Understanding Visual Representation of Linear Models: A Comparison of Real Students and ChatGPT

Alice Xu (alicex@g.ucla.edu)
Department of Psychology, University of California, Los Angeles
1285 Psychology Building, Los Angeles, CA 90095 USA

Ji Y. Son (json2@calstatela.edu)
Department of Psychology, California State University,
5151 State University Drive, Los Angeles, CA 90032 USA

James W. Stigler (stigler@ucla.edu)
Department of Psychology, University of California, Los Angeles

Abstract

The rise of ChatGPT has sparked interest among educators in integrating it into teaching and learning. However, effective teaching requires a deep understanding that allows instructors to use multiple representations to support comprehension and address students' misconceptions. Before relying on ChatGPT as a teaching tool, it is crucial to assess its ability to interpret multiple representations. This study evaluates ChatGPT's understanding of the fundamental statistical concept "data = model + error," which underpins a number of statistical analyses in introductory statistics courses. Through tasks involving graphical representations, we qualitatively examined GPT models' understanding and compared their strengths and limitations to those of students. The results showed that while ChatGPT demonstrated competence in certain areas, it also exhibited misunderstandings—some resembling students' and others unique to the models themselves.

Keywords: statistics and data science education; multiple representations; large language models (LLMs); ChatGPT

Introduction

The advent of ChatGPT has sparked widespread public interest in the educational potential of large language models (LLMs) (Chung, 2023; Zhao et al., 2023), including statistics and data science education (Tu et al., 2024). Across various domains, LLMs like GPT-4 and DeepSeek-V3 have demonstrated capabilities that rival or exceed those of human learners on certain benchmark assessments (Achiam et al., 2023; Hurst et al., 2024; Liu et al., 2024). Such findings raise the exciting possibility of integrating LLMs, such as ChatGPT, into educational contexts as interactive instructors, offering personalized, on-demand support tailored to individual learning needs.

However, for ChatGPT to be an effective teacher, it needs to know more than just what is required to pass a test. Teaching is not just about transmitting knowledge; it is about helping students actively construct their understanding. Good teachers are able to identify and address students' misconceptions about a domain (Modell et al., 2005). This requires pedagogical content knowledge, the skill to represent the subject matter in ways that make it comprehensible to others (Shulman, 1986). Although ChatGPT surpasses human performance on many benchmark exams, this does not necessarily mean it can guide students in constructing their own understanding. Doing so would require ChatGPT to draw on representations students already know to help them draw connections to new ideas, which necessitates a deep understanding of the subject matter. However, research has shown that ChatGPT may still struggle to effectively address students' misconceptions (Wardat et al., 2023).

In complex domains, understanding is not just about knowing isolated ideas—it requires the ability to map concepts onto other concepts and across multiple representations (Ainsworth, 1999). For example, in statistics and data science, deep understanding often depends on the ability to connect multiple representations, such as graphs, equations, and verbal descriptions (Estrella, 2018; Fries et al., 2021). Novices often focus on surface features rather than underlying principles, which makes it difficult for them to translate insights across different representational forms (Ainsworth, 1999; Kozma, 2003). For example, Cooper and Shore (2008) found that when undergraduate students were asked to judge variability using histograms, half incorrectly relied on the varying heights of the bars rather than the distribution of data values. Such difficulties may stem from the fact that visual features can be misleading. In a histogram, the height of the bars does not indicate variability, whereas the heights of boxes in a vertical box plot do. This contrast illustrates how similar visual features can represent different underlying concepts. As a result, expert guidance may be needed to help students appreciate how different representations relate to one another and what they reveal about the underlying concepts.

There is great potential for ChatGPT to offer students of statistics timely, adaptive support if it can accurately understand and explain how different representations interconnect. If not, it may risk perpetuating student misconceptions. While existing studies suggested that ChatGPT can competently explain statistical terminology (Ellis & Slade, 2023), its understanding of core statistical concepts across different representations has yet to be rigorously examined. With the recent improvements to ChatGPT's multimodal capabilities, including "vision,"

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

researchers now have the opportunity to assess its understanding of statistical concepts in a more nuanced manner, for example, by asking ChatGPT questions that incorporate different representations.

To effectively evaluate ChatGPT's capabilities, it is also critical to understand where students struggle to make sense of different representations, as this information can guide our discussions about ChatGPT's educational potential and its specific role in helping students overcome those challenges. In particular, we need studies that ask the same question, in the same format, to ChatGPT and students, and then compare their knowledge, conceptual understanding, and misconceptions.

In this study, we focus on learners' and ChatGPT's understanding of "data = model + error", a core concept that unifies a number of statistical analyses covered in introductory statistics courses (e.g., *t*-tests, ANOVA, simple regression) that can be depicted in a variety of representations (e.g., graphically, with algebraic notation). This concept encapsulates the foundational idea of statistical modeling: any distribution (the data) can be decomposed into two parts—a systematic component (the model) that formalizes the relationship between a predictor and outcome variable, and a random component (the error) that accounts for the variability not captured by the model. This concept was deliberately chosen as the focus of this study because modeling is a cornerstone for understanding and applying statistical principles in real-world scenarios (Rodgers, 2010; Son et al., 2021).

As part of a larger project including questions of multiple representations, this paper specifically examines questions involving data visualizations—one of the most important and widely used forms of representation in statistics (Unwin, 2020). We begin by assessing students' understanding and misconceptions they bring to these tasks. We then ask ChatGPT to respond to the same questions and compare its answers with those of students. This qualitative comparison allows us to pinpoint where ChatGPT's reasoning aligns with students' and where it diverges, which will inform us about what unique strengths and weaknesses ChatGPT might possess. Ultimately, this work aims to guide future studies investigating whether and how ChatGPT can foster the development of flexible, transferable knowledge in learners of statistics and data science.

Methods

Participants

The study involved 23 college students enrolled in an introductory-level statistics course at a public university. We offered participants a small extra credit for their volunteer participation. By the time students participated in the study, they had already been introduced to the relevant concepts as covered in the course.

Design and Procedures

We sent a Qualtrics survey to students. Students completed the survey online without time constraints. The survey consisted of nine questions, each with multiple sub-questions. Most of these were open-response questions designed to capture nuanced understanding and students' strategies for solving the problems.

We compared the performance of the GPT-4 and GPT-4o with that of students by presenting the same questions to both. The questions were delivered to the models via the ChatGPT online platform. When a question included an image, we uploaded it alongside the text prompt. Due to formatting limitations, some minor design elements from the survey (e.g., bolded text) were omitted in the prompts given to ChatGPT. Each question or set of questions was delivered in a single chat, mimicking the flow experienced by the students during the survey when encountering page breaks. In a few cases, we added brief introductory notes to the prompts to help ChatGPT answer the questions without requesting further clarification, which can be found in the results section.

To fairly evaluate the models' performance, we asked each question to each model five times in separate chats, as generative models can produce varying answers to the same prompts. The repeated trials allowed us to assess the models' "average" performance.

Results

Below, we summarize students' responses to each question and compare them to the corresponding answers generated by GPT models.

Q1. Identifying Components of "Data = Model + Error" on the Graph

In "data = model + error," "data" represents the observed values of the outcome variable, "model" represents the model prediction generated by the statistical model, and "error" (i.e., residuals) represents the deviations between the observed data and the model's predictions.

The first set of questions asks students to identify "data," "model," and "error" from different data visualizations (See Figure 1). Since different visualizations can represent the same components in visually distinct ways or different components in similar ways, correctly identifying them across various visualizations demonstrates a flexible understanding of each component.

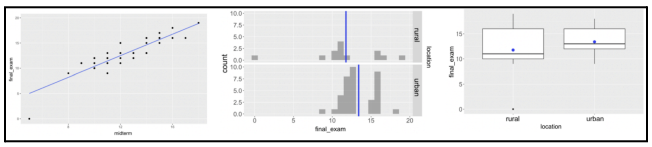

Figure 1: Identify "Data," "Model," and "Error" in Scatter Plot, Faceted Histogram, and Boxplots (Left, Middle, Right)

Scatter Plot In the scatterplot, the "data" is represented by black dots, with the value on the outcome variable depicted as the y-value. The "model" is represented by the blue line, the algebraic function that predicts the outcome variable (y) based on the explanatory variable (x). The "error" is represented by the vertical distances between the black dots and the blue line.

Seventy-four percent of students correctly identified all three components on the scatterplot. All but one correctly identified the black dots as the data. The exception was a student who described the data as the "*whole graph*" and also misidentified the model and error, referring to them as the "*points*" and the "*outlier*," respectively.

Two other students indicated that they could not see the model on the graph, saying instead that it should be "*the average of all points*" or "*the mean… a horizontal straight line of one number.*" These responses align with their explanation of another question asked early in the survey, where they interpreted "data = model + error" as referring to the empty model[1]. This indicates that their understanding of the model might be limited to the empty model. Interestingly, one of them identified error as the distance between each dot and the invisible horizontal line, demonstrating an understanding of the decomposition of data into model and error, but failed to map this idea onto the model depicted in this visualization.

Additionally, three more students identified the error as the outlier while correctly identifying the data and the model. Notably, none of the five students had previously mentioned "*outlier*" in their earlier responses to explain the phrase "data = model + error." This suggests that the concept of error might be the hardest to reconcile with the broader idea of "data = model + error."

GPT models consistently identified and described the three components of the graph correctly. One difference between how GPT models and students described the error is that GPT models specifically mentioned that the error is the "*vertical distance*" between the data points and the model line. In contrast, most students referred to it simply as the "*distance*" or "*gap*" without specifying its vertical nature. For a more precise assessment, educators could ask students to draw the distance on a graph which would clarify whether students truly understand how error was represented.

Faceted Histogram In faceted histograms, the individual data points are not displayed. Instead, the "data" is represented by the gray bars, with the width along the x-axis indicating the range of observed values within each bin. The "model" is depicted by the blue vertical lines, which represent the mean of the outcome variable for each group. Since the exact values of individual data points are not presented, the "error" can be roughly inferred as the horizontal distance between the blue line and the gray bars. If students identified the "data" as the bars and the "error" as the distance between the bar and the blue line—without explicitly discussing the nuance that individual data points are not visible—we still considered their answers correct, as our focus is on their understanding of the relationship between the three components.

Sixty-one percent of students correctly identified all three components. 78% identified the data as the grey bars, and most of the rest incorrectly referred to it as "*count*" or "*frequency*." In addition, one student described it using two values, corresponding to the maximum values on the x-axis for each group. For the model, 91.3% recognized it as the blue lines, while one student described it as "*the two different graphs between the rural and urban*" and another believed there was no model on the graph. Regarding error, 70% of students described it as the distance between the bars and the blue line. Four other students identified the error as the outlier. Two mentioned there was no explicit representation of error, noting, for example, that "*Error is not represented on this graph since it is a histogram that only displays the distribution of data, not the fit of a model to data.*" The student who incorrectly identified the data as the maximums of the two groups, again provided two numbers for the error, calculating each as the maximum minus the model for each group.

GPT-4o correctly identified the data as the gray bars, the model as the blue vertical lines, and the error as the deviation of the gray bars from the blue line in each case. Contrary to GPT-4o, GPT-4 demonstrated a much more limited recognition of the blue vertical lines as representing the model. In 80% of the responses, GPT-4 did not discuss the blue lines explicitly. In one instance, it said, "*…a model might be shown as a line or curve that fits through or represents the trends in the data. Here, there are no such lines or curves*," reflecting a narrower interpretation that aligns more with how models are typically visualized in scatter plots. Because of the complementary nature of the model and error, GPT-4 consequently failed to recognize the error components on faceted histograms, while it has no issue identifying the data represented in the bars.

Box Plots Compared to the histograms, boxplots provide an even more concise summary of the data distribution by showing the five-number summary (minimum, first quartile, median, third quartile, and maximum). This summarization introduces ambiguity when interpreting what constitutes the data (i.e., actual data points versus summary statistics) in the visualization. In the context of our question asking "data," "model," and "error" together, "data" should be interpreted as the observed values, and the ideal answer should acknowledge its "invisibility," as boxplots only display summary statistics rather than raw data points. The "model" is seen as the blue dots, which represent the group means. Since the data points are not visible, "error" is also not directly observable. If we could locate the actual data points

[1] An empty model is a model that does not include any predictors. The mean of the outcome variable is the model prediction for all observations. In the class taken by the students, the empty model is introduced as the first statistical model to help them build a foundational understanding of "data = model + error."

along the y-axis, the "error" would be the vertical distance between each data point and the respective group mean.

Since students may interpret what constitutes "data" in different ways, we considered their answers correct if they could both correctly identify the "model" and express the idea that "error" was not visible due to the invisibility of "data." Based on this criterion, only 21.7% of students answered the question correctly. Specifically, we observed various responses regarding where data was seen on the graph (whole plot, 17%; boxes, 26%; boxes and whiskers, 30%; blue dots which represent the means, 9%; some other statistics, such as the horizontal lines, 9%; black dot which represent the outlier, 9%; no representation, 13%). However, there was greater agreement on what represents the model: 70% of students correctly identified the model as the blue dot.

When it comes to the error, 26% of students claimed that the outliers represented the error. Interestingly, some of them did not identify outliers as errors in the previous plots. The reason could be that those students hold an unclear understanding of what error truly represents. The explicit visibility of outliers in the boxplot (outside the box and whiskers) might have led them to associate errors with outliers. In the scatter plot and histogram, outliers are not distinctly separated from the rest of the data in such an obvious way. Those students' understanding of "data," "model," and "error" may be tied to the specific type of visual representation they are interpreting.

Across all attempts, GPT models consistently recognized the blue dots as models, though they often described them as "basic" or "simple" models. GPT once answered, "*...if we consider 'model' in a more sophisticated sense (e.g., a regression line or statistical prediction model), such a model is not explicitly shown in this plot. This is a simple visualization summarizing the data by location [the predictor variable], not a full model with predicted outcomes or residual analysis.*" While a regression model with a quantitative predictor might yield a greater range of predictions and, in that sense, appear more "complex," this two-group model still includes predictions, albeit in a simpler form. However, GPT models overlooked this fact by saying that there were no prediction outcomes. This suggests that GPT models might only consider more sophisticated models as "models" that serve the purpose of making predictions.

For the data component, GPT models often pointed out that while individual data points were not visible on the graph, the box and whisker plots provided information about the distribution of the data, such as medians and interquartile ranges. Regarding the error component, there was a notable difference between GPT-4 and GPT-4o. GPT-4 frequently highlighted the role of outliers in the description of error, often stating, for example: "*Outliers, if any are present, would further indicate deviations from the model that could be considered errors or noise in the data.*" This focus on outliers mirrored how students were more likely to interpret errors as outliers in boxplots than histograms or scatterplots. Conversely, GPT-4o mentioned outliers in the error description only once, demonstrating a more contextually grounded interpretation of "error" that aligns with its statistical meaning—variability or deviation from the model—rather than a broader or more colloquial understanding of error as a "*mistake*" or "*anomaly*," as seen in GPT-4.

Q2. Same Model, Different Visualizations

After asking students to identify the "data," "model," and "error" in each type of data visualization, we presented all three graphs together (see Figure 1 again), and asked, "*Do any of them depict the same statistical model?*" The correct answer is the faceted histogram and the boxplots, as both depict the same relationship between the same two variables and visualize the model prediction at the group means.

Only 65% of students answered this question correctly. Among the four students who incorrectly selected all three graphs as representing the same model, three provided similar reasoning, stating that all the graphs either "*produced a prediction*" or "*represented variability in the data*." However, they failed to understand that for the statistical model to be the same, they must use the same predictor(s). Another student said all graphs used the same predictor variable.

For those who selected that none of the graphs depicted the same statistical model, their reasoning appeared also problematic. One student incorrectly reasoned that none of the axes matched across the graphs. Another argued that even if the variables were the same, the graphs were of different types and showed different data points. However, the same model can be represented in different graphs, and the graphs were generated using the same data. The third student failed to identify any statistical model in the histogram.

Finally, one student selected the scatterplot and the histogram to depict the same statistical model. Their reasoning was that "*the data 'spread' looks very similar; it looks as though the same number of values for each x-axis is present.*" This likely reflects a focus on surface-level similarities between the scatter plot and histogram, where both show data spread along the x-axis. Taken together, the incorrect responses highlight that some students' understanding of statistical models was fragile and tied to the surface features of graphs.

Surprisingly, GPT models performed worse than students on this question, with GPT-4 scoring 60% correctness and GPT-4o only 20%. Their responses reflected misunderstandings similar to those of the students. For example, they were overly reliant on the surface features of the different visualizations, as noted, "*Graph B and C both categorize data by location but present it differently; B shows frequency distributions while C provides summary statistics via a box plot.*" Additionally, there are instances where they failed to identify models in the faceted histogram and the boxplots, similar to how they answered earlier questions, such as: "*Graph B uses histograms, while*

Graph C uses boxplots, both focusing on a descriptive comparison rather than predictive modeling."

This question demands the synthesis of information from multiple sources and the drawing of conclusions based on logical reasoning, tasks that pose significant challenges for models primarily designed to process and generate text (Liu et al., 2023). While GPT-4 and GPT-4o have been trained with both text and images, they may still struggle with the higher-order reasoning required to effectively compare the elements and deduce meaningful conclusions. This limitation underscores the need to enhance the models' capacity to reason through visual data. Such improvements could significantly expand their applicability in answering complex data-related questions, thereby broadening their utility in fields that heavily rely on data visualization.

Q3. Identifying b_0 and b_1 on the Graph

Understanding the slope and intercept is crucial in statistics, as they are key to interpreting the linear relationship and have numerous practical applications. However, many studies conducted with middle- and high-school students have revealed persistent conceptual difficulties with these concepts (Davis, 2007; Hattikudur et al., 2012; Knuth, 2000). Even college students struggled to write an equation from a graph, particularly when asked to interpret the equation within a real-world scenario (Cho & Nagle, 2017).

In the standard GLM notation for a two-parameter linear model ($Y_i = b_0 + b_1 \times X_i + e_i$), b_0 represents the y-intercept (the predicted value of Y_i when X_i is 0), and b_1 is the slope of the line in the Cartesian coordinate system. We displayed a scatter plot overlaid with a linear regression line within a real-world context (i.e., label the actual variable names rather than generic notations like x and y). We asked students to "*estimate the values of b_0 and b_1 for the statistical model*" represented by the blue line. Students need to understand that b_0 is the y-intercept and that b_1 is the slope, then find these values using the graphical information provided.

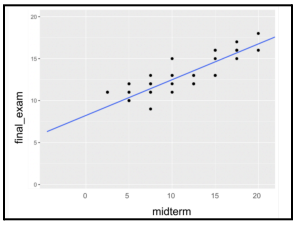

Figure 2: Estimate b_0 and b_1 from the Graph

To add more challenge, we selected a graph where the x-axis does not start at x = 0 (see Figure 2), which helps differentiate between students who determine the y-intercept heuristically based on where the line intersects the left-most boundary of the graph and those who understand that the y-intercept corresponds to the value of y when x = 0.

Intercept (b_0) Since we did not label specific coordinates, the slope and intercept values can only be roughly estimated visually from the graph. We considered answers within the range of 7.5 to 10 correct, as the graph shows that when x = 0, y falls between 7.5 and 10 according to the grid's labeled numbers. Surprisingly, only 22% of students' answers met this criterion. The most frequent incorrect answers were around 5 and 6 (44%), which correspond to where the line meets the left-most boundary of the graph. This suggests that they may have used a heuristic, incorrectly interpreting this boundary as the location of x = 0. When asked to explain their reasoning, 48% of students explained that they identified b_0 as the y-intercept, while another 22% indicated that they located the y-intercept where x = 0. Intriguingly, even when students claimed to use x = 0 as a reference, they did not always answer correctly. Among the incorrect strategies, two reported using the "*start*" point of the line, another two referred to where the line "*crossed*" the y-axis, and one identified b_0 as the mean of all data points.

Although it is possible to estimate the y-intercept by calculating the equation from two dots on the line, most students pursued the visual estimation approach. However, it seemed that many did not pay attention to a key feature of the y-intercept, that it relies on where x = 0, making their answers prone to error.

GPT models[2] performed no better—and in some cases, worse—than students. The performance was similar across GPT-4 and GPT-4o. Across a total of 10 trials between the two models, only once did GPT-4o provide a close answer. While GPT models had no difficulty recognizing that b_0 represents the y-intercept, they frequently (70% of the time) stated that the y-intercept was 5, mirroring the incorrect answers provided by many students.

To investigate this further, we examined the models' explanations to determine whether their mistakes stemmed from a similar heuristic—mistakenly assuming the left border of the graph represents the y-axis. For instance, in one explanation, GPT-4o stated: "*When X = 0, the blue line appears to intersect the y-axis near 5. This suggests $b_0 \approx 5$.*" A notable difference between GPT-4 and GPT-4o's approaches was their strategy for solving the problem. GPT-4 consistently selected two points on the graph, calculated the line equation, and then substituted x = 0 to solve for the y-intercept—even when one of the initially chosen points was already at x = 0. In contrast, GPT-4o pursued the visual approach as students. Despite GPTs' precise calculations, both models failed to select the correct points from the graph, leading to poor overall performance, demonstrating their limitation in "vision."

Slope (b_1) All but one student identified b_1 as the slope of the line or provided some form of definition or calculation, such as "*rise over run.*" However, students rarely documented which two points they used in the calculation, which makes it difficult to determine whether errors in

[2] We added the phrase "using the information on the graph" to the prompt because, without it, ChatGPT often requested additional details like specific coordinates. This directive was included to provide necessary context upfront and avoid unnecessary back-and-forth.

responses were due to error in point selection, reading coordinates, or calculation of the slope.

To facilitate fair comparisons between human participants and GPT models, we accepted answers in the range of 0.3 to 0.7, as the actual slope is approximately 0.5. Only 30% of students answered met this criterion. One possible reason for the low accuracy rate could be the use of an incorrect y-intercept in their calculations. While we found no direct evidence supporting this speculation—students' responses were often brief—one detailed answer from a successful student shed light on their approach. This student explained: "*I used the slope formula to calculate b1 given that I had a good estimate of the y-intercept and the line seems to have crossed at (10, 12.5)*." This suggests that students using similar strategies might have been influenced by errors in their initial estimate of the y-intercept, leading to inaccuracies in their slope calculations.

According to the same criterion, GPT models appeared to outperform humans, with 50% of their answers falling within this acceptable range. However, because GPT models typically provided detailed explanations, it was easier to identify exactly what went wrong in their reasoning. In many cases, their correct answers seemed to result more from chance than accuracy. For instance, GPT-4o once stated, "*The blue line seems to pass approximately through (0, 5) and (20, 15)*," even though the line was far from these two points. GPT models struggled to extract the coordinates of the points through which the line passed, and their ability to read coordinates seemed worse than that of students.

Q4. Comparing Model Fitness on Graph

When a model fits the data better, it produces predictions with smaller residuals, resulting in a smaller Sum of Squares Error. This means the model's predictions are generally closer to the actual values on a scatter plot. We presented students with two scatter plots (see Figure 3) containing the same data points but differing in the positioning of the model predictions—one with smaller errors and the other with larger errors.

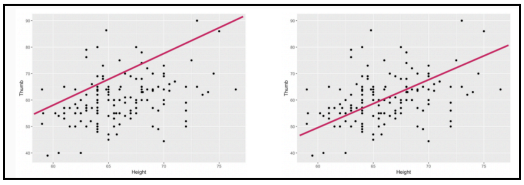

Figure 3: Compare Model Fitness

All students except one answered this question correctly, demonstrating their understanding of how errors are represented on the graph. E.g., "*The model on the left has more error. This is because the majority of the data points are further from the model, resulting in more error. The model on the right seems to balance the data points evenly, which means there is less error.*"

GPT-4o also achieved a 100% accuracy rate, however, GPT-4 only 20%. While both models pointed out that the closer the points to the regression line, the better the model prediction, GPT-4 often judged that the graph with larger dispensation (the one on the right) was tighter around the line, again likely due to its limitation in "vision."

Discussion

In this paper, we compared students' and GPT models' answers to four questions about the "data = model + error" involving reading and interpreting data visualizations. We found that while GPT models exhibit certain strengths, they also face notable limitations that align with challenges observed in student understanding.

We observed that mapping the "data = model + error" concept to visualizations remains challenging for some students (Q1). GPT-4o outperformed both GPT-4 and students in identifying key components of the framework, indicating basic ability to read graphs and improvements between the two GPT versions. However, when tasked with more complex problems, such as comparing models (Q2 and 4), GPT models performed no better—and sometimes worse—than students. These struggles may be attributed to the models' limited reasoning abilities when working with graphs and making comparisons. Moreover, GPT models demonstrated limitations, particularly in extracting precise details from graphs, such as identifying the coordinates of points (Q3), a finding consistent with previous studies (Verma et al., 2024).

Intriguingly, GPT models sometimes exhibited the same misconceptions as students. For example, both groups used similar incorrect heuristics to determine the y-intercept on a graph and frequently associated "error" with "mistake" when interpreting box plots. The shared misunderstandings suggest that LLMs may reflect limitations found in students' educational experiences, likely due to the training data mirroring typical educational content. For instance, students are rarely exposed to graphs without clearly defined axes or origins in their algebra class, which could lead both students and LLMs to develop similar errors. The parallels between GPT model errors and student misconceptions highlighted an opportunity for improving educational materials.

Our findings suggest that it is premature to conclude that GPT-4 and GPT-4o possess a solid understanding of the concepts we tested. While these models could produce fluent and plausible responses, they struggled with accurately interpreting visualizations. This aligns with findings in other domains showing that GPT models lack deep conceptual understanding (Wardat et al., 2023; Wheeler & Scherr, 2023). It's important to note that our study focused on just four questions targeting a specific statistical concept. As such, it provides only a narrow window into the models' capabilities. Further research is needed to assess how well these models handle more complex statistical reasoning and the ability to transfer knowledge across different contexts. If students and educators are not aware of LLMs' limitations, there is a risk that these tools will be misused in learning contexts, potentially reinforcing misunderstandings.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., . . . Zoph, B. (2023). *GPT-4 Technical Report.* arXiv:2303.08774v6. https://doi.org/10.48550/arXiv.2303.08774

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*(2-3), 131-152. https://doi.org/10.1016/S0360-1315(99)00029-9

Cho, P., & Nagle, C. (2017). Procedural and conceptual difficulties with slope: An analysis of students‟ mistakes on routine tasks. *International Journal of Research in Education and Science (IJRES), 3*(1), 135-150.

Chung, K. L. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Cooper, L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education, 16*(2), 1. https://doi.org/10.1080/10691898.2008.11889559

Davis, J. D. (2007). Real-world contexts, multiple representations, student-invented terminology, and y-intercept. *Mathematical Thinking and Learning, 9*(4), 387-418. https://doi.org/10.1080/10986060701533839

Ellis, A. R., & Slade, E. (2023). A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education. *Journal of Statistics and Data Science Education, 31*(2), 128–133. https://doi.org/10.1080/26939169.2023.2223609

Estrella, S. (2018). Data representations in early statistics: Data sense, meta-representational competence and transnumeration. In A. Leavy, M. Meletiou, & E. Paparistodemou (Eds.), *Statistics in early childhood and primary education—Supporting early statistical and probabilistic thinking* (pp. 239–256). Singapore: Springer.

Fries, L., Son, J. Y., Givvin, K. B., & Stigler, J. W. (2021). Practicing connections: A framework to guide instructional design for developing understanding in complex domains. *Educational Psychology Review, 33*(2), 739–762. https://doi.org/10.1007/s10648-020-09561-x

Hattikudur, S., Prather, R. W., Asquith, P., Alibali, M. W., Knuth, E. J., & Nathan, M. (2012). Constructing graphical representations: Middle schoolers' intuitions and developing knowledge about slope and y‐intercept. *School Science and Mathematics, 112*(4), 230-240. https://doi.org/10.1111/j.1949-8594.2012.00138.x

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., . . . Malkov, Y. (2024). *GPT-4o System Card.* arXiv:2410.21276v1. https://doi.org/10.48550/arXiv.2410.21276

Knuth, E. J. (2000). Student understanding of the Cartesian connection: An exploratory study. *Journal for Research in Mathematics Education, 31*(4), 500-514.

Kozma, R. (2003). The material features of multiple representations and their cognitive and social affordances for science understanding. *Learning and Instruction, 13*(2), 205–226. https://doi.org/10.1016/S0959-4752(02)00021-X

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., … & Piao, Y. (2024). *Deepseek-v3 technical report*. https://doi.org/10.48550/arXiv.2412.19437

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of chatgpt and gpt-4. arXiv:2304.03439v3. https://doi.org/10.48550/arXiv.2304.03439

Modell, H., Michael, J., & Wenderoth, M. P. (2005). Helping the learner to learn: the role of uncovering misconceptions. *The American Biology Teacher, 67*(1), 20-26. https://doi.org/10.2307/4451776

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*(1), 1–12. https://doi.org/10.1037/a0018326

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14. https://doi.org/10.3102/0013189X015002004

Son, J. Y., Blake, A. B., Fries, L., & Stigler, J. W. (2021). Modeling first: Applying learning science to the teaching of introductory statistics. *Journal of Statistics and Data Science Education, 29*(1), 4-21. https://doi.org/10.1080/10691898.2020.1844106

Tu, X., Zou, J., Su, W., & Zhang, L. (2024). What Should Data Science Education Do With Large Language Models?. *Harvard Data Science Review, 6*(1). https://doi.org/10.1162/99608f92.bff007ab

Unwin, A. (2020). Why Is Data Visualization Important? What Is Important in Data Visualization? *Harvard Data Science Review, 2*(1). https://doi.org/10.1162/99608f92.8ae4d525

Verma, A., Mukherjee, K., Potts, C., Kreiss, E., & Fan, J. E. (2024). Evaluating human and machine understanding of data visualizations. In *Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46)*.

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(7), em2286. https://doi.org/10.29333/ejmste/13272

Wheeler, S., & Scherr, R. E. (2023). ChatGPT reflects student misconceptions in physics. In *Proceedings of the Physics Education Research Conference (PERC)* (pp. 386-390).

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X.,

Liu, Z., … & Wen, J. R. (2023). *A survey of large language models.* arXiv:2303.18223v15. https://doi.org/10.48550/arXiv.2303.18223