Classification-Regression for Chart Comprehension

Matan Levy[1†], Rami Ben-Ari[2], and Dani Lischinski[1]

[1] The Hebrew University of Jerusalem, Israel
[2] OriginAI, Israel

In the supplementary material we elaborate on some results shown in the main paper, as well as present new ones. We start with further elaboration on the characteristics of Chart Question Answering (CQA) benchmarks in Sec. 2, showing the shortcomings of previous public datasets such as FigureQA [15] and DVQA [13] by examples. Next, we elaborate on PlotQA [20] question type distribution, presenting the richness, realistic lingual relations, as well as regression requirement in this dataset. In Sec. 4 we show further justifications for the new accuracy metric. We show additional explainability results in Sec. 5 to strengthen the reasoning process in our results. Next, we show our results on DVQA, emphasizing the shortcomings of this dataset to showcase our method. In Sec. 7, we show a new experiment for language robustness and compare our CRCT model with the PReFIL [14] which showed high performance in previous benchmarks. Finally, In Sec. 8, we run our model on a newly generated example (not from PlotQA) as a single demo case.

1 Model Architecture

In this section we provide some equations in order to further clarify our model descriptions in the paper. Let us denote the Query, Key and Value for each branch at certain block as $Q_v, K_v, V_v \in \mathbb{R}^{n_v \times d}$ and $Q_t, K_t, V_t \in \mathbb{R}^{n_t \times d}$ corresponding to the visual and textual branch respectively. Each branch is attended by the other, as the following:

$$z_t = attn_d(Q_v, K_t, V_t) := softmax(\frac{Q_v {K_t}^T}{\sqrt{d}})V_t \in \mathbb{R}^{n_t \times d} \tag{1}$$

$$z_v = attn_d(Q_t, K_v, V_v) := softmax(\frac{Q_t {K_v}^T}{\sqrt{d}})V_v \in \mathbb{R}^{n_v \times d} \tag{2}$$

The co-encoder output is followed by a regular self-attention encoder, namely:

$$\forall i \in \{t, v\} \quad O_i = attn_d(Q(z_i), K(z_i), V(z_i)) \tag{3}$$

Note that Q_t, Q_v are exchanged, to allow interaction between different modalities. Then the outputs of each branch, O_t, O_v, are fed to the next co-transformers block in their proper branch. Finally, the resulting h_{v0}, h_{w0} pooling tokens from

[†] Part of this research was conducted at IBM Research AI, Israel.

the last layer are used for predicting if the concatenated answer is aligned (C) and the answer numeric value, R:

$$Loss_{CLS} = BCELoss(M_{CLS}(h_{w0} * h_{v0}), C) \tag{4}$$

$$Loss_{REG} = L1(M_{REG}([h_{w0}; h_{v0}]), R) \tag{5}$$

We train our model with the combined loss:

$$Loss = \lambda_1 \cdot Loss_{CLS} + \lambda_2 \cdot Loss_{REG} \tag{6}$$

We find $\lambda_1 = \lambda_2 = 1$ to be effective.

2 Characteristics of public CQA datasets

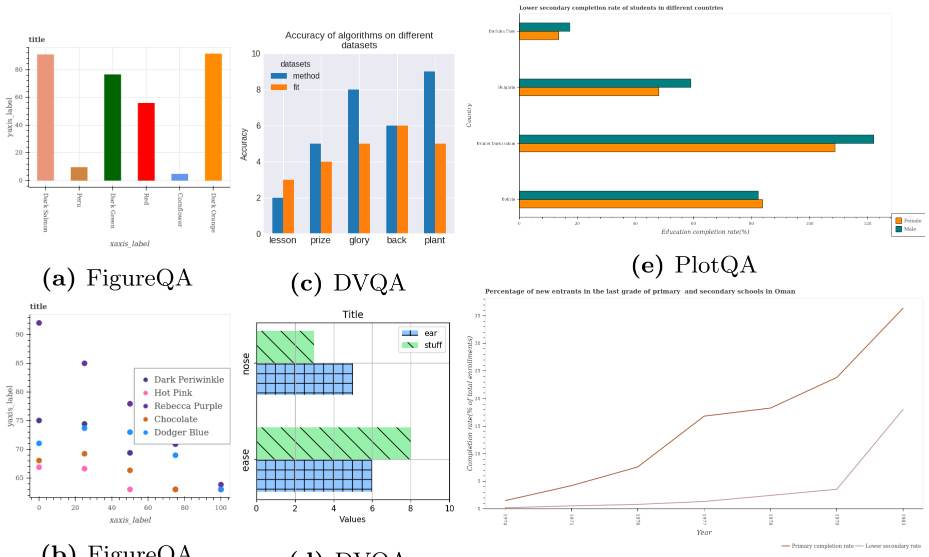
In this section we elaborate on the characteristics of different chart datasets previously used for Chart Question Answering (CQA) methods and further justify the choice of PlotQA as our main benchmark dataset. To this end, we present examples of charts from FigureQA, DVQA, and PlotQA in Fig. 1. This figure demonstrates the fixed templates and degenerate lingual forms used in two previous datasets. In FigureQA, the title, x-axis label and y-axis label are fixed in all the charts. Additional template pattern in FigureQA includes, the legend markers of the plots (*e.g.*, bars or scatter plots) named after their color. This pattern of naming is redundant throughout the entire dataset and is strictly used in the associated questions (see Fig. 1a & 1b). DVQA alleviates part of these shortcomings, yet with random words used as legend or bar labels, as shown in Fig. 1c & 1d. DVQA is limited to a single chart type and further introduces degenerated lingual forms that are unlikely to appear in a realistic chart. Note for instance, the word “Title” appearing as the title of the chart in Fig. 1d.

In Fig. 1e & 1f we show examples from the PlotQA dataset. To the best of our knowledge this is the most realistic dataset publicly available to date. In addition to it’s size and diversity (see Table 1, main paper) it is the only dataset that satisfies all the following terms: 1) Publicly available; 2) Fully annotated to train a detector; 3) Includes multiple chart types; 4) Charts with natural language patterns and relations; 5) Questions that demand regression.

These dataset characteristics are strongly related to the performance drop that was recently reported on PlotQA dataset in [20], and discussed in the paper. We further discuss additional factors, such as lack of regression required questions in previous benchmarks, in the main paper. In Tab. 1 we summarize the performance of several recently published methods against the existing datasets.

3 PlotQA Data Distribution

The PlotQA dataset suggests two benchmarks, which we refer to as PlotQA-D1 and PlotQA-D2 (see the paper for chart and Q&A breakdown). Both datasets



(a) FigureQA

(c) DVQA

(e) PlotQA

(b) FigureQA

(d) DVQA

(f) PlotQA

Fig. 1: Examples of charts from the FigureQA, DVQA, and PlotQA datasets. FigureQA charts - (a) and (b) lack any diversity in title and axis labels as well as the plot labels. In DVQA - (c) and (d) Random phrases and words are used in the chart text, resulting in lack of natural semantic relations between the different textual elements. These drawbacks are addressed in PlotQA, where the charts are taken from real world data, as shown in (e) and (f). Zoom in for better visibility.

share the same chart images. However, PlotQA-D1 is a subset of PlotQA-D2, with the latter having ×3.5 more Q&As. Tab. 2 shows the question type distributions for each benchmark with Fig. 2 depicting distributions of question templates in each question category, Structural (S), Data Retreival (D) and Reasoning (R). PlotQA-D1 introduces a relatively uniform distribution over the question templates, while PlotQA-D2 distribution is strongly skewed by a large number of questions requiring regression (with non-integer answers). PlotQA-D2 was designed to showcase the capability of a method on handling regression, a highly practical task and a strong shortcoming of previous datasets. The results reported in the paper demonstrate that CRCT outperforms previous methods on *both* of these benchmarks.

4 Accuracy Metric

In Fig. 3a we graphically visualize the dependency of the error tolerance on the ground truth value for the error ratio measure, in contrast to a fixed tolerance in the tick-based error, as suggested in our paper. Note the vanishing of the tolerance as the true value goes to zero (and vice versa).

Table 1: Accuracy of different methods on existing datasets. Note the significant drop in accuracy on PlotQA dataset (PlotQA-D). * our evaluation of PReFIL method on PlotQA-D

Method / Dataset	FigurQA-D	DVQA-D	LeafQA-D	PlotQA-D
PReFIL [14]	93.26	96.4	-	**10.36***
STL-CQA [27]	-	97.43	92.22	-
LEAF-QA [5]	81.15	72.8	67.42	-
PlotQA-M [20]	-	58.78	-	**22.52**

Table 2: Distribution over different question categories in PlotQA benchmarks

Data Ver.	Structural	Data Retrieval	Reasoning
PlotQA-D1	30.41%	24.01%	45.58%
PlotQA-D2	4.3%	13.74%	81.96%

This bias in the ratio based measure drives the errors to accumulate near zero as we show in Fig 3b. This figure presents a comparison between the error ratio measure and the suggested tick based error, for CRCT on PlotQA-D1, showing the bias in ratio based tolerance. We observe a relatively uniform error distribution on the ticked based alternative, as desired.

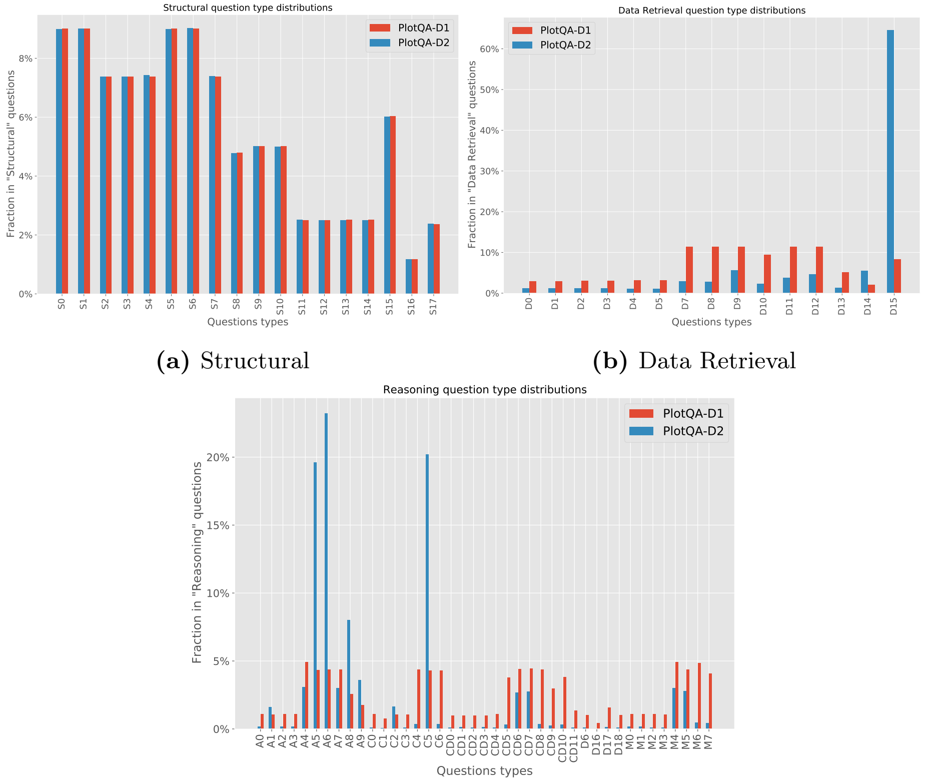
5 Additional Explainability Examples

In this section we show more visualization examples on CRCT explainability using *Captum* visualization tool (see Sec. 7 in the paper). All examples are drawn from the test set. In Fig. 4 we present a case with two line-plots in a chart. Note how the model attends to the correct plot among the two (hot bounding boxes) when asked about the *revenue*.

Fig. 5 shows an example of semantic understanding. Asked about the *intersection*, CRCT mostly attends to the two intersection points in the chart.

Fig. 6 shows another multi-plot chart. Here, the model correctly finds the *private credit* line plot as more influential to the question asked. Furthermore, asking about the *average* value drives CRCT to attend to all the plot elements corresponding to the *private credit* label.

In the next example in Fig. 7 we show a bar chart. Although the bars are very close in their heights (values), the relevant bar, with the minimum value gets the highest attention, leading to the correct answer.



(a) Structural

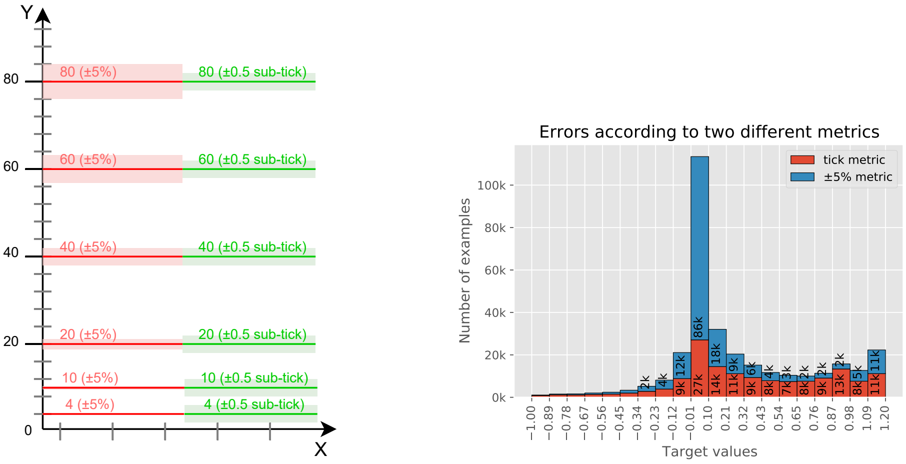
(b) Data Retrieval

(c) Reasoning

Fig. 2: PlotQA-D1 and D2 question type distributions. While in Structural questions the distribution is similar, in Data Retrieval and Reasoning questions, PlotQA-D2 is skewed towards few specific templates, which require a regression answer.

6 Result on DVQA

DVQA dataset is limited by 1) Single chart type (bar charts), 2) Lack of natural lingual text in the chart (see Sec. 2), 3) Answers appearing as a-priori known classes, eliminating the need for regression. This dataset further lacks the important legend marker annotation needed to train our detector. The importance of this object is clearly shown in the explainability examples in Sec. 5 (and in the paper), where legend markers are frequently highlighted, allowing CRCT to correspond to the correct plot/bar in the chart. The results of our CRCT model are shown in Table 3. Despite the limitations above, and errors involved in our heuristic annotation, we achieve a reasonable performance of 82.14%, ranked 3rd, on this benchmark and far beyond PlotQA-M that achieves 57.99% . Note that to showcase the strong limitation and existing performance saturation on



(a) A visual comparison between values and their error ranges, according to ±5% ratio metric (red) and ±1/2 sub-tick metric (green). The error tolerance of the error ratio measure depends on the ground truth value.

(b) Distribution of regression errors by two different metrics, the ±5% tolerance, in blue, and our suggested tick metric (Sec. 5 in the paper), in red. Note the peak in errors near zero, while the fixed tick based tolerance results nearly uniform distribution.

Fig. 3: Ratio vs Sub-tick metric.

DVQA we evaluate PReFIL, that reaches almost perfect performance on DVQA (96.37%), on the new PlotQA dataset (see results in the paper).

Table 3: Results on DVQA dataset. $CRCT_p$ indicates the CRCT model with a detector that was trained with *partial* bounding box annotations.

Method	SANDY [13]	PlotQA-M [20]	LEAF-Net [5]	$CRCT_p$	PReFIL [14]	STL-CQA [27]
Accuracy	56.48	57.99	72.72	82.14	96.37	**97.35**

7 Language Robustness

In this section we demonstrate the robustness under natural lingual variations of the CRCT model. Our transformer based model allows initialization with pre-trained BERT [7] on a large language corpus such as Wikipedia. CRCT is further trained downstream on all the textual elements in the chart, without any heuristics, such as string replacements (see the paper). This is in contrast to previous

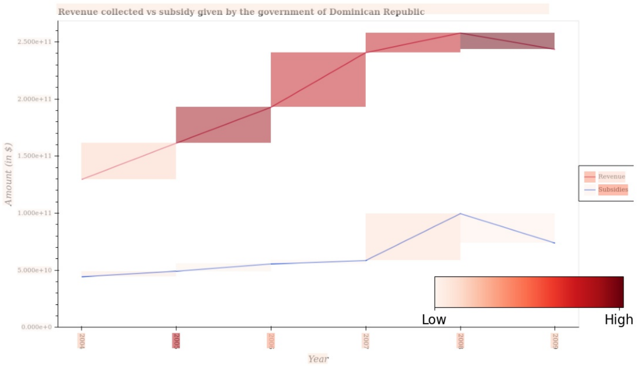


Fig. 4: Explainability visualizations for a PlotQA test sample. Q: *Is the amount of revenue collected in 2005 less than that in 2008?* ground truth: Yes. CRCT: Yes. Note the high attention on the correct plot between the two. The font sizes are from the dataset source.

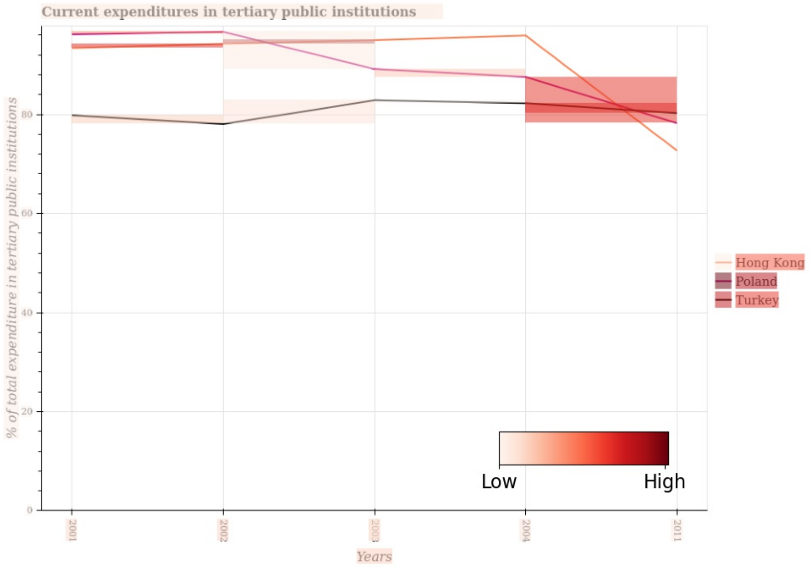


Fig. 5: Explainability visualizations for a PlotQA test sample. Q: *How many lines intersect with each other?*, Ground truth: 3. CRCT: 3. Note the hot spots at the intersection points.

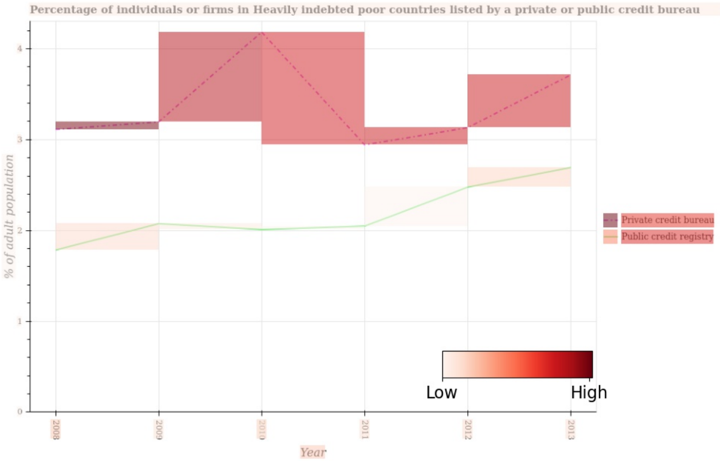


Fig. 6: Explainability visualizations for a PlotQA test sample. Q: *What is the average percentage of firms listed by* ***private*** *credit bureau per year?* Ground truth: 3.379. CRCT: 3.295 (Error: -2.49%). Note how the model attends the correct plot among the two, with ”hot” bounding boxes over all the plot due to *average* request.

methods, often using LSTM, based only on the chart dataset vocabulary. We further compare the CRCT robustness with PReFIL [14], where a LSTM is used for question encoding. In Tables 4, 5 and 6 we present question rephrasing on test figures. Each variation is a new manual phrasing of the original template. Note that the original template is the only one appearing in train set. Tables 4-6 show that while on the template question PReFIL gives the correct answer (indicated in green), it is mostly wrong (indicated in red) after question rephrasing. CRCT however is more robust to phrasing for various question types *e.g.* data retrieval, and regression.

8 New Generated Example

We conclude by showing a result from an experiment in our study in Fig. 8. To this end, we create a new chart showing our accuracy result compared to PReFIL. We now pose the following question to the model: *In 2.5% tolerance error, what is the difference between the accuracy of CRCT and PReFIL?*. Although this figure was not part of the PlotQA dataset, we obtained an answer that deviates the true result only by 0.47%. The robustness of CRCT is further illustrated here on handling unknown initials of the corresponding methods.

Table 4: Question rephrasing, for Fig. 5. Each variation was manually rephrased and never seen in train, except the original version. Note the high sensitivity of PReFIL to different question phrasings.

Var.	Question	CRCT	PReFIL
Original	What is the label or title of the X-axis ?	Years	Years
#1	What's the name of the X-axis?	Years	40
#2	What is the label or title of the horizontal axis?	Years	0
#3	What is the x label of the plot?	Years	2011
#4	The x-label of the figure?	2004	0
#5	What's the figure's x-axis label?	Years	2011
#6	Give me the x-axis label	Years	2011
#7	What the x-axis represents?	No	% of total expenditure in tertiary public institutions
#8	What is the label of X?	Years	2011
#9	X-label?	2011	2011

Table 5: Question rephrasing, for Fig. 1e. Each variation was manually rephrased and never seen in train, except the original version. Note the high sensitivity of PReFIL to different question phrasings.

Var.	Question	CRCT	PReFIL
Original	How many different coloured bars are there ?	2	2
#1	How many bar colors are there?	2	2
#2	How many colors of bars can you see?	2	1
#3	Coloured bars?	No	Bolivia
#4	How many colors paints each group of bars?	2	Bolivia
#5	How many colors are there?	2	2
#6	How many different bars exists in each group?	2	0
#7	Colors in each group?	No	0
#8	How many colors?	2	0
#9	Give me the size of each group of bars	Bolivia	2

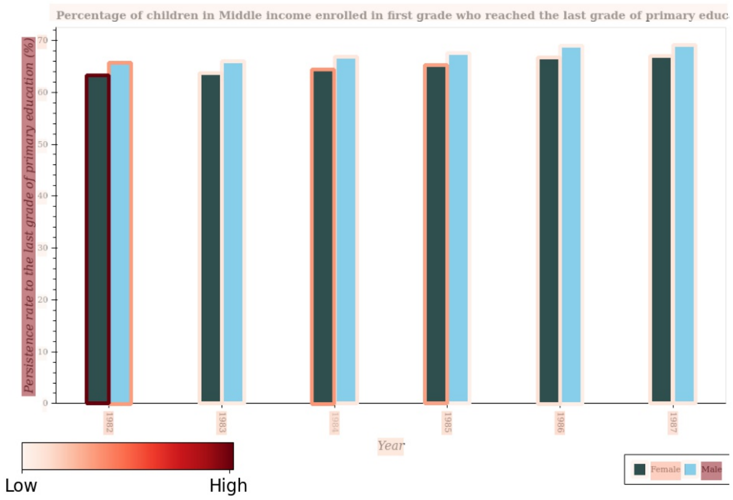


Fig. 7: Explainability visualizations for a PlotQA test sample. Q: *In which year was the persistence rate of female students minimum?*, Ground truth: 1982. CRCT: 1982. For better visibility, we overlay the visualization as colored bounding box around the bars. Note how green bars related to *Female* achieve higher attention with the correct bar receiving the highest attention.

Table 6: Question rephrasing, for Fig. 1f. Each variation was manually rephrased and never seen in train, except the original version. Note that this question requires regression. In every variation therefore the $\langle R \rangle$ token was chosen in CRCT's hybrid prediction head, leading to the regression value shown as an answer. The values in green and red are correct and wrong answers respectively. Values in blue present the deviation from the true value. Note the high sensitivity of PReFIL to different question phrasings

Var.	Question	CRCT	PReFIL
Original	Across all years, what is the maximum completion rate in primary schools ?	36.618 (+0.57%)	35 (−3.87%)
#1	What's the maximum primary completion?	36.625 (+0.59%)	0 (−100%)
#2	What is the maximal rate of primary school completion, over the years?	36.305 (−0.288%)	No
#3	Across all years, what is the maximum primary school completion rate?	36.613 (+0.56%)	0 (−100%)
#4	Over the years, what is the highest primary school completion rate?	36.502 (+0.25%)	No
#5	What is the maximum completion rate in primary schools across all years?	36.635 (+0.618%)	0 (−100%)
#6	In primary schools, what is the highest completion rate across all years?	36.53 (+0.33%)	3 (−91.76%)
#7	The maximum completion rate in primary schools is what - across all years?	36.622 (+0.58%)	0 (−100%)
#8	Give me the maximum rate of primary completion over the graph	33.581 (−7.77%)	5 (−86.27%)
#9	Average the primary completion rate	15.065 (−58.624%)	No

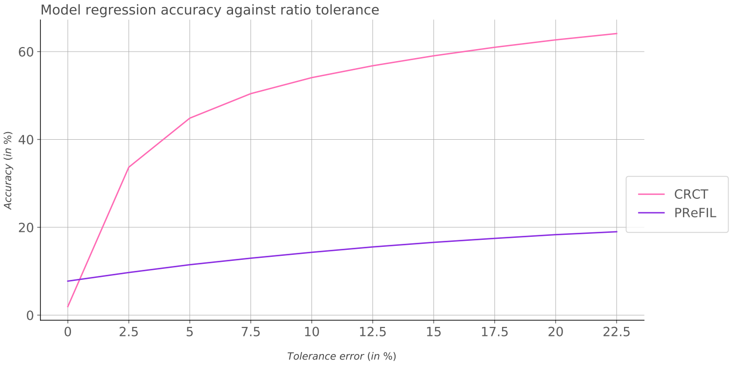


Fig. 8: We insert into CRCT a result from our paper showing the regression accuracy of our model against PReFIL. We pose the following question: *In 2.5% tolerance error, what is the difference between the accuracy of **CRCT** and **PReFIL**?*. Ground truth: 23.978. CRCT: 23.865 (-0.47%).

References

1. Acharya, M., Kafle, K., Kanan, C.: TallyQA: Answering Complex Counting Questions. In: AAAI (2019)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: ICCV (2015)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: ICLR (2015)
4. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In: CVPR (2021)
5. Chaudhry, R., Shekhar, S., Gupta, U., Maneriker, P., Bansal, P., Joshi, A.: LEAF-QA: Locate, Encode & Attend for Figure Question Answering. In: WACV (2020) 4, 6
6. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual Dialog. In: CVPR (2017)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT (2019) 6
8. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: CVPR 2017 (2017)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
11. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation (1997)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: CVPR (2017)
13. Kafle, K., Cohen, S., Price, B., Kanan, C.: DVQA: Understanding Data Visualizations via Question Answering. In: CVPR (2018) 1, 6
14. Kafle, K., Shrestha, R., Cohen, S., Price, B., Kanan, C.: Answering questions about data visualizations using efficient bimodal fusion. In: WACV (2020) 1, 4, 6, 8
15. Kahou, S.E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., Bengio, Y.: FigureQA: An Annotated Figure Dataset for Visual Reasoning. In: ICLRW (2018) 1
16. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Reynolds, J., Melnikov, A., Lunova, N., Reblitz-Richardson, O.: PyTorch Captum. https://github.com/pytorch/captum (2019)
17. Leino, K., Sen, S., Datta, A., Fredrikson, M., Li, L.: Influence-Directed Explanations for Deep Convolutional Networks. In: ITC (2018)
18. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In: NeurIPS. pp. 13–23 (2019)
19. Malinowski, M., Fritz, M.: A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In: NeurIPS (2014)
20. Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: PlotQA: Reasoning over Scientific Plots. In: WACV (March 2020) 1, 2, 4, 6
21. Miech, A., Zhukov, D., Alayrac, J., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019)

22. Pasupat, P., Liang, P.: Compositional Semantic Parsing on Semi-Structured Tables. In: ACL (Jul 2015)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Meila, M., Zhang, T. (eds.) ICML (2021)
24. Santoro, A., Raposo, D., Barrett, D.G.T., Malinowski, M., Pascanu, R., Battaglia, P.W., Lillicrap, T.: A simple neural network module for relational reasoning. In: NIPS (2017)
25. Schwartz, I., Yu, S., Hazan, T., Schwing, A.G.: Factor Graph Attention. In: CVPR (2019)
26. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards VQA Models That Can Read. In: CVPR (2019)
27. Singh, H., Shekhar, S.: STL-CQA: Structure-based Transformers with Localization and Encoding for Chart Question Answering. In: EMNLP (Nov 2020) 4, 6
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. In: NeurIPS (2017)
29. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
30. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked Attention Networks for Image Question Answering. In: CVPR (2016)
31. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering (2021)
32. Zou, J., Wu, G., Xue, T., Wu, Q.: An Affinity-Driven Relation Network for Figure Question Answering. 2020 ICME (2020)