases of hallucination, inaccurate reasoning, and misevaluation of the depth in students' responses. These cases may negatively impact students' learning, and consequently affect trust and uptake of the feedback. Researchers have emphasized the role of domain experts like teachers [3, 19] in evaluating MLLMs and adding well-defined task contexts and criteria (e.g., scoring thresholds, example scores). Results present opportunities for designing learning analytics systems that allow for iterative evaluation and modification MLLMs' assessment output. Examples from Human-Computer Interaction design include interactive interfaces for prompt refinement and automated evaluations, based on user-defined criteria [16]. Including specific criteria for responses' clarity and depth might improve MLLMs' ability to evaluate students' scientific knowledge.

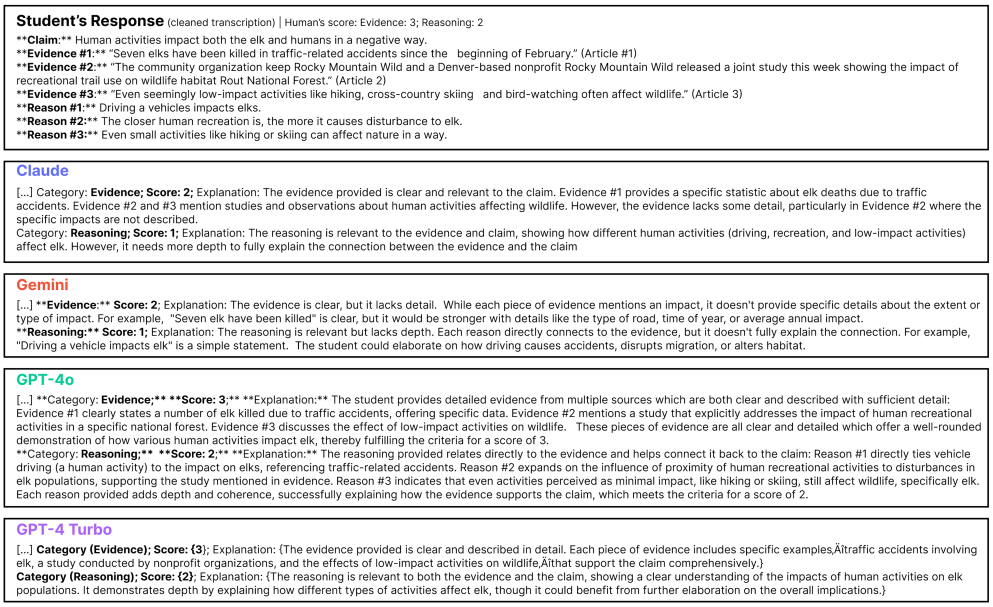


**Figure 4: Examples of disagreements (Claude; Gemini) between the MLLMs and human evaluators**

The limitations of the current work can guide future research. First, the dataset is small and focused on one grade level and domain (environmental science). Future research can construct a larger corpus of multimodal assessments spanning grades and domains. Second, researchers can include fine-grained rubrics in the MLLM prompts, to correlate the modalities with different aspects of science understanding. Third, future work should discuss ethical considerations, such as raising awareness among students and teachers about data protection. Another promising direction is to examine the models' capacity to provide feedback in real time, students' uptake of feedback, and impact on learning.

## 6 Conclusion

This research presents first steps toward using MLLMs to provide automated feedback on open-ended, multimodal science assessments. Such feedback can help students to deepen science understanding through utilizing various modalities and provide snapshots of learning to inform instruction. MLLMs provided with well-defined examples of student submissions and scoring explanations tended to perform better than those without examples. We identify areas of improvement for future research, including calibrating the models' reasoning with human evaluators, checking for hallucination, and designing learning analytics systems that allow for iterative evaluation and prompt refinement.

## Acknowledgments

This work is supported by the National Science Foundation, grant #2422286.

## References

[1] Kate T. Anderson, Steven J. Zuiker, Gita Taasoobshirazi, and Daniel T. Hickey. 2007. Classroom Discourse as a Tool to Enhance Formative Assessment and Practise in Science. *International Journal of Science Education* 29, 14 (Nov. 2007), 1721–1744. https://doi.org/10.1080/09500690701217295 Publisher: Routledge.

[2] Brian R Belland, Krista D Glazewski, and Jennifer C Richardson. 2011. Problem-based learning and argumentation: Testing a scaffolding framework to support middle school students' creation of evidence-based arguments. *Instructional Science* 39 (2011), 667–694.

[3] Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. 2024. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *arXiv preprint arXiv:2401.00832* (2024).

[4] Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)* 21 (2009), 5–31. Publisher: Springer.

[5] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Zheyi Chen, Liuchang Xu, Hongting Zheng, Luyao Chen, Amr Tolba, Liang Zhao, Keping Yu, and Hailin Feng. 2024. Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. *Computers, Materials & Continua* 80, 2 (2024).

[8] Xiuliang Duan, Dating Tan, Liangda Fang, Yuyu Zhou, Chaobo He, Ziliang Chen, Lusheng Wu, Guanliang Chen, Zhiguo Gong, Weiqi Luo, et al. 2024. Reason-and-Execute Prompting: Enhancing MultiModal Large Language Models for Solving Geometry Questions. In *ACM Multimedia 2024.*

[9] Meghan Rector Federer, Ross H Nehm, John E Opfer, and Dennis Pearl. 2015. Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Research in Science Education* 45 (2015), 527–553.

[10] Erin Marie Furtak. 2023. *Formative Assessment for 3D Science Learning: Supporting Ambitious and Equitable Instruction.* Teachers College Press. Google-Books-ID: vRPNEAAAQBAJ.

[11] Erin Marie Furtak, Katharina Kiemer, Ruhan Kizil Circi, Rebecca Swanson, Vanessa de León, Deb Morrison, and Sara C. Heredia. 2016. Teachers' formative assessment abilities and their relationship to student learning: findings from a four-year intervention study. *Instructional Science* 44, 3 (June 2016), 267–291. https://doi.org/10.1007/s11251-016-9371-3

[12] Scott E. Grapin and Laura Ascenzi-Moreno. 2024. Expansive assessment of expansive abilities: Teachers' perspectives and practices with multimodal and translanguaged content assessments. *Learning and Instruction* 92 (Aug. 2024),

101925. https://doi.org/10.1016/j.learninstruc.2024.101925
[13] Scott E Grapin and Lorena Llosa. 2022. Multimodal tasks to assess English learners and their peers in science. *Educational assessment* 27, 1 (2022), 46–70. Publisher: Taylor & Francis.
[14] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2256–2264.
[15] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27036–27046.
[16] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
[17] Che-Yu Kuo and Hsin-Kai Wu. 2013. Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. *Computers & Education* 68 (Oct. 2013), 388–403. https://doi.org/10.1016/j.compedu.2013.06.002
[18] J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* (1977), 363–374.
[19] Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence* 6 (2024), 100213. Publisher: Elsevier.
[20] Hee-Sun Lee, Gey-Hong Gweon, Trudi Lord, Noah Paessel, Amy Pallant, and Sarah Pryputniewicz. 2021. Machine Learning-Enabled Automated Feedback: Supporting Students' Revision of Scientific Arguments Based on Data Drawn from Simulation. *Journal of Science Education and Technology* 30, 2 (April 2021), 168–192. https://doi.org/10.1007/s10956-020-09889-7
[21] Richard Lehrer and Leona Schauble. 2015. The development of scientific thinking. *Handbook of child psychology and developmental science* 2, 7 (2015), 671–714. Publisher: Wiley Hoboken, NJ.
[22] Haiying Li, Janice Gobert, and Rachel Dickler. 2017. Automated Assessment for Scientific Explanations in On-Line Science Inquiry. *International Educational Data Mining Society* (2017). Publisher: ERIC.
[23] Stephanie Link, Mohaddeseh Mehrzad, and Mohammad Rahimi. 2022. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning* 35, 4 (2022), 605–634. Publisher: Taylor & Francis.
[24] Ou Lydia Liu, Joseph A Rios, Michael Heilman, Libby Gerard, and Marcia C Linn. 2016. Validation of automated scoring of science assessments. *Journal of Research in Science Teaching* 53, 2 (2016), 215–233. Publisher: Wiley Online Library.
[25] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255* (2023).
[26] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*. Springer, 387–402.
[27] Katherine L McNeill and Joseph S Krajcik. 2011. Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing. *Pearson* (2011).
[28] Elizabeth Birr Moje. 2007. Chapter 1 Developing Socially Just Subject-Matter Instruction: A Review of the Literature on Disciplinary Literacy Teaching. *Review of Research in Education* 31, 1 (March 2007), 1–44. https://doi.org/10.3102/0091732X07300046001 Publisher: American Educational Research Association.
[29] Ashlyn E. Pierson, Douglas B. Clark, and Corey E. Brady. 2021. Scientific modeling and translanguaging: A multilingual and multimodal approach to support science learning and engagement. *Science Education* 105, 4 (2021), 776–813. https://doi.org/10.1002/sce.21622 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sce.21622.
[30] Ashlyn E. Pierson, D. Teo Keifert, Sarah J. Lee, Andrea Henrie, Heather J. Johnson, and Noel Enyedy. 2023. Multiple Representations in Elementary Science: Building Shared Understanding while Leveraging Students' Diverse Ideas and Practices. *Journal of Science Teacher Education* 34, 7 (Oct. 2023), 707–731. https://doi.org/10.1080/1046560X.2022.2143612 Publisher: Routledge _eprint: https://doi.org/10.1080/1046560X.2022.2143612.
[31] Angi Shelton, Andrew Smith, Eric Wiebe, Courtney Behrle, Ruth Sirkin, and James Lester. 2016. Drawing and writing in digital science notebooks: Sources of formative assessment data. *Journal of Science Education and Technology* 25 (2016), 474–488. Publisher: Springer.
[32] Andy Smith, Samuel Leeman-Munk, Angi Shelton, Bradford Mott, Eric Wiebe, and James Lester. 2018. A multimodal assessment framework for integrating student writing and drawing in elementary science learning. *IEEE Transactions on Learning Technologies* 12, 1 (2018), 3–15. Publisher: IEEE.
[33] Yishen Song, Liming Guo, and Qinhua Zheng. 2024. Measuring scientific inquiry ability related to hands-on practice: An automated approach based on multimodal data analysis. *Education and Information Technologies* (2024), 1–31.
[34] Judith Stanja, Wolfgang Gritz, Johannes Krugel, Anett Hoppe, and Sarah Dannemann. 2023. Formative assessment strategies for students' conceptions—The potential of learning analytics. *British Journal of Educational Technology* 54, 1 (2023), 58–75. https://doi.org/10.1111/bjet.13288 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13288.
[35] Kok-Sing Tang. 2023. Distribution of Visual Representations Across Scientific Genres in Secondary Science Textbooks: Analysing Multimodal Genre Pattern of Verbal-Visual Texts. *Research in Science Education* 53, 2 (April 2023), 357–375. https://doi.org/10.1007/s11165-022-10058-6
[36] Xiaoyi Tian, Amogh Mannekote, Carly E Solomon, Yukyeong Song, Christine Fry Wise, Tom Mcklin, Joanne Barrett, Kristy Elizabeth Boyer, and Maya Israel. 2024. Examining LLM Prompting Strategies for Automatic Evaluation of Learner-Created Computational Artifacts. In *Proceedings of the 17th International Conference on Educational Data Mining*. 698–706.
[37] Claudia von Aufschnaiter and Alicia C. Alonzo. 2018. Foundations of formative assessment: Introducing a learning progression to guide preservice physics teachers' video-based interpretation of student thinking. *Applied Measurement in Education* 31, 2 (April 2018), 113–127. https://doi.org/10.1080/08957347.2017.1408629 Publisher: Routledge _eprint: https://doi.org/10.1080/08957347.2017.1408629.
[38] Hanna Wanselin, Kristina Danielsson, and Susanne Wikman. 2022. Analysing Multimodal Texts in Science—a Social Semiotic Perspective. *Research in Science Education* 52, 3 (June 2022), 891–907. https://doi.org/10.1007/s11165-021-10027-5
[39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
[40] Christopher D Wilson, Kevin C Haudek, Jonathan F Osborne, Zoë E Buck Bracey, Tina Cheuk, Brian M Donovan, Molly AM Stuhlsatz, Marisol M Santiago, and Xiaoming Zhai. 2024. Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching* 61, 1 (2024), 38–69. Publisher: Wiley Online Library.
[41] Rachel E Wilson and Leslie U Bradbury. 2016. The pedagogical potential of drawing and writing in a primary science multimodal unit. *International Journal of Science Education* 38, 17 (2016), 2621–2641. Publisher: Taylor & Francis.
[42] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9, 1 (2023), 1.
[43] Xiaoming Zhai, Peng He, and Joseph Krajcik. 2022. Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching* 59, 10 (2022), 1765–1794. Publisher: Wiley Online Library.
[44] Mengxiao Zhu, Ou Lydia Liu, and Hee-Sun Lee. 2020. The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education* 143 (Jan. 2020), 103668. https://doi.org/10.1016/j.compedu.2019.103668
[45] Zichen Zhu, Yang Xu, Lu Chen, Jingkai Yang, Yichuan Ma, Yiming Sun, Hailin Wen, Jiaqi Liu, Jinyu Cai, and Yingzi Ma. 2024. Multi: Multimodal Understanding Leaderboard with Text and Images. *arXiv preprint arXiv:2402.03173* (2024).