



Applying Generative Artificial Intelligence to Critiquing Science Assessments

Ha Nguyen¹ · Jake Hayward²

Accepted: 30 October 2024
© The Author(s) 2024

Abstract

High-quality science assessments are multi-dimensional. They promote disciplinary practices, core ideas, cross-cutting concepts, and science sense-making. In this paper, we investigate the feasibility of using generative artificial intelligence (GenAI), specifically multimodal large language models (MLLMs), to annotate and provide improvement ideas for K-12 science assessments. The AI-generated annotations critique how the assessments align with the three dimensions of the Next Generation Science Standards (NGSS) and suggest ideas to elicit students' science sense-making. We outline our process with various prompting strategies: few-shot and zero-shot learning (prompting with and without examples), chain of thought (eliciting the MLLM's reasoning), and sampling strategies (outputting high or low level of randomness). Overall, the AI annotations can reason about the alignment between the assessments and NGSS dimensions and overlap with annotations from K-12 educators. Annotations generated with few-shot learning generally score higher overall and provide more details than zero-shot prompts. Further, interviews with science teachers reveal that the MLLM annotations can support teachers' reflection on instructional practices and assessment revision. We discuss the application of MLLMs to develop three-dimensional science assessments.

Keywords Multimodal large language models (MLLMs) · Artificial intelligence (AI) · Science education · Assessment · K-12

Introduction

In the United States, the Next Generation Science Standards (NGSS) emphasize the development of student competencies across three dimensions: *science and engineering practices* used by professionals in the fields, *disciplinary core ideas* encompassing key knowledge within science disciplines, and *cross-cutting concepts* across disciplines (National Research Council, 2014). High-quality assessments need to measure each of these dimensions and invite students to apply the dimensions in tandem, to simulate professionals' approaches

to scientific tasks (Fulmer et al., 2018; NGSS Lead States, 2013; Penuel et al., 2019).

Assessments should also promote equitable learning and ensure that all students can access science practices and have authentic opportunities to demonstrate understanding (Grapin et al., 2023). One way to facilitate equity is to incorporate multimodality into the assessment tasks—combining written, oral, and visual representations—to elicit different modes of science sense-making (Pierson et al., 2021; Shelton et al., 2016; Smith et al., 2018). Such assessments illuminate the complexity of science and literacy practices that might have been hidden in text-focused assessments (Grapin & Llosa, 2022; Heritage, 2021). Including compelling and relevant science phenomena can further support sense-making (Fine & Furtak, 2020; Penuel et al., 2019).

In practice, however, landscape studies of existing science assessments find few examples that target the three NGSS dimensions in tandem (Wertheim et al., 2016). Additionally, assessment design does not always consider equity (Rodriguez, 2015). To address these challenges, educators can employ tools to annotate science assessments before

✉ Ha Nguyen
ha.nguyen@unc.edu

Jake Hayward
a02298038@usu.edu

¹ School of Education, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

² Department of Instructional Technology & Learning Sciences, Utah State University, Logan, UT, USA

classroom implementation. Educators can then use insights from the annotations to refine the assessment design. In this paper, we explore the application of generative artificial intelligence (GenAI), specifically multimodal large language models (MLLMs), to annotate science assessments. MLLMs can interpret multimodal input (e.g., assessments that combine text and images) to generate textual output. The models have demonstrated human-like capacity in responding to questions and showing scientific and mathematical reasoning (Lu et al., 2023; Yang et al., 2023). We report on an exploratory study applying a MLLM (specifically GPT-4 Vision) with different model prompting strategies to answer the research questions (RQs):

RQ1: How well do different prompting strategies help to create assessment annotations that align with NGSS standards, support equity and science sense-making, and inform instructional practices?

RQ2: How do educators perceive the AI annotations, which are developed with specific instructional contexts and examples to support standards-aligned and equitable science learning?

To answer RQ1, we compare annotations of K-12 science assessments generated by a MLLM (GPT-4 Vision) and by educators. The assessments span grade levels (elementary, middle, and high school) and science domains (life science, physical science, and earth and space sciences). To achieve the best results, we include various prompting strategies: few-shot and zero-shot learning (prompting with or without examples), context (detailing the assessments' context and purpose), chain of thought (prompting the model's reasoning), and sampling strategies (outputting low and high level of randomness in the responses). We manually score the AI annotations for the extent to which they correctly evaluate the alignment between the assessments and the target NGSS dimensions, overlap with educators' feedback, suggest detailed improvement ideas, and include accurate details. For RQ2, we interview science teachers from different grade levels, to explore how they perceive and suggest improvement ideas for the AI annotations. Our findings contribute insights for employing emerging MLLMs to support the development of high-quality science assessments.

Literature Review

Science Assessments for Learning

Vision of assessments for learning, as opposed to assessment of learning, emphasizes that the assessments should elicit evidence about specific aspects of student performance to inform instruction (Black & William, 1998; Mislevy et al.,

2003). This vision calls for a system of assessments, combining short performance tasks and instructionally embedded assessments, to support immediate adjustments and long-term instructional shifts (Osborne et al., 2015).

Regardless of the assessment type, high-quality science assessments should measure competencies across the *three dimensions of science learning*, as proposed by the NGSS (Brown, 2017; National Research Council, 2014). These dimensions include the practices that engineers and scientists conduct in their work, the core ideas within science and engineering disciplines, and the cross-cutting concepts across disciplines (National Research Council, 2014). The dimensions can be placed on a learning progression, defined as coherent and increasingly sophisticated understanding of core ideas over time and across grade levels (Kaldaras et al., 2021; NGSS Lead States, 2013). Assessments that consider learning progression should be appropriate to the grade levels and generate evidence that pinpoints learning areas to focus subsequent instruction (William, 2010).

Beyond the assessment content, it is critical to consider *how* students are assessed. *Equity* should remain a focus in assessment design (Fine & Furtak, 2020; Rodriguez, 2015). Traditionally, the conceptualization of equity within science assessment has focused on *access*, for example, providing sufficient resources for every student to perform the tasks (Guzman-Orth et al., 2021), avoiding jargons (Rivera & Stansfield, 2004), and integrating linguistic and visual scaffolds (Lyon, 2013; Siegel et al., 2014). Emergent research has highlighted the need to move from accommodating students' access to supporting *sense-making* (Gravin et al., 2023). Such research defines equity beyond access, to promote science interest, identity, and learning achievement; broaden conceptualizations of science; and envision science in connection to social movements (Philip & Azevedo, 2017; National Academies of Sciences Engineering and Medicine, 2022). To support these expansive conceptualizations of equity, the assessments should introduce authentic task scenarios (Penuel et al., 2019). They should integrate multiple sense-making modes, to elevate the local, personal, and science-related values that students bring to the classroom (Brown, 2017; Fine & Furtak, 2020; Rodriguez, 2015). Finally, assessments should be linked to students' prior knowledge, provide multiple representations and modalities to make sense of science content, and elicit students' reasoning (Herman, 2013; National Research Council, 2012).

While there exists consensus around what constitutes high-quality science assessments, developing such assessments is nontrivial. A review of 51 science assessments across grade levels reveals that most evaluate one NGSS dimension at a time, instead of integrating multiple dimensions (Wertheim et al., 2016). Furthermore, equity is not always explicit in assessment design (Rodriguez, 2015). We propose that having a tool to critique and modify

assessments before classroom deployment may support educators to iterate upon how they can design and adapt assessments. Such tools can help educators to articulate which NGSS learning goals they are targeting, reflect on how the assessment tasks align with these goals, and review suggestions from high-quality assessments that purposefully target the NGSS dimensions in tandem (Fulmer & Polikoff, 2014; Fulmer et al., 2018; McElhaney et al., 2018; Wertheim et al., 2016). We explore the use of multimodal large language models to perform this task at scale.

Multimodal Large Language Models

Advances in large language models (LLMs) present opportunities to develop AI systems capable of solving complex natural language tasks (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023). These systems have been used to generate assessments adaptive to learners' knowledge, automatically score student submissions, and develop teaching simulations in education settings (Doughty et al., 2024; Lee et al., 2023; Lu & Wang, 2024; Markel et al., 2023). LLMs demonstrate sophisticated understanding of language, offer accurate evaluations that overlap with human annotations, and show applicability in a range of contexts without extensive training data (Latif & Zhai, 2024; Steiss et al., 2024).

MLLMs extend LLMs to account for multimodal input, with promising results in vision-language understanding (Dai et al., 2024; Pichai & Hassabis, 2023; Yang et al., 2023; Zhu et al., 2023). MLLMs have demonstrated human-level capacities in generating image captions, answering questions, analyzing emotions, and reasoning with scientific knowledge given visual cues (Li et al., 2024; Wang et al., 2024). The models can understand and follow multilingual instruction, explain its reasoning, and engage in multi-turn exchanges with users (Yang et al., 2023). In education contexts, GPT-4 Vision performs well in math reasoning tasks that incorporate graphs and geometry (Lu et al., 2023). Researchers have proposed applying these models to analyze multimodal data, including behaviors, discourse, and emotions in learning interactions, to gauge students' engagement and respond to learning needs (Gan et al., 2023).

Despite these promising findings, AI-generated output may include inaccuracies, lack contextual knowledge, and perpetuate bias from the models' training data (Gunjal et al., 2024; Yang et al., 2023). These models may "hallucinate" and describe elements that do not exist in the input (Zhou et al., 2023). Thus, there exists a need for prompt engineering, defined as tuning instruction for LLMs and MLLMs, to improve their performance in specific contexts (Lee et al., 2023; Liu et al., 2023; Wei et al., 2022; Zhao et al., 2024).

Prompt Engineering

Several prompt engineering strategies for LLMs have been extended to MLLMs (Mu et al., 2024; Wang et al., 2022; Yang et al., 2023). We review four prompt engineering aspects that have been applied to education settings for LLMs and MLLMs (Lee et al., 2024; Table 1). First, providing a few examples of the desired output (*few-shot* learning) can make responses more consistent with researchers' expectations (Brown et al., 2020). To illustrate, Swanson et al. (2021) developed Story Centaur, an LLM-integrated co-writing system that generated new text and provides creative suggestions, based on the examples that users provided. In using LLMs to generate feedback for students' programming code, Nguyen and Allan (2024) found that prompts with example feedback resulted in more accurate results than those without examples.

Second, studies have illustrated the importance of *context*, or specifying information about learning theories and rubrics for feedback in the prompts (Ali et al., 2023; Lee et al., 2024). Lee et al. (2024) tested this strategy in generating evaluations of students' answers to science assessments with LLMs (GPT-3.5, GPT-4). The authors found that providing detailed assessment descriptions and grading rubrics, along with examples of human-generated feedback, resulted in evaluations consistent with human scoring. Similarly, Nguyen et al. (2023) found that providing LLMs with multiple data sources, including the grading context and rubric, resulted in AI-generated feedback that could identify misconceptions in students' mathematical reasoning.

Third, utilizing LLMs and MLLMs for complex tasks requires logical reasoning (*chain of thought*). Prompting LLMs to engage in "a series of intermediate natural

Table 1 Prompt engineering strategies

Strategy	Definition
Few-shot learning	Include examples of the desired output in the prompts to LLMs/MLLMs
Specify context	Detail information about the context of the task, such as the rubrics or learning theories guiding the AI feedback
Chain of thought	Prompt the AI to explain its reasoning, e.g., "think step by step."
Sampling	Vary model's hyperparameters that affect response diversity, e.g., nucleus/greedy sampling

language reasoning steps that lead to the final output” (Wei et al., 2022, p. 2) can improve the models’ outcomes. For example, asking the models to “think step by step” can illuminate their reasoning path to arrive at the solution (Kojima et al., 2022). This strategy has shown promising performance in mathematics reasoning (Imani et al., 2023), question answering (Wang et al., 2024), and automatic scoring of science assessments (Lee et al., 2024). It has also been used with MLLMs to generate successful results with input combining text and images (Yang et al., 2023).

Finally, researchers have considered *sampling strategies* that affect the response diversity (Dai et al., 2024; Holtzman et al., 2020; Su et al., 2022). Building on prior work (Lee et al., 2024), we consider *greedy decoding* and *nucleus sampling*. During the output generation process, greedy decoding selects the token with the highest probability at each step, whereas nucleus sampling (or top-p sampling) samples from the top-p most probable tokens (Holtzman et al., 2020). Both sampling strategies can produce useful insights in our context. Greedy decoding generally results in more accurate and consistent output, which is useful in assessment scoring (Lee et al., 2024; Lin & Chen, 2023). Meanwhile, nucleus sampling introduces randomness and can result in more creative responses (Nichols et al., 2020) to guide assessment redesign.

In sum, applying GenAI technologies, including LLMs and MLLMs, requires prompt engineering strategies with considerations for example output, context, probing for reasoning, and sampling strategies. Researchers have examined these prompting strategies in science education with text-based models, such as using GPT-4 and GPT-3.5 (iterations of OpenAI’s LLMs) to automatically score students’ open-ended submissions (Lee et al., 2024). We extend this research to explore using different strategies for vision-language understanding tasks, specifically leveraging MLLMs

to critique science assessments involving text and visual elements.

Materials and Methods

The Task Annotation Project in Science (TAPS)

Comparing AI-generated and human annotations is a common approach to evaluating the technology’s performance (Nguyen & Allan, 2024; Nguyen et al., 2023; Steiss et al., 2024). In this research, we compared the MLLM-created annotations with publicly available, educators’ annotations of nine science assessments (42 questions) from the TAPS (NGSS, n.dc). The project selected open-source science assessments across grade levels and science domains (Achieve, n.d). Annotation effort involved 40 expert reviewers, including classroom assessment developers, teachers, and administrators. Each assessment was pre-screened by the annotators, to ensure that it was coherent and comprehensible from a student’s perspective; included a central phenomenon; could only be answered by using information in the task; and required students to use science and engineering practices (SEPs), disciplinary core ideas (DCIs), cross-cutting concepts (CCCs), and reasoning. Once selected for review, each assessment was scored by three TAPS reviewers and discussed until reviewers reached consensus. The scoring included several dimensions: Equity, Scenario, Sense-making, NGSS SEPs, DCIs, CCCs, and Assessment Purpose (Table 2). These dimensions were created through an iterative process involving content experts (practitioners and researchers with expertise in science teaching and learning, assessment, and equity), pilot implementation in states and districts for professional learning use cases, and refinement (NGSS, n.db).

Table 2 TAPS annotations dimensions (adapted from Achieve, nd)

Dimension	Definition
What is assessed	
SEPs	Include opportunities to demonstrate understanding and application of science and engineering practices at the appropriate grade level
DCIs	Include opportunities to demonstrate disciplinary core ideas at the appropriate grade level
CCCs	Include opportunities to use cross-cutting concepts for science sense-making at the appropriate grade level
How it is assessed	
Equity	Support diverse learners. Assessments are authentic and personally meaningful; provide choices of engagement and multiple ways to demonstrate learning; include opportunities for students to monitor understanding and progress
Scenario	Provide high-quality scenarios that drive three-dimensional assessments; feature uncertainty to help students understand what they are addressing and why it matters
Sense-making	Provide opportunities to reason about the central phenomena and demonstrate understanding across multiple NGSS dimensions
Assessment purpose	Highlight how the task features are linked to the intended assessment use

We selected this dataset for several reasons. First, the data was well-aligned with our conceptualization of high-quality assessments. It was pre-screened by TAPS educators to integrate three-dimensional and equitable science learning. Second, it involved rigorous review and consensus reaching among reviewers with expertise in the NGSS, ensuring the quality of the human annotations (Achieve, nd). Third, each assessment focused on eliciting multiple modes of sense-making through text instruction, graphs, images, screenshots from videos, and models. This multimodal input provided a setting to leverage MLLMs.

Rubric to Evaluate Prompts

We developed a rubric to score the MLLMs' performance (Table 3). The rubric components were informed by a literature review about considerations in educational contexts for AI-generated information around accuracy, overlap with human's evaluation, and usefulness to inform instructional practices (Lee et al., 2024; Nguyen et al., 2023; Singhal et al., 2023; Steiss et al., 2024; Wang & Demszky, 2023). The rubric components were drafted by the researchers based on the literature review and refined in two discussions.

We evaluated *Standards*, referring to the extent to which the annotations correctly stated the alignment between the

assessment items and the DCIs, SEPs, and CCCs in the NGSS. We selected this component, because aligning assessment tasks and NGSS dimensions remained a challenge in science assessment design (Fulmer et al., 2018; Wertheim et al., 2016). We scored each of the three dimensions (DCIs, SEPs, CCCs) on a scale of 1 (no standard alignment) to 4 (standard alignment and accurate evaluation of whether the assessments are grade appropriate).

A second component was *Overlap with Educators*, defined as the alignment between the MLLM and TAPS annotations. This component followed prior work that has evaluated the quality of LLM-generated feedback, compared to human feedback, to assess the models' effectiveness in education settings (Estévez-Ayres et al., 2024; Lee et al., 2024). We scored this component on a scale of 1 (no overlap) to 5 (complete agreement with educators in the TAPS annotations).

The next two components addressed concerns in leveraging MLLMs that the models may generate inaccurate information (Yang et al., 2023) and their suggestions do not always generate useful insights to inform instruction (Wang & Demszky, 2023). We measured *Accuracy* on a scale of 1 (completely inaccurate) to 5 (completely accurate). This component explored whether the feedback included accurate details or added incorrect information that was not present in

Table 3 Rubric to evaluate MLLM's annotations and inter-rater agreement (Cohen's κ)

Dimension	Score
Standard (similar scoring scale for SEPs, DCIs, CCCs) $\kappa = 1$ overall (1 for each scale: SEPs, DCIs, CCCs) Exact agreement, 100% (100% for each scale) Within 1-point, 100% (100% for each scale)	1: Does not provide an accurate explanation of how the assessment item aligns with the NGSS dimension 2: Provides a partially accurate explanation of how the assessment item aligns with the NGSS dimension. Incorrect assessment of whether and how the item is grade appropriate 3: Provides an accurate explanation of how the assessment item aligns with the NGSS dimension. Incorrect assessment of whether and how the item is grade appropriate 4: Provides an accurate explanation of whether and how the assessment item aligns with the NGSS dimension. Correct assessment of whether and how the item is grade appropriate
Overlap with Educators $\kappa = 0.79$ Exact agreement, 90% Within 1-point, 100%	1: Does not overlap with TAPS annotations from experienced educators 2: Mostly does not overlap with TAPS annotations or overlaps with the categories in TAPS annotations (e.g., scenario, sense-making, SEPs) but not the critique 3: A mix of annotations that overlap and do not overlap with TAPS 4: Most of the annotation overlaps; one piece of the annotation does not 5: All pieces of the annotation overlap with TAPS critiques
Details $\kappa = 0.66$ Exact agreement, 80% Within 1-point, 100%	1: Does not reference details from the assessment and does not give concrete steps 2: Minimal reference to the assessment or lacks concrete next steps 3: A mix of detailed and vague suggestions or suggestions do not reference details from the assessment 4: Mostly provides concrete, actionable steps, with references from the assessment 5: Provides concrete, actionable next steps and consistently references details from the assessment
Accuracy $\kappa = 0.65$ Exact agreement, 80% Within 1-point, 100%	1: Annotation is inaccurate and unrelated to details from the assessment 2: Annotation is mostly inaccurate 3: Annotation is accurate, but more than one detail is inaccurate 4: Annotation is mostly accurate, but one detail is inaccurate 5: All pieces of the annotation are accurate

the input. We further evaluated *Details*, or how the feedback provided clear direction for improvement. Scoring ranged from 1 (vague details) to 5 (clear directions, with references to details within the assessments).

After deciding on the rubric, two researchers collaboratively coded four assessment items and identified examples for the scoring categories over three discussions. Then, the researchers separately coded 12% of the sample to establish inter-rater agreement (Cohen's κ range 0.65–1). The first author coded the rest of the data and identified uncertain items for discussion.

Prompt Development (RQ1)

Phase 1: Testing Prompt Strategies

We developed the prompts for the MLLM in two phases (Fig. 1). In phase 1, we experimented with a subset of the TAPS assessment (10 items from three assessments; one assessment from each grade level) using a combination of strategies: zero-shot and few-shot, greedy and nucleus

sampling, and chain of thought (CoT) and no CoT. We made API calls in Python to OpenAI's GPT-4 Vision (GPT-4V) in February 2024 (max tokens = 1000). Input for the MLLM included the prompt instruction, along with image of the assessment (Fig. 2).

For few-shot learning, we provided the MLLM with one example annotation of a different question from the same assessment. We also tested with nucleus sampling (temperature = 0.9, top- p = 0.95; following Lee et al., 2024) and greedy decoding (temperature = 0). For some prompts, we added instruction for CoT so the model could “think step by step” (Kojima et al., 2022; Lee et al., 2024). For all prompts, we included the assessment context, specifically stating the grade levels and the NGSS dimensions that the assessments targeted. Note that we only mentioned the target NGSS learning standards and did not include the content of NGSS or TAPS annotations (other than the examples in few-shot prompts). This allowed us to examine whether the MLLM was already aware of NGSS and included any inaccuracy. Figure 2 outlines the prompt components. All prompts included assessment context, task instruction, and

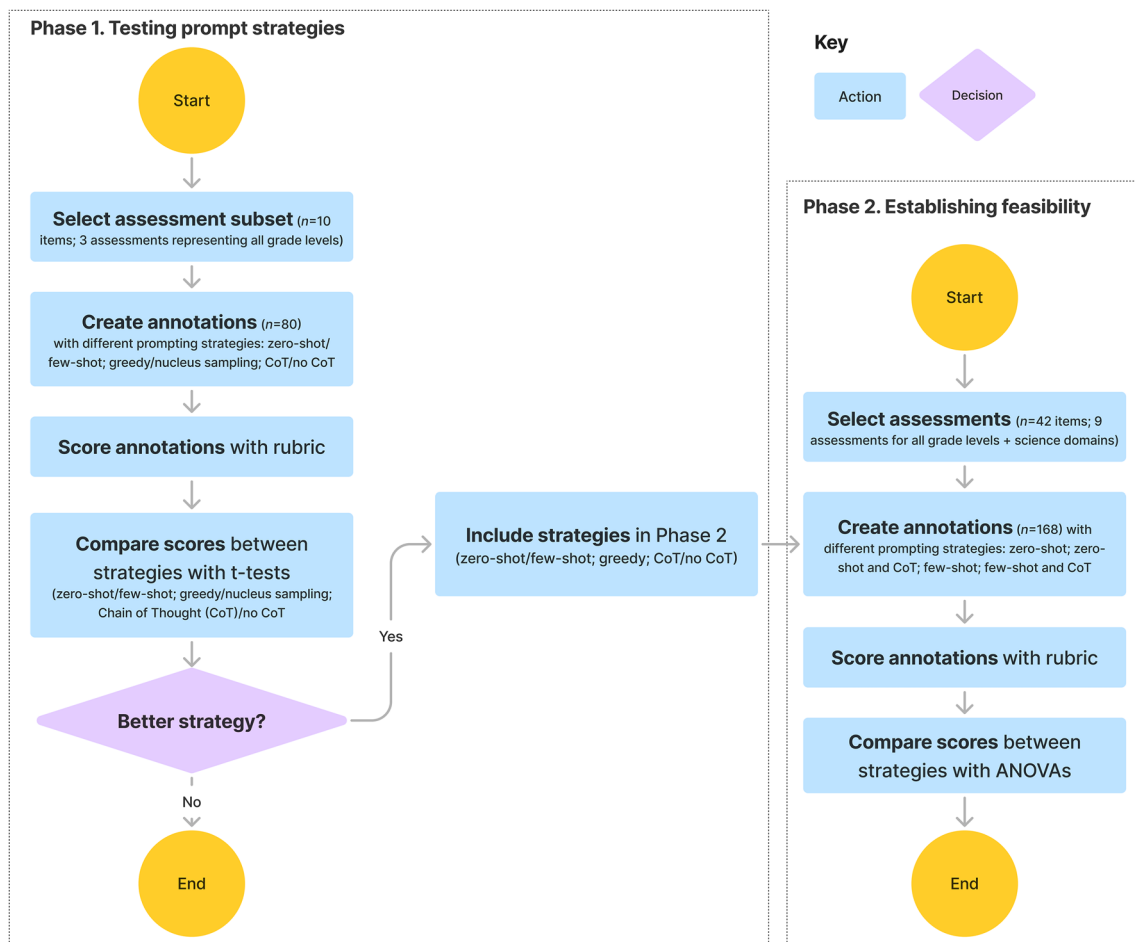


Fig. 1 Prompt development procedures

CONTEXT

USER: <image of assessment>

You will critique a formative science assessment for middle school grades 6-8 using the following categories. Account for the text, visuals, and graphs within the assessment in your critique:

1. Equity: explain how the assessment supports or does not support a wide range of diverse students. The assessment should be authentic and meaningful to students, and allow multiple ways to demonstrate learning.
2. Scenarios: explain how assessment provides information to elicit performances and sense-making.
3. Science and engineering practices: explain how task supports the targeted Next Generation Science Standards' science and engineering practices and is grade appropriate.
4. Disciplinary core ideas: explain how task supports the targeted Next Generation Science Standards' disciplinary core ideas and is grade appropriate.
5. Cross-cutting concepts: explain how task supports the targeted Next Generation Science Standards' crosscutting concepts and is grade appropriate.
6. Sense-making: explain how the assessment invites students' reasoning about phenomena and problems.
7. Assessment purpose: explain whether it is clear how the task features connect to intended assessment use.

The assessment targets the following standards:

1. Science and engineering practices: Planning and Carrying Out Investigations.
2. Disciplinary core ideas: 6-8 PS3.A Definitions of Energy.
3. Cross-cutting concept: cause and effect.

TASK

First, state whether Question 2 captures the targeted science and engineering practices, disciplinary core ideas, and cross-cutting concepts at the middle school level, or are not sophisticated enough and are better for elementary schools. Explain why or why not.

Second, provide a critique of Question 2 and suggest specific ideas for improvement using the provided categories and targeted standards. Your critique only needs to include the most relevant categories and should state how Question 2 builds on or is related to the other questions. State if the included visuals are relevant to the assessment purpose and why. Use the following format: Category: (idea for improvement).

CHAIN OF THOUGHT

Let's think step by step.

FEW-SHOT LEARNING

Here is an example critique of Question 1:

Question 1 targets the disciplinary core ideas at the target grade level. Successfully answering this question requires that students use their understanding that different substances are made up of different particles, and that those differences can influence changes in temperature when energy is added. This most closely connects to the 6-8 PS3.A DCI element "the temperature of a system is proportional to the average kinetic energy and potential energy per atom or molecule...the details of this relationship depend on the type of atom or molecule and the interactions among the atom in the material...". The cross-cutting concepts and science and engineering practices are not explicitly addressed.

Ideas for improvement:

Scenario: While the scenario is somewhat specific, it is not clear to students why this is a worthwhile investigation, limiting how well the task will engage all students. **Equity:** The concept of different kinds of sand might not be accessible to all students without further support, such as pictures to illustrate the differences between black and white sand.

Fig. 2 Prompt engineering components. *Notes.* While we included each assessment as a single image, it is also possible to provide multi-part prompts, e.g., separating the textual instruction and images within the assessment

Table 4 Prompt combinations

Prompt	Few-shot	Chain of thought (CoT)	Greedy	Included in phase 2
Zero-shot greedy			x	x
Zero-shot nucleus				
Few-shot greedy	x		x	x
Few-shot nucleus	x			
Zero-shot CoT greedy		x	x	x
Zero-shot CoT nucleus		x		
Few-shot CoT greedy	x	x	x	x
Few-shot CoT nucleus	x	x		

assessment image. The prompts varied in whether they included CoT and few-shot learning. Table 4 lists the prompt combinations. Appendix A provides example GPT-4V's output and rubric scoring.

We ran *t*-tests comparing the total scores across rubric dimensions for feedback with zero-shot versus few-shot learning, greedy versus nucleus sampling, and CoT versus no CoT ($n = 40$ per comparison group). We found no difference between zero-shot and few-shot learning, M zero-shot = 21.63; $SD = 2.93$; M few-shot = 22.70; $SD = 2.67$; $t(78) = 1.71$, $p = 0.09$, or between CoT and no CoT, M

CoT = 22.75; $SD = 2.62$; M no CoT = 21.55; $SD = 2.99$; $t(78) = 1.91$, $p = 0.06$. Annotations generated with greedy decoding generally scored higher than nucleus sampling, M greedy = 22.95; $SD = 2.28$; M nucleus = 20.38; $SD = 2.34$; $t(78) = 4.98$, $p < 0.001$.

The research team's open-ended notes revealed that responses with nucleus sampling tended to be more verbose and more likely to focus on the wrong SEPs, DCIs, and CCCs. For example, an annotation (zero-shot nucleus) suggested the following to align an assessment item with DCIs: "To improve, it should be linked to the impact of human activity on the environment (ESS3.C: Human Impacts on Earth Systems)." This annotation focused on a different DCI within earth and space science (ESS3C) than the targeted DCI within physical science (PS2.B). Given these results, we applied greedy decoding for all prompts in phase 2.

Phase 2: Establishing Feasibility

In this phase, we experimented with four prompt combinations: zero-shot with context, few-shot with context, zero-shot with context and CoT, and few-shot with context and CoT. We tested these strategies with 42 questions from nine assessments (three assessments per grade level; including the original questions from phase 1) for a total of 168 annotations. We selected assessments that spanned science

domains and required open-ended responses to invite for sense-making. The assessments asked students to explain their thinking through text, oral presentation, drawing, and modeling. Appendix B describes the selected assessments. We used the same rubric (Table 3) to evaluate the annotations. Additionally, for each annotation, the coders wrote brief notes to explain their scoring. An example note reads: “Recognizes SEP at the correct grade level but does not overlap with educators about equity (that the assessment as is recognizes students’ thinking). Introduces inaccuracy that assessment lacks visuals.”

Collecting Educators’ Feedback (RQ2)

To elicit additional feedback about the quality of the annotations, we conducted semi-structured interviews with six educators. The interviews took place on Zoom, were audio-recorded, and lasted 45 min on average. Interview transcripts were automatically generated and cleaned by the researchers. Table 5 lists the participants’ backgrounds. Participants were eligible for the interviews if they taught science in states that used or adopted the NGSS. Participants received a \$25 gift card as compensation (Institutional Review Board, #24–1675). We recruited participants through convenience sampling, based on established research partnerships related to AI integration in education. All participants had experience with GenAI tools like text-based ChatGPT but not with MLLMs.

The interviews included three parts. First, we asked participants to describe their practice with formative science assessments. Participants both used existing assessments to align with state standards and developed their own. Second, participants selected a TAPS assessment from their grade levels and science domains. They critiqued how closely the assessments were aligned with their instructional practices and articulated what they might change about the assessments. Third, we explained the idea behind the AI annotation tool and asked participants what features such a tool might include. Then, we shared the MLLM

annotations of the same assessments that participants critiqued. We prompted participants to make sense of the AI annotations (e.g., “What do you think about this part of the annotation?”) and suggest ideas to make the annotations useful for their instructional practices. Appendix C lists the interview protocol.

Analytic Strategies

To answer the RQ1 around the feasibility of applying MLLMs to annotate science assessments, we reported descriptive statistics for the four prompting strategies. We ran analyses of variance (ANOVAs) to determine if there existed differences between prompting strategies for the total score (across rubric categories) and scores within each category (Standards, Overlap with Educators, Details, and Accuracy). Additionally, we ran *t*-tests to determine if there existed differences in scores for prompts generated with zero-shot versus few-shot learning, and with and without CoT. These findings contribute insights into effective prompting strategies to critique science assessments. Additionally, we used researchers’ open-ended notes on the output to contextualize the quantitative findings, specifically highlighting areas that the MLLM performed well or under-performed.

For RQ2 about educators’ perspectives, we conducted thematic analysis of the interview transcripts (Braun & Clarke, 2012; Byrne, 2022). The interview protocol allowed us to examine potential overlap between participants’ critiques of science assessments and the AI annotations, what they identified as useful and not useful about the annotations, and suggestions for improvement. We first familiarized ourselves with the data by reading the transcripts and identifying interesting excerpts. We then collaboratively generated descriptive codes (e.g., annotations lack understanding of learning objectives), formed themes based on the codes, and reviewed, refined, and defined the themes through three discussions.

Table 5 Teachers’ demographics and teaching experiences

Name	Grade	Domain	Race/ethnicity	Gender	US location	Years teaching	Years teaching science
Katie	Elementary	Earth science	White	Female	West	31	8
Hannah	Middle	Life science	White	Female	West	3	3
Amy	Middle	Physical science	White	Female	West	8	8
Mia	High	Physical science	Asian	Female	Southwest	31	31
Michael	High	Physical science	White	Male	West	20	7
Allan	High	Life science	White	Male	Southwest	14	14

Notes. Names are pseudonyms. Domain, the science domain of the assessments that teachers critiqued

Results

Comparing Model Results (RQ1)

Overall, we found significant differences in the total scores for prompt types, $F(3, 164) = 3.17$, $p = 0.03$. Post hoc Tukey HSD tests suggested differences between few-shot with CoT and zero-shot without CoT, $t = 2.77$, $p = 0.03$. We found no difference between prompt types for Standards, $F(3, 164) = 0.11$, $p = 0.95$, Overlap with Educators, $F(3, 164) = 2.27$, $p = 0.08$, or Accuracy, $F(3, 164) = 0.97$, $p = 0.41$, but significant differences for Details, $F(3, 164) = 3.23$, $p = 0.02$. Table 6 lists the descriptive statistics of the prompt conditions.

We further compared annotations using prompts with zero-shot versus few-shot, and with and without CoT (Fig. 3). Annotations generated with few-shot scored higher than those with zero-shot learning overall, few-shot: $M = 22.44$, $SD = 1.63$; zero-shot: $M = 21.70$, $SD = 1.63$, $t(166) = 2.93$, $p < 0.001$, as well as for Details, few-shot: $M = 3.35$; $SD = 0.57$; zero-shot: $M = 3.09$; $SD = 0.53$, $t(166) = 3.12$, $p = 0.002$. Those generated with CoT did not significantly

differ from those without, CoT: $M = 22.20$, $SD = 1.61$; no CoT: $M = 21.94$, $SD = 1.72$; $t(166) = 0.98$, $p = 0.33$. These results indicate that prompts that included example annotations performed better overall and added details to improve the assessments, compared to prompts without examples.

Finally, for robustness check, we ran ANOVAs to determine if the total scores—as a proxy of the annotations' quality—differed by grade levels (elementary, middle, and high school). We did not find difference, $F(2, 165) = 0.59$, $p = 0.56$.

Understanding of NGSS Dimensions and Learning Progressions

Most MLLM annotations scored high on Standards alignment overall (M range 11.39–11.49, max score 12, Table 6). They received scores of 3 or 4 (on a scale of 1–4) for each NGSS dimension (SEP, DCI, CCC), meaning that the annotations provided correct explanations of how assessment tasks aligned with standards. To illustrate, an annotation (few-shot, no CoT) critiques an assessment item for its alignment with the DCI:

Table 6 Descriptive statistics of prompt types, M (SD)

Prompt type	Standards (max 12)	Overlap educators (max 5)	Details (max 5)	Accuracy (max 5)	Total (max 27)
Zero-shot ($n = 42$)	11.39 (0.83)	2.59 (0.77)	3.07 (0.65)	4.51 (0.64)	21.56 (1.76)
Few-shot ($n = 42$)	11.49 (0.78)	2.85 (0.76)	3.34 (0.57)	4.63 (0.54)	22.32 (1.62)
Zero-shot, CoT ($n = 42$)	11.44 (0.81)	2.78 (0.85)	3.10 (0.37)	4.51 (0.55)	21.83 (1.50)
Few-shot, CoT ($n = 42$)	11.46 (0.78)	3.05 (0.86)	3.37 (0.58)	4.68 (0.52)	22.56 (1.64)

Notes. Bolded, highest values in each scoring category. Range 1–4 for each Standards dimension (DCI, CCC, SEP; total range 1–12 for Standards), range 1–5 for the other dimensions

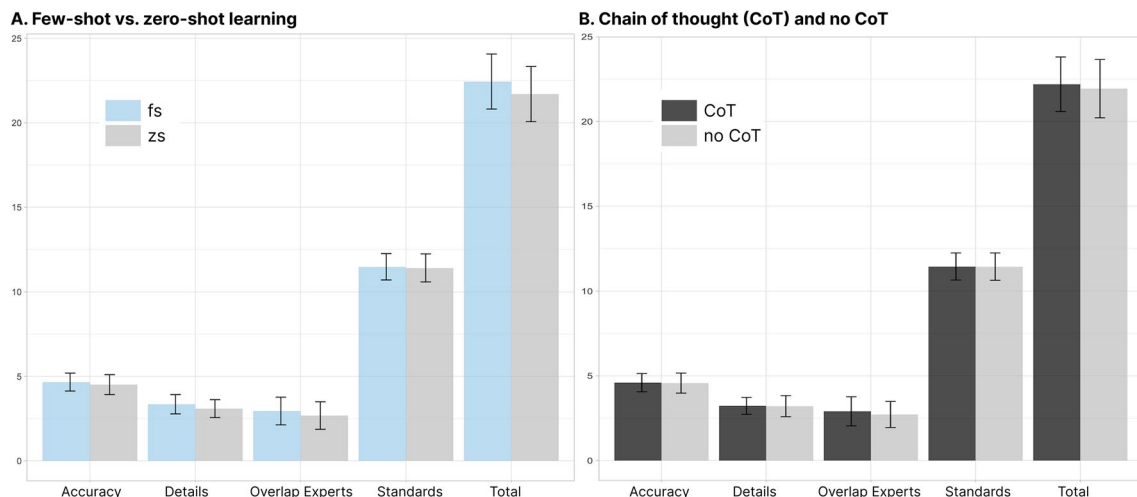


Fig. 3 Comparing results across prompt conditions. Panel A: few-shot and zero-shot; Panel B: Chain of Thought and no CoT. *Notes.* Bars represent mean, and error bars represent standard deviations

It touches on the disciplinary core idea of types of interactions (PS2.B) by exploring how magnets can interact with oil mixed with a powder [...] The question could be improved by explicitly connecting the phenomenon to the core idea of magnetic interactions.

This critique overlaps with TAPS educators' critique that the question does not target the DCI directly. The human annotation reads: "The scoring rubric suggests that this question assesses more of the DCI ... These are not assessed here because students do not bring an understanding of how distance influences the size of forces" This annotation by experienced educators provided more rationale than the MLLM; it specified how the task did not require comprehensive understanding of the DCI.

We observed that the AI annotations might lack specific rationales for why an item was or was not appropriate for the target grade level. Consider a TAPS (human) annotation for a life science assessment at the high school level: "While this does still require sense-making to connect their understanding of how populations change over time in response to environmental changes, this is relatively superficial and more appropriate to middle school (MS.LS.4C)." The annotation highlights how the scenario will likely only invoke DCIs at lower grades and cites the appropriate standards. In comparison, an example AI annotation (zero-shot, CoT) states:

Question 6 does capture the targeted science and engineering practices, disciplinary core ideas, and cross-cutting concepts to some extent. The question touches on the core idea of common ancestry and diversity but could be improved by explicitly asking students to explain how the data supports or refutes the idea of common ancestry.

This annotation tangentially aligns with the TAPS educators' insights that the DCIs were only partially addressed. However, it does not recognize the nuances in learning progressions across grade levels to critique the sophistication of the assessment.

Overlap with Educators' Feedback

Overall, the annotations were somewhat aligned with educators' ideas to improve the assessments (M range 2.59–3.05, out of 5). They noted ways to deepen SEPs, DCIs, or CCCs, and provided appropriate suggestions to create more relevant task scenarios and promote equity and students' sense-making. For instance, a suggestion to improve SEPs (*italicized*) reads: "The questions support the practice of *analyzing and interpreting data*, but they could better scaffold the process of *engaging in argument from evidence* by providing more explicit instructions on how to use the data to support a claim." This suggestion aligned with the TAPS (human)

annotations, which also highlighted how the SEP could be strengthened:

This requires students to make sense of the provided information using two dimensions: DCIs and SEPs. [...] Students had to *analyze and interpret simple data* (part of a 3-5 SEP), and *use provided information as evidence* to support claims (part of a 6-8 SEP) [...] Because the claim is supported by a single piece of evidence, this use of SEP is closer to middle school level than high school.

Regarding task scenario, in multiple cases, GPT-4V correctly evaluated when a task scenario was too broad or disconnected from students' experiences. To illustrate, the model provided the following critique for a middle school assessment:

Scenario: The question does not provide a specific scenario or context other than the general topic of natural hazards. This might limit the ability of students to relate the data to real-world situations. Idea for improvement: Include a brief scenario that contextualizes the data, such as a recent natural disaster that the students might be familiar with, to make the data more relevant and engaging.

This critique showed overlap with the TAPS annotations, which stated that the scenario did not provide a "clear phenomenon or puzzling problems," which might "limit how students engage with the task" and provide sense-making.

Notably, an area of misalignment between the MLLM and TAPS annotations was in the conceptualization of equity. The MLLM-generated annotations tended to associate equity with the accessibility of the learning materials. They often encouraged assessments to use accessible language, avoid jargons, and apply appropriate visual cues and activity scaffolds, for instance, to make sure "all students have the background knowledge and skills to interpret graphs." These suggestions showed sensitivity to different learner populations. However, they were still limited compared to the TAPS annotations. Beyond access, the TAPS evaluations by experienced science educators defined equity through opportunities for sense-making. They highlighted instances when an assessment item provided or did not provide choices to support learners' agency and multiple modes of sense-making. An example TAPS annotation might state: "The scenario is written with simple language but is text-heavy without any additional supports for engagement or understanding, which might impede some students' ability to connect with the scenario." The MLLM's annotations rarely noted these cases. Such misalignment potentially points to different conceptualizations of equity—as access or as sense-making to build on learners' own resources, knowledge, and practice (Fine & Furtak, 2020; Grapin et al., 2023).

Level of Details in Feedback

Generally, the annotations provided details to improve the assessments across prompting strategies (M range 3.07–3.37, out of 5). However, we found few cases that linked the critiques to specific details within the assessments (scores of 4 or 5). An excerpt that received a score of 3 might suggest improvement ideas on a high level:

Sensemaking - The assessment invites reasoning about the phenomenon of matter movement in ecosystems. To enhance sensemaking, students could be asked to explain how human actions can disrupt movements and what can be done to mitigate such effects.

In comparison, one with a score of 4 for the same item cited the assessment task (creating a model) as an opportunity to deepen sense-making:

The questions ask students to create models. They could better invite reasoning about phenomena by asking students to make predictions or inferences based on the model they create. For example, prompt students to predict what might happen to the ecosystem if a non-living element were removed or altered significantly.

Vague annotations that lack relevant details from the assessments can lead to confusion or distrust of the feedback (Hall et al., 2023; Hattie & Timperley, 2007). They might not result in timely follow-up actions to revise the assessments (Shernoff et al., 2017).

Feedback Accuracy

Finally, we evaluated the extent to which the annotations might include inaccuracies, given that GenAI models (including LLMs and MLLMs) are prone to hallucination and can make up details that do not exist in the input (Yang et al., 2023; Zhou et al., 2023). Across prompting strategies, the annotations did not frequently add details that did not exist in the inputs (M accuracy range 4.51–4.68; out of 5). The most common inaccuracy was when the annotations misstated that the assessments were lacking certain details (e.g., scenario, visuals, graphs, question wording), even though these elements already existed in the original assessments.

Educators' Perceptions of MLLM Annotations (RQ2)

We reported on the main themes (*italicized*) from teachers about how the annotations could inform instructional practices (RQ2). We first analyzed teachers' independent critiques of the science assessments, before viewing the AI annotations. Teachers focused on the *alignment with standards and students' understanding* in evaluating the

assessments. All participants scanned for standard alignment, to check which performance goals the assessment targeted and whether the tasks were framed to assess those goals. Four participants checked tasks' presentation to make sure they were grade appropriate and supported students' understanding of the tasks (e.g., vocabulary, use of graphs, illustration, and text). Additionally, two participants commented on the task scenarios and emphasized that these scenarios should be based on what students were familiar with (e.g., through problem-based learning or lab experiences). These practices aligned well with the Standards, Sense-making, and Scenario components in the TAPS dimensions and our design of the MLLM feedback (Table 2).

Next, we invited teachers to review the MLLM annotations. We used annotations generated with few-shot learning and CoT as the best performing strategy. We identified two affordances of the annotations: supporting *reflection on instructional practices* and providing *improvement ideas*. The annotations helped educators to reflect on how assessments might promote students' thinking. Katie (elementary, math and science teacher) remarked:

It [the annotation] is saying that we should ask the kids to test or revise their models based on evidence and reasoning. I like the critical thinking aspect ... really emphasize evidence and investigative approach.

Relatedly, the annotations led to generative inquiries, which we defined as questions that invited teachers to reflect on their instructional intent and students' understanding to develop new insights and actions. Consider the following quote from Mia (high school, chemistry):

Teachers always want feedback. You can ask another teacher, but feedback can start a conversation and have people think. What do we need to cover? What can we expect for answers? What do students need to know? This makes me think about how I can make students understand.

Further, the annotations provided *improvement ideas* to revise assessments. We observed this theme in four interviews. Michael (high school, physics) stated:

I like the sensemaking prompt to stretch students' thinking. "Encourage students to make predictions about what might happen" [reading out loud AI annotations] – this could be an extension for students who need to be stretched. This brings differentiation values.

Similarly, Mia shared that the annotations suggested ideas to embed learning progressions. "I might expect them to know how to draw the model. For Honors and AP, the questions can extend to being able to explain what is happening and adding data from labs." Teachers commented that the AI suggestions for rewording the questions (e.g., emphasizing

“phenomena” and “data practices”) helped to align the assessments with NGSS and state standards.

While the annotations were promising, teachers also identified areas for improvement, specifically a *lack of contextual understanding*. They highlighted that the annotations lacked understanding of teachers’ pedagogical approaches, lesson objectives, and the learning activities leading to the assessments. For example, Hannah (middle school, life science) commented: “It is saying all this about cultural awareness. That’s cool but is that the goal of the lesson?” Three participants mentioned a desire to have interactive exchanges with the AI (instead of just receiving static annotations), to generate more targeted feedback. Allan (high school, life science) emphasized that:

It needs a conversation about how the teacher teaches. I would want to talk [to AI] and change the prompt, I just really want to focus on genetics today. Do you have comments on that? – and then follow up, Great! What about increasing student discussion?

Two participants suggested an additional feature for the AI to have knowledge of students’ understanding of the target concepts, to simulate students’ responses to the assessments. “What do students still need help with and how do we fix that?” — Katie wondered. This feature can help teachers to plan backward and identify areas to focus instruction.

Discussion

We examine the feasibility of applying MLLMs (GPT-4V) to generating annotations to improve the quality of science assessments. Across prompting strategies, the AI-generated annotations generally provide coherent explanation of how the assessments align with the NGSS, somewhat overlap with K-12 educators’ annotations, and provide detailed improvement ideas without adding inaccurate elements to the responses (RQ1). Further, the MLLM annotations aligned with teachers’ critiques in their focus on standards, task scenarios, and sense-making. They can support teachers’ reflection on instructional practices and assessment revision (RQ2).

Our findings around the comparative performance of different prompting strategies contribute insights into how to prompt MLLMs. We find that few-shot learning—providing MLLMs with example annotations—can improve the overall quality and better ground the annotations in assessment details. Further, prompts with few-shot learning, chain of thought, and contextual information tended to score the highest across rubric categories.

While these results are promising, some AI annotations include inaccurate details that might cause confusion or lead to distrust (Hall et al., 2023). Further, the annotations are not

always aligned with educators’ insights, particularly regarding grade appropriateness, equity conceptualization, and instructional contexts. It is possible that the MLLM reflects its training data, which might overly associate equity with access. Recent analyses of NGSS-aligned instructional materials find a stronger focus on ensuring inclusion, representation, and accessibility than broadening conceptualizations of science and supporting sense-making (National Academies of Sciences Engineering and Medicine, 2022; Tzou et al., 2021). Additionally, prompting strategies that require examples like few-shot learning can be laborious, affecting educators’ adoption of the strategies. To address these challenges, we discuss how to improve the response quality and how we might apply MLLMs in science education.

Improving Response Quality

The areas of under-performance call for further incorporation of human insights into prompt development and MLLM applications (Wang & Demszky, 2023). Building on findings about the effectiveness of context and few-shot learning, we conjecture that adding instruction that associates equity with both access and science sense-making can better align the AI-generated annotations with educators’ visions. Providing the MLLMs with additional examples, such as paired assessment-critique of grade appropriateness, might improve their performance.

In addition, conversing with MLLMs in multi-turn exchanges could improve their output. In practice, users could send an initial prompt with the assessment context and task instruction. As suggested by the teacher interviewees, users can then follow up to clarify the models’ suggestions, explain their learning contexts and teaching approaches, or probe into specific scoring criteria. To demonstrate, consider the following exchanges where we followed up with GPT-4V to expand its explanation of equity (Table 7). The follow-up response shows a more aligned focus with the equity definition.

Implications for Education

Developing science assessments that target NGSS dimensions and equity in tandem is complex (Wertheim et al., 2016). Related work has used LLMs and MLLMs to generate questions for education assessments (Lee et al., 2023) or automatically score students’ responses (Lee et al., 2024). We demonstrate another use case to improve existing assessments that educators might already plan to implement. While not perfect, the AI annotations may provide relevant insights to help educators iterate upon their assessment design. It is encouraging that the generated annotations could provide concrete suggestions with considerations for the NGSS, task scenarios, equity, and

Table 7 Multi-turn exchanges with GPT-4V

Prompt 1	Critique a formative science assessment using the following categories [...] (see Fig. 2)
GPT-4V's response	Equity: The assessment could include more context or scaffolding to support students who might not be familiar with the concept of magnetic attraction or who have not had experiences with oil spills. This could include a pre-lesson or a hands-on activity to explore magnetism
Prompt 2	Beyond scaffold, we define equity as supporting students to understand the task context and promote students' sense-making. How does the question support equity in this sense?
GPT-4V's response	<p>Question 1 does a few things well:</p> <ul style="list-style-type: none"> - It uses a real-world problem (oil spills) that is likely to be familiar or at least understandable to most students, which can engage them and make the science feel relevant to their lives - The question is framed in a way that encourages students to think about the properties of materials and their interactions, which is a fundamental aspect of scientific inquiry <p>Improvement idea: Provide scaffolding to support sense-making</p> <ul style="list-style-type: none"> - To help students make sense of the phenomenon, the question could include prompts that guide them through the reasoning process ... asking students to describe what they think will happen before the experiment is conducted and then to reflect on their predictions afterward

sense-making. Findings from teacher interviews showed that the annotations led to reflection and inquiries for teachers and provided suggestions to differentiate instruction and assessment tasks.

We concur with prior work that has emphasized the importance of *AI literacy for educators* (Chiu et al., 2024; Steiss et al., 2024). AI literacy refers to technical knowledge (e.g., how AI works); applications of the technology (e.g., how to prompt MLLMs); and understanding of AI's capacities, limitations, and ethics (Chan & Hu, 2023; Chiu et al., 2024). Surveys of educators' attitudes toward emerging GenAI technologies have uncovered both excitement and cautiousness around how these technologies could be used effectively and fairly in instruction (Chiu, 2023; Hall et al., 2023). Critical engagement with the AI-generated annotations—including evaluating their accuracy, clarifying instructional objectives, identifying relevant areas, and ignoring irrelevant suggestions—helps to improve the tools' contextual understanding and supports their uses to develop science assessments.

Pedagogical knowledge also plays a role in effective uses of the annotations. Findings indicate that the MLLMs tend to perform better when provided with sample annotations (few-shot learning). Arguably, this prompting strategy is the most labor intensive, as it requires teachers to possess pedagogical knowledge to critique assessments. A promising model is to foster collaboration among teachers (Forkosh Baruch & Gadot, 2021) to curate high-quality assessments and example annotations (NGSS, n.da). Further, instructional coaching can provide teachers with models for creating and using formative assessments (Snodgrass Rangel et al., 2017). These examples can be included in the prompts to the MLLMs. Once provided with example responses, the MLLMs can generate useful outputs for additional questions and assessments to support assessment design at scale.

Limitations and Future Work

The limitations of the current paper provide directions for future work. First, we present an exploratory analysis of the MLLMs' feasibility in annotating assessments. We leveraged an existing dataset (the TAPS annotations) with specific focus on NGSS dimensions and equity. Future research can use the same approach to investigate MLLMs' performance in other assessment contexts and explore creative uses, such as using these models to simulate how students with different levels of understanding might respond to the assessments.

Second, pedagogical decisions are highly dependent on instructional settings. It is critical to examine how the annotations augment teachers' understanding of high-quality assessments, and what changes teachers might make based on the AI-created annotations. While we conducted exploratory interviews with a small number of teachers, follow-up studies can include a larger, more representative sample of educators as they apply the annotations to instruction.

Third, to improve the feedback quality, researchers and educators can consider other prompt engineering strategies and incorporate human expertise. For example, we did not find significant difference in annotations generated with and without CoT. This prompting strategy has been commonly applied in mathematics problem-solving and complex decision-making that can be broken down into subproblems (Feng et al., 2024). Refining the annotation prompts to specify the decision-making steps and illustrating these steps in few-shot examples could improve the performance of CoT prompts (Wei et al., 2022). Further, researchers can explore how to refine and extend the prompts' components (e.g., context, task, few-shot examples) to other instructional contexts.

Finally, we only involved one-off exchanges and did not iteratively prompt GPT-4V to refine its responses. Future

research should examine how well MLLMs perform in multi-turn exchanges and with more extensive prompt tuning.

Conclusions

Our research provides evidence of the feasibility of applying MLLMs to critiquing science assessments combining textual and visual inputs. This would give educators immediate ideas to improve assessment design, to promote three-dimensional and equitable science learning. The annotations did not differ in performance across grade levels, adding to the models' usability. Findings reveal important areas for improvement, particularly critiques of grade appropriateness, conceptualization of equity, and understanding of instructional contexts. To apply emerging AI tools effectively, we note the need for further prompt engineering. We call for promoting AI literacy and pedagogical knowledge for educators, to design effective strategies to prompt the tools and critically engage with the generated information to inform science instruction.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10956-024-10177-x>.

Author Contribution Ha Nguyen: conceptualization, methodology, software, formal analysis, investigation, writing — original draft, visualization, project administration, funding acquisition. Jack Hayward: investigation, writing — review and editing.

Funding This study was supported by a National Science Foundation grant, #2422286.

Data Availability The Task Annotation Project in Science (TAPS) dataset is available at <https://www.nextgenscience.org/task-annotation-project-science>. The interview data are protected and are not available due to institutional review board agreement.

Declarations

Ethical Approval This study was performed in line with the ethical standards of the institutional review board. The study was approved by the university IRB (#24–1675).

Informed Consent Participants have provided written consent for the analysis and reporting of their anonymized data.

Consent to Participate The study has obtained written consent from all participants to take part in the study protocols.

Consent for Publication All participants have consented to publish their anonymized data.

Competing Interests The authors declare no competing interests.

Statement Regarding Research Involving Human Participants The research involves human participants.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Achieve. (n.d). *Task annotation project in science: Methodology*. Retrieved February 1, 2024, from https://www.achieve.org/files/sites/default/files/methodology_02142019%20%284%29.pdf
- Ali, F., Choy, D., Divaharan, S., Tay, H. Y., & Chen, W. (2023). Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering. *Learning: Research and Practice*, 9(2), 135–147.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
- Braun, V., & Clarke, V. (2012). *Thematic analysis*. American Psychological Association.
- Brown, J. C. (2017). A metasyntesis of the complementarity of culturally responsive and inquiry-based science education in K-12 settings: Implications for advancing equitable science teaching and learning. *Journal of Research in Science Teaching*, 54(9), 1143–1173.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56(3), 1391–1412.
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1), 43.
- Chiu, T. K. (2023). The impact of generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*, 1–17.
- Chiu, T. K., Ahmad, Z., Ismailov, M., & Sanusi, I. T. (2024). What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open*, 6, 100171.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240), 1–113.
- Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., & Hoi, S. (2024). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.
- Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., & Sakr, M. (2024, January). A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming

- education. In *Proceedings of the 26th Australasian Computing Education Conference* (pp. 114–123).
- Estévez-Ayres, I., Callejo, P., Hombrados-Herrera, M. Á., Alario-Hoyos, C., & Delgado Kloos, C. (2024). Evaluation of LLM tools for feedback generation in a course on concurrent programming. *International Journal of Artificial Intelligence in Education*, 1–17.
- Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., & Wang, L. (2024). Towards revealing the mystery behind chain of thought: A theoretical perspective. *Advances in Neural Information Processing Systems*, 36.
- Fine, C. G. M., & Furtak, E. M. (2020). A framework for science classroom assessment task design for emergent bilingual learners. *Science Education*, 104(3), 393–420.
- Forkosh Baruch, A., & Gadot, R. (2021). Social curation experience: Towards authentic learning in preservice teacher training. *Technology, Knowledge and Learning*, 26(1), 105–122.
- Fulmer, G. W., & Polikoff, M. S. (2014). Tests of alignment among assessment, standards, and instruction using generalized linear model regression. *Educational Assessment, Evaluation, & Accountability*, 26(3), 225–240. <https://doi.org/10.1007/s11092-014-9196-z>
- Fulmer, G. W., Tanas, J., & Weiss, K. A. (2018). The challenges of alignment for the Next Generation Science Standards. *Journal of Research in Science Teaching*, 55(7), 1076–1100.
- Gan, W., Qi, Z., Wu, J., & Lin, J. C. W. (2023, December). Large language models in education: Vision and opportunities. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 4776–4785). IEEE.
- Grapin, S. E., & Llosa, L. (2022). Multimodal tasks to assess English learners and their peers in science. *Educational Assessment*, 27(1), 46–70.
- Grapin, S. E., Pierson, A., González-Howard, M., Ryu, M., Fine, C., & Vogel, S. (2023). Science education with multilingual learners: Equity as access and equity as transformation. *Science Education*, 107(4), 999–1032.
- Gunjal, A., Yin, J., & Bas, E. (2024). Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* 38(16), 18135–18143.
- Guzman-Orth, D., Supalo, C. A., Smith, D. W., Lee, O., & King, T. (2021). Equitable STEM instruction and assessment: Accessibility and fairness considerations for special populations. *ETS Research Report Series*, 2021(1), 1–16.
- Hall, E., Seyam, M., & Dunlap, D. (2023). Identifying usability challenges in AI-based essay grading tools. In *International Conference on Artificial Intelligence in Education* (pp. 675–680). Cham: Springer Nature Switzerland.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Heritage, M. (2021). *Formative assessment: Making it happen in the classroom*. Corwin Press.
- Herman, J. (2013). Formative assessment for Next Generation Science Standards: A proposed model. In *Invitational research symposium on science assessment*.
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*. https://iclr.cc/virtual_2020/poster_rygGQyrFvH.html
- Imani, S., Du, L., & Shrivastava, H. (2023). MathPrompter: Mathematical reasoning using large language models. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.
- Kaldaras, L., Akaze, H., & Krajcik, J. (2021). Developing and validating Next Generation Science Standards-aligned learning progression to track three-dimensional learning of electrical interactions in high school physical science. *Journal of Research in Science Teaching*, 58(4), 589–618.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.
- Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 100210.
- Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, 1–33.
- Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 100213.
- Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., & Gao, J. (2024). LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.
- Lin, Y. T., & Chen, Y. N. (2023). LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)* (pp. 47–58).
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35.
- Lu, X., & Wang, X. (2024). Generative students: Using LLM-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale* (pp. 16–27).
- Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., & Gao, J. (2023, October). MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.
- Lyon, E. G. (2013). Learning to assess science in linguistically diverse classrooms: Tracking growth in secondary science preservice teachers' assessment expertise. *Science Education*, 97(3), 442–467.
- Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). GPTEach: Interactive TA training with GPT-based students. In *Proceedings of the Tenth ACM Conference on Learning@ Scale* (pp. 226–236).
- McElhaney, K. W., Zaidi, S., Gane, B. D., Alozie, N., & Harris, C. J. (2018). Designing NGSS-aligned assessment tasks and rubrics to support classroom-based formative assessment. In *NARST Annual International Conference, Atlanta, GA*.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1(1), 3–62.
- Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., & Luo, P. (2024). EmbodiedGPT: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36.
- National Academies of Sciences, Engineering, and Medicine. (2022). *Science and engineering in preschool through elementary grades: The brilliance of children and the strengths of educators*. National Academies Press. Retrieved March 10, 2024, from <https://nap.nationalacademies.org/catalog/26215/science-and-engineering-in-preschool-through-elementary-grades-the-brilliance>
- National Research Council. (2012). A framework for K-12 science for education: Practices, crosscutting concepts, and core ideas. Committee on a Conceptual Framework for New K-12 Science Education Standards.
- National Research Council. (2014). Developing assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig, and A.S. Beatty, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

- NGSS. (n.da). *Assessment Resources*. Retrieved February 1, 2024, from <https://www.nextgenscience.org/assessment-resources/assessment-resources>
- NGSS (n.db). Frequently asked questions. Retrieved February 1, 2024, from https://nextgenscience.org/sites/default/files/resource/files/Task%20tools_FAQ.pdf
- NGSS. (n.dc). *Task annotation project in science*. Retrieved February 1, 2024, from <https://www.nextgenscience.org/task-annotation-project-science>
- NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. The National Academies Press.
- Nguyen, H., & Allan, V. (2024). Using GPT-4 to provide tiered, formative code feedback. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 958–964).
- Nguyen, H. A., Stec, H., Hou, X., Di, S., & McLaren, B. M. (2023). Evaluating ChatGPT's decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning* (pp. 278–293). Cham: Springer Nature Switzerland.
- Nichols, E., Gao, L., & Gomez, R. (2020). Collaborative storytelling with large-scale neural language models. In *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games* (pp. 1–10).
- Osborne, J.; Pecheone, R.; Quinn, H.; Holthuis, N.; Schultz, S.; Wertheim, J.; and Martin, M. (2015). A system of assessment for the Next Generation Science Standards in California: A discussion document. Retrieved from snapgse.stanford.edu December 1, 2015.
- Penuel, W. R., Turner, M. L., Jacobs, J. K., Van Horne, K., & Sumner, T. (2019). Developing tasks to assess phenomenon-based science learning: Challenges and lessons learned from building proximal transfer tasks. *Science Education*, 103(6), 1367–1395.
- Philip, T. M., & Azevedo, F. S. (2017). Everyday science learning and equity: Mapping the contested terrain. *Science Education*, 101(4), 526–532.
- Pichai, S., & Hassabis, D. (2023). Introducing Gemini: our largest and most capable AI model. *Google*. Retrieved February 12, 2024.
- Pierson, A. E., Clark, D. B., & Brady, C. E. (2021). Scientific modeling and translanguaging: A multilingual and multimodal approach to support science learning and engagement. *Science Education*, 105(4), 776–813.
- Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment*, 9(3–4), 79–105.
- Rodriguez, A. J. (2015). What about a dimension of engagement, equity, and diversity practices? A critique of the Next Generation Science Standards. *Journal of Research in Science Teaching*, 52(7), 1031–1051.
- Shelton, A., Smith, A., Wiebe, E., Behrle, C., Sirkin, R., & Lester, J. (2016). Drawing and writing in digital science notebooks: Sources of formative assessment data. *Journal of Science Education and Technology*, 25, 474–488.
- Sherhoff, E. S., Lekwa, A. J., Reddy, L. A., & Cocco, C. (2017). Examining teachers' attitudes and experiences with coaching to inform research-based practice: An iterative developmental design study. *Journal of Educational and Psychological Consultation*, 27(4), 459–485.
- Siegel, M. A., Menon, D., Sinha, S., Promyod, N., Wissehr, C., & Halverson, K. L. (2014). Equitable written assessments for English language learners: How scaffolding helps. *Journal of Science Teacher Education*, 25(6), 681–708.
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., & Natarajan, V. (2023). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.
- Smith, A., Leeman-Munk, S., Shelton, A., Mott, B., Wiebe, E., & Lester, J. (2018). A multimodal assessment framework for integrating student writing and drawing in elementary science learning. *IEEE Transactions on Learning Technologies*, 12(1), 3–15.
- Snodgrass Rangel, V., Bell, E. R., & Monroy, C. (2017). A descriptive analysis of instructional coaches' data use in science. *School Effectiveness and School Improvement*, 28(2), 217–241.
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894.
- Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., & Collier, N. (2022). A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35, 21548–21561.
- Swanson, B., Mathewson, K., Pietrzak, B., Chen, S., & Dinalescu, M. (2021). Story centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 244–256).
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Tzou, C., Bang, M., & Bricker, L. (2021). Commentary: Designing science instructional materials that contribute to more just, equitable, and culturally thriving learning and teaching in science education. *Journal of Science Teacher Education*, 32(7), 858–864.
- Wang, R., & Demszyk, D. (2023). Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 626–667).
- Wang, Z., Li, M., Xu, R., Zhou, L., Lei, J., Lin, X., & Ji, H. (2022). Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35, 8483–8497.
- Wang, L., Hu, Y., He, J., Xu, X., Liu, N., Liu, H., & Shen, H. T. (2024, March). T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence* 38(17), 19162–19170).
- 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.
- Wertheim, J., Osborne, J., Quinn, H., Pecheone, R., Schultz, S., Holthuis, N., & Martin, P. (2016). An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS. *Palo Alto, CA: Stanford NGSS Assessment Project Team (SNAP)*.
- Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. *Handbook of formative assessment*, 18–40.
- Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). The dawn of MLLMs: Preliminary explorations with GPT-4V (ision). *arXiv e-prints*, arXiv:2309.
- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1–38.
- Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., & Yao, H. (2023, November). Analyzing and mitigating object hallucination in large vision-language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
- Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.