# Previous Report

## 1 Comparison of models

We first compare the results of GPT-3.5 and GPT-4 using one round of prompts for building static views. Similarly, we do not compare the GPT-3.5 and GPT-4 for multiple coordinated views because GPT-3.5 can barely finish these tasks. Note that we do not use multiple rounds of prompts in this comparison, as we find that VSR@1 is a strong indicator of performance. We refer the readers to section 2.3 for results with multiple rounds of interactions.

**Overall performance.** We find that the two LLMs exhibit different performance patterns across libraries, as shown in table 1. GPT-4 achieves the highest VSR@1 using ECharts for both the static and interactive views, while GPT-3.5 performs the best using D3. Additionally, GPT-4 outperforms GPT-3.5 in every single scenario, but the performance gap varies. For example, for interactive views, GPT-4 shows significant progress on ECharts and Vega-Lite (from 24.25% to 70.73% using ECharts), but only marginal improvements on D3 (from 36.36% to 40.91%). A possible explanation is that OpenAI might adopt a focused data enhancement that specifically targets areas where GPT-3.5 shows weakness, leading to improved performance of GPT-4. In this section, we will further investigate the performance difference on libraries and the impact of fine-tuning on focused data.

**Case analysis.** When investigating the visualization results in detail, we find that GPT-3.5 may exhibit a special type of visual representation error, that it might visualize the data with a wrong type of chart. This issue only occurred when using GPT-3.5 and D3. In contrast, this kind of mismatch was rare in the codes generated by GPT-4, indicating a higher level of alignment between the task instructions and the generated visualization codes.

Table 1: The VSR@1 performance of GPT-3.5 and GPT-4 on generating individual views using D3, ECharts and Vega-Lite.

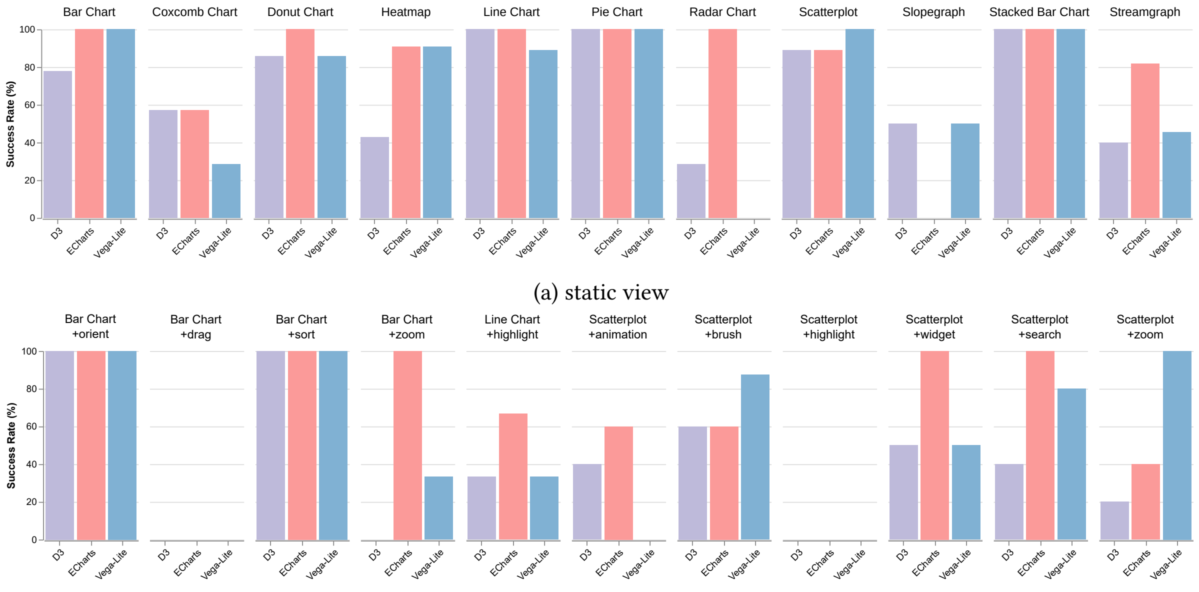
| | Static View | | | Interactive View | | |
|---|---|---|---|---|---|---|
| VSR@1 | D3 | ECharts | Vega-Lite | D3 | ECharts | Vega-Lite |
| GPT-3.5 | 61.54% | 58.62% | 53.33% | 36.36% | 24.25% | 23.08% |
| GPT-4 | 71.79% | 87.18% | 74.36% | 40.91% | 70.73% | 52.27% |

## 2 Performance Across Tasks

This section mainly focuses on the performance of the model GPT-4.

### 2.1 Static View

**The success rate of GPT-4 on generating static views seems to be correlated with the popularity of chart types**, as shown in the first row of fig. 1. For the four basic charts (bar charts, line charts, scatterplots, and line charts), GPT-4 achieved 100% VSR@1 in most tasks (8 out of 12) using the three libraries. For the remaining ones that were not perfectly successful, the VSR@1 was higher than 77.78%. For other more advanced charts (donut charts, stacked bar charts, and heatmaps), GPT-4 also delivered satisfactory performance, with the VSR@1 higher than 80% in 8/9 tasks. The only exception was generating heatmaps using D3, which had a VSR@1 of 42.86%.



(a) static view

(b) interactive view

Figure 1: The one-round success rate (VSR@1) for GPT-4 in generating static views and interactive views using D3, ECharts, and Vega-Lite. The first row shows the results of static views, and the second row shows those of interactive views.

However, the less common charts (coxcomb charts, streamgraphs, radar charts, slopegraphs) clearly posed a challenge for GPT-4. For these charts, the VSR@1 was relatively stable across tasks using D3 (between 28.57% to 57.14%), while more fluctuated using ECharts (0% for slopegraphs and 100% for radar charts). The VSR@1 using Vega-Lite was lower (between 0% and 50%).

**The success rate might rely on the libraries as well.** The diverse success rates across chart types may be attributed to both the training data and the nature of the libraries. The most commonly used charts have more training samples and receive better support from the libraries, leading to high success rates for all libraries. For the less common charts, GPT-4 never completely failed using D3, which might be due to its flexibility. However, the flexibility also introduces extra coding complexity, constraining the success rate using D3. The other two libraries were easier to use but less customizable, leading to larger variation in VSR@1. This can be further confirmed by the following analysis of error types.

Table 2: Distributions of error types for GPT-4 in generating static views using D3, Echarts, and Vega-Lite.

| Error Type | Visual Representation Error | Compilation Failure | Unsupported Features |
|---|---|---|---|
| D3 | 86.36% | 4.55% | 9.10% |
| ECharts | 20.00% | 10.00% | 70.00% |
| Vega-Lite | 55.00% | 5.00% | 40.00% |

**The distributions of error types vary across libraries.** As shown in table 2, most of the errors (86.36%) using D3 were related to visual representation, while only a small portion (9.1%) was due to unsupported features. In contrast, ECharts exhibited more errors (70%) because of unsupported features and only 20% of visual representation errors. Vega-Lite had similar percentages between these two error types. Examining the failure cases, the errors using D3 were typically misaligned elements, such as axes, legends, or labels that are improperly positioned. This kind of error was less common for libraries based on specifications, but it might be more difficult for GPT-4 to add features that were not supported by these libraries.

## 2.2 Interactive View

**The success rates depend on the task difficulties and libraries used.** This observation is similar to those on the static views. GPT-4 achieved 100% VSR@1 in the common tasks such as sorting and customizing the orientation of bar charts using all three libraries, but completely failed on the rare tasks such as making the bars draggable in bar charts, or highlighting the selected points on the axes of scatterplots.

Other than the extreme cases (100% or 0% VSR@1), GPT-4 usually performed well using ECharts, with VSR@1 of higher than 60% in 3/4 tasks. Vega-Lite was close in most cases except for the animation which was not supported. D3 did not completely fail in all these tasks, but it was constantly outperformed by the other two in most cases.

Table 3: Distributions of error types for GPT-4 in generating interactive views using D3, ECharts, and Vega-Lite.

| Error Type | Interactive Error | Visual Rep. Error | Compilation Failure | Unsupported Features |
|---|---|---|---|---|
| D3 | 65.38% | 34.62% | 0.00% | 0.00% |
| ECharts | 41.67% | 33.33% | 0.00% | 25.00% |
| Vega-Lite | 28.57% | 4.76% | 4.76% | 61.90% |

**The error type distributions still correlate with the libraries used.** As shown in table 3, all errors were related to interaction and visual representation using D3, which might be due to its coding complexity. Slightly different from the results of static views was that the unsupported feature error became a less critical factor for ECharts but the most prominent one for Vega-Lite. This suggests that Vega-Lite's declarative approach, while simplifying visualization creation, also sacrifices flexibility in customizing interactions.

## 2.3 Multiple Coordinated Views

For simplicity, GPT-4 was prompted to create an interface consisting of two views for each task.

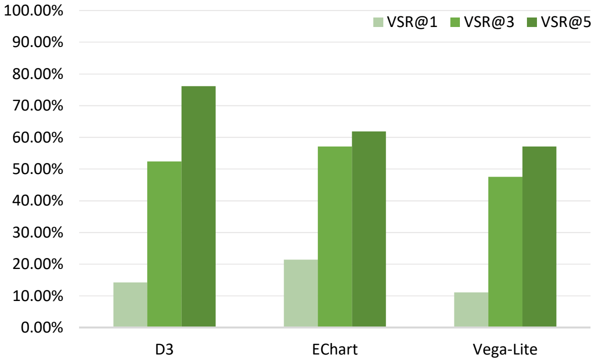


Figure 2: The success rate within 1, 3, and 5 rounds of interactions to multiple coordinated views using D3, ECharts, and Vega-Lite.

**The tasks to build multiple coordinated views benefit more from the interaction.** For all three libraries, significant improvements were observed. Compared to tasks to build the interactive views, these tasks had lower success rates in the first round but ended up with similar or even higher success rates in round five.

This demonstrated that the follow-up prompts were useful for correcting the errors, and the marginal improvement found in the interactive views might be due to that GPT-4 was approaching its limits without exemplar codes.

**The improvement might vary across libraries.** D3 exhibited the largest improvement from VSR@1 of 14.28% to VSR@5 of 76.19%. Falling behind ECharts at VSR@1 and VSR@3, it successfully surpassed ECharts after five rounds of interactions. This might still be attributed to the flexibility of D3.

## References