VIS2022

Using Large Language Models to Generate Engaging Captions for Data Visualizations

Ashley Liew, Rochester Institute of Technology
Klaus Mueller, Stony Brook University

Captions in Technical Papers are Often Boring

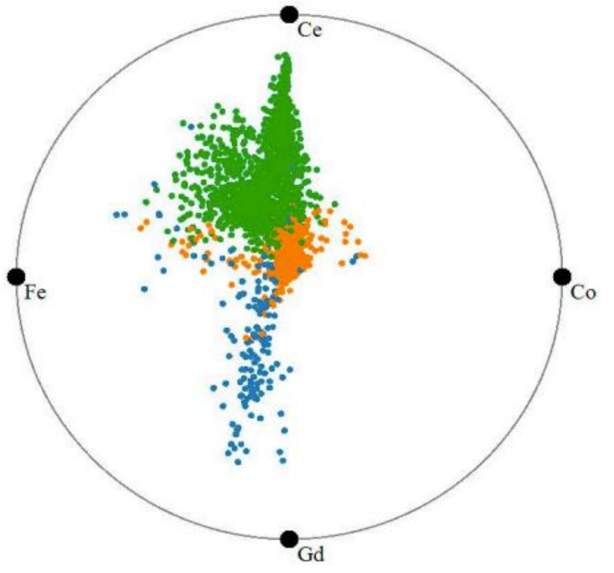


Figure 1. The battery dataset visualized with RadViz.

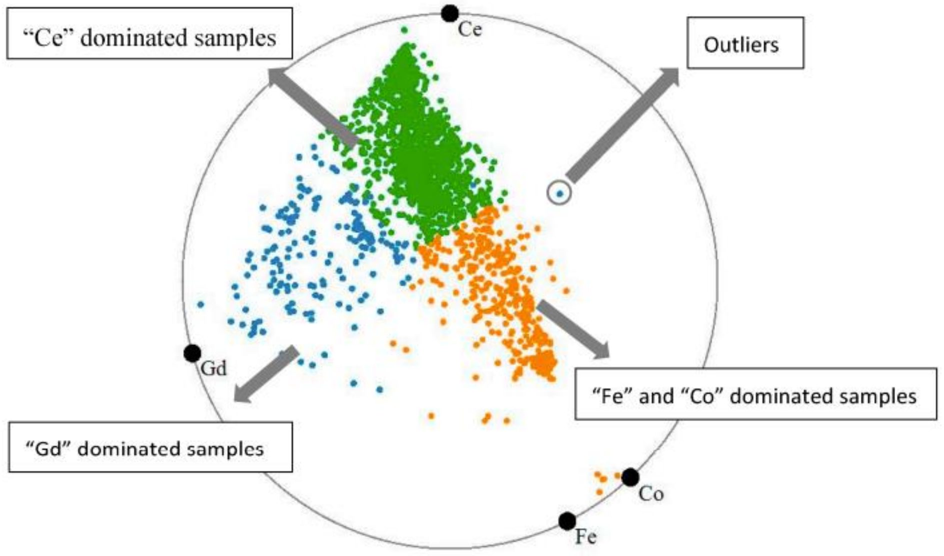


Figure 7. The battery data set visualized with RadViz Deluxe.

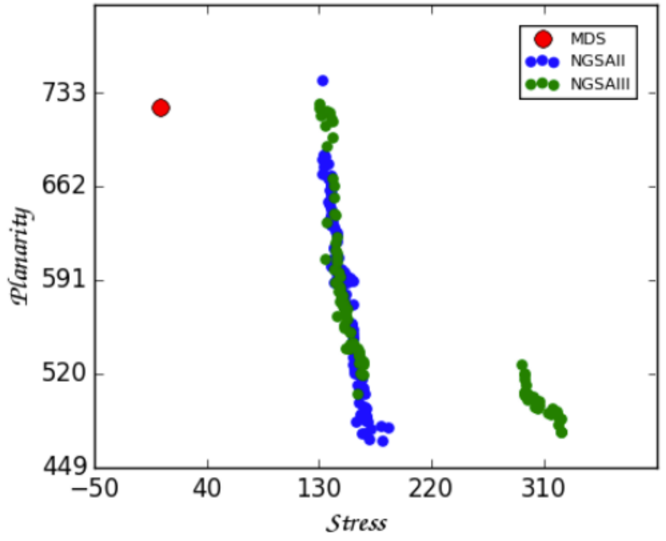


Figure 2: Pareto-Optimal graph representing the comparasation between Metric MDS, NSGA-II and NSGA-III for World Soccer Dataset

- Geek speak, for the most part
- No interesting observations of what we can actually learn from the data
- How does the visualization connect to facts from the data domain?

S. Cheng, W. Xu, K. Mueller "RadViz Deluxe: An Attribute-Aware Display for Multivariate Data" Processes, 2017
B. Ghai, A. Mishra, K. Mueller "Visualization of Multivariate Data with Network Constraints using Multi-Objective Optimization" VIS 2017

People are Very Interested in Reading Captions

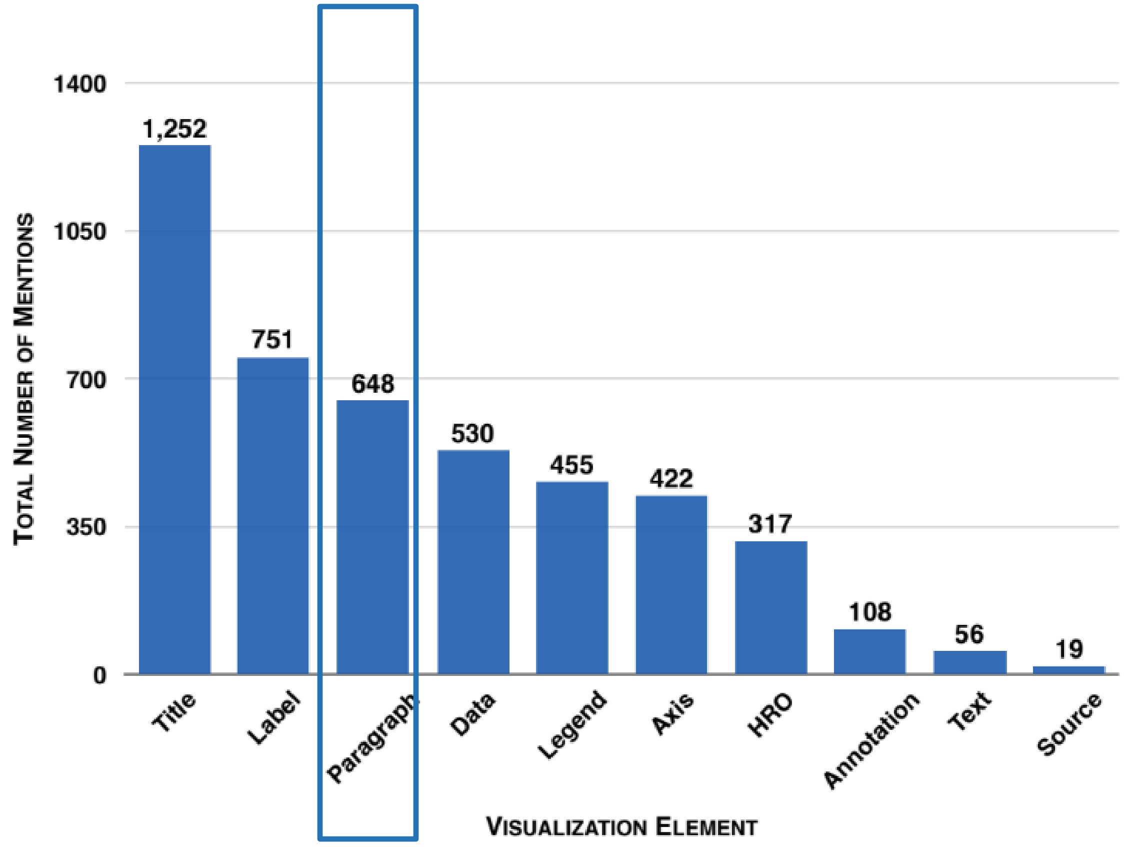
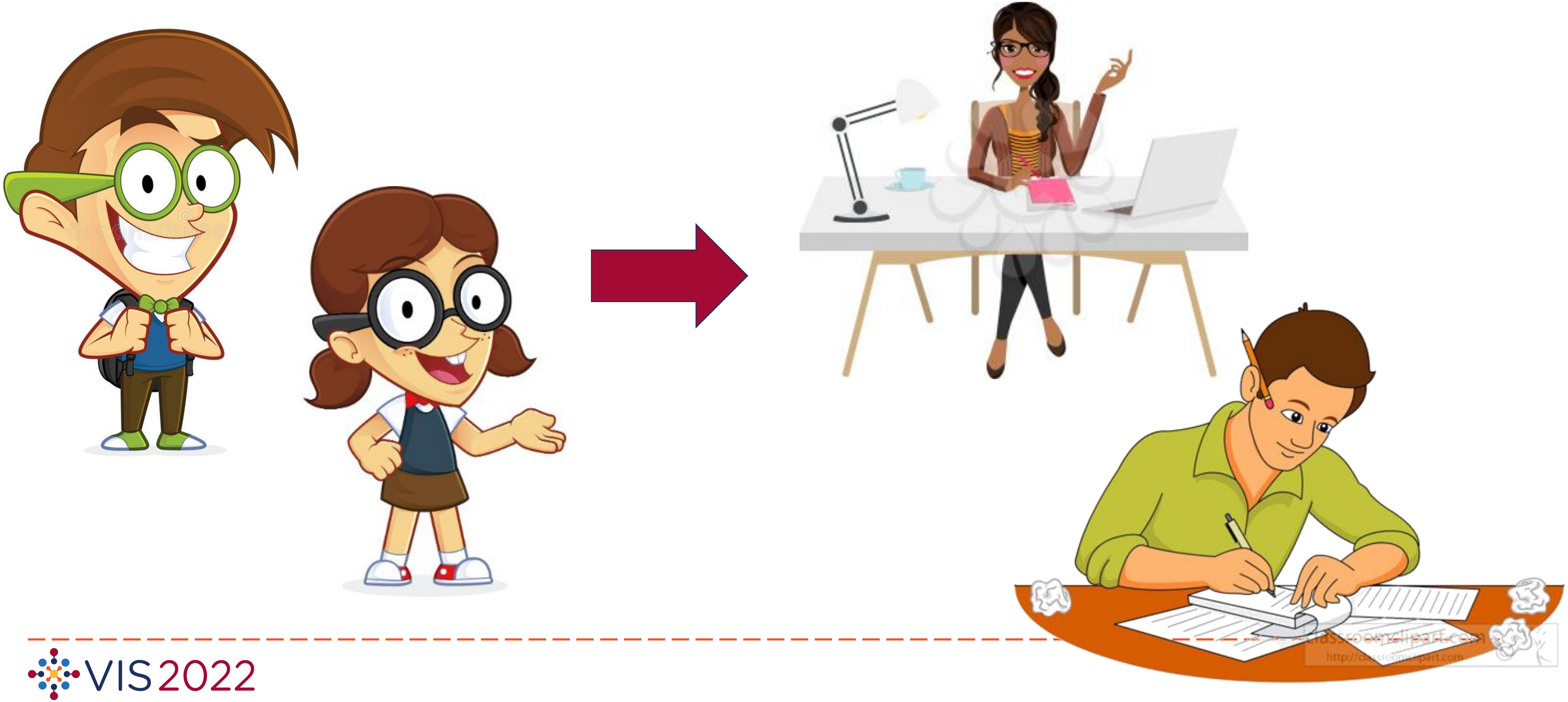


Fig. 8. The total number of mentions of specific visualization elements in the participant-generated recall text descriptions. Textual elements received the most mentions overall, and specially the title received the most mentions across all visualizations.

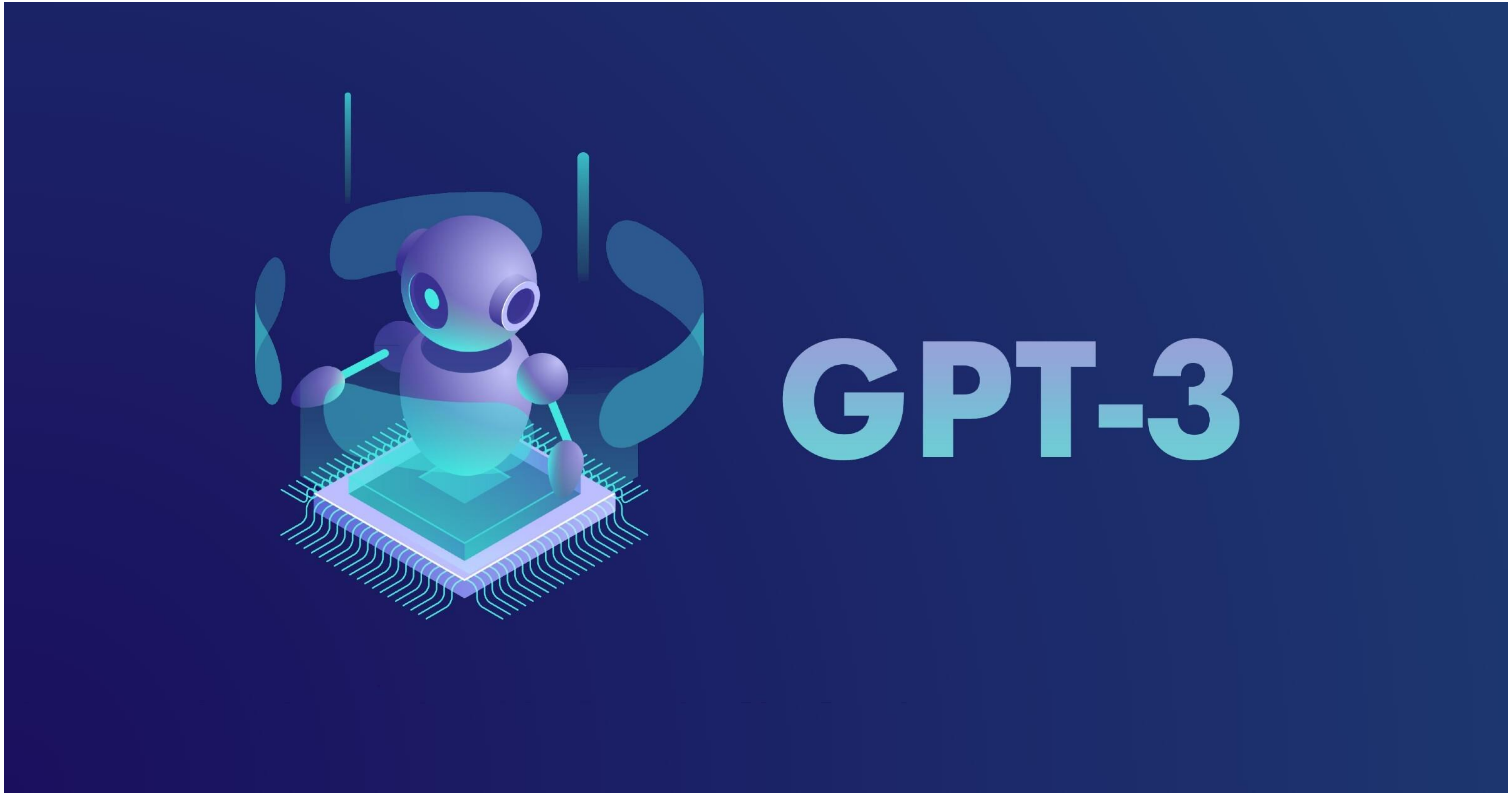
- Paragraph is the 3rd most mentioned element of a visualization
- After title and label, but before data!!

- See Borkin et al. "Beyond memorability: Visualization recognition and recall." IEEE Transactions on Visualization and Computer Graphics, 2015

So How Can Geeks Like Us Become Cool Journalists?



Large Language Models to The Rescue



GPT-3 – 3rd Gen Generative Pre-trained Transformer

- A language prediction model – think auto-complete on steroids

- Works in a ”next-token” generative process
- After submitting the input prompt, GPT-3 tokenizes it and calculates the probabilities of candidate tokens that would continue off the last token

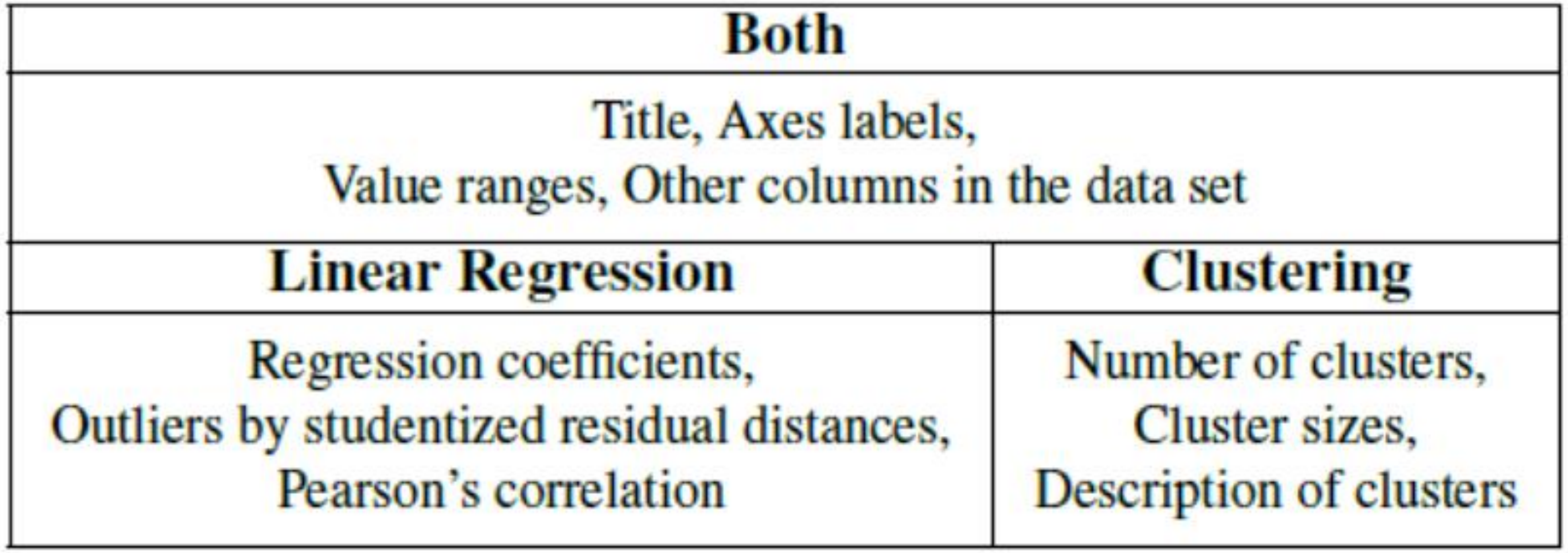
- Number of tokens includes input prompt and determines cost and time
- Hence a prompt must include just enough detail to get the desired text
- This is called prompt engineering

Own Initial Experiments

- We found that GPT-3 does not fare well with numeric tasks, such as math operations and aggregation of given data points

- Hence, the prompt will need to contain post-analysis measurements

- Information provided



Both	
Title, Axes labels, Value ranges, Other columns in the data set	
Linear Regression	**Clustering**
Regression coefficients, Outliers by studentized residual distances, Pearson's correlation	Number of clusters, Cluster sizes, Description of clusters

Developed Three Tiers of Prompts

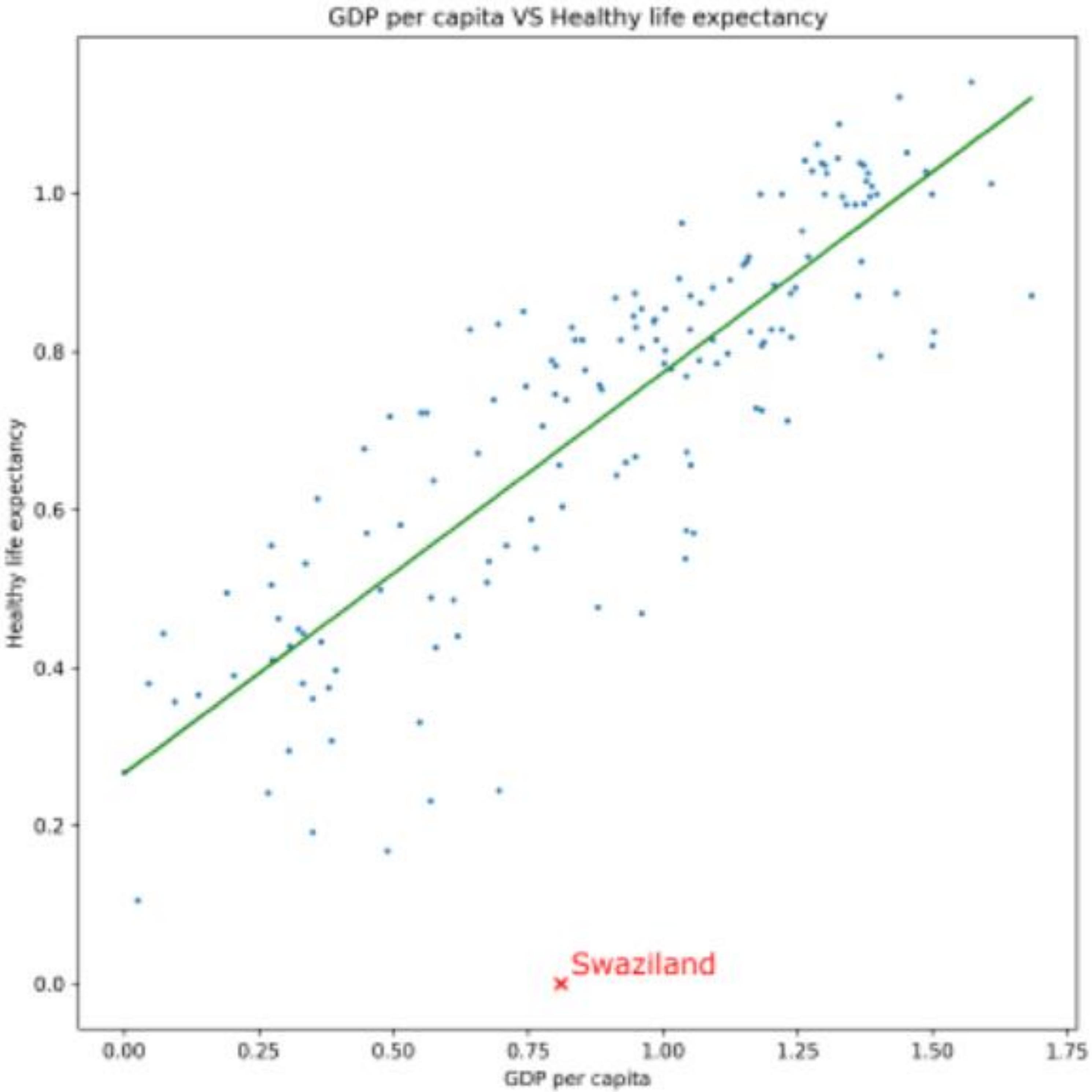
- Tier 1: template only (with metadata) – no human effort needed

- Tier 2: template + instruction (sentence describing specific talking points) – some human effort needed

- Tier 3: template + instruction + Q&A (user-stated modifications, questions, and clarifications) – considerable human effort needed

Example

- Visualization to be captioned
- GDP per capita vs. healthy life expectancy



Tier 1 – Prompt

(Template auto-filled with measurements and all data variables for context)

Generate an engaging caption for a scatter plot titled GDP per capita VS Healthy life expectancy with the x-axis labeled as GDP per capita and the y-axis labeled as Healthy life expectancy. Other columns from the data set include Social support, Perceptions of corruption, Generosity, Overall rank, Score, Country or region, and Freedom to make life choices. The range of GDP per capita is 0.0 to 1.684. The range of Healthy life expectancy is 0.0 to 1.141. The linear regression intercept is 0.27 and the slope is 0.51. The correlation coefficient is 0.84. Outliers found are Swaziland which had a lower healthy life expectancy than would be expected of its GDP per capita.

Tier 1 – Generated Caption

The higher the GDP per capita, the higher the healthy life expectancy!

(picked up on the positive correlation, somewhat witty, but good start)

Tier 2 – Prompt

(Prompt from before, plus the following text)

….The caption should include information explaining causes of the large positive correlation and why there is an outlier in detail.

Tier 2 – Generated Caption

There is a strong positive correlation between GDP per capita and Healthy life expectancy. A country's GDP per capita is indicative of the average income of its citizens and the overall wealth of the country. A higher GDP per capita generally means that citizens have more disposable income, which can be used to purchase goods and services that improve their health and wellbeing. The outlier in this data is Swaziland, which has a lower healthy life expectancy than would be expected of its GDP per capita. This is likely due to the high prevalence of HIV/AIDS in the country, which has a significant impact on the health of the population.

(now the bot turned into a seasoned data journalist!)

Tier 3 – Prompt

(A few more added prompt extensions, each met with a response)

Are there any other reasons why Swaziland has a lower healthy life expectancy?

Why does Swaziland have poor sanitation?

What is the reason for Swaziland’s poor nutrition?

Tier 3 – Final Generated Caption

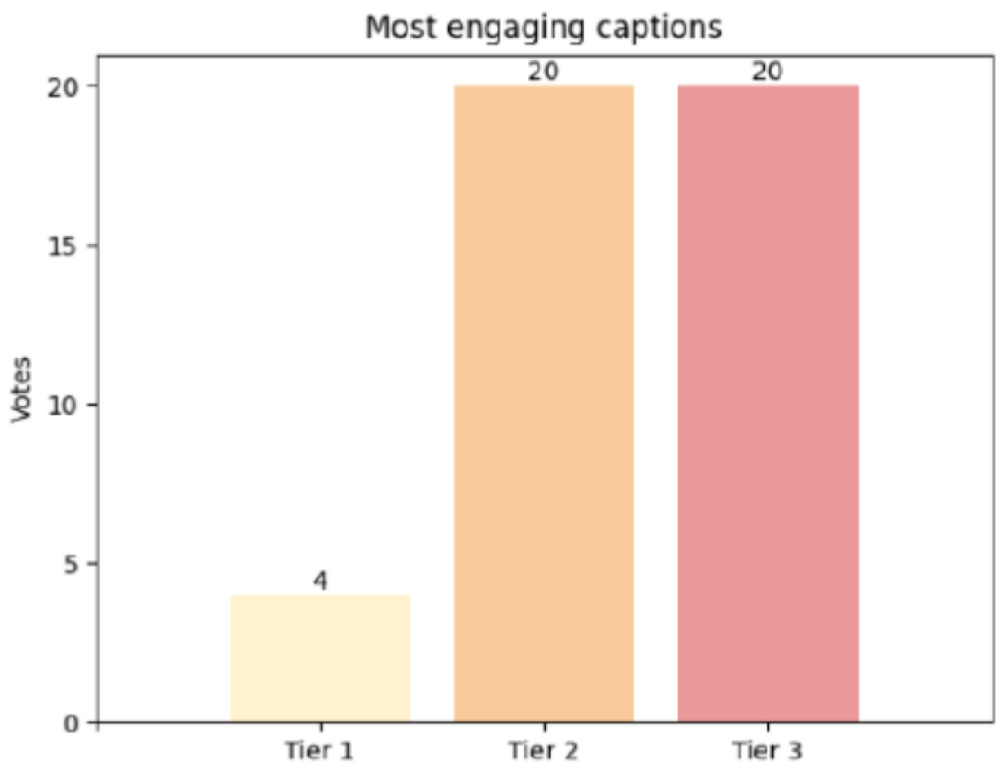
There is a strong positive correlation between GDP per capita and Healthy life expectancy. A country's GDP per capita is indicative of the average income of its citizens and the overall wealth of the country. A higher GDP per capita generally means that citizens have more disposable income, which can be used to purchase goods and services that improve their health and wellbeing. The outlier in this data is Swaziland, which has a lower healthy life expectancy than would be expected of its GDP per capita. This is likely due to the high prevalence of HIV/AIDS in the country, as well as other factors such as poor access to healthcare, sanitation, and nutrition. Swaziland's poor nutrition is likely due to a combination of factors, including poverty, food insecurity, and a lack of access to nutritious foods.

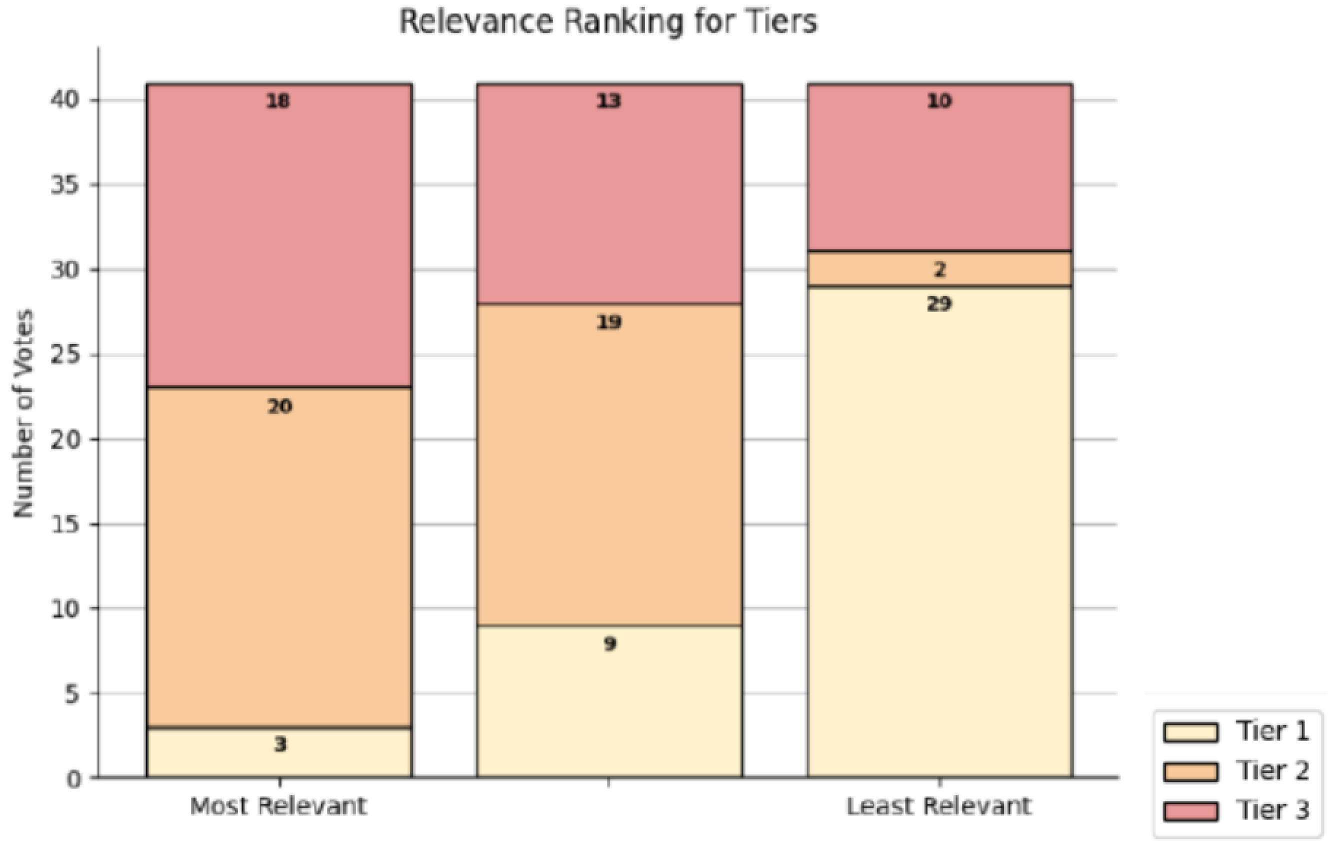
Observations

- Setting Temperature parameter to 1 can create more diverse responses
- Sometimes the bot returns confused responses, like claiming that size is the same even though the data attribute area increases
- Or the bot will not elaborate more even when prompted to do so
- See paper for two more insightful examples

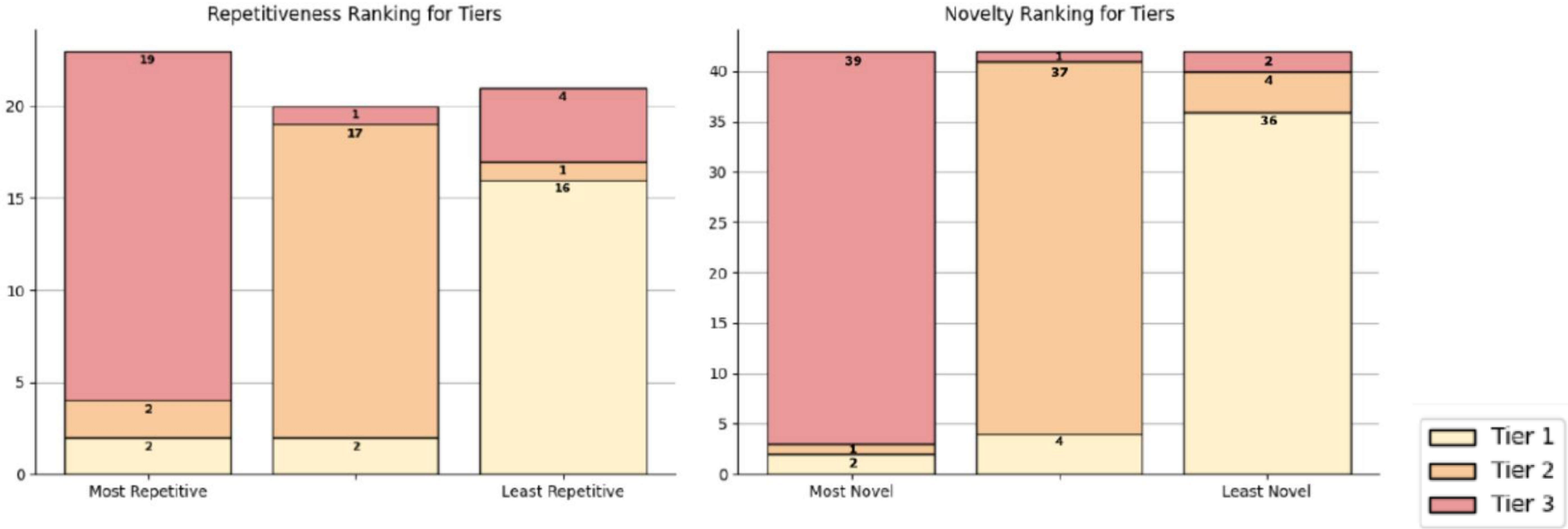
User Study

- 11 participants were shown 4 visualizations in turn and all 3 tier captions
- They were asked to rank them





User Study

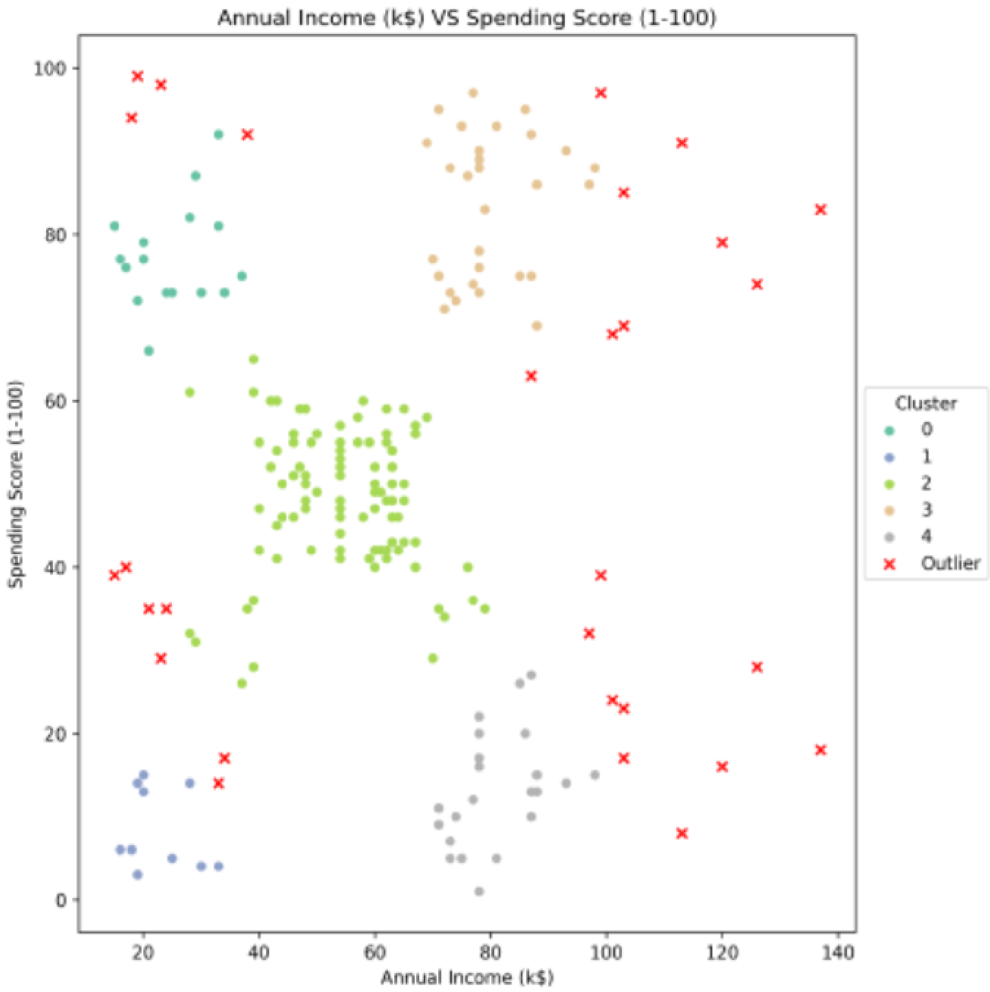

Repetitiveness Ranking for Tiers
0
5
10
15
20
Most Repetitive
Least Repetitive
19
2
2
1
17
2
4
1
16
Novelty Ranking for Tiers
0
5
10
15
20
25
30
35
40
Most Novel
Least Novel
39
1
2
1
37
4
2
4
36
Tier 1
Tier 2
Tier 3

Funding: Data + Computing = Discovery! REU site, sponsored by the NSF under grant 1950052.

Questions?

(Tier 3 caption for visualization on the right)

The clusters in this scatter plot represent different spending habits of customers based on their annual income. The largest cluster, in the middle of the plot, represents customers with average incomes and average spending scores. The other clusters represent customers with low incomes and low spending scores, low incomes and high spending scores, high incomes and low spending scores, and high incomes and high spending scores. The cause of the clusters is likely due to the fact that customers with higher incomes tend to spend more money than customers with lower incomes. However, there are some outliers in the data, such as the cluster of customers with low incomes and high spending scores, or the cluster of customers with high incomes and low spending scores. these outliers could be due to a variety of factors, such as customers with high incomes who are thrifty, or customers with low incomes who make impulsive purchases.


Annual Income (k$) VS Spending Score (1-100)
Spending Score (1-100)
Annual Income (k$)
Cluster
Outlier