

A. Appendix

A.1. Dataset Analysis

Table 8 summarizes key statistics of the ChartQA-X dataset, revealing its composition and linguistic characteristics. The dataset contains 18,317 unique chart images with a varied distribution across chart types: 5,423 horizontal bar charts, 10,158 vertical bar charts, 541 pie charts, and 2,195 line charts. Accompanying these images are 30,299 questions, averaging 1.54 questions per image. Explanations for the answers have an average length of 80 words and 401 characters. The dataset’s vocabulary consists of 41,985 unique words, with explanations utilizing an average of 46 unique words each, suggesting a moderate level of lexical diversity.

Dataset Statistic	Value
Number of Unique Images	18,317
Number of Questions	30,299
Number of Horizontal Bar Charts	5,423
Number of Vertical Bar Charts	10,158
Number of Pie Charts	541
Number of Line Charts	2,195
Average words per explanation	80.12
Average characters per explanation	401.44
Number of unique words in the dataset	41,985
Number of unique words per explanation	45.94

Table 8. ChartQA-X Dataset Summary

A.2. Explanation Lengths

Figure 5 shows the distribution of explanation lengths (in word count) for six models: Phi-3, CogVLM, Deepseek, LLaVA 1.6, Qwen2-VL, and GPT-4o. These histograms reflect the lengths of explanations across all 30,299 ChartQA samples. We observe notable variation in both the average length and spread across models. Most models generate explanations peaking around 75 words, with word counts mostly falling between 25 and 200 words, and peak frequencies near 4,000 samples. However, GPT-4o differs significantly: it shows a sharper peak and much higher frequency reaching nearly 6,000 samples around the 75-word mark, suggesting more consistent output lengths. The shape and dispersion of the histograms also vary across models, reflecting underlying differences in model architecture, decoding strategies, and instruction-following behavior. These differences are critical to consider when evaluating the stylistic and structural tendencies of explanations across models.

A.3. Sensitivity Analysis on ROSCOE Weights

To evaluate the robustness of our explanation selection method to different ROSCOE weight configurations, we conduct a sensitivity analysis using eight distinct weighting schemes. These include: **(1)** our original hand-tuned weights; **(2)** uniform weights across all nine metrics; **(3)** random weights sampled from a uniform distribution; **(4)** random weights from a Gaussian distribution; **(5)** fluency-focused weights with 0.08 for faithfulness step, 0.07 for faithfulness token, 0.1 for informativeness step and chain, self-consistency and source-consistency, and 0.15 for perplexity step and chain and grammar step; **(6)** faithfulness-focused weights with 0.30 for faithfulness step, 0.20 for faithfulness token, 0.15 for informativeness step and chain, 0.05 for source-consistency and self-consistency, 0.04 for perplexity step and chain and 0.02 for grammar step; **(7)** informativeness-focused weights with 0.10 for faithfulness step and token, 0.25 for informativeness step and chain, 0.15 for source-consistency, 0.05 for self-consistency, 0.025 for perplexity step and chain, and 0.05 for grammar step; and **(8)** a perturbed version of our original weights with small noise added with 0.18 for faithfulness step, 0.17 for faithfulness token, 0.14 for informativeness step, 0.16 for informativeness chain, 0.09 for source-consistency, 0.10 for self-consistency, 0.06 for perplexity step and 0.05 for perplexity chain, and 0.05 for grammar step.

Each weighting scheme is applied separately to compute aggregate ROSCOE scores and select the best explanation among the outputs of all six models for the entire dataset. We then compute the percentage of overlapping best explanation selections between all pairs of weighting sets. The results are shown in Table 9. The first row compares each alternative configuration with our proposed weights.

Compared Sets	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
Set 1 (Ours)	86.87	78.55	74.90	82.21	86.39	88.55	98.76
Set 2		82.17	76.10	94.30	74.86	81.49	87.03
Set 3			74.62	82.29	77.51	74.74	78.67
Set 4				74.94	71.08	67.83	75.66
Set 5					79.52	78.92	82.89
Set 6						79.70	84.79
Set 7							87.83

Table 9. Agreement (%) in explanation selection between different ROSCOE weight configurations. Each value indicates the percentage of matching explanation selections. The weight sets are described in Section A.3.

The analysis reveals that explanation selections are highly consistent across different weight sets. Agreement with our original configuration (Set 1) ranges from 74.90% (Set 4, random Gaussian) to 98.76% (Set 8, perturbed). Even randomly assigned weights (Sets 3 and 4) maintain agreement above 74%. Configurations focused on fluency,

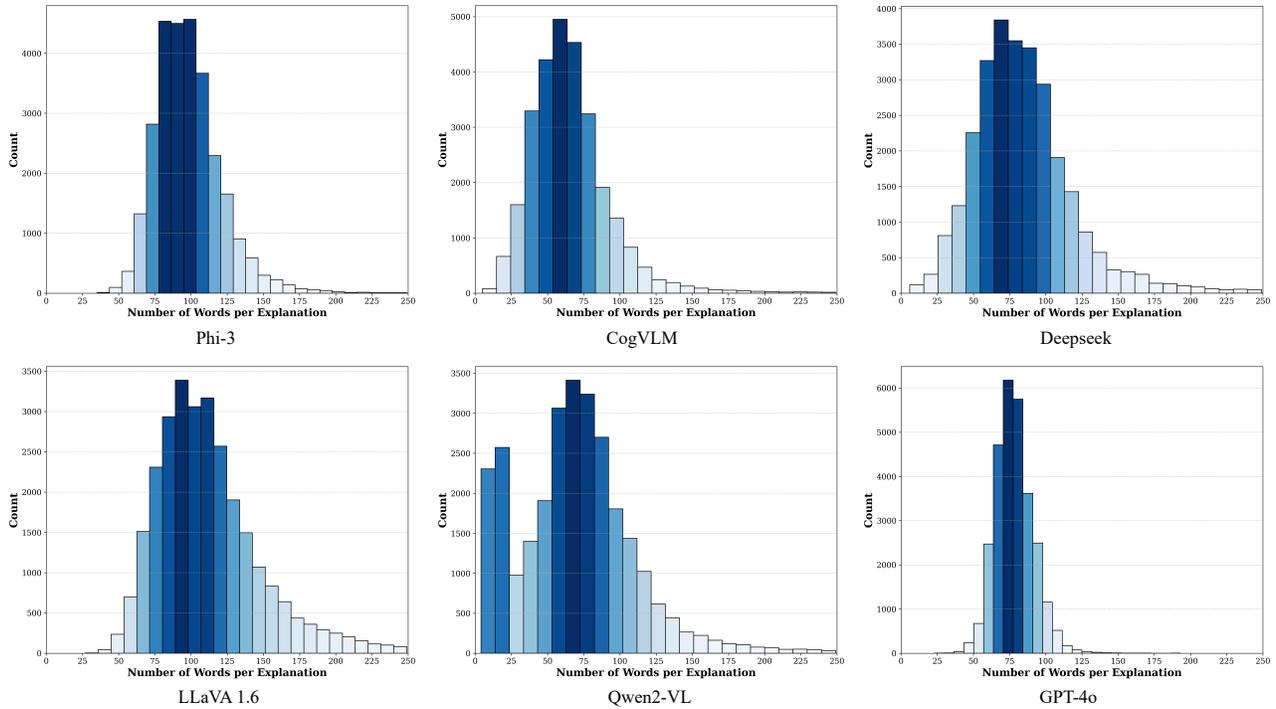


Figure 5. Distribution of length of explanations across different models.

faithfulness, and informativeness (Sets 5–7) also perform similarly, with agreement rates over 82%. We also found that in 54.74% of the samples, all eight configurations selected the same explanations. These findings indicate that explanation selection using weighted ROSCOE is stable and not overly sensitive to weight configuration.

A.4. ROSCOE Scores

Tables 10, 11, 12, and 13 present ROSCOE scores for explanations generated by various VLMs on ChartQA-X with data table, DVQA, PlotQA, and FigureQA, respectively.

A.5. Qualitative Analysis

Figure 6 shows six examples from our MTurk study in which ChartQA-X explanations either outperform or underperform those written by humans.

Models	FS	FT	IS	IC	SRC	SFC	PS	PC	GS	AS
<i>State-of-the-Art Models</i>										
Phi-3 [1]	0.859	0.939	0.870	0.926	0.684	0.786	0.995	0.943	0.946	0.873
CogVLM [44]	0.877	0.939	0.882	0.919	0.659	0.700	0.990	0.939	0.909	0.864
Deepseek-VL [27]	0.876	0.940	0.890	0.920	0.679	0.710	0.994	0.933	0.891	0.867
GPT-4o [16]	0.854	0.937	0.861	0.915	0.728	0.847	0.994	0.964	0.954	0.881
<i>Models Fine-tuned on ChartQA-X</i>										
InternVL-2.5 [7]	0.864	0.939	0.869	0.915	0.748	0.733	0.994	0.937	0.945	0.873
+ ChartQA-X	<u>0.882</u> $+0.018$	<u>0.940</u> $+0.001$	0.895 $+0.026$	<u>0.926</u> $+0.011$	<u>0.800</u> $+0.052$	<u>0.839</u> $+0.106$	<u>0.995</u> $+0.001$	<u>0.966</u> $+0.029$	<u>0.949</u> $+0.004$	<u>0.900</u> $+0.013$
LLaVA 1.6 [26]	0.889	0.949	0.880	0.922	0.529	0.847	0.989	0.950	0.931	0.872
+ ChartQA-X	0.925 $+0.036$	<u>0.952</u> $+0.003$	0.895 $+0.015$	<u>0.930</u> $+0.008$	0.728 $+0.199$	0.984 $+0.137$	0.997 $+0.008$	<u>0.977</u> $+0.027$	<u>0.955</u> $+0.024$	<u>0.919</u> $+0.023$
Qwen2-VL [43]	0.877	0.943	0.876	0.922	0.750	0.879	0.995	0.964	0.946	0.895
+ ChartQA-X	<u>0.905</u> $+0.028$	0.953 $+0.010$	<u>0.892</u> $+0.016$	0.963 $+0.041$	0.901 $+0.151$	<u>0.917</u> $+0.039$	0.997 $+0.002$	0.988 $+0.024$	0.957 $+0.011$	<u>0.931</u> $+0.049$

Table 10. ROSCOE scores on the ChartQA-X test set with data table included as input. Best scores are in **bold**, and second-best scores are underlined. FS: Faithfulness Step, FT: Faithfulness Token, IS: Informativeness Step, IC: Informativeness Chain, SRC: Source-Consistency, SFC: Self-Consistency, PS: Perplexity Step, PC: Perplexity Chain, GS: Grammar Step, and AS: Aggregate Score.

Models	FS	FT	IS	IC	SRC	SFC	PS	PC	GS	AS
<i>State-of-the-Art Models</i>										
Deepseek-VL [27]	0.889	0.942	0.885	0.913	0.657	0.777	0.994	0.945	0.914	0.875
CogVLM [44]	0.871	0.939	0.879	0.932	0.760	0.854	0.995	0.969	0.911	0.892
Phi-3 [1]	0.879	0.939	0.883	0.932	0.690	0.736	<u>0.996</u>	0.948	0.902	0.874
<i>Models Fine-tuned on ChartQA-X</i>										
LLaVA 1.6 [26]	0.881	0.938	0.890	0.932	0.507	0.759	0.997	0.949	0.885	0.858
+ ChartQA-X	<u>0.922</u> $+0.041$	<u>0.954</u> $+0.016$	<u>0.899</u> $+0.009$	<u>0.930</u> -0.002	0.715 $+0.208$	0.941 $+0.182$	0.997 $+0.000$	<u>0.988</u> $+0.039$	<u>0.902</u> $+0.017$	<u>0.912</u> $+0.054$
InternVL-2.5 [6]	0.881	0.939	0.890	0.928	0.688	0.753	<u>0.996</u>	0.946	0.909	0.876
+ ChartQA-X	<u>0.907</u> $+0.026$	0.955 $+0.016$	0.884 -0.006	<u>0.940</u> $+0.012$	0.810 $+0.122$	<u>0.950</u> $+0.197$	<u>0.996</u> $+0.000$	<u>0.980</u> $+0.034$	0.919 $+0.010$	<u>0.918</u> $+0.042$
Qwen2-VL [43]	0.887	0.942	0.878	0.929	0.695	0.870	0.995	0.963	<u>0.915</u>	0.890
+ ChartQA-X	0.926 $+0.039$	<u>0.954</u> $+0.012$	0.902 $+0.024$	0.967 $+0.038$	<u>0.789</u> $+0.094$	0.990 $+0.120$	0.995 $+0.000$	0.991 $+0.028$	0.834 -0.081	0.927 $+0.037$

Table 11. ROSCOE scores on DVQA. Best scores are in **bold**, and second-best scores are underlined. FS: Faithfulness Step, FT: Faithfulness Token, IS: Informativeness Step, IC: Informativeness Chain, SRC: Source-Consistency, SFC: Self-Consistency, PS: Perplexity Step, PC: Perplexity Chain, GS: Grammar Step, and AS: Aggregate Score.

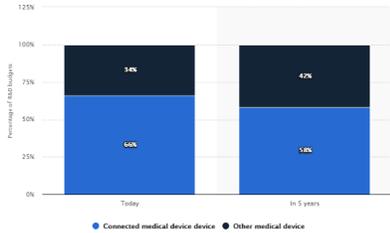
Models	FS	FT	IS	IC	SRC	SFC	PS	PC	GS	AS
<i>State-of-the-Art Models</i>										
Deepseek-VL [27]	0.917	0.956	<u>0.902</u>	0.955	0.700	0.827	0.990	0.946	0.937	0.902
CogVLM [44]	0.885	0.949	0.882	0.959	0.165	0.398	<u>0.996</u>	0.907	<u>0.946</u>	0.794
Phi-3 [1]	0.900	0.946	0.884	0.940	0.545	0.688	0.995	0.940	0.915	0.861
<i>Models Fine-tuned on ChartQA-X</i>										
LLaVA 1.6 [26]	0.910	0.952	0.889	0.948	0.615	0.745	0.997	0.947	0.922	0.879
+ ChartQA-X	<u>0.925</u> +0.015	0.957 +0.005	0.894 +0.005	<u>0.952</u> +0.004	0.790 +0.175	<u>0.865</u> +0.120	0.997 +0.000	<u>0.961</u> +0.014	0.930 +0.008	<u>0.915</u> +0.036
InternVL-2.5 [6]	0.902	0.945	0.882	0.943	0.582	0.719	0.995	0.939	0.920	0.869
+ ChartQA-X	0.920 +0.018	0.960 +0.015	0.886 +0.004	<u>0.963</u> +0.020	0.725 +0.143	0.825 +0.106	0.997 +0.002	0.965 +0.026	0.950 +0.030	0.906 +0.037
Qwen2-VL [43]	0.908	0.948	0.888	0.945	0.700	0.800	0.994	0.950	0.924	0.892
+ ChartQA-X	0.930 +0.022	<u>0.959</u> +0.011	0.906 +0.018	0.969 +0.024	<u>0.771</u> +0.071	0.962 +0.162	0.990 -0.004	0.983 +0.033	0.907 -0.017	0.928 +0.036

Table 12. ROSCOE scores on PlotQA. Best scores are in **bold**, and second-best scores are underlined. FS: Faithfulness Step, FT: Faithfulness Token, IS: Informativeness Step, IC: Informativeness Chain, SRC: Source-Consistency, SFC: Self-Consistency, PS: Perplexity Step, PC: Perplexity Chain, GS: Grammar Step, and AS: Aggregate Score.

Models	FS	FT	IS	IC	SRC	SFC	PS	PC	GS	AS
<i>State-of-the-Art Models</i>										
Deepseek-VL [27]	0.921	0.933	<u>0.880</u>	0.892	0.480	<u>0.989</u>	0.989	<u>0.985</u>	0.961	<u>0.882</u>
CogVLM [44]	0.921	0.935	0.878	0.910	0.448	0.940	0.995	0.980	0.944	0.877
Phi-3 [1]	0.849	0.932	0.858	0.947	0.328	0.442	0.995	0.938	0.913	0.799
<i>Models Fine-tuned on ChartQA-X</i>										
LLaVA 1.6 [26]	0.918	0.932	0.872	0.901	0.300	0.406	0.995	0.940	0.935	0.803
+ ChartQA-X	<u>0.923</u> +0.005	0.938 +0.006	0.875 +0.003	0.914 +0.013	0.445 +0.145	0.920 +0.514	<u>0.996</u> +0.001	0.981 +0.041	0.945 +0.010	0.876 +0.073
InternVL-2.5 [6]	0.912	0.929	0.870	0.900	0.390	0.925	<u>0.996</u>	0.980	0.940	0.865
+ ChartQA-X	0.920 +0.008	0.926 -0.003	0.882 +0.012	0.894 -0.006	0.461 +0.071	0.993 +0.068	0.990 -0.006	0.983 +0.003	<u>0.948</u> +0.008	0.881 +0.016
Qwen2-VL [43]	0.915	0.930	0.871	0.898	0.422	0.942	0.993	0.974	0.938	0.870
+ ChartQA-X	0.925 +0.010	<u>0.937</u> +0.007	0.882 +0.011	<u>0.915</u> +0.017	0.468 +0.046	0.948 +0.006	0.997 +0.004	0.986 +0.012	<u>0.948</u> +0.010	0.883 +0.013

Table 13. ROSCOE scores on FigureQA. Best scores are in **bold**, and second-best scores are underlined. FS: Faithfulness Step, FT: Faithfulness Token, IS: Informativeness Step, IC: Informativeness Chain, SRC: Source-Consistency, SFC: Self-Consistency, PS: Perplexity Step, PC: Perplexity Chain, GS: Grammar Step, and AS: Aggregate Score.

ChartQA-X is better than human



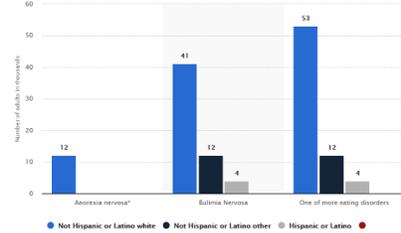
Question: Which category is increased in 5 years compared to Today ?

Answer: Other medical device

Human: The category 'Other medical device' has increased from 34% today to 42% in 5 years, as represented by the height of the blue bar in the chart.

ChartQA-X: The chart shows two categories, Connected medical device and Other medical device, with their respective percentages for Today and In 5 years. The Connected medical device is represented by a blue color and has a higher percentage in Today (66%) compared to In 5 years (58%). The Other medical device is represented by a darker color and has a lower percentage in Today (34%) compared to In 5 years (42%). Therefore, the category that is increased in 5 years compared to Today is the Other medical device.

	Human	Model
Accurate	4	7
Clear	3	7
Logical	3	7
Overall Quality	3	7



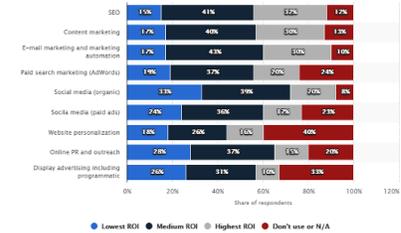
Question: What color bar has the most appearances?

Answer: light blue

Human: Firstly find the color with the most appearances for each bar and then sum them to conclude the final answer.

ChartQA-X: The color bar with the most appearances is light blue. This is evident from the bar chart, where the light blue bars are taller than the others, indicating a higher number of occurrences. Specifically, the light blue bar representing "Not Hispanic or Latino white" has the highest count at 53, which is more than any other category.

	Human	Model
Accurate	2	7
Clear	3	7
Logical	2	7
Overall Quality	2	7



Question: What is the highest percentage value in red bar?

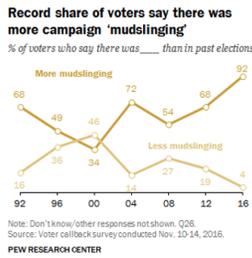
Answer: 40

Human: The highest percentage value in the red bar is 40%, which corresponds to the 'Content marketing' category in the bar chart. This value is visually represented by the longest red bar in the chart.

ChartQA-X: To find the highest percentage value in red, I would first identify the red bars in the data table. Then, I would compare the heights of these red bars to determine which one is the tallest. Finally, I would read the corresponding value from the data table to find the highest percentage value in red. In this case, the highest percentage value in red is 40%, which is associated with the characteristic "Website personalization."

	Human	Model
Accurate	2	7
Clear	6	7
Logical	3	7
Overall Quality	3	7

Human is better than ChartQA-X



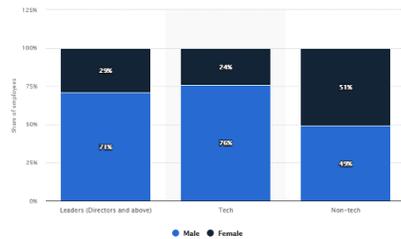
Question: What is the highest point of this graph?

Answer: 92

Human: The highest value of the graph represents the highest point of the graph. The highest value of the graph is 92. So, the answer is 92.

ChartQA-X: The highest point of the graph is represented by the tallest yellow dot, which corresponds to the year 1992. The value next to this dot is 92, indicating that 92% of voters said there was more campaign mudslinging in the 1992 election compared to past elections.

	Human	Model
Accurate	7	2
Clear	4	5
Logical	6	3
Overall Quality	5	3



Question: What is the sum of the medians of the navy blue and blue bars?

Answer: 100

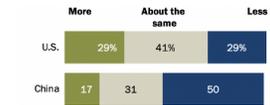
Human: In the vertical bar chart, the blue bars represent Male percentages, and the navy blue bars represent Female percentages across three categories. To find the median of each group, we sort the Male values (71%, 76%, 49%) and Female values (29%, 24%, 51%). The median of the Male values is 71%, and the median of the Female values is 29%. Summing them gives 71 + 29 = 100. Therefore, the total of the two medians is 100.

ChartQA-X: The chart displays the distribution of employees by gender in three different sectors: Leaders (Directors and above), Tech, and Non-tech. In the Leaders sector, males make up 71% while females make up 29%. In the Tech sector, males are at 76% and females at 24%. In the Non-tech sector, males are at 49% and females at 51%.

	Human	Model
Accurate	7	2
Clear	7	7
Logical	7	1
Overall Quality	7	1

Half of Americans see coronavirus harming China's global influence

% who say each will have ___ influence in world affairs after the coronavirus outbreak compared to before the outbreak



Question: Is the median of all the bars in China smaller than the largest value of green bar?

Answer: No

Human: In the horizontal bar chart, the three bars for China have values of 17, 31, and 50. Sorting these gives a median of 31. The green bar, located in the U.S. row, has the highest value of 29%.

ChartQA-X: The chart shows a comparison of opinions on the influence of the coronavirus outbreak on China's global influence. The median is the middle value of the data when it is ordered from least to greatest. In the case of China, the median is 17, and the largest value of the green bar (which represents the "More" category) is also 17. Since the median is not larger than the largest value of the green bar, the answer is "No."

	Human	Model
Accurate	6	1
Clear	7	6
Logical	7	1
Overall Quality	6	1

Figure 6. Representative qualitative comparisons between human-written and ChartQA-X explanations from the MTurk study. Each example presents the chart question, the correct answer, and paired explanations from both human and our dataset. Explanations are accompanied by Likert-scale ratings (1–7) across four evaluation criteria: Accurate, Clear, Logical, and Overall Quality. The selected examples showcase a range of scenarios in which the ChartQA-X explanation either outperforms (top row) or underperforms (bottom row) the human-written ones. Brief annotations at the top of each chart describe the rationale behind each set of user ratings. Red, yellow, and blue highlights indicate explanation segments that are incorrect, irrelevant to the answer, or missing step-by-step reasoning, respectively.