Supplementary Materials:
Does visualization help AI understand data?

Victoria R. Li*, Johnathan L. Sun*, Martin Wattenberg

IEEE VIS 2025

We conducted a brief additional analysis of the model's failure mode behavior (Sec. 1), and show the complete set of boxplots describing model dataset description length and statistical term use (Sec. 2). We also release all code used, including for dataset generation, experimental execution, model output classification, analysis, and visualization. We put all these documents, along with the datasets containing model responses in full, in the publicly available repository: github.com/johnathansun/lvlm-vis-data-understanding.

1 Failure Analysis

We further investigate GPT and Claude's incorrect outputs within each of our three task settings. We generate confusion matrices for both LVLMs with task subtlety on the y axis and model response classifications on the x axis to show at-a-glance shifts in failure mode behavior.

1. **Clustering:** We used Gemini-as-a-judge to identify the exact number of clusters a model response described, which is compared against the true number of clusters (Figs. 1 and 2)
2. **Parabola:** We classify whether model descriptions included both specific and broad, only specific, only broad, or neither specific nor broad terms in their descriptions of a dataset trend (Figs. 3 and 4). The keywords we searched for are shown in Tab. 1.

Specific Terms	[“parabola”, “parabolic”, “quadratic”, “concave”, “opens down”]
Broad Terms	[“nonlinear”, “non-linear”, “isn't * linear”, “not * linear”, “curved”, “polynomial”]

Table 1: The terms we match to classify if a model is specific or broad in describing a parabolic trend.

3. **Outlier:** We use Gemini-as-a-judge to evaluate whether the exact true outlier is identified in the response. We also classify if other or no points are described as anomolous across subtlety conditions (Figs. 5 and 6).

For both GPT and Claude, the distribution of failure mode responses in the clustering and outliers tasks under the “Data & Blank” condition appear more similar to the “Data Only” condition than the “Data & Wrong” condition. This result indicates that not only visual input but visual information may be necessary to modify model behavior under some conditions.

However, GPT responses under the “Data & Blank” condition are more similar to the “Data & Wrong” responses than the “Data Only” responses for parabola identification (Fig. 3). This contrasts with both models' behavior on the other two tasks and Claude responses in the same parabolic task, in which “Data & Blank” is more similar to “Data Only.” This result suggests that the “Data & Blank” condition can help elucidate whether the effect of any visual input itself may vary jointly between models and tasks, possibly systematically impacting the output distribution.

In particular, for the parabola task, GPT responses under the “Data & Blank” condition, especially for 3-7 points beyond the vertex, are much less likely to reference either broad or specific keywords compared to the “Data Only” condition, revealing one scenario in which the blank image may degrade the quality of model responses (Fig. 4). Note that our keywords lists (Tab. 1) may not be comprehensive, but nevertheless indicate that responses may shift focus to other salient features in dataset descriptions with a blank image. The number of Claude responses including broad keywords also drops significantly more than GPT responses for data with 3-7 points beyond the vertex, further emphasizing that the effect of the “Data & Wrong” condition can also differ between models (Fig. 3).

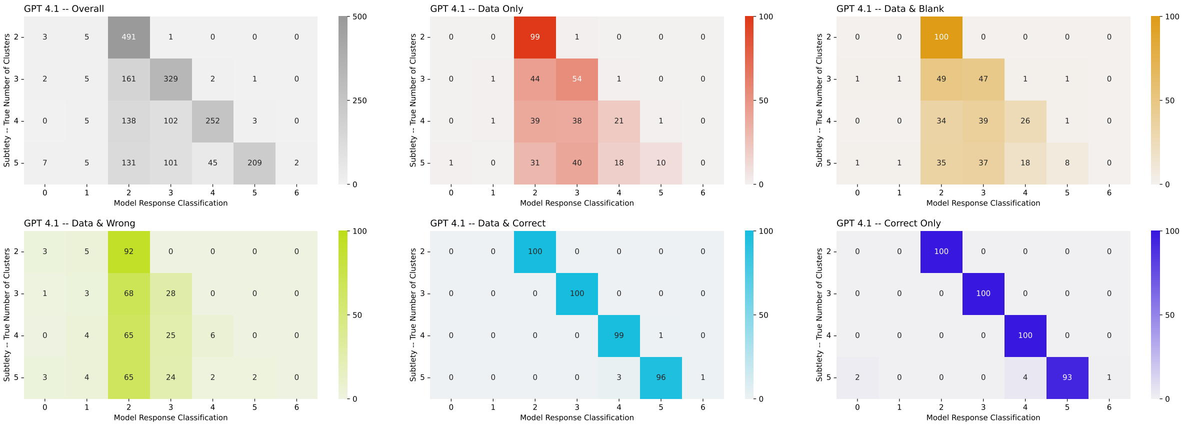


Figure 1: The Gemini-judge based classification of GPT responses across subtleties for the clustering task.

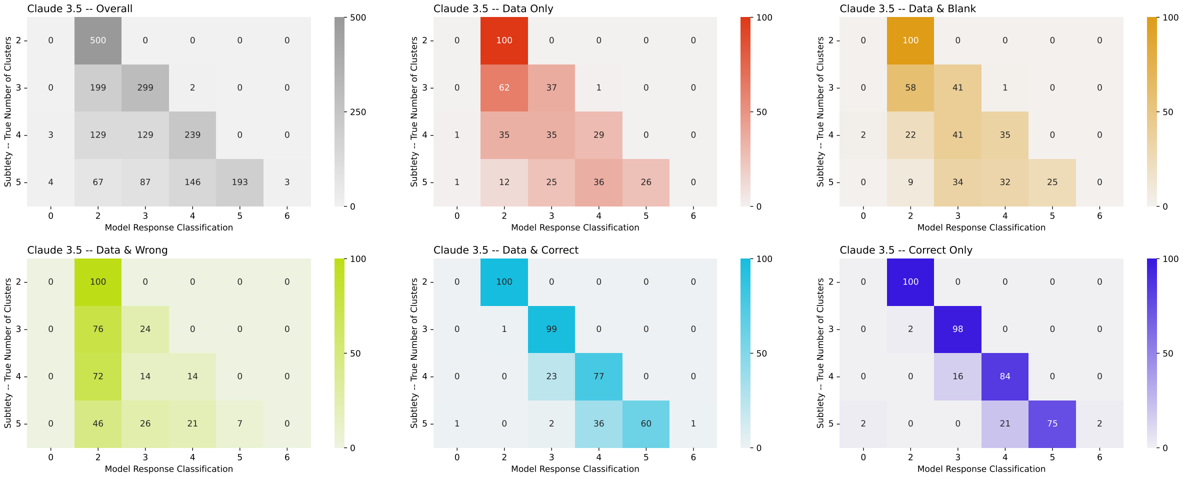


Figure 2: The Gemini judge-based classification of Claude responses across subtleties for the clustering task.

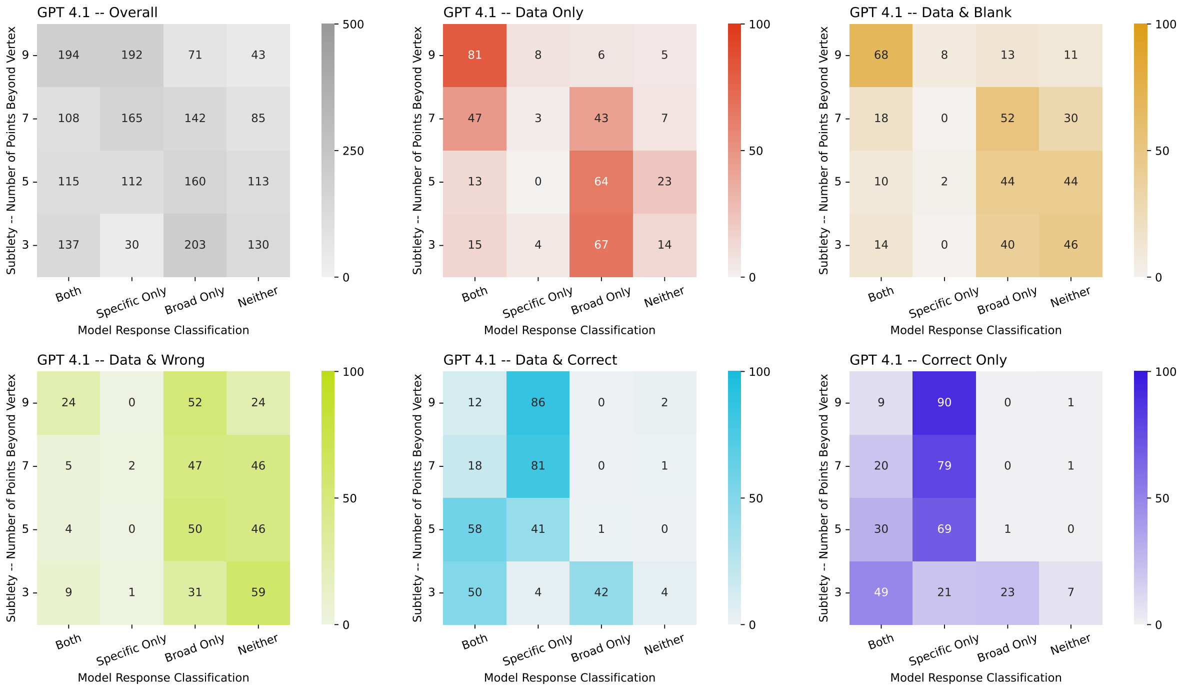


Figure 3: The keyword-based classification of GPT responses across subtleties for the parabola task. See Tab. 1 for the particular terms we used to classify whether a model description was specific or broad.

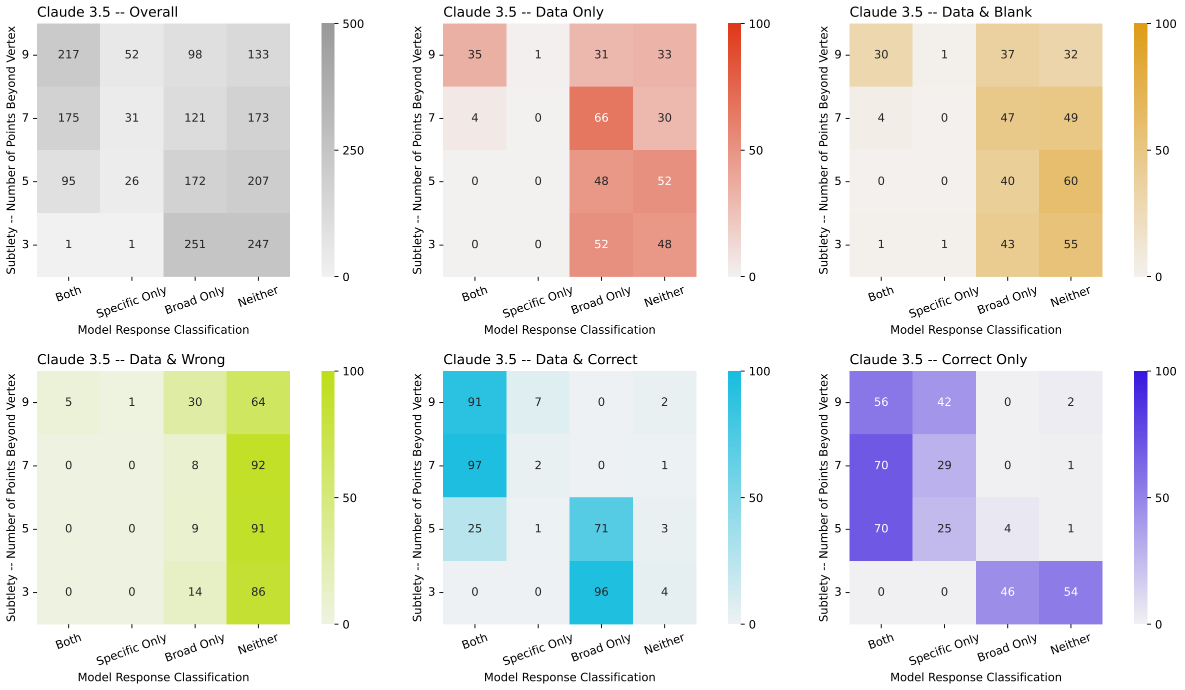


Figure 4: The keyword-based classification of Claude responses across subtleties for the parabola task. See Tab. 1 for the particular terms we used to classify whether a model description was specific or broad.

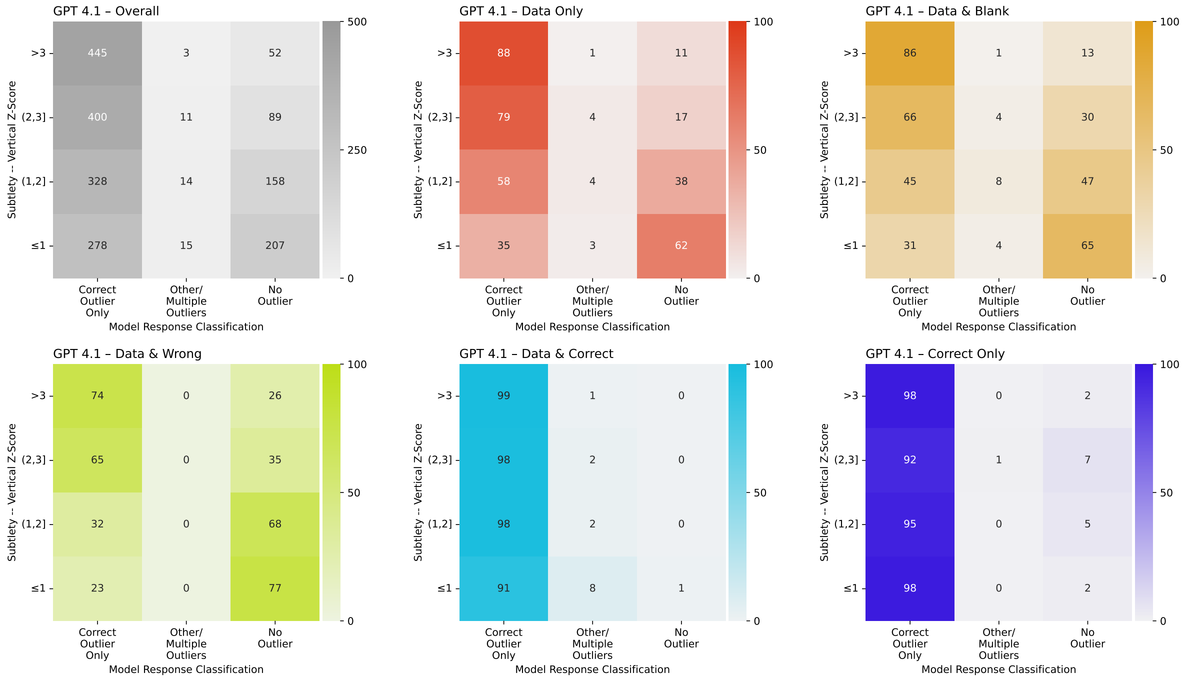


Figure 5: The Gemini judge-based classification of GPT responses across subtleties for the outlier task. We evaluate if only the correct outlier is mentioned as anomalous. "No Outlier" counts cases where no points are identified as outliers. "Other/Multiple Outliers" involves cases where at least one outlier is identified, including when the correct outlier is pointed out with others as anomalous in the model dataset description.

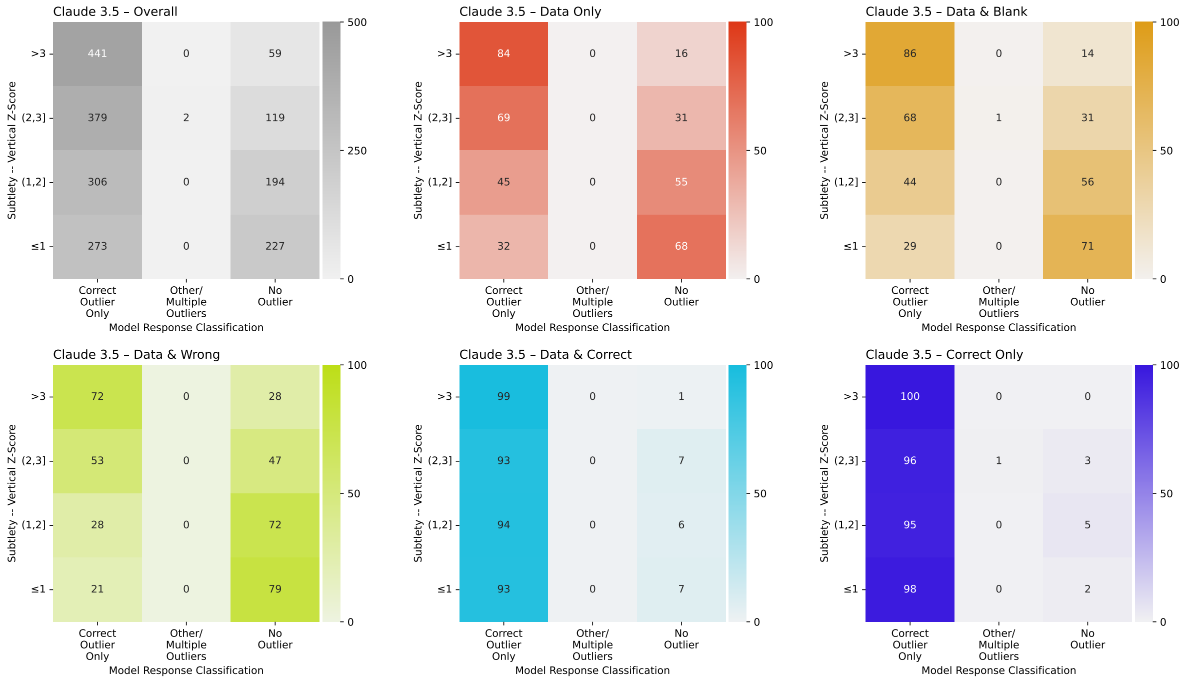


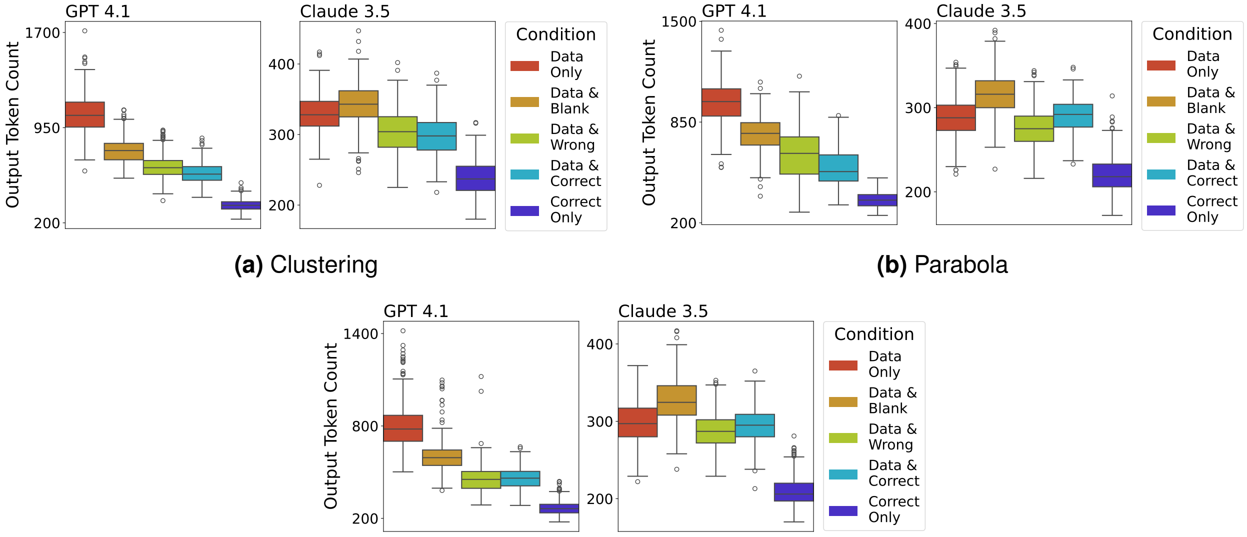
Figure 6: The Gemini judge-based classification of GPT responses across subtleties for the outlier task. The categories are defined in the same way as Fig. 5.

2 Overall

Across tasks, as mentioned in the main body of the paper, we also compute the output token count and the models' use of statistical terms in their dataset descriptions. The specific statistical terms we counted model use of were: "mean", "min", "max", "average", "variance", "standard deviation", "correlation", "range", and "domain," which is broad-ranging but possibly incomprehensive.

Across both GPT and Claude in all tasks, we observe that:

- The number of tokens per response is higher under conditions without a visualization than under conditions with the correct scatterplot, indicating any visual input, but particularly an accurate data visualization, increases dataset description concision. In general, the "Data Only" condition has the longest responses, and "Correct Only" the shortest (Fig. 7).
- The proportion of statistical terms appearing in model responses generally decreases when the correct visualization is added to the model prompt. Notably, for GPT 4.1, especially on the clustering and outlier tasks, the "Data & Wrong" condition contains the greatest proportion of statistical terms. This observation indicates models may conduct more numerical analysis when faced with conflicting information even when they do not explicitly call out a difference (Fig. 8).
- The raw number of statistical terms in model responses show a decrease in the total count of statistical terms in model responses given a correct scatterplot. The difference in models' use of statistical terms under the "Data & Blank", "Data & Wrong" and "Data & Correct" conditions indicate that seeing specifically a correct visualization decreases model reliance on the included raw dataset (Fig. 9).

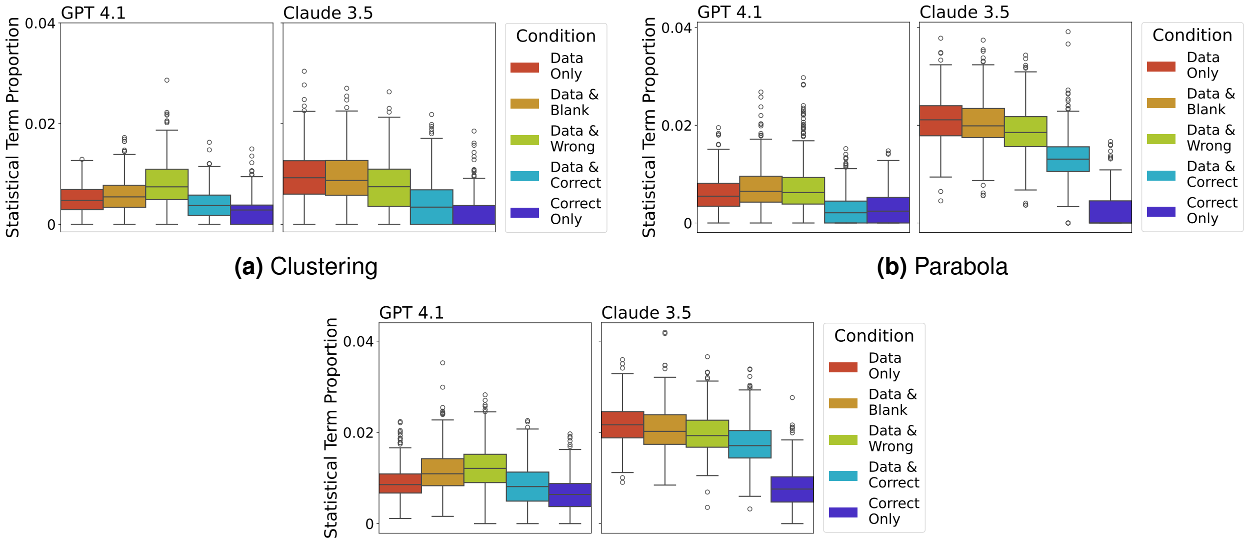


(a) Clustering

(b) Parabola

(c) Outlier

Figure 7: Distribution of the number of tokens per response for each task and condition.

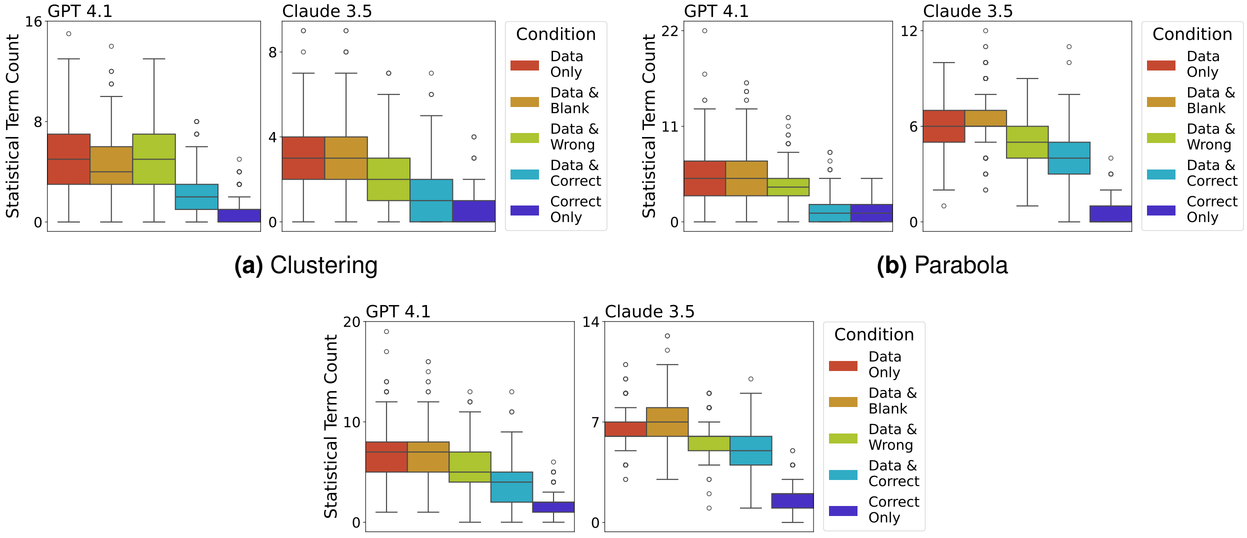


(a) Clustering

(b) Parabola

(c) Outlier

Figure 8: Distribution of normalized term count—the number of times statistical terms appear in a response divided by the response token length—for each task and condition.



(a) Clustering

(b) Parabola

(c) Outlier

Figure 9: Distribution of statistical term count per response for each task and condition.