Proceedings of Machine Learning Research 1:1–28, 2024 FM-EduAssess at NeurIPS 2024 Workshop

Gemini Pro Defeated by GPT-4V: Evidence from Education

Gyeonggeon Lee GYEONGGEON.LEE@NIE.EDU.SG
Department of Natural Sciences and Science Education, National Institute of Education, Singapore, Singapore
Lehong Shi LS77437@UGA.EDU
National GENIUS Center, Department of Workforce Education and Instructional Technology, University of Georgia, Athens, GA, USA
Ehsan Latif EHSAN.LATIF@UGA.EDU
AI4STEM Education Center, National GENIUS Center, University of Georgia, Athens, GA, USA
Xiaoming Zhai* XIAOMING.ZHAI@UGA.EDU
AI4STEM Education Center, National GENIUS Center, Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA, USA

Abstract

This study compared the classification performance of Gemini Pro and GPT-4V in educational settings. Employing visual question answering (VQA) techniques, the study examined both models' abilities to read text-based rubrics and then automatically score student-drawn models in science education. We employed both quantitative and qualitative analyses using a dataset derived from student-drawn scientific models and employing NERIF (Notation-Enhanced Rubrics for Image Feedback) prompting methods. The findings reveal that GPT-4V significantly outperforms Gemini Pro in terms of scoring accuracy and Quadratic Weighted Kappa. The qualitative analysis reveals that the differences may be due to the models' ability to process fine-grained texts in images and overall image classification performance. Even adapting the NERIF approach by further de-sizing the input images, Gemini Pro seems not able to perform as well as GPT-4V. The findings suggest GPT-4V's superior capability in handling complex multimodal educational tasks. The study concludes that while both models represent advancements in AI, GPT-4V's higher performance makes it a more suitable tool for educational applications involving multimodal data interpretation.
Keywords: Gemini, GPT-4V, Vision, Education, and Automatic Scoring

1. Introduction

The keynote and education methods have been in step with the concurrent technologies (Organisation for Economic Co-operation and Development, 2019). Nowadays, artificial intelligence (AI) technologies based on deep neural networks are getting closer to realizing the long-pursued AI in education (AIEd) initiative (Luckin et al., 2016; Hwang et al., 2020). Especially in the 2020s, the development and release of large language models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are facilitating the integration of AI technologies into educational research and practice due to their high capabilities of natural language processing, understanding and reasoning, and generation for various tasks such as teaching

* corresponding author

material preparation, data augmentation, item generation, and automatic scoring (Lo, 2023; Fang et al., 2023; Lee et al., 2023d). Particularly, user-friendly LLMs such as ChatGPT have hugely impacted the global trend of AI usage in research and everyday lives by enabling conversational interactions between humans and machines based on natural languages.

What even sputters the innovations in AIEd research is the realization of visual question answering (VQA), which enables machines to answer the user's natural language-based query on a given image (Antol et al., 2015), which can assist in learning and teaching. VQA has the potential in educational settings when using visual data to interpret students' ideas on the learning content or classroom dynamics (Lee et al., 2023b). For example, teachers may feed a computer with a student-drawn model to explain refraction, and the computer can recognize the lights and angles and provide feedback for improvements(Wang et al., In press). VQA, in principle, suggests the possibility of multimodal interaction between humans and computers, enabling multimodal learning and assessment with AI (Lee et al., 2023b).

GPT-4V, released on September 2023 by Open AI (2023d), has provided the global users with the affordance of VLM, as an extended module of ChatGPT and GPT-4. Since their release, ChatGPT and GPT-4 have notably influenced educational research and applications due to the high capability of natural language processing, understanding, and generation for various educational tasks and applications. Much research (Dempere et al., 2023; Lo, 2023; Ausat et al., 2023) suggested the effectiveness of ChatGPT in various educational tasks, such as aiding automated assessment, suggesting educational materials, and facilitating personalized learning. For example, Zhai (2022) conducted a piloting study of user experiences, focusing on the effectiveness of using ChatGPT as a tool to support writing a research paper, suggesting its potential and benefit in processing natural language-based tasks. However, the recent release of GPT-4V expanded its educational potential to facilitate learning and assessments featuring multimodalities, such as text and image (Lee and Zhai, 2023; Lee et al., 2023b). For example, Lee and Zhai (2023) tested GPT-4V's performance of automatic scoring student-drawn models by processing problem image and textual context with rubric and reported its multinomial classification accuracy as mean =.51, SD = .037. While this level of accuracy is remarkable, effort is needed to improve it before this method can be used in classrooms.

The release of Gemini Pro, an extended module of Bard, by Google DeepMind in December 2023 provided an emergent opportunity to fill this gap. Google DeepMind claimed that Gemini "is built from the ground up for multimodality - reasoning seamlessly across text, images, video, audio, and code" and is the first model to outperform human experts on MMLU (massive multitask language understanding) (Google, 2023). Google Bard and Gemini Pro were released only a few months later than ChatGPT and GPT-4V, which shows the technical competition for state-of-the-art AI services. Google Bard and Gemini Pro have also been tested and used for reasoning, answering knowledge-based questions, solving math problems, translating between languages, generating code, and acting as instruction-following agents through benchmarks Google (2023). Akter et al. (2023a) performed an extensive study to test the capabilities and functionalities of Gemini and compared it against GPT-3.5 turbo and found that Gemini underperforms GPT-3.5 turbo in reasoning, generating code, and solving math problems. Waisberg et al. (2023) provided a side-by-side comparison of Bard (underlying Gemini-pro) with ChatGPT for its application in ophthalmology. McIntosh et al. (2023) provided a comprehensive survey on their transformative impacts of Mixture of Experts (MoE), and multimodal learning and suggested the speculated advancements towards Artificial General Intelligence Goertzel (2014); Latif et al. (2023a). Furthermore, Microsoft researchers Liu and Chen (2023) performed a comprehensive evaluation of the performances of GPT-4V and Gemini Pro using VQA online dataset Chen et al. (2023) and reported that the average accuracy of GPT-4V and Gemini is 0.53 and 0.42, respectively. If Gemini Pro's performance is also promising for educational tasks, it could be used for a variety of purposes while competing with GPT-4V. Also, since there have been scarce studies on the use of Gemini Pro for education, exploring prompt engineering methods to elicit its full potential will also contribute to the AIEd research field.

However, as known to the authors, no studies have examined the eligibility of Gemini Pro in educational settings, not to mention the comparison of its performance with GPT-4V in dealing with VQA tasks. In this situation, with competing technologies struggling for SOTA, we suggest that comparing their applicability and effectiveness to educational studies is of timely need and could give a direction to the ongoing AIEd initiative regarding which VLM model could best serve educational tasks for researchers and practitioners. In this study, we answered to the following research questions. 1. How are the scoring performance of GPT-4V and Gemini Pro? 2. What are the characteristics of GPT-4V and Gemini Pro? 3. How Gemini Pro's performance on an educational task can be improved?

2. Gemini Pro vs GPT-4V

The development of Large Multimodal Models (LMMs) aims to expand upon the capabilities of Large Language Models (LLMs) by integrating multi-sensory skills to achieve even stronger general intelligence, thus supporting more natural human-computer interactions Yang et al. (2023). LMMs have one shared embedding space to integrate and process multiple data modalities, such as text, image, audio, video, 3D, etc. GPT-4V and Gemini Pro, the latest LMMs, can take images and/or text as inputs to perform various language, vision, and vision-language tasks, such as language translation and coding Devlin et al. (2018), image recognition Yu et al. (2023), object localization Yang et al. (2023), visual question answering Li et al. (2023), and visual dialogue Zhu et al. (2023). Furthermore, in the context of education, LLMs can be used for automatic scoring (Latif and Zhai, 2023) by applying chain-of-thought (Lee et al., 2023a). A close examination of the two state-of-the-art models in terms of their architecture, training approach, training datasets, performance, capabilities, safety, and applications can illuminate the usability and affordances of the two models, especially in education.

2.1. Architecture

GPT-4V, a product of OpenAI, is built on a transformer-based model Vaswani et al. (2017) designed to understand context and meaning through relationships in sequential data. This architecture is known for its ability to handle complex language and image-processing tasks Open AI (2023a). In contrast, Gemini Pro, developed by Google, is a large multimodal model that goes beyond processing text and images to include audio and video inputs. This broader range of input types suggests a more versatile and comprehensive approach to multimodal learning Google (2023). The architectural comparison can be seen in Fig. 1, as OpenAI has not released the exact architecture, we have anticipated its architecture based on the available informationOpen AI (2023c).

2.2. Training Approach

The training regime of GPT-4V includes advanced techniques such as Reinforcement Learning from Human Feedback (RLHF) Christiano et al. (2017). This approach refines the model's output based on human input, ensuring more accurate and contextually relevant responses. Additionally, GPT-4V employs a loss prediction method grounded in Power Law Kaplan et al. (2020), utilizing scaling laws to optimize training efficiency Henighan et al. (2020). On the other hand, Gemini Pro is trained using Google's tensor processing unit TPUv4 Jouppi et al. (2023), which offers high computational power. Its extended 32K context length allows for the processing of much larger chunks of data at once, potentially leading to deeper insights and understanding in complex tasks Google (2023). TPUs are trained to support 32k context length, employing efficient attention mechanisms (e.g. multi-query attention Shazeer (2019)). TPUv4 accelerators are deployed in "SuperPods" of 4096 chips, each connected to a dedicated optical switch, which can dynamically reconfigure 4x4x4 chip cubes into arbitrary 3D torus topologies in around 10 seconds Jouppi et al. (2023).

2.3. Training Datasets

GPT-4V is a visual extension Open AI (2023b) of GPT-4 and has been trained with bulk on online image data along with various sources, including books, journals, code, and other text formats, which were included in the GPT-4 training dataset Open AI (2023a). Thanks to intensive training, the AI model can now process and produce language and images that nearly match human-generated stuff. Because it is a generative model, GPT-4V can create new content in response to image and textual inputs. It can quickly and effectively analyze and understand the image using the inbuilt image decoding modules and massive amounts of textual data because of its dual-transformer design. It is particularly good at translating languages given in an image, for example, making cross-language communication easily written in an image. It may also produce various artistic content types and offer educational answers to user inquiries.

On the other hand, Gemini models are trained on a dataset that is both multimodal and multilingual. Their pertaining dataset uses data from web documents, books, and code and includes image, audio, and video data. They use the SentencePiece tokenizer Kudo and Richardson (2018) and find that training the tokenizer on a large sample of the entire training corpus improves the inferred vocabulary and subsequently improves model performance.

2.4. Performance

Regarding performance, GPT-4V has shown significant improvements over previous models in visual input tasks. It has been reported to be more adept at generating factual responses and less likely to produce disallowed content, making it a more reliable choice in sensitive applications Yang et al. (2023); Zhou et al. (2023); Open AI (2023b). In education, researchers found that GPT-4V is capable of understanding scoring rubrics and scoring students' drawn models to science phenomena with a certain degree of accuracy Lee and Zhai (2023).

With its sophisticated multimodal reasoning capabilities, Gemini understands and synthesizes insights from various inputs, including complex mathematical and scientific data Google (2023). Google claimed that the newly released Gemini models notably achieved a milestone by reaching human-expert performance on the widely researched massive multitask language understanding exam benchmark. Furthermore, it enhanced the SOTA across all of the 20 multimodal benchmarks we investigated Google (2023).

However, the outcomes from Gemini Pro are not consistent. For example, Google's report compared the performance of Gemini with GPT-4V for their ability to understand images. They evaluated the models on four capabilities: high-level object recognition using captioning, fine-grained transcription using tasks, chart understanding, and multimodal reasoning using tasks Akter et al. (2023a). The model is instructed to provide short answers aligned with the specific benchmark for zero-shot QA evaluation. All numbers are obtained using greedy sampling and without external optical character recognition tools. Results reported in Google's technical report have shown that GPT-4V outperformed Gemini Pro for all benchmarks with, on average, 7% higher accuracy. We also have presented the results in Table 1 by leveraging Google's permission to replicate and use the plots for research purposes. However, it is widely acknowledged that the capabilities of AI models, such as Gemini Pro and GPT-4V, in analyzing multimodal inputs and reasoning depend on context-specific tasks, which can provide a comprehensive qualitative explanation of their performance. Hence, in this study, we aim to analyze the performance of Gemini Pro and GPT-4V for educational tasks by providing assessment items, examples, rationales, and student responses in a condensed image and prompt models to categorize student responses based on the rationale provided in the examples.

2.5. Capabilities

GPT-4V's capabilities extend to processing both text and image inputs, making it versatile in applications like content creation, language translation, and educational tools. It has been integrated into various platforms, demonstrating its adaptability in different use cases Open AI (2023c); Lyu et al. (2023). Gemini Pro, with its ability to handle a wider range of input types, is designed for both heavy-duty cloud applications and on-device solutions. This flexibility indicates a focus on scalability and accessibility in various environments Google (2023).

2.6. Safety

Safety and alignment are key concerns in AI development, and both OpenAI and Google's Gemini team highlight the concerns on their website. GPT-4V has shown a significant reduction in its likelihood to respond to disallowed content requests and an increase in producing factual responses. These improvements are crucial for maintaining ethical standards in AI interactions Open AI (2023b) and reducing potential AI biases (Latif et al., 2023b). Gemini Pro has undergone extensive bias and toxicity analysis, with Google collaborating with external experts to identify and mitigate potential risks. Such efforts indicate the increasing importance placed on the ethical development of AI Google (2023).

2.7. Applications in Education

Since its release, GPT-4V has been broadly integrated with existing technologies, ranging from integrating with Microsoft Bing (Yang et al., 2023) for enhanced search capabilities to collaborating with Duolingo for language learning advancements (Gimpel et al., 2023). These partnerships demonstrate GPT-4V's utility in improving user experience and knowledge management across domains Wu et al. (2023).

(Senkaiahliyan et al., 2023) demonstrated the capabilities of GPT-4V for clinical education through medical image interpretation and found that GPT-4V can identify and explain medical images but cannot provide safe clinical decisions and diagnostics. Similarly, Xu et al. (2023) evaluated GPT4-V's capabilities for ophthalmological studies and reported 63% accuracy of GPT4-V in diagnosing ocular images. For automatic scoring by providing problem image and textual context with rubric, GPT-4V was able to achieve 51% accuracy for science based assessments (Lee and Zhai, 2023). Additionally, a comprehensive survey on multi-modality of AI for Education (Lee et al., 2023b) focused on GPT-4V's capabilities to revolutionize education technology and challenges of stepping forward to artificial general intelligence.

Likewise, Gemini Pro, being integrated into products like Google Bard and Pixel, enhances reasoning, planning, and writing capabilities. Its availability through the Gemini API[1] makes it a valuable resource for developers and enterprise customers, showcasing its potential applicability in education technology Google (2023). Furthermore, Akter et al. (2023a) reported Gemini-Pro's capabilities for solving math problems with a high accuracy of 69.67% for the GSM8K dataset (Cobbe et al., 2021), but we did not find any such study as of December 23rd, 2023 on science educational assessment which again emphasize the significance of our study.

Overall, both GPT-4V and Gemini Pro represent significant advancements in the field of AI and language models. While they share some commonalities in terms of their multimodal capabilities, their differences in architecture, training methodologies, performance, safety measures, and applications illustrate AI technology's diverse and evolving landscape. A comprehensive comparison details are also presented in Table 2. These models push the boundaries of what AI can achieve and raise important considerations for their ethical and practical implementation in various sectors. However, GPT4-V has shown higher image understanding performance than Gemini-pro, as evidenced by Table. 1.

3. Methods

3.1. Materials and Dataset

This study reanalyzed student-created scientific models from a dataset derived from a primary study Zhai et al. (2022). The items, formulated by the NGSA team (Harris et al., 2024), are designed to align with the *NGSS* (NGSS Lead States, 2013) performance expectations. They are part of a three-dimensional assessment strategy, integrating disciplinary core ideas, cross-cutting concepts, and science and engineering practices.

In our experiment involving six items, we selected 100 test cases for each item, using random sampling. The test datasets maintained a balanced distribution across three profi-

1. https://ai.google.dev/docs

ciency levels: 34 cases for 'Proficient,' 33 for 'Developing,' and 33 for 'Beginning,' with each category constituting one-third of the cases.

3.2. Experimental Design

This study conducted three experiments to answer the corresponding research questions. First, in the quantitative study, we compared the image classification performance of GPT-4V and Gemini Pro for automatic scoring of student-drawn models for science phenomena. We provided the two VLMs with the same prompt and image input for each task during this experiment. The performance metrics are quantitatively reported. Second, in the qualitative study, we heuristically explored the prompt that increases Gemini Pro's image classification performance. Third, we downsized the image input and examined how the performance of Gemini Pro changes.

3.3. Prompt Design

For the prompt design, we adopted the NERIF (Notation-Enhanced Rubric Instruction for Few-shot Learning) method from Lee and Zhai (2023). In the NERIF, a user first writes a prompt with components essential to the task. After that, validation cases are used to confirm whether the prompt achieves the user intended for the task. If not, the *Notation-Enhanced Scoring Rubric* is introduced/revised in the prompt, which combines *human experts' scoring rules*, *scoring rules aligned with proficiency levels*, and *instructional notes* for better scoring. The validation and revision of the prompt are repeated until the improvement of the machine's performance reaches saturation.

For the first experiment, we employed the prompt used in Lee and Zhai (2023), which used GPT-4V to analyze combined visual and text questions for automatic scoring, with slight revision. The input image and prompt used in Lee and Zhai (2023) consisted of seven components, as shown in Figs. 2-4: (1) *Role* that designates ChatGPT's role as a science teacher that scores student-drawn model, (2) *Task* that explains what ChatGPT is requested to do, (3) *Problem context* that ChatGPT has to retrieve from an image, (4) *Notation-Enhanced Scoring Rubrics*, (5) Nine human scoring *Examples* for few-shot learning (3 for 'Proficient,' 3 for 'Developing,' and 3 for 'Beginning' cases), (6) *Models drawn by students* as test cases, and (7) Temperature/top_p = 0/0.01 as hyper-parameters. In Lee and Zhai (2023), (3) and (5) were given in the first attached image (left of Fig. 2; Fig. 3), (6) in the second attached image (right of Fig. 2), and others in the text (Fig. 4)

The prompt has been slightly changed in this study, since there was an additional constraint as Gemini Pro was limited to processing up to one image as input, different from GPT-4V, which could process up to four drawn models simultaneously (as of Dec 12, 2023). To accommodate both LMMs, we changed the input protocol by merging (3), (5), and (6) into one image (Fig. 2). The structure of image input and text prompt was consistent throughout the Tasks. The individual image given to the two VLMs was 2,935 (width) × 3,515 (height) pixels and about 1MB size large, throughout the tasks.

For the second experiment, we started from asking the VLMs "what do you see in the given image?" with the input image. After that, we asked them to "tell me about how the 'PROBLEM CONTEXT' is given in the attached image, in detail" to check whether they appropriately retrieve information from the image.

For the third experiment, in the situations where VLM(s) fail to proceed with the given Tasks (in the first experiment), we tried various strategies to make these done with the VLM(s), including breaking down the images and text prompt into smaller compartments.

4. Findings

The details of the two models' scoring outcomes and potential reasons are presented in the following subsections.

4.1. Scoring Performance on Student Drawn Models: Gemini Pro vs. GPT-4V

4.1.1. Scoring Accuracy

We found that the response patterns of GPT-4V and Gemini Pro differed substantially. GPT-4V returned image classification results for all six Tasks. However, Gemini Pro produced image-classifying responses as instructed only for Task 42. For other Tasks, Gemini Pro produced various alternative forms of responses, rather than classification. Therefore, GPT-4V successfully scored students' drawn models for 600 cases, while Gemini Pro only scored around 100 cases. This indicates Gemini Pro's incapability of processing a large image aggregated with both text prompts and image information. The scoring performance of the two VLMs is presented in Table 3. Note that GPT-4V's performance is presented for every Task - in contrast, Gemini Pro's performance is presented only for Task 42, and all others are presented as 'NA' (Non-Available).

Specifically, the mean accuracy of GPT-4V on the image classification Tasks was M = .48, SD = .06. Also, the mean precision was M = .50 (SD = .09), recall M = .46 (SD = .05), and F1 M =.43 (SD = .06). The category-wise accuracy was highest for 'Beginning' cases (M = .67; SD = .12), followed by 'Developing' (M = .58; SD = .16) and 'Proficient' cases (M = .26, SD = .19).

For Gemini Pro, we only received the scoring accuracy for Task 42. The mean accuracy was M = .3 (SD = NA). Also, the average precision was M =.3 (SD = NA), recall .30 (SD = NA), and F1 .3 (SD = NA). Note that the accuracy was less than .33, the expected value of random response in a trinomial classification task. The category-wise accuracy was highest for 'Developing' cases (M = .44), followed by 'Proficient' (M = .26) and 'Beginning' cases (M = .21).

To sum up, the number of successful production of the anticipated answer type (600 for GPT-4V versus 100 for Gemini Pro), and the classification accuracy (.48 for GPT-4V and .30 for Gemini Pro, which means GPT-4V shows 60% higher accuracy than Gemini Pro) quantitatively show that GPT-4V's VQA performance on automatic scoring task is superior than that of Gemini Pro.

4.1.2. Quadratic Weighted Cohen's Kappa

Although scoring accuracy provides a measure to understand the performance of Gemini Pro on understanding the scoring rubric and automatic scoring of drawn models compared to GPT-4V, this measure does not reflect that machine-human disagreements differ, some disagreement are more severe than others. For example, misscoring a "Proficient"-level student-drawn model as "Beginning" is more severe than as "Developing," and thus should receive more penalty. To further understand this difference between the two VLMs, we compared the confusion matrices of automatic scoring for Task 42 between Gemini Pro and GPT-4V (Table 4) and calculated the Quadratic Weighted Cohen's Kappa,

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where

$$P_o = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{N}$$

and

$$P_e = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} e_{ij}}{N^2}$$

In these formulas:

- k is the number of rating categories.
- $w_{ij} = (i-j)^2$ is the weight for the disagreement between categories i and j, representing quadratic weighting.
- x_{ij} is the observed count of ratings in the cell corresponding to category i by rater 1 and category j by rater 2.
- e_{ij} is the expected count under chance agreement, calculated as (row total of $i \times$ column total of $j)/N$.
- N is the total number of ratings.

As presented in Table 3, the Quadratic Weighted Kappa of GPT-4V on Task 42 was .37. And that for other tasks spanned from .26 to .50, with M = .37 and SD = .09, which can be considered as 'Fair' to 'Moderate' level (Landis and Koch, 1977). In contrast, the Quadratic Weighted Kappa of Gemini Pro on Task 42 was -.14. Note that since Kappa aims to correct chance agreement, value 0 indicates that all agreements are by chance. A negative value suggests that the agreement is worse than guessing. This could be because Gemini misinterpreted the information and made systematically. (see Table 4

To sum up, the scoring accuracy, quadratic weighted Cohen's Kappa and confusion matrix show that GPT-4V's image processing capability for automatic scoring is superior to Gemini Pro.

4.2. Qualitative Characteristics of Scoring by Gemini Pro vs. GPT-4V

To uncover the divergent performance between Gemini Pro and GPT-4V, we qualitatively analyzed the scoring patterns of GPT-4V and Gemini Pro with the NERIF prompting methods. Note that we not only present what the two LLM's returned for the complete prompt and input image, but also the results of heuristic examination of their VQA performance.

4.2.1. Finding 1: Gemini Pro fails to recognize fine-grained texts printed in the image | GPT-4V succeeds

We gave the problem context and example (left of Fig. 2) as one input image and asked the two VLMs, "What do you see in the given image?". The outcomes from GPT-4V and Gemini Pro are presented in Fig. 5. GPT-4V responded, "The image you've uploaded appears to be an educational material that explains the diffusion of red dye water at different temperatures. ..." which is correct information about the given image (left of Fig. 5). It also correctly described the locations and details of the problem context, nine examples, rationale for proficiency, three proficiency levels, etc. In contrast, Gemini Pro responded that "The given image is a poster showing various examples of a "rad dye illusion" (4:042.03-602). The poster is divided into two sections: "EXAMPLE" and "KEY"." (right of Fig. 5). This falsely describes the information in the given image. The image is not a poster; The image says about "Red dye diffusion (ID#: 042.03-e02)", not "Rad dye illusion (4:042.03-602)". Also, the image is divided into "PROBLEM CONTEXT" and "EXAMPLE," not "EXAMPLE" and "KEY."

This outcome shows that Gemini Pro fails to precisely recognize fine-grained texts in the image and generates false information, while GPT-4V succeeds with the same image.

4.2.2. Finding 2: When failing to recognize the image, Gemini Pro often considers the image as a scientific poster

We gave the same image as the above and asked the two VLMs, "Tell me about how the 'PROBLEM CONTEXT' is given in the attached image, in detail" to see whether they retrieve the information as requested by the user. The scoring outcomes from GPT-4V and Gemini Pro are presented in Fig. 6, which are similar to Fig. 5. GPT-4V correctly retrieved the problem context, an "experiment about red dye diffusion" conducted by "Shawn," in the situation where "the red dye will diffuse differently in each dish based on the water's temperature" (left of Fig. 6). In contrast, Gemini Pro returned, "the PROBLEM CONTEXT is given by the title of the poster, "Rad dye illusion".," which is wrong. Also, it fabricated non-existent information such as "the text ... reads: The Rad dye illusion is a visual phenomenon in which a series of colored dots appear to be moving in a circular direction, ...," and also falsely said that "the images on the poster show ... Rhodamine B, Fluorescein, Malachite green, ..." It is noteworthy that Gemini Pro recognized the given image as a scientific poster and generated hallucinated information based on it (Figs. 5-6). This implies that its training dataset could have included scientific posters.

4.2.3. Finding 3: Gemini Pro fails to retrieve a random example from few-shot examples | GPT-4V succeeds

We gave GPT-4V and Gemini Pro an image for Task 42, similar to Fig. 2 with three 'Beginning' student-drawn models on the right side. The automatic scoring results of GPT-4V and Gemini Pro are presented in Fig. 7. Notably, GPT-4V strictly followed the instruction that requires it to choose a random one out of the nine examples, and correctly retrieved that "Example 3 is judged "Proficient"." It had further correctly retrieved in the image why example 3 is labeled as "Proficient." After that, it scored the three 'Beginning' examples correctly. In contrast, although Gemini Pro seems to say something about an example, it did not specify which example it retrieved, and thus there was no evidence it retrieved one. After that, it scored the three 'Beginning' examples as 'Proficient,' 'Developing,' and 'Developing,' which were all mis-scored. Lee and Zhai (2023) experimentally reported that GPT-4V is likely to score student-drawn models when it had been provided with few-shot examples and retrieved one of them, which explains how GPT-4V succeeded in this qualitative case. If this is also the case for Gemini Pro, its failure to retrieve a certain example from the image input could be one of the reasons for its low scoring accuracy.

4.3. Limited Improvement of Gemini Pro's Performance on the Automatic Scoring Task

Given that Gemini Pro could not process the fine-grained, large-size input image as well as GPT-4V, we broke down the input image to reduce the complexity of the input image prompt and expected to improve Gemini Pro's scoring performance. We applied this strategy to Task 53, one of the previously unsuccessful Tasks. We reduced the number of few-shot learning examples from nine to three (a set of one example for each of the 'Beginning,' 'Developing,' and 'Proficient' category; Fig. 8). We provided Gemini Pro with the input image without test cases (Left of Fig. 8) and asked "What do you see in the given image?" and "Tell me about how the 'PROBLEM CONTEXT' is given in the attached image, in detail." For the former question, it correctly responded, "The image you sent to me is ... to explain the interaction of water molecules when water is heated." (Top of Fig. 9). It successfully explained the location of the problem context, three examples, and 'Rationale for Proficiency." However, it responded that example 1 is a 'Beginning' and example 3 is 'Proficient,' which are both incorrect. For the latter question, it correctly stated that "the problem context in the attached image is given by series of instructions task ask the reader to construct a model ..." (Bottom of Fig. 9) . However, it incorrectly responded that "the image then shows two examples." This is even inconsistent with the very above answer, which correctly identified three examples given in the same input image. When we tried the same questions using an image with nine few-shot examples for comparison, Gemini Pro started malfunctioning and responded that the image is a "poster" again. This result implies that Gemini Pro's scoring performance was improved with less information in and less pixel size of the input image. However, it still failed to precisely retrieve information from the image. Also, when we tried to get automatic scoring results for one student-drawn image by providing it with Fig. 8, Gemini Pro did not return any scoring result. In summary, reducing the input image size improved Gemini Pro's outcome for retrieving information from the image, but this change was insufficient for automatic scoring tasks.

5. Discussion

The study aimed to compare Gemini Pro and GPT-4V in educational settings, particularly in scoring student-drawn models using NERIF. Major findings highlighted GPT-4V's superior accuracy in image classification compared to Gemini Pro and its adeptness at processing detailed text in images, evidenced in scoring students' drawn models. The study also uncovered the nuance of Gemini's scoring performance, which accounts for the lower performance compared to GPT-4V, including failing to recognize fine-grained texts printed in the image, often considering the image as a scientific poster, and failing to retrieve a random example from few-shot examples. Even adapting NERIF for Gemini Pro was not improved enough to be comparable to GPT-4V. The findings contribute to the literature in several aspects.

First, the findings contribute significantly to the existing literature by showcasing GPT-4V's remarkable scoring accuracy, distinguishing it from Gemini Pro, and setting new benchmarks. This performance aligns with trends noted in the literature (Zhai et al., 2020b), which highlighted the evolving precision of AI in the automatic scoring of constructed response assessments. GPT-4V's ability to accurately score complex student-drawn models aligns with the findings of Lee and Zhai (2023), who underscored the potential of AI in enhancing visual assessments. Furthermore, unlike earlier automatic scoring systems that primarily focused on textual data (Zhai et al., 2020a), GPT-4V integrates advanced image processing, marking a significant leap in the capability of AI tools. This study not only corroborates the growing efficacy of AI in educational settings but also extends it by demonstrating practical applications in multimodal formats of student responses. Such advancements address some of the limitations discussed in earlier research and the challenges in accurately assessing non-textual student work (Zhai et al., 2022; Lee et al., 2023c).

The findings of this study, highlighting the nuanced limitations of Gemini's scoring performance compared to GPT-4V, contribute notably to uncovering the mechanisms of the newly released Gemini Pro in image analysis. The findings overall seem inconsistent with the major report of Google (2023). Gemini's challenges in recognizing fine-grained text within images and its tendency to misclassify images as scientific posters resonate with concerns raised in earlier research. For instance, studies by Akter et al. (2023b) pointed out similar difficulties faced by Gemini in discerning detailed textual elements in complex visual contexts. Furthermore, Gemini's failure to effectively utilize few-shot visual examples reflects the limitations discussed by Fu et al. (2023), highlighting the challenges of interpreting images with a large number of elements because of its concision approach. This study's examination of these specific shortcomings not only corroborates the observations from previous research but also provides a concrete comparison of how different AI models handle complex educational tasks. Such insights are invaluable for the ongoing development of more sophisticated and context-aware AI tools in education, as suggested by the work of Latif et al. (2023a), who emphasized the need for AI systems to better adapt to the nuanced and varied nature of educational content and assessment methodologies.

The study's findings on the limitations of adapting NERIF for Gemini Pro, which did not significantly enhance its performance to match that of GPT-4V, offer several contributions to the existing literature on AI in educational contexts. This suggests the challenges of integrating specific AI frameworks into existing models, often finding that such adaptations do not always yield the expected improvements in performance. Furthermore, the results highlighted the complexity of AI systems in education, noting that modifications like NERIF require careful calibration to align with the intricacies of educational content. The inability of Gemini Pro to reach the accuracy level of GPT-4V, even with NERIF, underscores that AI advancements in education depend not just on the initial capabilities of the AI model but also the the intrinsic design of prompts. This study thereby adds a nuanced perspective to the discussion on the limits of augmenting AI systems with additional frameworks and suggests the need for a more holistic approach to AI development in educational settings. The findings highlight the intricate balance between AI adaptability and the inherent design of educational AI systems, contributing to a deeper understanding of how AI can be effectively tailored and utilized in complex educational assessments.

6. Conclusions

This study presented a comprehensive comparison between Gemini Pro and GPT-4V in educational settings, focusing on their ability to score student-drawn models using NERIF. The findings highlighted GPT-4V's superior accuracy in image classification and its proficiency in processing detailed text in images, demonstrating its potential for enhancing multimodal assessments in education. The study revealed that even with adaptations to NERIF, Gemini Pro could not match GPT-4V's performance, emphasizing the complexities in AI model adaptation and the importance of intrinsic design and initial capabilities. These results contribute significantly to the literature on AI in education, suggesting the need for more sophisticated, context-aware AI tools. This study adds to the understanding of AI's role in educational assessment, indicating directions for future research and development in this rapidly evolving field.

Acknowledgments

This study was funded by the National Science Foundation (NSF) (Award no. 2101104). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. The authors thank the NGSA team and the researchers involved in the parental study (Zhai, He, and Krajcik, 2022) and those who coded the student-drawn models. The authors specifically thank Joon Kum, who helped developing the prompts and making predictions on the test data.

References

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*, 2023a.

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. An in-depth look at gemini's language abilities. *eprint: 2312.11444*, 2023b.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

Abu Muna Almaududi Ausat, Berdinata Massang, Mukhtar Efendi, Nofirman Nofirman, and Yasir Riady. Can chat gpt replace the role of the teacher in the classroom: A fundamental analysis. *Journal on Education*, 5(4):16100–16106, 2023.

Chongyan Chen, Mengchen Liu, Noel Codella, Yunsheng Li, Lu Yuan, and Danna Gurari. Fully authentic visual question answering dataset from online communities. *arXiv preprint arXiv:2311.15562*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Juan Dempere, Kennedy Prince Modugu, Allam Hesham, and Lakshmana Ramasamy. The impact of chatgpt on higher education. *Dempere J, Modugu K, Hesham A and Ramasamy LK (2023) The impact of ChatGPT on higher education. Front. Educ*, 8:1206936, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Luyang Fang, Gyeong-Geon Lee, and Xiaoming Zhai. Using gpt-4 to augment unbalanced data for automatic scoring, 2023.

Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*, 2023.

Henner Gimpel, Kristina Hall, Stefan Decker, Torsten Eymann, Luis Lämmermann, Alexander Mädche, Maximilian Röglinger, Caroline Ruiner, Manfred Schoch, Mareike Schoop, et al. Unlocking the power of generative ai models and systems such as gpt-4 and chatgpt for higher education: A guide for students and lecturers. Technical report, Hohenheim Discussion Papers in Business, Economics and Social Sciences, 2023.

Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

C. J. Harris, J. S. Krajcik, and J. W. Pellegrino. *Creating and using instructionally supportive assessments in NGSS classrooms.* NSTA Press, 2024.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Gwo-Jen Hwang, Haoran Xie, Benjamin W. Wah, and Dragan Gašević. Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1:100001, 2020.

Norm Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, et al. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings.

In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–14, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

J. Richard Landis and Gary G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374, 1977.

Ehsan Latif and Xiaoming Zhai. Automatic scoring of students' science writing using hybrid neural network. *arXiv preprint arXiv:2312.03752*, 2023.

Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. Agi: Artificial general intelligence for education. page arXiv:2304.12479, 2023a. doi: 10.48550/arXiv.2304.12479. URL https://arxiv.org/abs/2304.12479.

Ehsan Latif, Xiaoming Zhai, and Lei Liu. Ai gender bias, disparities, and fairness: Does training data matter? *arXiv preprint arXiv:2312.10833*, 2023b.

Gyeong-Geon Lee and Xiaoming Zhai. Nerif: Gpt-4v for automatic scoring of drawn models, 2023.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. Applying large language models and chain-of-thought for automatic scoring. *arXiv preprint arXiv:2312.03748*, 2023a.

Gyeong-Geon Lee, Lehong Shi, Ehsan Latif, Yizhu Gao, Arne Bewersdorff, Matthew Nyaaba, Shuchen Guo, Zihao Wu, Zhengliang Liu, Hui Wang, Gengchen Mai, Tiaming Liu, and Xiaoming Zhai. Multimodality of ai for education: Towards artificial general intelligence. *arXiv preprint arXiv:2312.06037*, 2023b. doi: 10.48550/arXiv.2312.06037. URL https://doi.org/10.48550/arXiv.2312.06037.

Jaeyong Lee, Gyeong-Geon Lee, and Hun-Gi Hong. Automated assessment of student hand drawings in free-response items on the particulate nature of matter. *Journal of Science Education and Technology*, 32:549–566, 2023c. doi: 10.1007/s10956-023-10042-3. URL https://doi.org/10.1007/s10956-023-10042-3.

Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jewoong Moon, and Hyeoncheol Kim. Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, 2023d. doi: 10.1007/s10639-023-12249-8. URL https://doi.org/10.1007/s10639-023-12249-8.

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, and Min Zhang. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*, 2023.

Mengchen Liu and Chongyan Chen. An evaluation of gpt-4v and gemini in online vqa. *arXiv preprint arXiv:2312.10637*, 2023.

Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410, 2023. doi: 10.3390/educsci13040410. URL https://doi.org/10.3390/educsci13040410.

Rosemary Luckin, Wayne Holmes, Mark Griffiths, and Lauri B. Forcier. *Intelligence Unleashed: An Argument for AI in Education*. Pearson Education, 2016.

Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. Gpt-4v (ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*, 2023.

Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N Halgamuge. From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*, 2023.

NGSS Lead States. *Next Generation Science Standards: For States, By States*. National Academies Press, 2013.

Open AI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023a.

Open AI. Gpt-4v(ision) system card. 2023b.

Open AI. Gpt-4v(ision) technical work and authors. 2023c.

Open AI. Chatgpt can now see, hear, and speak. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak, September 2023d. Published on September 25, 2023.

Organisation for Economic Co-operation and Development. Oecd future of education and skills 2030: Oecd learning compass 2030 – a series of concept notes, 2019. OECD Publishing.

Senthujan Senkaiahliyan, Augustin Toma, Jun Ma, An-Wen Chan, Andrew Ha, Kevin R An, Hrishikesh Suresh, Barry Rubin, and Bo Wang. Gpt-4v (ision) unsuitable for clinical care and education: A clinician-evaluated assessment. *medRxiv*, pages 2023–11, 2023.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. Google's ai chatbot "bard": a side-by-side comparison with chatgpt and its utilization in ophthalmology. *Eye*, pages 1–4, 2023.

C. Wang, X. Zhai, and Shen J. *Applying Machine Learning to Assess Paper-Pencil Drawn Models of Optics*. Oxford University Press, UK, Oxford, In press.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Pusheng Xu, Xiaolan Chen, Ziwei Zhao, Yingfeng Zheng, Guangming Jin, Danli Shi, and Mingguang He. Evaluation of a digital ophthalmologist app built by gpt4-v (ision). *medRxiv*, pages 2023–11, 2023.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1), 2023.

Haotian Yu, Yushan Fu, Bo Jiang, Pan Fan, Ling Shen, Lin Gao, Haomin Lu, Yinmi Zhang, Su Dong, Jin Lv, et al. Applications of gpt-4 for accurate diagnosis of retinal diseases through optical coherence tomography image recognition. 2023.

Xiaoming Zhai. Chatgpt user experience: Implications for education. *Available at SSRN 4312418*, 2022.

Xiaoming Zhai, Kevin C Haudek, Lehong Shi, Ross Nehm, and Mark Urban-Lurain. From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, 57(9):1430–1459, 2020a. doi: 10.1002/tea.21658.

Xiaoming Zhai, Yue Yin, James W Pellegrino, Kevin C Haudek, and Lehong Shi. Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, 56(1):111–151, 2020b. ISSN 0305-7267.

Xiaoming Zhai, Peng He, and Joseph Krajcik. Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, 59(10):1765–1794, 2022. ISSN 0022-4308.

Peilin Zhou, Meng Cao, You-Liang Huang, Qichen Ye, Peiyan Zhang, Junling Liu, Yueqi Xie, Yining Hua, and Jaeboum Kim. Exploring recommendation capabilities of gpt-4v (ision): A preliminary case study. *arXiv preprint arXiv:2311.04199*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Appendix A. Appendix / supplemental material

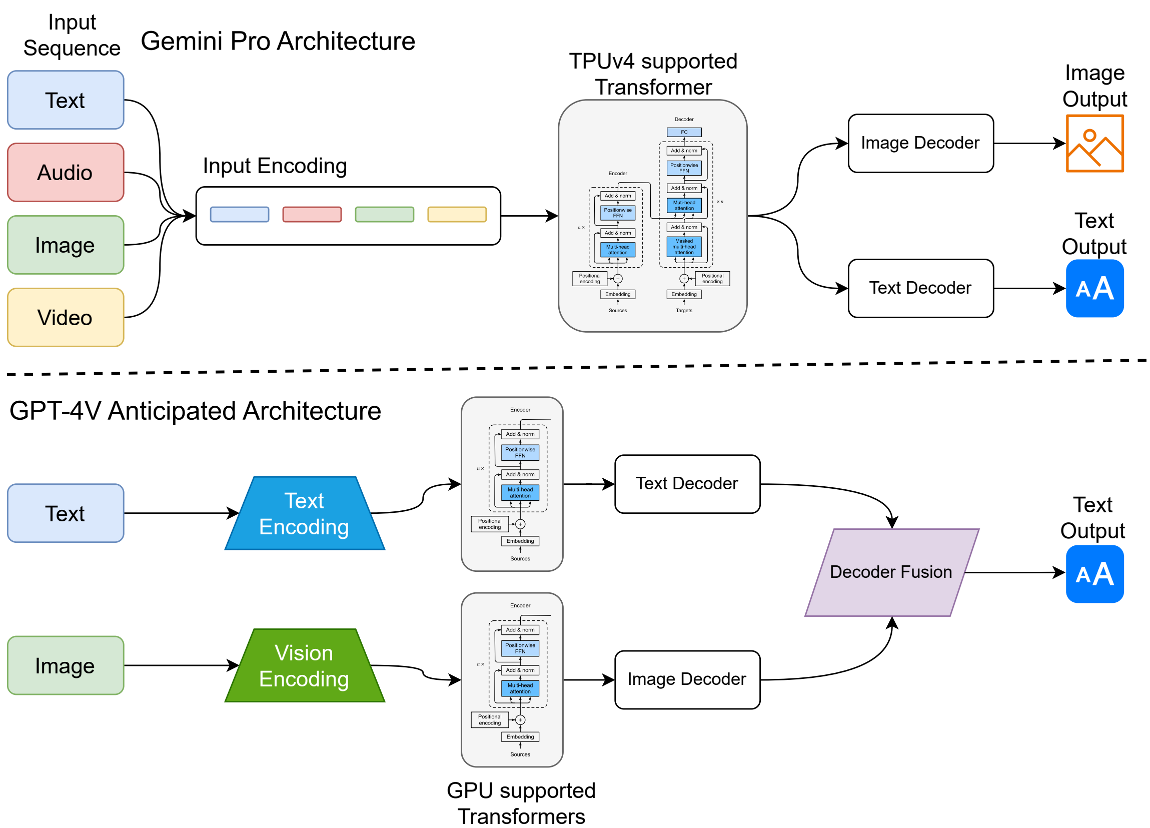

Figure 1: Architectural Comparison of Gemini Pro and GPT-4V

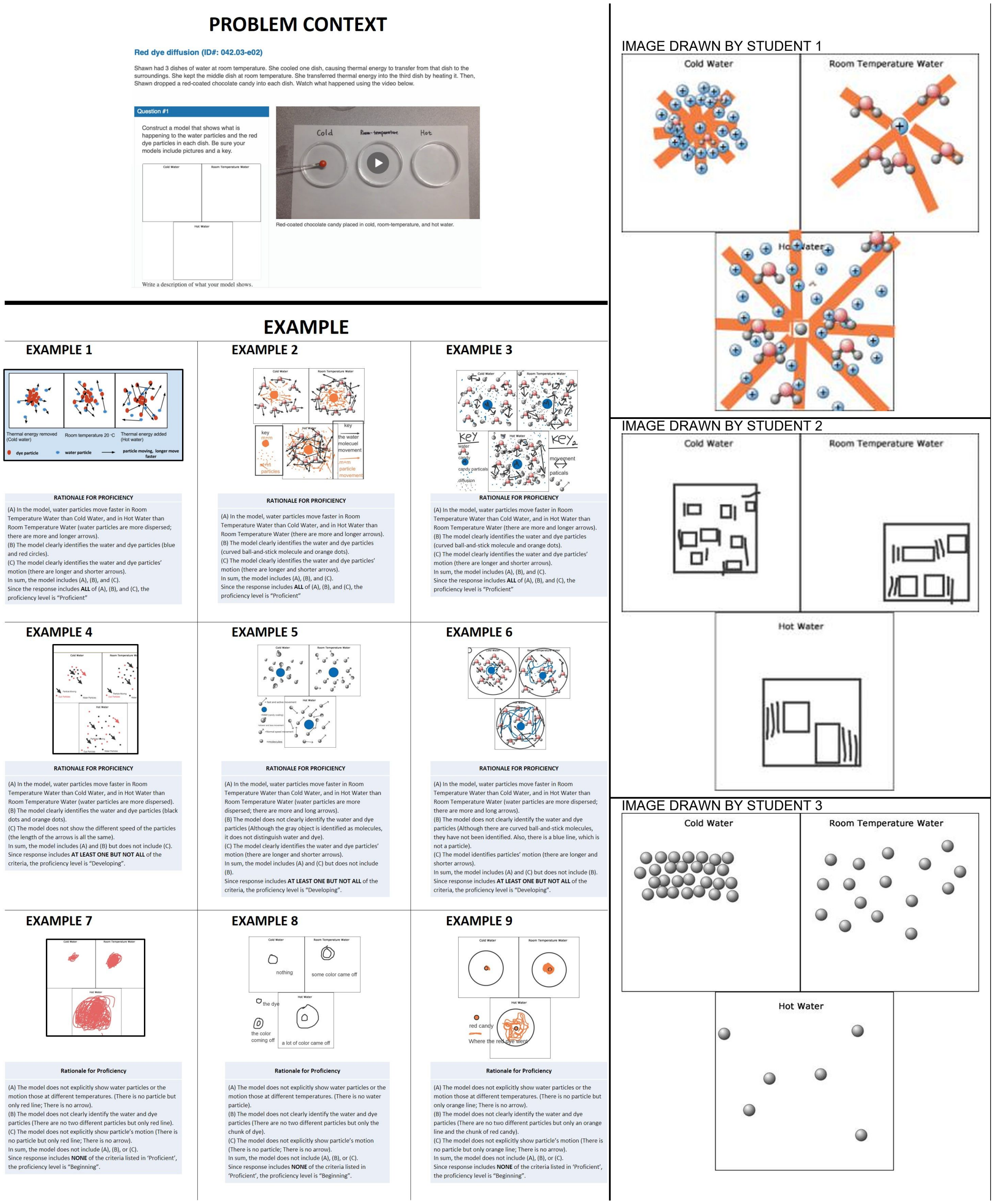

Figure 2: Example Input Image from Task 42

Table 1: Image understanding comparison between Gemini Pro and GPT-4V (Results reorganized from (Google, 2023))

Task	Gemini Pro	GPT-4V
MMMU (val)	47.9%	**56.8%**
TextVQA (val)	74.6%	**78.0%**
DocVQA (test)	88.1%	**88.4%**
ChartQA (test)	74.1%	**78.5%**
InfographicVQA (test)	75.2%	**75.3%**
MathVista (testmini)	45.2%	**49.9%**
AIZ2D (test)	73.9%	**78.2%**
VQA v2 (test-dev)	71.2%	**77.2%**

Table 2: Comprehensive Feature Comparison of Gemini-Pro and GPT-4V

Feature	Gemini-Pro	GPT-4V
Model Type	Decoder-only Transformer	Autoregressive Transformer
Parameter Size	280 billion	1000 billion
Training Data	Google's internal datasets	OpenAI's publicly available datasets
Inference Hardware	TPUs (Tensor Processing Units)	GPUs (Graphics Processing Units)
Context Length	Supports 32K tokens	Supports up to 8K tokens
Safety Measures	Extensive bias and toxicity analysis	Improved safety and alignment over predecessors
Performance	Sophisticated multimodal reasoning	Advanced reasoning and problem-solving
Applications	Integrated into Bard, Pixel, and accessible via API	Integrated with Microsoft Bing, Duolingo, etc.
Real-world feedback	Rigorous testing with external experts	Continuous improvements based on user feedback
Multimodal Capabilities	Image, audio, video, and text	Text and image inputs
Fine-Tuning Capabilities	Tailored versions for different platforms	Fine-tuning with RLHF techniques
Applications in Education	Online article finding and summarization for research, Domain-specific content extraction from internet	Automatic Scoring, Feedback system, paper writing, enhance creativity

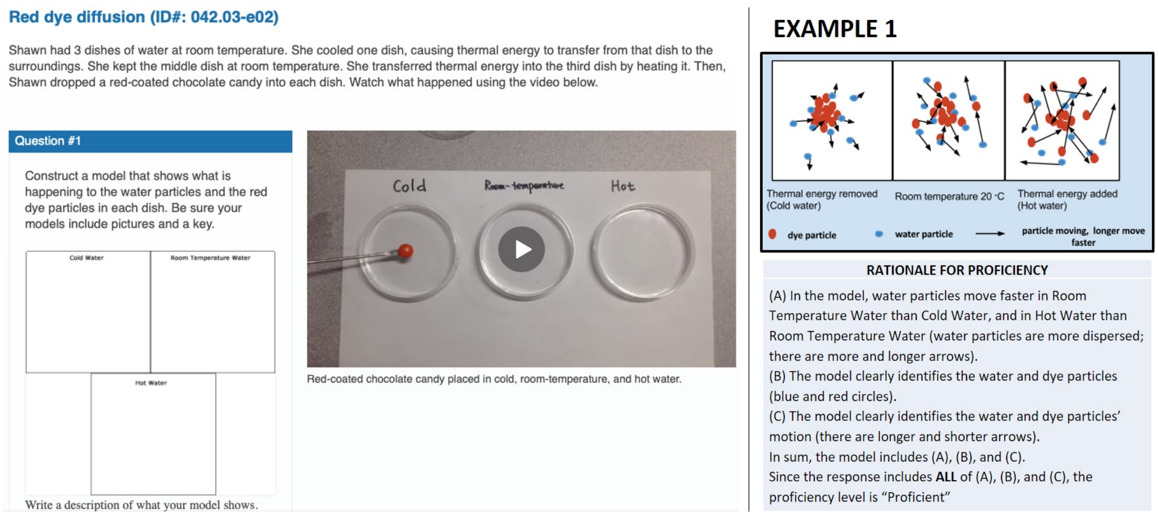

Figure 3: Example of Problem Context and Example from Task 42

Table 3: Image Classification Performance of GPT-4V and Gemini Pro on the NERIF Tasks

LLM	Task	Accuracy	Precision	Recall	F1	KappaQW	Acc_Beg	Acc_Dev	Acc_Prof
GPT-4V	42	0.47	0.50	0.47	0.46	0.37	0.62	0.56	0.24
	44	0.44	0.60	0.44	0.38	0.26	0.88	0.31	0.12
	45	0.57	0.59	0.57	0.55	0.44	0.82	0.56	0.32
	48	0.41	0.35	0.41	0.36	0.29	0.65	0.56	0.03
	53	0.46	0.53	0.46	0.45	0.38	0.56	0.63	0.21
	57	0.54	0.57	0.54	0.54	0.50	0.65	0.59	0.38
	Mean	0.48	0.50	0.46	0.43	0.37	0.67	0.58	0.26
	SD	0.06	0.09	0.05	0.06	0.09	0.12	0.16	0.19
Gemini Pro	42	0.30	0.30	0.30	0.30	-0.14	0.21	0.44	0.26
	44	NA	NA	NA	NA	NA	NA	NA	NA
	45	NA	NA	NA	NA	NA	NA	NA	NA
	48	NA	NA	NA	NA	NA	NA	NA	NA
	53	NA	NA	NA	NA	NA	NA	NA	NA
	57	NA	NA	NA	NA	NA	NA	NA	NA
	Mean	0.3	0.3	0.30	0.3	-0.14	0.21	0.44	0.26
	SD	NA	NA	NA	NA	NA	NA	NA	NA

Table 4: Confusion Matrices of GPT-4V and Gemini Pro on Task 42

True Label	GPT-4V's Prediction			Gemini Pro's Prediction		
	Beginning	Developing	Proficient	Beginning	Developing	Proficient
Beginning	21	13	0	7	14	13
Developing	8	18	6	11	14	7
Proficient	7	19	8	12	13	9

Prompt

ROLE
: You will be a science teacher who categorizes student responses to science items for proficiency.

TASK
: You are going to categorize each of three '''IMAGES DRAWN BY STUDENTS''' that model why a particular phenomenon occurs in the given '''PROBLEM CONTEXT.''' Your categorization will depend on the '''RUBRIC'''. You will also be given nine '''EXAMPLE''' responses that are categorized for proficiency.
1. By (1) briefly explaining the '''PROBLEM CONTEXT''', '''RUBRIC''' and (2) retrieving 'RATIONAL FOR PROFICIENCY' of one random example (1-9) from '''EXAMPLE''', you will show that you are ready to categorize '''IMAGES DRAWN BY STUDENTS'''.
2. Please provide your categorization of '''IMAGES DRAWN BY STUDENTS''' strictly following the format as like 'RATIONALE FOR PROFICIENCY' in the '''EXAMPLE'''.

PROBLEM CONTEXT
: See the image in the attachment.

RUBRIC
- Component (A): WATER molecules move slow when the water is cold, faster when at room temperature, and fastest when the water is hot.
- Component (B): The key identifies water and dye PARTICLES
- Component (C): The key identifies particle's motion (faster/slower).
- Summary: The scorer MUST summarize which of EACH specific component is included in the student drawing, before judging its proficiency.
- Proficiency: The student proficiency will be "Proficient" if the drawing includes ALL of (A), (B), and (C); "Developing" if the drawing includes AT LEAST ONE BUT NOT ALL of (A), (B), and (C); "Beginning" if the drawing includes NONE of (A), (B), and (C).
* Notes
- The three boxes labeled "Cold Water", "Room Temperature Water", and "Hot Water", are templates for the responses. These themselves DO NOT mean that the student has attempted to depict the motion of the particles based on temperature. Therefore, a scorer should focus on what is drawn in each box, rather than the label of each box.
- LINES OR REGIONS ARE NOT PARTICLE OR MOLECULES.
- Simply showing expanding regions or lines does NOT mean that particle motion (faster/slower) is expressed.
- Component (A) has to focus on WATER molecules's movement, NOT dye.
- You can say Component (B) is found when there are two kinds of distinct PARTICLES in the images.

EXAMPLE
: See the image in the attachment.

IMAGES DRAWN BY STUDENTS
: See the image in the attachment.

TEMPERATURE
: 0

TOP_P
: 0.01

Figure 4: Example Prompt from Task 42

Answer

ChatGPT

The image you've uploaded appears to be an educational material that explains the diffusion of red dye in water at different temperatures. It is structured to provide a problem context followed by examples of student work with varying levels of proficiency.

At the top, the PROBLEM CONTEXT" details an experiment where three dishes of water at different temperatures (cold, room-temperature, and hot) have a red-coated chocolate candy placed in them. Below this description, there is a section for the student to write a description of what their model shows.

Following this, there are nine examples labeled from “EXAMPLE1” to “EXAMPLE 9”, each depicting a different student's model of the diffusion process, along with a 'RATIONALE FOR PROFICIENCY" that explains why each example has been given a certain proficiency level such as "Proficient". "Developing", or "Beginning". The models use different symbols to represent water and dye particles, and arrows to indicate movement and energy transfer. The rationales critique each model based on several criteria, such as whether the model clearly identifies water and dye particles, if the length of the arrows is appropriate, and whether the model includes all necessary components to accurately reflect the diffusion process.

The examples are visual representations, using drawings and symbols, to convey the molecular motion and energy transfer between the particles in water at different temperatures. This is a common teaching method in science education to help students understand abstract concepts such as diffusion and thermal energy transfer.

Gemini

The given image is a posters showing various examples of "rad dye illusion" (4:042.03-602. The poster is divided into two sections: “EXAMPLE" and KEY”.

In the "EXAMPLE" section, there are nine images of molecules with different structures. In each image, there is a central atom or group of atoms surrounded by other atoms or groups of atoms. The surrounding atoms are connected to the central atom by lines.

In the "KEY" section, there is a single image of a molecule with a different structure than the molecules in the EXAMPLE" section. The molecule in the "KEY" section has a ring structure, with each atom in the ring connected to two other atoms in the ring.

The text at the top of the poster. "PROBLEM CONTEXT", suggests that the poster is intended to be used as part of a problem-solving activity. The viewer is likely being asked to identify the molecule in the "KEY" section that is different from the molecules in the EXAMPLE" section.

Here is a more detailed description of each of the images on the poster:

Example 1: A central carbon atom surrounded by four hydrogen atoms.
Example 2: A central carbon atom surrounded by two hydrogen atoms and two chlorine atoms.
Example 3: A central carbon atom surrounded by two hydrogen atoms, one oxygen atom, and one chlorine atom.
Example 4: A central carbon atom surrounded by four other carbon atoms.
Example 5: A central carbon atom surrounded by three other carbon atoms and one hydrogen atom.
Example 6: A central carbon atom surrounded by two other carbon atoms and two hydrogen atoms.
Example 7: A central carbon atom surrounded by one other carbon atom, one hydrogen atom, and two oxygen atoms.
Example 8: A central carbon atom surrounded by two other carbon atoms, one hydrogen atom, and one chlorine atom.
Example 9: A central carbon atom surrounded by three other carbon atoms and one oxygen atom.

Key: A ring structure with four carbon atoms and two oxygen atoms.

I hope this helps!

Figure 5: Answers from GPT-4V and Gemini Pro When Asked "What Do You See in the Given Image?" (Left: GPT-4V; Right: Gemini Pro)

Answer

ChatGPT

The image depicts an educational setup titled "PROBLEM CONTEXT," relating to an experiment about red dye diffusion. It presents a scenario where an individual named Shawn has three dishes of water at different temperatures: one cold, one at room temperature, and one hot. Into each dish, a red-coated chocolate candy is dropped.
The intent is to observe what happens using a video provided in the document, which presumably shows the diffusion of the red dye in water at different temperatures.

Below the description, there is a section with the heading "Question #1" which asks students to construct a model that shows what is happening to the water particles and the red dye particles in each dish. It suggests that the model should include pictures and a key.

The problem context is intended to help students understand how temperature affects the movement of particles in a fluid medium, with the expectation that the red dye will Diffuse differently in each dish based on the water's temperature.

The image also contains a section with multiple examples, labeled from "EXAMPLE 1" to 'EXAMPLE 9." Each example seems to be a student's attempt to model the particle movement, accompanied by a rationale for proficiency, where the effectiveness of each attempt is evaluated based on how well it meets certain criteria. The criteria assess whether the models show water and dye particles' motion, whether the particle types are clearly identified, and the length of the arrows representing particle movement.

Examples 1,2, and 3 are marked as "Proficient," meaning they have effectively met the expectations. Examples 4,5. and 6 are marked as "Developing," indicating that they meet some but not all of the criteria. Lastly, Examples 7,8, and 9 are marked as "Beginning," which implies that these attempts do not meet the necessary criteria.

The proficiency levels and the rationales given for each example provide feedback on the models created by the students, serving as a guide for what is expected in the students' responses to Question #1.

Gemini

In the attached image, the PROBLEM CONTEXT is given by the title of the poster. "Rad dye illusion" and the text below it, which reads:

The Rad dye illusion is a visual phenomenon in which a series of colored dots appear to be moving in a circular direction, even though they are actually stationary. This illusion is caused by the way that the human brain processes visual information.

The poster also includes a number of images of different types of molecules. These images are used to illustrate the different types of molecules that can be used to create the Rad dye illusion.
The poster's title and text provide the overall context for the problem, while the images provide specific examples of the molecules that can be used to create the illusion.
Here is a more detailed explanation of the problem context, based on the information given in the image: ' he Rad dye illusion is a type of optical illusion that is created by the way that the human brain processes visual information. When we look at the image of the Rad dye illusion, our brains interpret the dots as being in motion.
However, the dots are actually stationary.
The R3d dye illusion can be created with a variety of different types of molecules. The molecules that are used to create the illusion must have certain properties, such as the ability to absorb and emit light at specific wavelengths.
The images on the poster show a variety of different types of molecules that can be used to create the Rad dye illusion. Some examples of these molecules include:

• Rhodamine B • Fluorescein • Malachite green • Crystal violet • Methylene blue

These molecules are all able to absorb and emit light at specific wavelengths. When these molecules are placed in a solution, they will absorb light at one wavelength and emit light at another wavelength. This difference in wavelengths is what causes the Rad dye illusion.

The Rad dye illusion is a fascinating example of how the human brain can be tricked by visual information. The illusion is also a valuable tool for scientists, who use it to study the way that the brain processes visual information.

I hope this explanation is helpful. Please let me know if you have any other questions.

Figure 6: Scoring Outcomes from GPT-4V and Gemini Pro When Asked "Tell me about how the 'PROBLEM CONTEXT'" (Left: GPT-4V; Right: Gemini Pro)

Answer

ChatGPT

1 The **PROBLEM CONTEXT** involves understanding how water molecules behave at different temperatures The context is related to a classic experiment where dye is added to water at different temperatures to observe the diffusion rate, which illustrates molecular motion. The rubric for assessing student proficiency in modeling this phenomenon is based on three components:

- **Component (A)** assesses whether the student's drawing shows water molecules moving slower in cold water, faster at room temperature, and fastest in hot water.
- **Component (B)** evaluates if the student's drawing clearly distinguishes between water and dye particles.
- **Component (C)** looks for an indication of the panicles' motion (faster or slower) in the student's drawing.

The proficiency level is determined by whether the student's drawing includes all three components (Proficient), at least one but not all (Developing), or none (Beginning).

2. Let's select a random example to retrieve the rationale for proficiency. I will take **EXAMPLE 3**:

- **Rationale for Proficiency**: The student's drawing in Example 3 is judged "Proficient" because it includes all required components It shows that water molecules move at different speeds at different temperatures (A), distinct panicles for water and dye (B), and indicates the motion of these particles (C).

Now, I'll proceed to categorize the **IMAGES DRAWN BY STUDENTS**:

- **Image Drawn by Student 1**:
  - Summary: The student's drawing includes distinct shapes for water in each temperature condition but does not distinguish between water and dye particles. There is no indication of the particles' motion.
  - Proficiency: The drawing is "Beginning" because it does not include any of the required components (A), (B), or (C).
- **Image Drawn by Student 2**:
  - Summary: The student's drawing uses lines emanating from a central point to indicate motion. However, lines or regions are not considered particles or molecules, and there is no distinction between water and dye particles.
  - Proficiency. The drawing is "Beginning" because it does not correctly depict any of the components (A), (B), or (C).
- **Image Drawn by Student 3**:
  - Summary: The student's drawing includes cartoonish representations of M&Ms with faces that express the freezing and melting process but does not differentiate between water and dye panicles The particle's motion is not clearly depicted.
  - Proficiency: The drawing is "Beginning" because it includes none of the components (A), (B), or (C).

This categorization is based on the provided rubric and the examples for proficiency, ensuring that the student's understanding of molecular motion in relation to temperature is accurately assessed.

Gemini

PROBLEM CONTEXT

This is a science experiment that demonstrates how the temperature of water affects the movement of its molecules. The three boxes labeled Cold Water. Room Temperature Water, and Hot Water represent different temperature conditions, in the Cold Water box, the water molecules are moving slowly. In the Room Temperature Water box, the water molecules are moving faster, in the Hot Water box, the water molecules are moving fastest

RUBRIC

The student's drawing will be categorized as follows:

- Proficient: The drawing includes ALL of the following components:
  - Component (A). WATER molecules move slow when the water is cold, faster when at room temperature, and fastest when the water is hot.
  - Component (B). The key identifies water and dye PARTICLES
  - Component (C): The key identifies particle's motion (faster/slower)
- Developing: The drawing includes AT LEAST ONE BUT NOT ALL of the following components:

--- (Omitted) ---

- Beginning: The drawing includes NONE of the following components:

--- (Omitted) ---

EXAMPLE

In the example the student has correctly identified the three components of the rubric.

- Component (A) The student has drawn the water molecules moving faster in the Mot Water box than in the Cold Water box.
- Component (B). The student has used different symbols to represent the water molecules and the dye molecules.
- Component (C): The student has used arrowheads to show the direction of motion of the molecules.

Image 1:

- Component (A): Yes
- Component (B): Yes
- Component (C): Yes

Rationale for Proficiency: The student's drawing is proficient because it includes ALL of the components required by the rubric The student has correctly identified that the water molecules are moving faster In the Hot Water box than in the Cold Water box. The student has also used different symbols to represent the water molecules and the dye molecules, and has used arrowheads to show the direction of motion of the molecules.

Image 2:

- Component (A): No
- Component (B): Yes
- Component (C): Yes

Rationale for Proficiency: The student's drawing is developing because it includes TWO of the three components required by the rubric. The student has used different symbols to represent the water molecules and the dye molecules, and has used arrowheads to show the direction of motion of the molecules. However, the student has not correctly identified that the water molecules are moving foster in the Hot Water box than in the Cold Water box.

Image 3:

- Component (A): Yes
- Component (B): No
- Component (C): Yes

Rationale for Proficiency: The student's drawing is developing because it includes ONE of the three components required by the rubric. The student has used different symbols to represent the water molecules and the dye molecules, but has not used arrowheads to show the direction of motion of the molecules, nor has the student correctly identified that the water molecules are moving faster in the Hot Water box than in the Cold Water box.

I hope this is helpful!

Figure 7: Scoring Outcomes from GPT-4V and Gemini Pro When Given the Full Image and Prompt Input (Left: GPT-4V; Right: Gemini Pro)

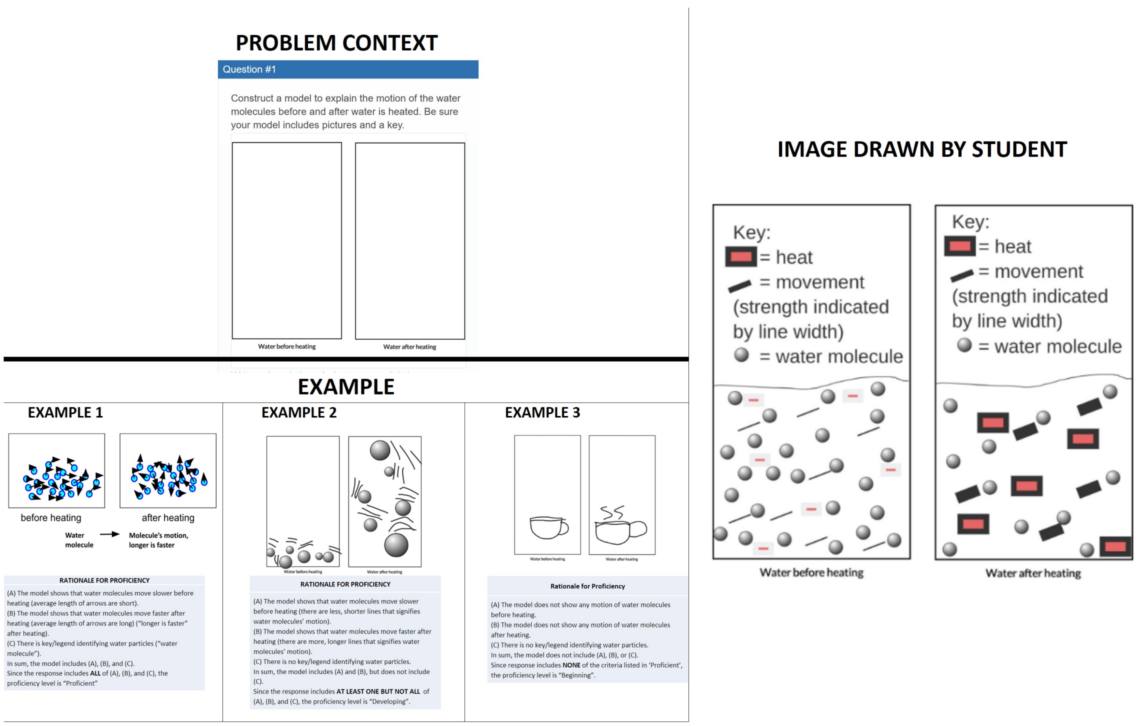

Figure 8: Example of Simplified Input Image from Task 53

Answer

Gemini

The image you sent me is a diagram that shows a model to explain the interaction of water molecules when water is heated. The top of the diagram is labeled 'Problem Context' and includes the question: "Construct a model to explain the motion of the water molecules before and after water is heated. Be sure your model includes pictures and a key."

The diagram then shows three example models, labeled Example 1. Example 2. and Example 3. Each example model shows two pictures, one labeled "before heating" and one labeled after heating. The pictures are made up of arrows, which are presumably meant to represent the water molecules. The arrows in the "before heating" pictures are shorter than the arrows in the "after heating" pictures. There is also a key at the bottom of the diagram that explains that the arrows represent water molecules and that the length of the arrows represents the speed of the molecules.

Below the examples, there is a section labeled Rationale for Proficiency.' This section includes three criteria for a proficient model:

1. The model shows that water molecules move slower before heating.
2. The model shows that water molecules move faster after heating.
3. The model includes a key/legend identifying water particles.

The diagram then assesses each of the example models against these criteria. Example 1 is rated as "Beginning because it does not meet any of the criteria. Example 2 is rated as "Developing ‘ because it meets criteria 1 and 2, but not 3. Example 3 is rated as "Proficient" because it meets all three criteria.

Overall, the image is a diagram that shows how to create a model to explain the motion of water molecules before and after water is heated. The diagram includes three example models, as well as a set of criteria for evaluating the proficiency of the models.

I hope this description is helpful! Let me know if you have any other questions.

The problem context in the attached image is given by a series of instructions that ask the reader to construct a model to explain the motion of water molecules before and after water is heated. The instructions also specify that the model should include pictures and a key.

The image then shows two examples of models that could be used to explain the motion of water molecules. The first example shows water molecules as simple circles with arrows attached to them. The arrows are short and point in different directions, which is meant to show that the water molecules are moving slowly before they are heated.

The second example shows water molecules as the same simple circles, but with longer arrows that point in more or less the same direction. This is meant to show that the water molecules are moving faster after they have been heated.

The image also includes a rationale for proficiency, which explains why the first example is not considered to be a proficient model. The rationale states that the first example does not show any motion of water molecules before heating, and that it does not have a key or legend to identify the water particles.

In conclusion, the problem context in the attached image is given by a series of instructions that ask the reader to construct a model to explain the motion of water molecules before and after water is heated. The instructions also specify that the model should include pictures and a key.

I hope this helps! Let me know if you have any other questions.

Figure 9: Scoring from Gemini Pro with Simplified Input Image (Top: "What do you see in the given image?" GPT-4V; Bottom: “Tell me about how the ’PROBLEM CONTEXT’ is given in the attached image, in detail.")