A. Segmentation Annotation

Pies: The coordinates for three vertices of the slices are already provided as keypoints. It remains to approximate the arc of the circle via insertion of intermediate points, roughly 5 per radian. To achieve this, we compute the angles and radii of the two edge vertices with respect to the center, then linearly interpolate these quantities for the intermediate points.

Lines: As the provided keypoints are placed at line centers, no information regarding the thickness of a particular line is available. To address this while accounting for differing image sizes, we set annotation line thickness to 1% of the image height. While this parameter choice produces fairly accurate annotations in most cases, it does not necessarily correspond to the ground-truth line thickness shown in images. The edges of the created polygon are defined by line segments parallel to the ones in the provided line, placed at the line width distance apart from each other. Their endpoints (which are polygon vertices) are the intersection points of two adjacent pairs of lines. To account for acute bends in the line producing elongated 'spikes' in its segmentation mask, the line vertices at such bends are duplicated and shifted a minute distance apart.

Bars and others: The remaining chart components have rectangular annotations, for which the four vertices can be computed directly from the given parameters.

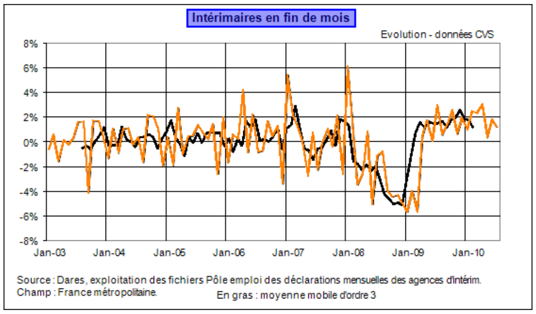
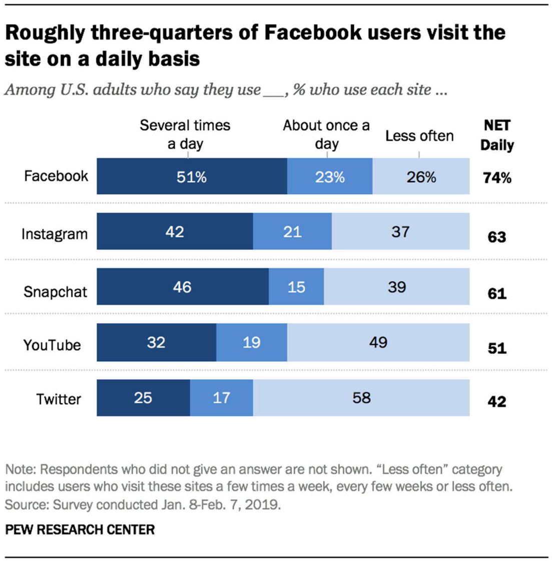


Figure 6. Missing line annotation example; orange annotation correctly labels one line, but the second line misses the point ground truth to be annotated.

Annotation Bias The ground truth annotation significantly impacts performance, often resulting in lower-than-expected performance levels, particularly noticeable in the observed performance discrepancy within the line category. Since each line chart may have varying line thickness, the machine-generated annotations may not accurately reflect the true thickness of the lines. We also observed that incorrect or missing line annotation occurs as shown in Figure 6. As a result, differences arise when the model predicts the actual line thickness. For instance, if a line is annotated with a thickness of 3 pixels and a length of 100 pixels, while the actual width value is 15 pixels, the model may predict the width as 15 pixels and the length as 100 pixels. As a result, the IOU (Intersection over Union) threshold is calculated as $(3 \cdot 100)/(15 \cdot 100) = 0.2$, even if the model's prediction is entirely correct. Similarly, due to labeling noise and the absence of language information incorporated into the model, the effectiveness of detecting **ValueAxisTitle** and **CategoryAxisTitle** categories is reduced.

B. OpenCQA Case Study

We examine an example from the OpenCQA test set, as shown in Figure 7, with the question: Describe the trend of Facebook users' usage of the site on a daily basis.



QDChart: The survey also finds that they use Facebook a site on a daily basis, with 51 % who use their site, while a quarter (25 %) say they use Twitter. The share of Facebook users use the site on a daily basis, with 17 % who say they use the site using a few ways, while a smaller share (26 %) say they use the site early daily basis

VLT5: About three-quarters of Facebook users (71%) say they visit the site several times a day, or several times a day. The roughly one-third (26%) say they visit the site at least once a day, or several times a day.

Gold: Roughly three-quarters of Facebook users (74%) visit the site daily, including about half who do so several times a day.

Figure 7. Texts in blue represent correct facts or values that can be detected from the chart, Texts in red represent wrong values.

C. Computational Cost

We train and evaluate our models using 1 NVIDIA A40. We trained ChartFormer on ExcelChart400K for 3 epochs in 48 hours. We trained QDChart on ChartQA for 20 epochs in 57 hours. We trained QDChart on OpenCQA for 200 epochs in 72 hours.

D. Supplementary materials

The performance gap between the human-written and machine-generated question splits is attributed to the difficulty of the questions. Of the 300 randomly selected human-written questions [28], 43% are classified as ”compositional,” meaning they involve at least two mathematical or logical operations, such as sums, differences, or averages. Current language models often struggle with such mathematical reasoning. In contrast, 86.64% of the machine-generated questions involve simply identifying values or labels from chart elements. As demonstrated in 7, all models show consistent performance differences between the human-written and machine-generated splits.

	Human		Machine	
Split	Charts	Questions	Charts	Questions
Training	3,699	7,398	15,474	20,901
Validation	480	960	680	960
Test	625	1,250	987	1,250
Total	4,804	9,608	17,141	23,111

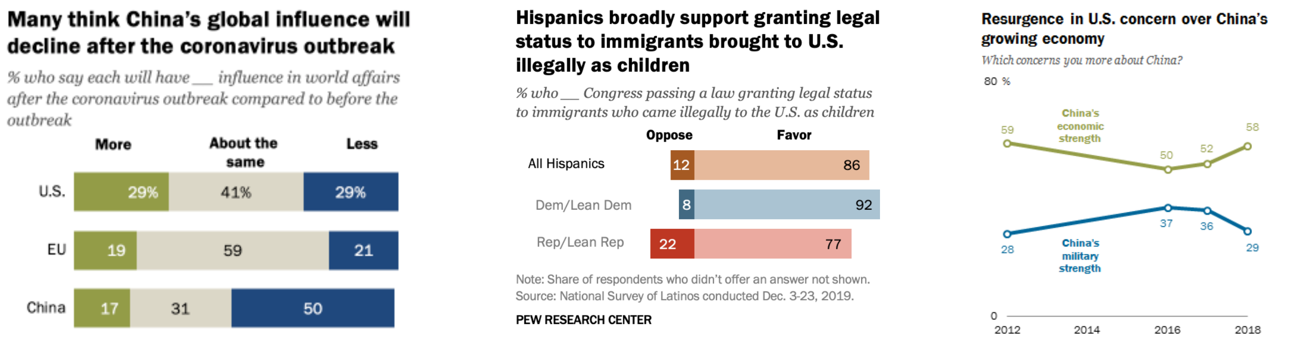
Table 6. Distribution of data in the ChartQA dataset.

QDCHART shows limited performance on questions requiring mathematical reasoning, struggling with operations like averages, medians, products, and sums. Three illustrative examples are provided in Figure 8.

We provide detailed performance comparison examples of all the models employed in chart element detection and classification. As exemplified in Figure 11, the left three models fail to predict nearly every line segment. In complex scenarios such as stacked bar charts and overlapping line charts, CHARTFORMER more accurately detects the correct number of components. Comparison examples are provided in Figures 13 and 12.

Model	**Human**	**Machine**	**Average**
Donut [16]	-	-	41.8
VisionTaPas [28]	29.6	61.4	45.5
TaPas [14]	28.7	53.8	41.3
T5 [34]	25.1	57.0	41.0
VL-T5 [5]	26.2	56.9	41.6
ChartReader [4]	-	-	52.6
Pix2Struct [18]	30.5	81.6	56.0
VL-T5$_{\text{pre}}$ [5]	40.1	63.6	51.8
VisionTaPas$_{\text{pre}}$ [28]	32.5	61.6	47.1
ChartT5 [49]	31.8	74.4	53.2
MatCha [21]	38.2	90.2	64.2
UniChart [29]	43.9	88.6	66.2
PaLI-17B [2]	30.4	64.9	47.6
LLaVA1.5-13B [22]	37.7	73.0	55.3
DEPLOT [20]	91.0	67.6	**79.3**
QDCHART-Donut	34.9	79.4	57.2
QDCHART-UniChart	44.7	88.5	**66.6**

Table 7. Comparison with baseline results on ChartQA.



Question (human): What's the average of all the values in the green bars (round to one decimal)?
Prediction: 17
Answer: 21.6

Question (human): What is the product of the smallest value on the blue bar and the smallest value on the upper bar?
Prediction: 18
Answer: 96

Question (human): What's the sum of median value of blue and green graph?
Prediction: 9
Answer: 87.5

Figure 8. Failure cases on ChartQA

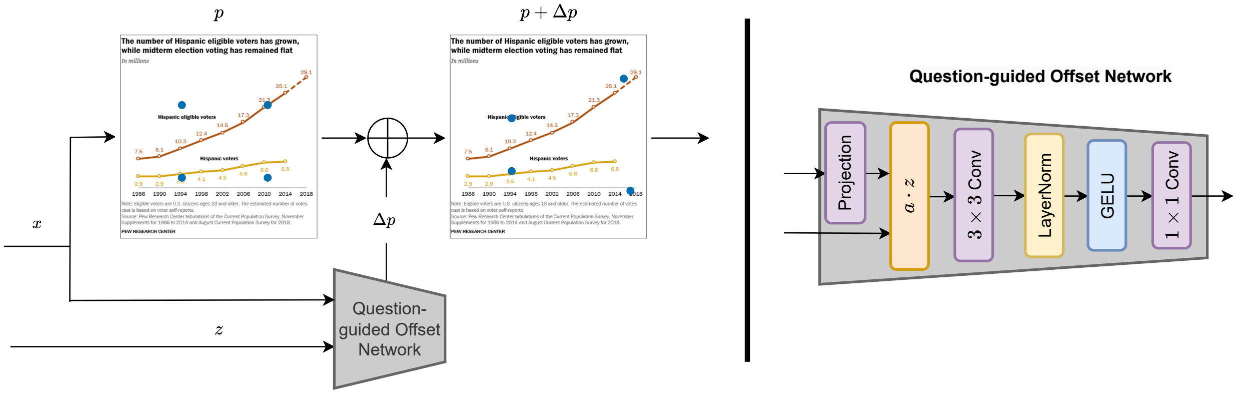
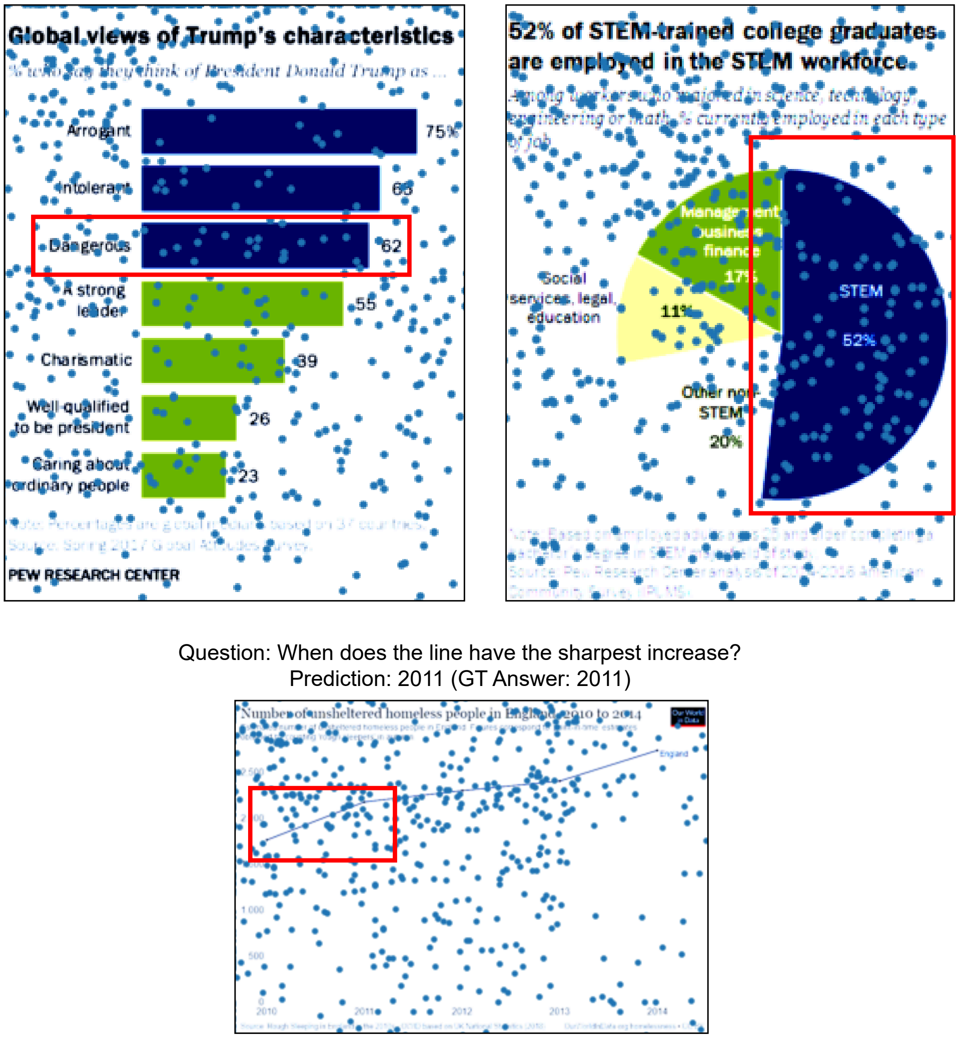


Figure 9. Question-guided Offset Network Flowchart



Question: What percent who think of President Donald Trump as Dangerous?
Prediction: 62 (GT Answer: 62)

Question: Is the largest segment greater than sum of all the other segments?
Prediction: Yes (GT Answer: Yes)

Question: When does the line have the sharpest increase?
Prediction: 2011 (GT Answer: 2011)

Figure 10. Visualization on question-guided deformed points

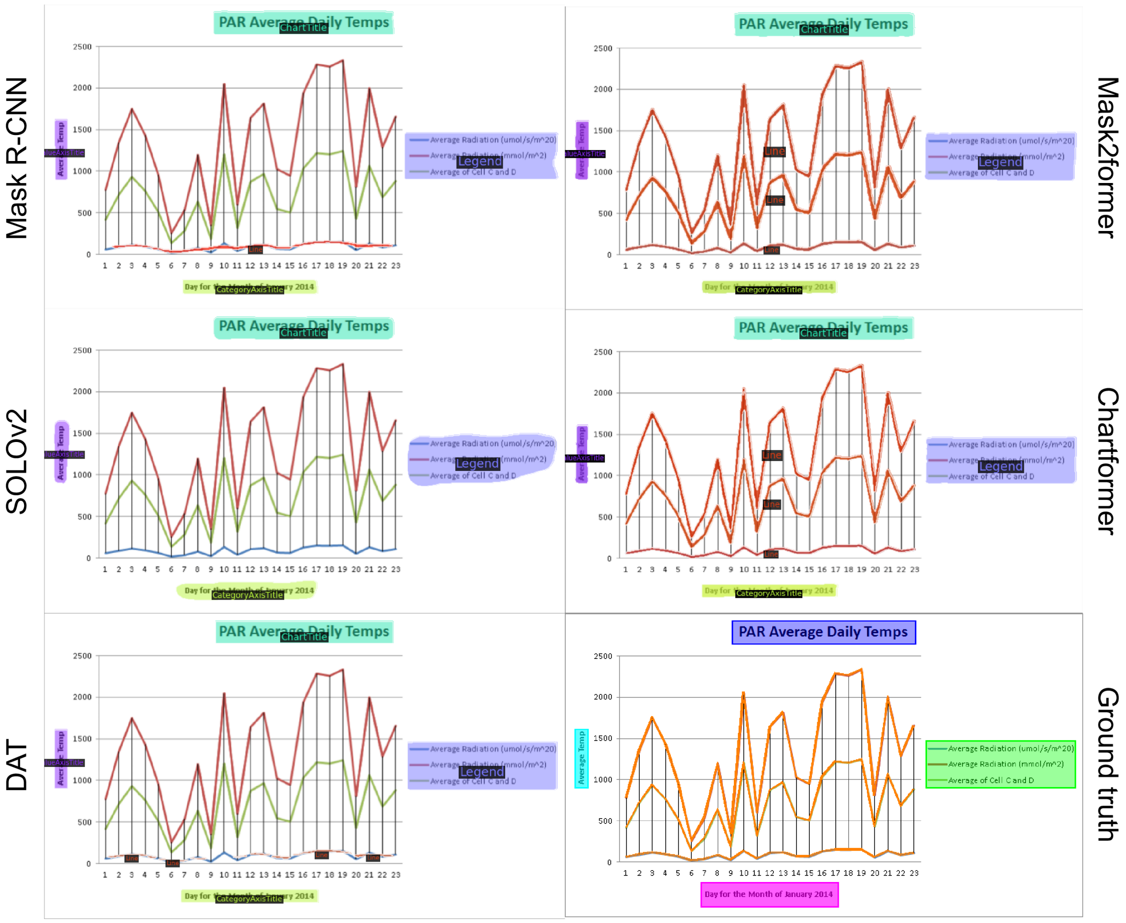


Figure 11. Line performance case study

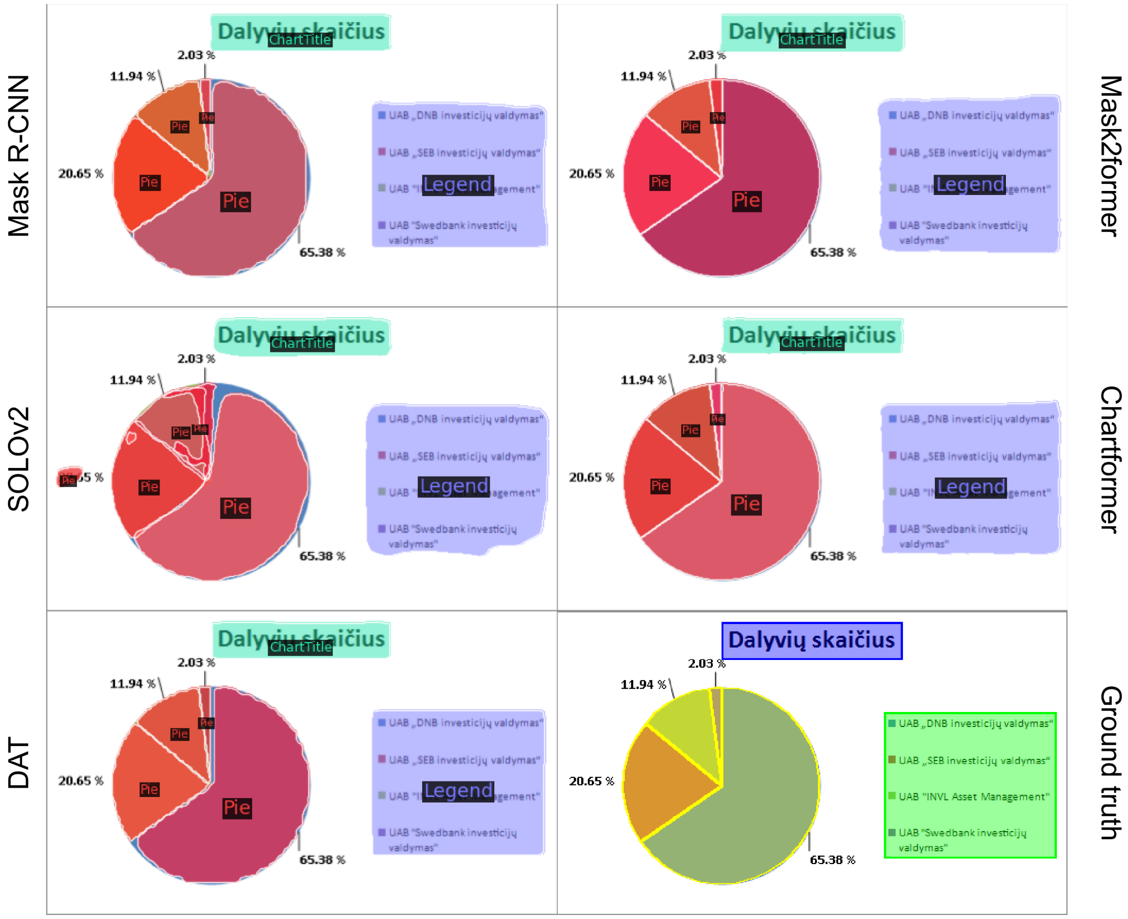


Figure 12. Pie performance case study

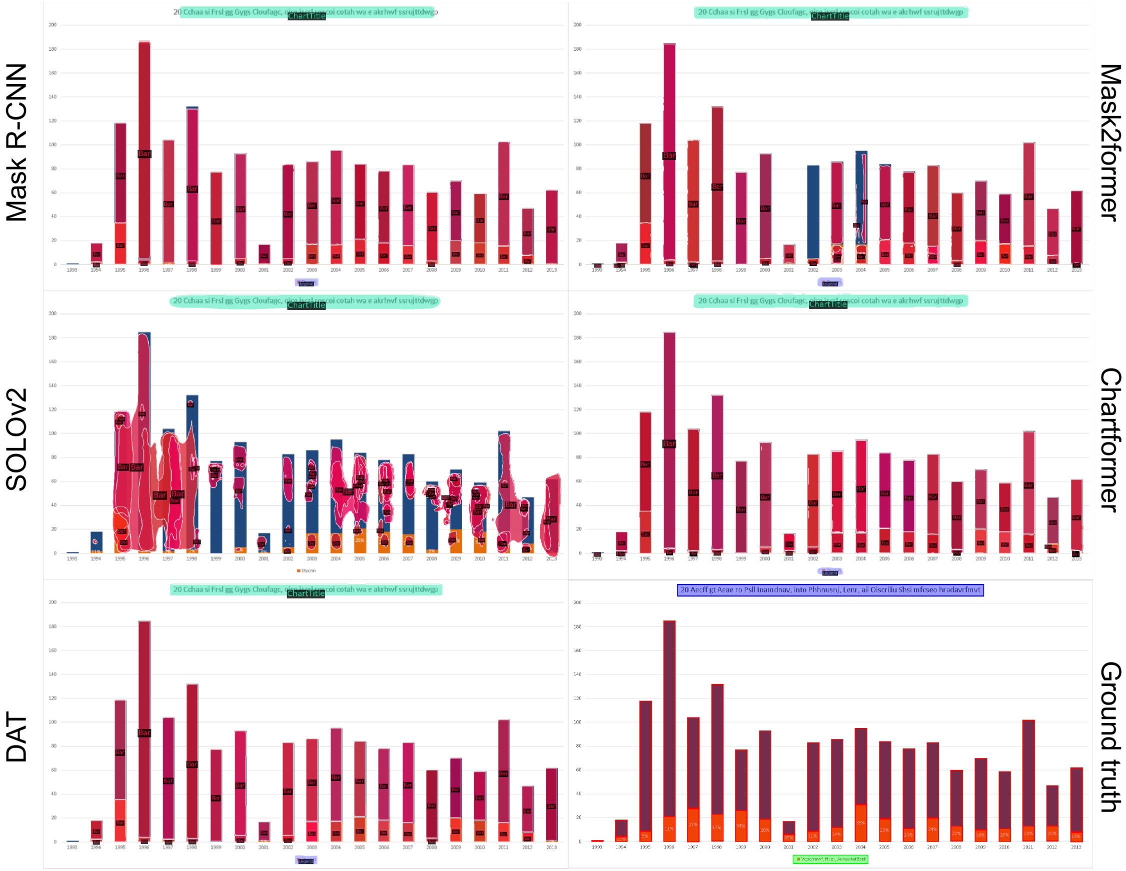


Figure 13. Bar performance case study