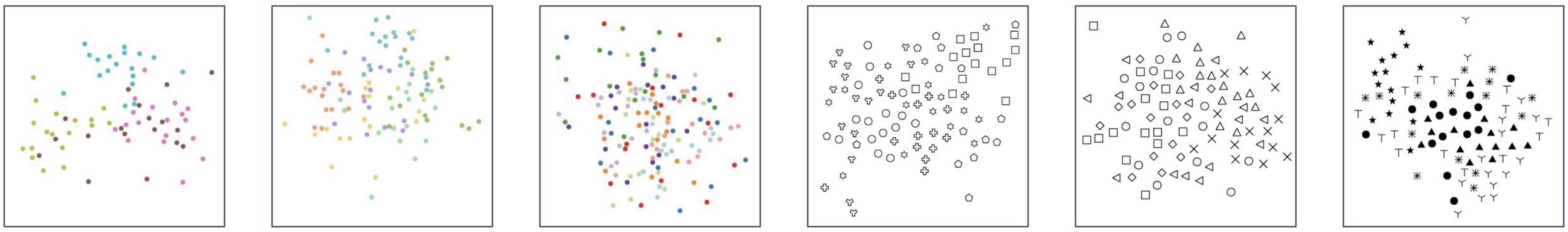
Examining the Capabilities of LLMs in Interpreting Categorical Encodings from Data Visualizations

Arran Zeyu Wang, Matt-Heun Hong, *and* ***Danielle Albers Szafir***

University of North Carolina at Chapel Hill

Can GPT-4 Achieve Human Performance in Categorical Scatterplots?

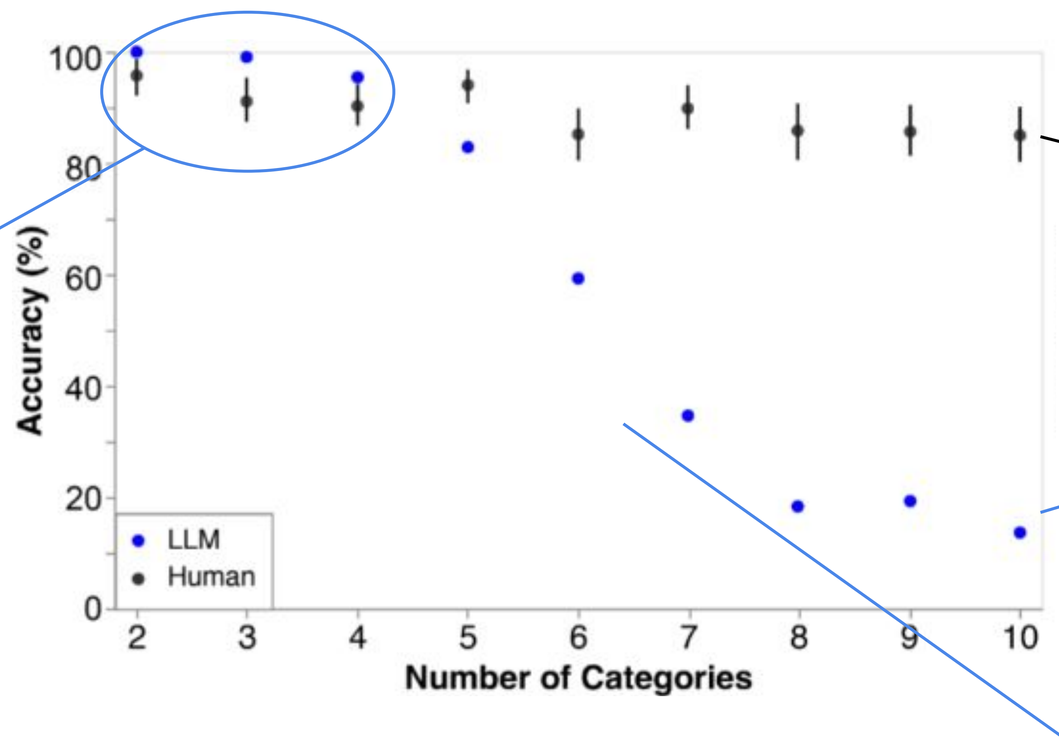


Task: Find a class with the **highest average y-value**. Encoded by color or shape[1-2].

Color Encodings

...

Nearly **100% accuracy** and outperformed human results for < 4 categories

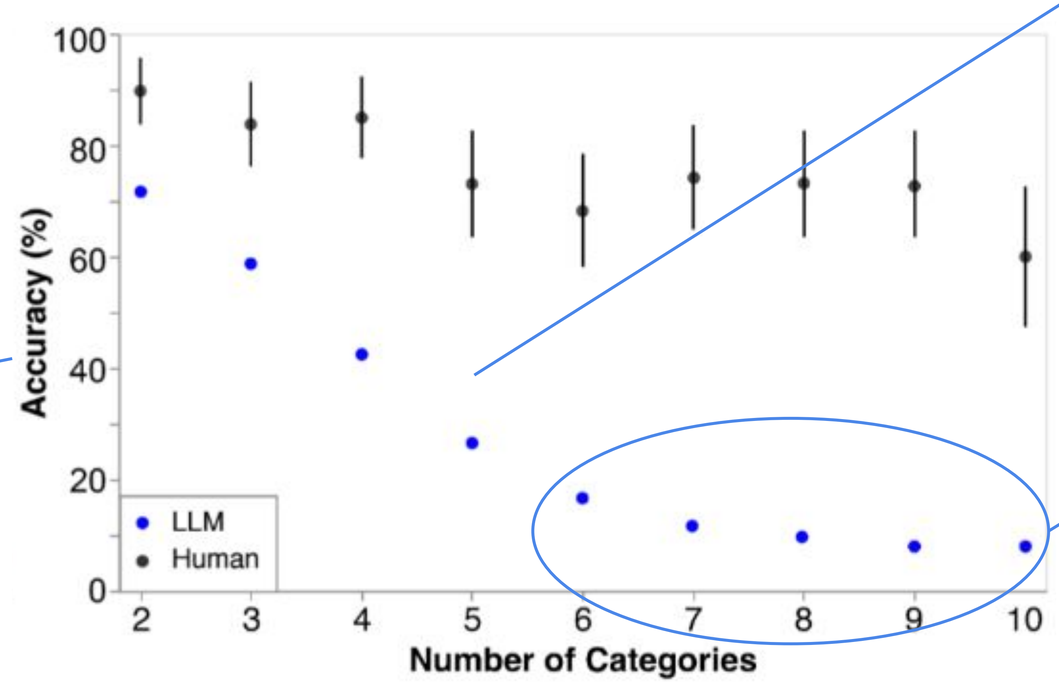


People maintained **80+%** acc., GPT-4 dropped to **20%**

Dramatically dropped with the category number increasing for both encodings

Shape Encodings

Accuracy comparison: **Human** vs **LLM**



GPT's performance on shapes is **significantly worse** than with color encodings

Worse than chance for > 6 categories

Summary

Benefits toward Design and Accessibility

GPT-4 can accomplish **accessibility best practices (*alt-text*)** for static charts with **a small number of color-coded categories** at near-perfect performance.

Pitfalls

Current LLMs still **struggle with visualizations having many categories**, achieving less than 10% accuracy compared to humans' 85% (color-coded) and 60% (shape-coded).

[1] Tseng et al. Measuring categorical perception in color-coded scatterplots. ACM CHI, 2023.
[2] Tseng et al. Shape it up: An empirically grounded approach for designing shape palettes. IEEE TVCG (Proc. IEEE VIS 2024), 2025.