An Eye for an AI: Evaluating GPT-4o's Visual Perception Skills and Geometric Reasoning Skills Using Computer Graphics Questions

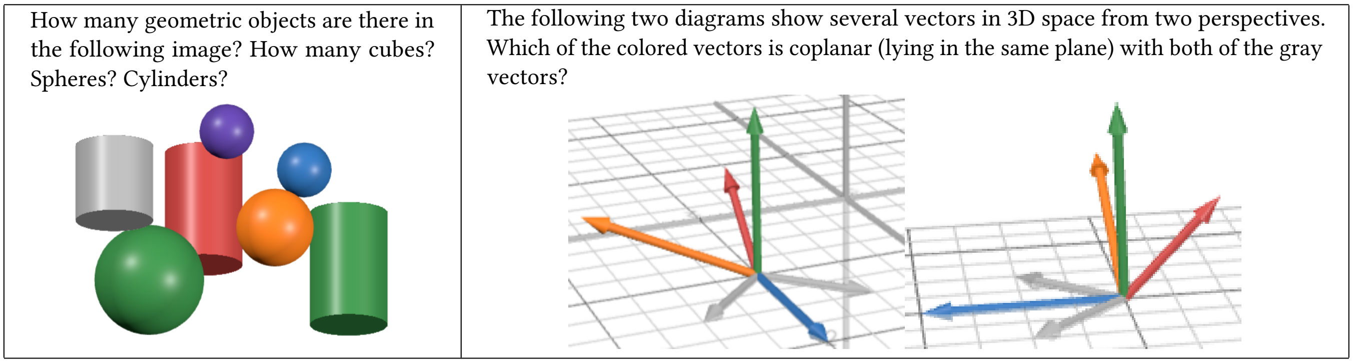
Tony Haoran Feng
University of Auckland
New Zealand
hfen962@aucklanduni.ac.nz

Paul Denny
University of Auckland
New Zealand
paul@cs.auckland.ac.nz

Burkhard C. Wünsche
University of Auckland
New Zealand
burkhard@cs.auckland.ac.nz

Andrew Luxton-Reilly
University of Auckland
New Zealand
a.luxton-reilly@auckland.ac.nz

Jacqueline Whalley
Auckland University of Technology
New Zealand
jacqueline.whalley@aut.ac.nz



How many geometric objects are there in the following image? How many cubes? Spheres? Cylinders?

The following two diagrams show several vectors in 3D space from two perspectives. Which of the colored vectors is coplanar (lying in the same plane) with both of the gray vectors?

Figure 1: Two questions from CG_EASY that GPT-4o struggled with (2 out of 10 responses are correct for the left question, and 0 out of 10 responses are correct for the right question)

Abstract

CG (Computer Graphics) is a popular field of CS (Computer Science), but many students find this topic difficult due to it requiring a large number of skills, such as mathematics, programming, geometric reasoning, and creativity. Over the past few years, researchers have investigated ways to harness the power of GenAI (Generative Artificial Intelligence) to improve teaching. In CS, much of the research has focused on introductory computing. A recent study evaluating the performance of an LLM (Large Language Model), GPT-4 (text-only), on CG questions, indicated poor performance and reliance on detailed descriptions of image content, which often required considerable insight from the user to return reasonable results. So far, no studies have investigated the abilities of LMMs (Large Multimodal Models), or multimodal LLMs, to solve CG questions and how these abilities can be used to improve teaching.

In this study, we construct two datasets of CG questions requiring varying degrees of visual perception skills and geometric reasoning skills, and evaluate the current state-of-the-art LMM, GPT-4o, on the two datasets. We find that although GPT-4o exhibits great potential in solving questions with visual information independently, major limitations still exist to the accuracy and quality of the generated results. We propose several novel approaches for CG educators to incorporate GenAI into CG teaching despite these limitations. We hope that our guidelines further encourage learning and engagement in CG classrooms.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Computer graphics**; • **Social and professional topics** → **Computing education**.

Keywords

Large Language Models, LLMs, Large Multimodal Models, LMMs, Visual Language Models, VLMs, Generative Artificial Intelligence, GenAI, GPT-4, GPT-4o, Visual Perception, Geometric Reasoning, Computer Graphics, Computing Education, Evaluation, Assessment

ACM Reference Format:
Tony Haoran Feng, Paul Denny, Burkhard C. Wünsche, Andrew Luxton-Reilly, and Jacqueline Whalley. 2024. An Eye for an AI: Evaluating GPT-4o's Visual Perception Skills and Geometric Reasoning Skills Using Computer Graphics Questions. In *SIGGRAPH Asia 2024 Educator's Forum (SA Educator's Forum '24), December 03–06, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3680533.3697064

SA Educator's Forum '24, December 03–06, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1136-7/24/12
https://doi.org/10.1145/3680533.3697064

1 Introduction

The advancement of GenAI (Generative Artificial Intelligence), especially LLMs (Large Language Models), has garnered global attention from the Computing Education research community [Denny et al. 2024b]. LLMs excel at generating solutions that are typical of many programming-focused computing courses [Denny et al. 2023; Finnie-Ansley et al. 2022, 2023; Savelka et al. 2023]. However, since LLMs can only process textual inputs, they perform poorly in tasks requiring image inputs and/or visual and geometric processing skills [Feng et al. 2024a], which are essential in solving questions in CG (Computer Graphics) [Rodrigues et al. 2021; Suselo et al. 2017].

LMMs (Large Multimodal Models), or VLMs (Visual Language Models), are extensions of LLMs that allow users to provide information in non-textual formats, such as images. An example is GPT-4o (GPT-4 Omni) [OpenAI 2024b], an LMM developed by OpenAI that allows image inputs. The release of LMMs opened many new opportunities. With image inputs, users can provide visual context to the GenAI model, making human-AI interactions easier, and questions requiring visual context can now be asked effortlessly. This also possibly allows LMMs to solve questions requiring visual and geometric reasoning skills, such as those in CG.

Investigating the performance of LMMs, such as GPT-4o, on CG questions can provide insight into decisions and opportunities related to teaching CG. Past research suggests that the poor performance of GPT-4 in CG questions limits students' ability to misuse it [Feng et al. 2024a], but also makes it harder for CG instructors to use GenAI for teaching (e.g., by providing formative feedback, creating practice questions [Feng et al. 2024b], generating explanations). Evaluating the capabilities and limitations of GPT-4o in the context of CG enables educators to make more informed decisions about integrating GenAI into their teaching.

In this work, we investigate the visual perception and geometric reasoning capabilities of GPT-4o by using two datasets of CG-related questions. We compare the visual processing capabilities of GPT-4o to its textual processing capabilities and outline implications and recommendations for CG educators. Our study aims to answer the following Research Question:

How well can GPT-4o solve Computer Graphics questions requiring visual perception and geometric reasoning skills?

2 Related Work

2.1 LLMs in Education

There has been substantial research on various GenAI models, with a prominent focus on LLMs. Past research showed impressive capabilities of LLMs in many subjects, such as reading comprehension [Brown et al. 2020], law [Katz et al. 2024], medicine [Liévin et al. 2024; Nori et al. 2023], and various other academic fields [AI4Science and Quantum 2023], thus bringing opportunities and challenges in many disciplines [Abd-Alrazaq et al. 2023; Tu et al. 2023; Yeadon and Hardy 2023]. In CS (Computer Science), LLMs achieved high performance in CS1 [Denny et al. 2023; Finnie-Ansley et al. 2022], CS2 [Finnie-Ansley et al. 2023], and programming-related MCQs (multiple-choice questions) [Savelka et al. 2023], often surpassing average student performance. Additionally, LLMs are capable of assisting CS educators by generating educational material [Leinonen et al. 2023; Liffiton et al. 2023; MacNeil et al. 2023], providing many new opportunities [Bernstein et al. 2024; Denny et al. 2024a]. Despite the wide variety of tasks that LLMs can complete, there are still many areas in which they exhibit limited performance, such as reasoning [Bang et al. 2023], visual programming [Singla 2023], and Parsons Problems (a programming exercise where students reorder shuffled code blocks) [Reeves et al. 2023].

Understanding the capabilities of LLMs for CG is an area of active research. We evaluated the performance of GPT-4 (text-only) on assessment questions used in an undergraduate introductory CG course and found that GPT-4 produced correct solutions to only 42.1% of the questions [Feng et al. 2024a]. Another study assessed GPT-4's ability to generate code for a Ray Tracing application, and the results demonstrated a similar performance compared to the previous study (42% accuracy) [Feng et al. 2024b].

2.2 Evaluations of LMMs

We theorize that the low performance of LLMs for CG questions is due to CG requiring extensive visual-based reasoning skills, and LLMs struggle with these tasks due to their textual nature and lack of visual training data [Feng et al. 2024a; Singla 2023]. LMMs allow users to provide visual context to the model directly. However, since LMMs are still relatively new, few studies have been conducted to measure their capabilities in various tasks.

Two early evaluation reports on GPT-4V (GPT-4 Vision [OpenAI 2024a], the predecessor of GPT-4o) showcased its capabilities on queries requiring visual contexts in a wide variety of settings, such as visual math reasoning and code generation [Wu et al. 2023b; Yang et al. 2023]. The results showed impressive visual-based reasoning skills of LMMs. Nevertheless, they often produce errors. Similar evaluation studies on more specialized areas showed that LMMs are somewhat capable of assisting in medical diagnoses [Wu et al. 2023a], map analysis [Xu and Tao 2024], and autonomous driving [Driessen et al. 2024; Wen et al. 2024]. However, the consensus remains that there are significant limitations to the capabilities of LMMs, and more development is needed before they can reliably support real-world applications.

In the context of education, GPT-4V has been compared with its text-only counterpart, GPT-4 Turbo, on a specialized medical examination, and no statistically significant differences between the results were found between the two models [Hirano et al. 2024], indicating that LMMs do not necessarily outperform LLMs. In a study more relevant to CS, the ability of GPT-4V to generate code based on UML diagrams was evaluated, and it was observed to perform well for simpler, single-class UML diagrams, but it failed to consistently generate correct code for more complex, multi-class UML diagrams [Antal et al. 2024].

Despite the mediocre performance of LMMs on some educational tasks, the use of LMMs or similar applications can increase student performance and interest [Zain et al. 2023]. Effectively leveraging this in educational settings may lead to similar positive impacts.

3 Methods

3.1 Overview

We investigate the current capabilities of GPT-4o on CG questions by 1) collecting and creating CG questions; 2) converting them into the JSON format accepted by GPT-4o; 3) fetching responses from

GPT-4o (through the OpenAI API) as attempts at answering the questions; 4) evaluating the correctness of the responses. We then interpret the results and make recommendations to CG educators about how LMMs can be used for CG teaching.

3.2 Step 1: Collecting Questions

Our first dataset derives from a previous study using GPT-4 [Feng et al. 2024a]. It contains 101 assessment questions used in a third-year introductory CG course, 68 of which are MCQs and 33 are programming questions. The questions are taken from the mid-semester tests and final exams of the 2022 and 2023 iterations of the course (i.e., four assessments in total). We refer to this dataset as CG_TEST in this paper. The topics covered in the questions include but are not limited to introductory Linear Algebra, introductory OpenGL, Colors and Lighting, Illumination and Shading, Texture Mapping, Ray Tracing, 3D Modelling, Parametric Curves and Surfaces, and Image Processing.

Each assessment is split into Theory and Programming parts. Theory parts consist of MCQs of four or more options. Programming parts consist of programming questions that often require students to write code snippets, which are then executed against pre-written test cases. If all test cases are passed, then the student is awarded all marks allocated for the question. Otherwise, no marks are awarded. No partial marks are given.

Of all 101 questions, 67 contain no images, and 34 contain images. Although many of the questions contain no images, almost all questions require visual perception and geometric reasoning intelligence as the course focuses heavily on developing these skills and contains highly visual concepts. Several example questions are listed throughout the paper.

Since the assessment questions are quite technical and specialized, we also want to investigate GPT-4o's ability to process visual information without using specialized knowledge and whether this makes a difference in performance. Hence, we also created a small dataset containing 10 basic image-based CG-related short-answer questions, which we refer to as CG_EASY in this paper. However, little to no CG background is required to answer these questions, and only common sense and a moderate amount of visual and geometric reasoning skills are needed. The questions involve identifying and counting geometric objects in a scene, light-surface interactions, basic 3D geometry, and basic 3D transformations (translations and rotations). Figure 1 shows two example questions from this dataset.

The two datasets used in this study are publicly available through the link provided in Section 7.

3.3 Step 2: Converting to JSON

GPT-4o allows for inputs in various formats, such as images from publicly accessible URLs, and combinations of multimodal content as single inputs, such as interweaving texts and images. Multimodal inputs follow the JSON format shown in Figure 2.

The questions collected from CG assessments contain texts, mathematical formulas, and images, but they are not in the format accepted for multimodal inputs. Hence, some preprocessing needs to occur before the questions can be processed by GPT-4o.

```
[
  { "type": "text", "text": "[TEXT]" },
  { "type": "image_url", "image_url": { "url": "[IMAGE_URL]", } },
  ...
]
```

Figure 2: GPT-4o multimodal input JSON format

Given is a plane 3x+2y-z=3 and a ray

$$p(t) = \begin{pmatrix}1\\0\\1\end{pmatrix} + t * \begin{pmatrix}-1\\c\\0\end{pmatrix}$$

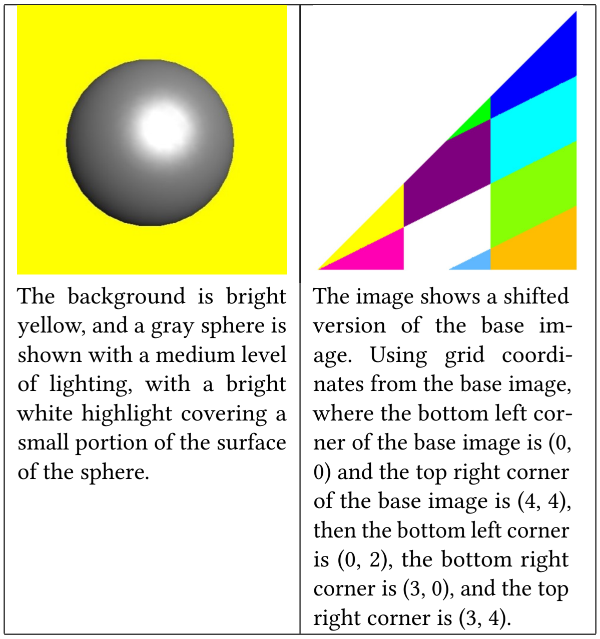
For what value of c is the ray parallel to the plane?
Select one:
a. c=0; b. c=1.5; c. c=0.5; d. c=1; e. c=-0.5

```
[
  {
      "type": "text", "text": """
Given is a plane 3x+2y-z=3 and a ray
$$p(t)=\begin{pmatrix}1\\0\\1\end{pmatrix}+t*\begin{pmatrix}
-1\\c\\0\end{pmatrix}.$$
For what value of c is the ray parallel to the plane?
Select one:
a. c=0; b. c=1.5; c. c=0.5; d. c=1; e. c=-0.5
      """
  },
]
```

Figure 3: An example assessment question containing formulas (top); the textual version of the question using LaTeX commands, in the GPT-4o text-only input format (bottom)

Texts can be copied directly into the *"text"* field corresponding to *"type": "text"*. Mathematical formulas are replaced by their corresponding LaTeX commands. An alternative method would be to replace formulas with their textual counterparts, for example, replacing "2×3" (written with LaTeX commands) with text symbols such as "2x3" and "2*3", but such replacements can be inconsistent. Hence, TeX commands are used to keep conversions consistent. Figure 3 shows an example of this conversion.

Although we can directly feed images to GPT-4o, we also want to investigate any differences in performance on image-based questions with its text-only capabilities. Past literature has used image descriptions to encompass all relevant information in the images, which is then fed as input to LLMs in place of the images [Yang et al. 2022]. Our previous study that evaluated CG questions also used this strategy before the publication of LMMs [Feng et al. 2024a]. Therefore, for every image in our dataset of assessment questions, we can replace it with a textual description of the image, at the level of detail a capable student could produce, containing all the information necessary to solve the corresponding question. Two examples are shown in Figure 4. Then, for every question containing images, we constructed two JSON objects in their corresponding input formats for GPT-4o: text-only using textual descriptions and multimodal using real images. An example is shown in Figure 5.



The background is bright yellow, and a gray sphere is shown with a medium level of lighting, with a bright white highlight covering a small portion of the surface of the sphere.

The image shows a shifted version of the base image. Using grid coordinates from the base image, where the bottom left corner of the base image is (0, 0) and the top right corner of the base image is (4, 4), then the bottom left corner is (0, 2), the bottom right corner is (3, 0), and the top right corner is (3, 4).

Figure 4: Two example images used in CG_TEST and their corresponding textual descriptions

3.4 Step 3: Fetching Responses

After the questions are converted to JSON objects, the data is sent to GPT-4o via the OpenAI API, to which the model responds with its answers. Each JSON object is sent 10 times, and 10 responses are received, which are treated as 10 independent attempts. For every question containing images, two separate JSON objects are sent (text-only version and multimodal version), and 20 responses are received for that question, 10 for each version. The model's temperature is set to 0.75, which is reported to perform well on previous, similar studies [Feng et al. 2024a; Pursnani et al. 2023]. The system message we use in this study is "You are a helpful assistant, and you are knowledgeable in Computer Graphics. When you answer a multiple-choice question, you state your selected option explicitly while providing a concise and accurate explanation.".

3.5 Step 4: Evaluating Correctness

The responses from the GPT-4o are then evaluated for correctness. No partial marks are given (responses are categorized as correct or incorrect).

For MCQs, responses are marked as correct if they state the correct option or the letter associated with the correct option. The responses usually contain explanations of their solutions, but they are not required to be considered correct.

For programming questions, responses are marked as correct if they contain the correct solution code that can be copied and pasted into the AAT (Automated Assessment Tool) used in the assessments and pass all test cases [Wünsche et al. 2018; Wünsche et al. 2019]. We allow for some deletions from the generated code solutions, such as boilerplate code which is often present in outputs, as boilerplate code is already supplied by the AAT.

Table 1: Percentages of correct responses from GPT-4 (text-only) [Feng et al. 2024a] and GPT-4o for questions in various categories (the higher percentage is marked in bold)

Category	GPT-4	GPT-4o
CG_TEST: All questions	42.1%	**50.1%**
CG_TEST: MCQs	53.5%	**62.6%**
CG_TEST: Programming (1 attempt)	27.1%	**31.8%**
CG_TEST: Programming (10 attempts)	**53.6%**	50.9%
CG_TEST: No images	60.0%	**67.9%**
CG_TEST: Images (textual descriptions)	**36.5%**	35.6%
CG_TEST: Images (real images)	N/A	**29.4%**
CG_EASY: All questions	N/A	**62.0%**

Furthermore, the accuracy is evaluated in two ways for programming questions: "1 attempt" and "10 attempts". Under the conditions of these assessments, students are allowed to submit their code solutions for programming questions as many times as they want with no penalties, and they obtain full marks for the question as long as one of their attempts passes all test cases [Wünsche et al. 2018]. The "10 attempts" marking scheme is used to mimic this setting. Alternatively, "1 attempt" mimics the setting where the student has only 1 attempt for each question, and it measures the expected score from 10 separate attempts. This is done by taking the average of the 10 responses for every question.

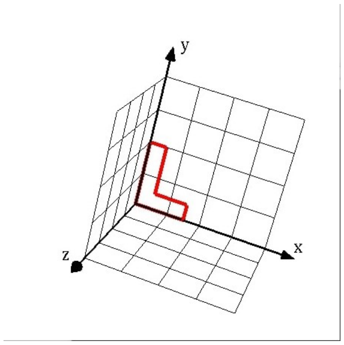
4 Results

For CG_TEST, out of all 1350 responses received from GPT-4o (67 text-only questions, 34 image-based questions converted to text-only using textual descriptions, 34 image-based questions using multimodal input, 10 responses each), 676 responses are marked as correct, which is 50.1% of all responses. Out of 800 responses to MCQs (56 text-only MCQs, two versions of 12 image-based questions), 501 responses are correct (62.6%). Of 550 responses to programming questions (11 text-only questions, two versions of 22 image-based questions), 175 responses are correct (31.8%). There are 55 groups of 10 responses to programming questions, each corresponding to one programming question, and 28 out of the 55 groups contained at least one correct response (50.9%). Of 670 responses to text-only questions (67 text-only questions), 455 are correct (67.9%). Of 340 responses to image-based questions using textual descriptions (34 image-based questions using textual descriptions), 121 are correct (35.6%). Of 340 responses to image-based questions using real images (34 image-based questions using real images), 100 are correct (29.4%). For CG_EASY, out of 100 responses from GPT-4o (10 image-based questions using real images), 62 are correct (62.0%). These results are summarized in Table 1.

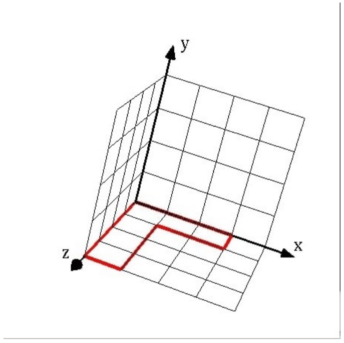
5 Discussion

5.1 Overall Results

Overall, GPT-4o answers around half of the queries from CG assessments correctly, i.e., it is not a reliable source of answers for CG assessments or specialized CG questions. A slightly higher accuracy is achieved for questions in the CG_EASY dataset. However, Given a function drawShape() which draws a wireframe representation of the letter "L" in the xy-plane as shown in the image below.



Please write OpenGL code to transform this shape such that you obtain the scene displayed in the image below:



IMPORTANT:
Please only use OpenGL transformations, e.g. glScalef, glTranslatef, glRotatef.
Please do NOT draw the shape itself - this is done automatically by the uploaded code.

```
[
  {
    "type": "text", "text": """
Given is a function drawShape() which draws a wireframe representation of the letter "L" in the xy-plane as shown in the image below.

Image description: A letter L is placed on the x-y plane. The vertices on the shape are p1 = (0, 0, 0), p2 = (0, 2, 0), p3 = (1.5, 0, 0).

Please write OpenGL code to transform this shape such that you obtain the scene displayed in the image below:

Image description: A letter L with twice the size is placed on the x-z plane. The vertices on the shape are p1 = (0, 0, 0), p2 = (0, 0, 4), p3 = (3, 0, 0).

IMPORTANT:
Please only use OpenGL transformations, e.g. glScalef, glTranslatef, glRotatef.
Please do NOT draw the shape itself - this is done automatically by the uploaded code.
    """
  },
]
```

```
[
  {
    "type": "text", "text": """
Given is a function drawShape() which draws a wireframe representation of the letter "L" in the xy-plane as shown in the image below.
    """
  }, {
    "type": "image_url", "image_url": { "url": """
https://raw.githubusercontent.com/TFPlusPlus/GPT-4V-vs.-CG/main/CG_Assessments_Images/2022b15-1.jpg
      """
    }
  }, {
    "type": "text", "text": """
Please write OpenGL code to transform this shape such that you obtain the scene displayed in the image below:
    """
  }, {
    "type": "image_url", "image_url": { "url": """
https://raw.githubusercontent.com/TFPlusPlus/GPT-4V-vs.-CG/main/CG_Assessments_Images/2022b15-2.jpg
      """
    }
  }, {
    "type": "text", "text": """
IMPORTANT:
Please only use OpenGL transformations, e.g. glScalef, glTranslatef, glRotatef.
Please do NOT draw the shape itself - this is done automatically by the uploaded code.
    """
  }
]
```

Figure 5: An example assessment question containing images (left); the textual version of the question using image descriptions, in the GPT-4o text-only input JSON format (middle); the multimodal version of the question using the real images, in the GPT-4o multimodal input JSON format (right)

it is still well below human performance, as these questions can be solved easily with common sense and minimal visual and geometric reasoning skills. This indicates a lack of consistency and reliability in GPT-4o's visual and geometric reasoning skills. Users should be mindful of these risks for questions requiring these skills.

The performance of GPT-4o on MCQs is higher than that on programming questions. This may be because MCQs are generally easier than programming questions, as they usually only involve one or two specific concepts. In contrast, programming questions are more complex and require more steps and critical thinking. Additionally, even if GPT-4o does not "know" the correct answer, it can still return a correct answer due to chance (25% chance for MCQs with 4 options, 20% for MCQs with 5 options).

GPT-4o performs much better on questions containing no images than those containing images. This indicates that despite possessing visual processing power, GPT-4o is still vastly superior at textual processing than visual processing, even if the question context requires visual and geometric reasoning skills. The question difficulty may also contribute to the difference in performance, as more difficult questions often require images to illustrate the context, so questions containing images may, on average, be more difficult than those containing none.

Comparing the results for image-based questions, queries using textual descriptions outperform those using real images. This suggests that prompts written by humans can distinguish important features in images that GPT-4o fails to do, and descriptions of these features may help improve the model's performance. Another

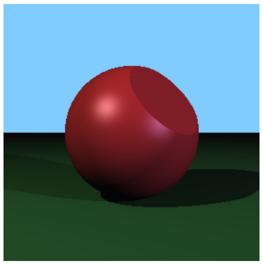


Figure 6: An image used in a programming question asking students to write code for intersecting a ray with a sphere cut by a plane. The function takes the ray's start point and direction as input and returns the intersection point.

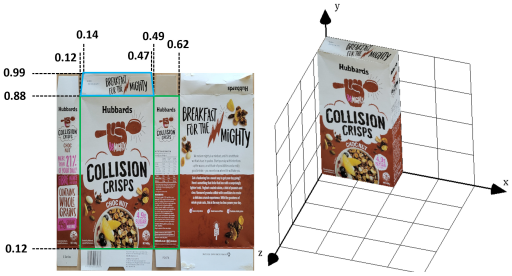


Figure 7: Two images used in a programming question asking students to map a texture image (left) onto a polygon mesh (right).

interpretation is that through describing the images, the human, instead of the model, does some of the visual processing, so the visual processing required for GPT-4o is reduced. Since GPT-4o is observed to underperform in visual processing tasks, this reduces the risk of GPT-4o making a mistake.

A similar past study evaluated the performance of GPT-4 (text-only) on the CG_TEST dataset [Feng et al. 2024a]. Comparing the results, we see an increase in the performance of "All questions", "MCQs", and "Contains no images". This indicates that GPT-4o has, on average, improved its capability to answer CG questions compared to its predecessor GPT-4, especially on textual questions. "Programming (1 attempt)" shows a slight increase in performance, but "Programming (10 attempts)" does not reach past performance. This suggests that GPT-4o is more consistently correct on questions that it is confident in, but the difference may simply be due to chance and is too small to be conclusive evidence. The performance of image-based questions using textual descriptions also does not improve, indicating a lack of improvement in visual reasoning skills from textual descriptions.

5.2 Common Characteristics of Responses

Please note that the behavior of the responses is dependent on the system message used with the query, and queries using different system messages may not show the following characteristics.

5.2.1 Varying lengths of explanations. The responses to the CG_TEST dataset are generally lengthy and detailed, with most explanations to questions reaching more than 10 lines of text and some even reaching 30 lines. Additionally, GPT-4o would often make mistakes but continue to elaborate along incorrect lines of thinking, which can confuse students and reduce learning. Common errors made by GPT-4o for this dataset are conceptual errors (e.g., using incorrect concepts, hallucinating false facts), mathematical errors (e.g., incorrectly substituting values into formulas, incorrectly expanding expressions, calculation errors), and logical errors (e.g., stating fallacious causal relationships).

The responses to CG_EASY are much shorter, typically only around 1-10 lines long, since solutions are often straightforward and do not require complex explanations (although the responses can still be incorrect).

5.2.2 Unnecessary code. For programming questions in the CG_TEST dataset, the generated code snippets can be unnecessarily long and contain large amounts of boilerplate code. For example, the solution to a 3D programming question (shown in Figure 8) is simply 3 lines of OpenGL code, calling the functions glRotatef(), glScalef(), glTranslatef() once each. However, most generated solutions for this question exceed 20 lines of code and include boilerplate code supplied by the AAT, such as a main() function. There are also many instances where the questions state that some classes and functions are already provided, but GPT-4o still includes those classes and functions with their full implementations, which leads to redefinition errors when the unmodified code solutions are executed, and these extra code snippets have to be manually removed when testing for correctness.

5.2.3 Failure to follow specific instructions. In addition to disregarding statements specifying the existing classes and functions, GPT-4o often fails to follow the correct syntax to call existing functions. For example, a common function supplied by the debugger is "dot(Vector v1, Vector v2)" for calculating the dot product of two vectors. The syntax for calling this function is always specified in the questions, but GPT-4o often fails to follow the syntax and instead writes "v1.dot(v2)". More than 40 responses out of the total 550 for programming questions include errors of this kind.

5.3 Specific Observations

5.3.1 Breakthroughs in difficult questions. The CG_TEST dataset contains several questions GPT-4 could not answer correctly in previous studies [Feng et al. 2024a,b]. One example is a programming question for ray tracing a cut sphere (shown in Figure 6), which GPT-4 answered incorrectly for all 30 attempts across two studies. In this study, GPT-4o answered this question correctly (with complete working code) in 1 out of the 10 responses for the textual description version and also 1 out of 10 for the image version. This is impressive since fewer than 5% of students could answer this question with unlimited attempts in an exam.

The CG_TEST dataset also contains a texture mapping programming question (shown in Figure 7), GPT-4 (text-only) could consistently solve this question, but this could be due to the textual descriptions of the images, and the visual processing that the human did by extracting the coordinates of the faces when describing

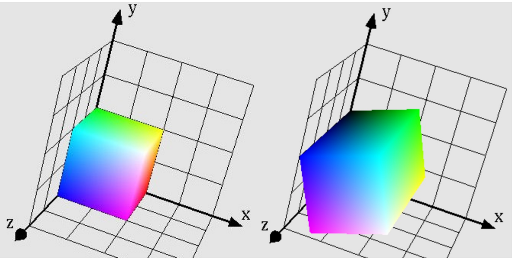


Figure 8: Two images used in a programming question asking students to transform a 3D color cube from its original position (left) to a new position (right) using OpenGL functions glScalef, glTranslatef, glRotatef. A textual description of the 3D transformation is also provided in the question.

the images. However, without the help of textual descriptions, GPT-4o can successfully extract the coordinates of the faces from the images and solve this question in 1 out of the 10 responses.

We acknowledge that the breakthrough performance on these two questions could potentially be only due to chance, but when combined with all other results, we suggest that over the past year, GPT-4 has improved (multimodal vs. text-only) in performance for CG questions.

5.3.2 *Challenges in answering questions in CG_EASY.*

GPT-4o can successfully solve 62.0% of the image-based questions in CG_EASY without any human assistance or textual descriptions for the images, which is certainly an impressive performance. However, there are still some questions that GPT-4o struggles with, such as the two questions shown in Figure 1.

For the question shown on the left of Figure 1, GPT-4o answers correctly in only 2 out of the 10 responses. In the 8 other attempts, GPT-4o states that there are either 8 or 9 objects in the image, and in most cases, it identifies 4 cylinders. However, when directly asking about the number of cylinders in the image, GPT-4o answers correctly 9 out of 10 times. We theorize that the added complexity of the question may have confused GPT-4o, and this reduction in performance may not be related to its visual perception skills.

For the question on the right of Figure 1, GPT-4o incorrectly answers "orange" in all 10 attempts. Although it sometimes states that the blue vector is also coplanar with the gray vectors, it always perceives the orange vector as coplanar, hence they cannot be marked as correct. From the results of this question, we suggest that although GPT-4o has modest visual perception skills, it still lacks geometric reasoning skills.

5.3.3 *Challenges in answering 3D transformation questions.*

A question type that GPT-4 and GPT-4o struggle with is programming questions related to 3D transformations, one of which is shown in Figure 8. All 40 attempts from this study and the previous study provide incorrect code solutions for this question, despite the correct solution only being 3 lines of code. This is further evidence that GPT-4 and GPT-4o lack geometric reasoning skills, which are essential in solving this question.

5.4 Implications

5.4.1 *GenAI models are unreliable in visual question-answering.*

The results of our study suggest that GPT-4o, or LMMs in general, may not reliably answer CG questions requiring visual perception skills and especially geometric reasoning skills. However, this does not mean that GenAI models cannot be used to improve learning. For example, CG educators can write or generate image descriptions for CG problems and ask students to evaluate the quality of the descriptions and/or improve the descriptions to enable GenAI models to solve the original problems. Additionally, LMMs are also useful for improving self-reflective practice [Kumar et al. 2024].

Conversely, CG educators should also raise student awareness of the limitations of GenAI for CG questions and the importance of critically evaluating the generated solutions. For CG educators who are opposed to the use of GenAI for teaching and learning purposes, since GPT-4o performs more poorly on image-based questions than textual questions, greater use of image-based questions may discourage students from using GenAI and encourage independent thinking and learning.

5.4.2 *A new exercise: Spot the error.*

An exercise for CG educators is to use incorrect AI-generated solutions to CG questions and ask students to find the errors in these solutions. This can simultaneously encourage students to reflect critically on their understanding of the topics and also raise awareness of the limitations of GenAI.

5.4.3 *Prompt engineering for more accurate and helpful responses.*

In this study, we directly used the question texts (after formatting) as prompts for the GenAI model. Recent research suggests that GenAI can achieve higher performance through better prompting strategies, such as splitting each question into smaller subquestions or asking GenAI to explain step by step [Denny et al. 2023; Kojima et al. 2022]. Different system messages can also be used to achieve different performances and characteristics. This could also be a good learning task for students, i.e., develop prompting strategies to solve complex CG questions.

6 Conclusion

In this study, we constructed two datasets of CG assessment and basic visual CG-related questions requiring varying degrees of visual perception skills and geometric reasoning skills. We evaluated the performance of GPT-4o on these two datasets. Although GPT-4o has improved in performance on visual questions compared to predecessor models, it still lacks the visual processing power to provide reliable academic support to CG students and, in general, real-world applications requiring visual understanding. We also described several common characteristics exhibited by GPT-4o in its responses and outlined various specific questions on which GPT-4o performed well or poorly. Finally, we suggested some implications for CG education and provided recommendations to CG educators on utilizing LMMs to improve CG teaching.

7 Resources

All images, textual descriptions, and JSON objects can be accessed through this link: https://github.com/TFPlusPlus/GPT-4V-vs.-CG.

References

Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, Javaid Sheikh, et al. 2023. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Medical Education* 9, 1 (2023), e48291.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361* (2023).

Gábor Antal, Richárd Vozár, and Rudolf Ferenc. 2024. Assessing GPT-4-Vision's Capabilities in UML-Based Code Generation. *arXiv preprint arXiv:2404.14370* (2024).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

Seth Bernstein, Paul Denny, Juho Leinonen, Lauren Kan, Arto Hellas, Matt Littlefield, Sami Sarsa, and Stephen Macneil. 2024. "Like a Nesting Doll": Analyzing Recursion Analogies Generated by CS Students Using Large Language Models. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (Milan, Italy) *(ITiCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 122–128. https://doi.org/10.1145/3649217.3653533

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

Paul Denny, Viraj Kumar, and Nasser Giacaman. 2023. Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language. In *Proc. of the 54th ACM Tech. Symp. on Computer Science Education V. 1*. 1136–1142.

Paul Denny, Juho Leinonen, James Prather, Andrew Luxton-Reilly, Thezyrie Amarouche, Brett A. Becker, and Brent N. Reeves. 2024a. Prompt Problems: A New Programming Exercise for the Generative AI Era. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (Portland, OR, USA) *(SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 296–302. https://doi.org/10.1145/3626252.3630909

Paul Denny, James Prather, Brett A Becker, James Finnie-Ansley, Arto Hellas, Juho Leinonen, Andrew Luxton-Reilly, Brent N Reeves, Eddie Antonio Santos, and Sami Sarsa. 2024b. Computing education in the era of generative AI. *Commun. ACM* 67, 2 (2024), 56–67.

Tom Driessen, Dimitra Dodou, Pavlo Bazilinskyy, and Joost De Winter. 2024. Putting ChatGPT vision (GPT-4V) to the test: risk perception in traffic images. *Royal Society Open Science* 11, 5 (2024), 231676.

Tony Haoran Feng, Paul Denny, Burkhard C. Wünsche, Andrew Luxton-Reilly, and Steffan Hooper. 2024a. More Than Meets the AI: Evaluating the performance of GPT-4 on Computer Graphics assessment questions. In *Proceedings of the 26th Australasian Computing Education Conference*. 182–191.

Tony Haoran Feng, Burkhard C. Wünsche, Paul Denny, Andrew Luxton-Reilly, and Steffan Hooper. 2024b. Can GPT-4 Trace Rays. In *Eurographics 2024 - Education Papers*, Beatriz Sousa Santos and Eike Anderson (Eds.). The Eurographics Association. https://doi.org/10.2312/eged.20241003

James Finnie-Ansley, Paul Denny, Brett A Becker, Andrew Luxton-Reilly, and James Prather. 2022. The robots are coming: Exploring the implications of openai codex on introductory programming. In *Proceedings of the 24th Australasian Computing Education Conference*. Association for Computing Machinery, 10–19.

James Finnie-Ansley, Paul Denny, Andrew Luxton-Reilly, Eddie Antonio Santos, James Prather, and Brett A Becker. 2023. My ai wants to know if this will be on the exam: Testing openai's codex on cs2 programming exercises. In *Proceedings of the 25th Australasian Computing Education Conference*. 97–104.

Yuichiro Hirano, Shouhei Hanaoka, Takahiro Nakao, Soichiro Miki, Tomohiro Kikuchi, Yuta Nakamura, Yukihiro Nomura, Takeharu Yoshikawa, and Osamu Abe. 2024. GPT-4 Turbo with Vision fails to outperform text-only GPT-4 Turbo in the Japan Diagnostic Radiology Board Examination. *Japanese J. of Radiology* (2024), 1–9.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A* 382, 2270 (2024), 20230254.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N Rafferty, John Stamper, et al. 2024. Supporting Self-Reflection at Scale with Large Language Models: Insights from Randomized Field Experiments in Classrooms. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 86–97.

Juho Leinonen, Arto Hellas, Sami Sarsa, Brent Reeves, Paul Denny, James Prather, and Brett A Becker. 2023. Using large language models to enhance programming error messages. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 563–569.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns* 5, 3 (2024).

Mark Liffiton, Brad E Sheese, Jaromir Savelka, and Paul Denny. 2023. Codehelp: Using large language models with guardrails for scalable support in programming classes. In *Proc. of the 23rd Koli Calling Int. Conf. on Computing Education Research*. 1–11.

Stephen MacNeil, Andrew Tran, Arto Hellas, Joanne Kim, Sami Sarsa, Paul Denny, Seth Bernstein, and Juho Leinonen. 2023. Experiences from using code explanations generated by large language models in a web software development e-book. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 931–937.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

OpenAI. 2024a. GPTV_System_Card.pdf. https://cdn.openai.com/papers/GPTV_System_Card.pdf. [Accessed 25-04-2024].

OpenAI. 2024b. Hello GPT-4o | OpenAI. https://cdn.openai.com/papers/GPTV_System_Card.pdf. [Accessed 31-07-2024].

Vinay Pursnani, Yusuf Sermet, Musa Kurt, and Ibrahim Demir. 2023. Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence* 5 (2023), 100183.

Brent Reeves, Sami Sarsa, James Prather, Paul Denny, Brett A Becker, Arto Hellas, Bailey Kimmel, Garrett Powell, and Juho Leinonen. 2023. Evaluating the performance of code generation models for solving Parsons problems with small prompt variations. In *Proc. of the 2023 Conf. on Innovation and Tech. in CS Education V. 1*. 299–305.

Rui Rodrigues, Teresa Matos, Alexandre Valle de Carvalho, Jorge G Barbosa, Rodrigo Assaf, Rui Nóbrega, António Coelho, and A Augusto de Sousa. 2021. Computer Graphics teaching challenges: Guidelines for balancing depth, complexity and mentoring in a confinement context. *Graphics and Visual Computing* 4 (2021), 200021.

Jaromir Savelka, Arav Agarwal, Marshall An, Chris Bogart, and Majd Sakr. 2023. Thrilled by your progress! Large language models (GPT-4) no longer struggle to pass assessments in higher education programming courses. In *Proc. of the 2023 ACM Conf. on International Computing Education Research-Volume 1*. 78–92.

Adish Singla. 2023. Evaluating ChatGPT and GPT-4 for Visual Programming. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2*. 14–15.

Thomas Suselo, Burkhard C. Wünsche, and Andrew Luxton-Reilly. 2017. The journey to improve teaching computer graphics: A systematic review. In *Proceedings of the 25th International Conference on Computers in Education (ICCE 2017). APSCE, Christchurch, New Zealand*. 361–366.

Xinming Tu, James Zou, Weijie J Su, and Linjun Zhang. 2023. What Should Data Science Education Do with Large Language Models? *arXiv preprint arXiv:2307.02792* (2023).

Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, MA Tao, Yingxuan Li, XU Linran, Dengke Shang, et al. 2024. On the Road with GPT-4V (ision): Explorations of Utilizing Visual-Language Model as Autonomous Driving Agent. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. 2023a. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909* (2023).

Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. 2023b. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:2310.16534* (2023).

Burkhard C. Wünsche, Zhen Chen, Lindsay Shaw, Thomas Suselo, Kai-Cheung Leung, Davis Dimalen, Wannes van der Mark, Andrew Luxton-Reilly, and Richard Lobb. 2018. Automatic assessment of OpenGL computer graphics assignments. In *Proceedings of the 23rd annual ACM conference on innovation and technology in computer science education*. 81–86.

Burkhard C. Wünsche, Edward Huang, Lindsay Shaw, Thomas Suselo, Kai-Cheung Leung, Davis Dimalen, Wannes van der Mark, Andrew Luxton-Reilly, and Richard Lobb. 2019. CodeRunnerGL - An Interactive Web-Based Tool for Computer Graphics Teaching and Assessment. In *Proceedings of the International Conference on Electronics, Information, and Communication (ICEIC 2019)*. IEEE, New York, NY, USA, 1–7. https://doi.org/10.23919/ELINFOCOM.2019.8706402

Jinwen Xu and Ran Tao. 2024. Map Reading and Analysis with GPT-4V (ision). *ISPRS International Journal of Geo-Information* 13, 4 (2024), 127.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 3081–3089.

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

Will Yeadon and Tom Hardy. 2023. The Impact of AI in Physics Education: A Comprehensive Review from GCSE to University Levels. *arXiv preprint arXiv:2309.05163* (2023).

Iffah NM Zain, Mohd AB Setambah, Mohd S Othman, and Mazarul HM Hanapi. 2023. Use of Photomath Applications in Helping Improving Students' Mathematical (Algebra) Achievement. *European Journal of Education and Pedagogy* 4, 2 (2023), 85–87.