Representing Charts as Text for Language Models: An In-Depth Study of Question Answering for Bar Charts

Victor Soares Bursztyn, Jane Hoffswell, Eunyee Koh, and Shunan Guo

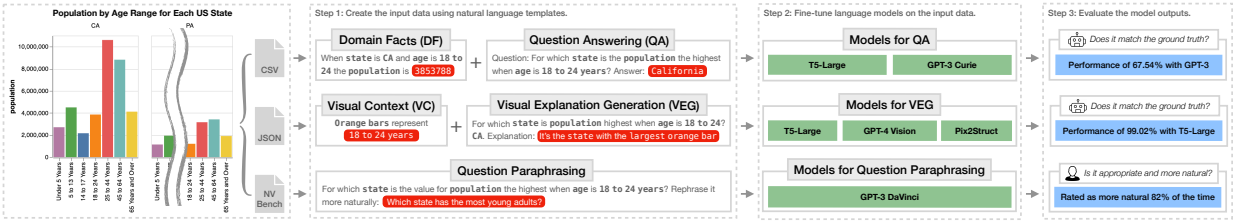

Fig. 1: We explore two main tasks related to chart-grounded Q&A: *question answering (QA)* and *visual explanation generation (VEG)*. QA leverages templated *domain facts (DF)* from the chart's CSV file, whereas VEG relies on *visual context (VC)* from its JSON file. In the first fine-tuning step, the charts' underlying text files are injected into the language models (LMs). We then fine-tune the QA and VEG steps on 90% of the charts, with 10% held out for testing during our evaluation in §4. To understand the robustness of our LMs to natural language variation, we also perform a *question paraphrasing* task to rephrase our template-generated questions more naturally.

Abstract—Machine Learning models for chart-grounded Q&A (CQA) often treat charts as images, but performing CQA on pixel values has proven challenging. We thus investigate a resource overlooked by current ML-based approaches: the declarative documents describing how charts should visually encode data (i.e., *chart specifications*). In this work, we use chart specifications to enhance language models (LMs) for chart-reading tasks, such that the resulting system can robustly understand language for CQA. Through a case study with 359 bar charts, we test novel fine tuning schemes on both GPT-3 and T5 using a new dataset curated for two CQA tasks: question-answering and visual explanation generation. Our text-only approaches strongly outperform vision-based GPT-4 on explanation generation (99% *vs.* 63% accuracy), and show promising results for question-answering (57–67% accuracy). Through in-depth experiments, we also show that our text-only approaches are mostly robust to natural language variation.

Index Terms—Machine Learning Techniques; Charts, Diagrams, and Plots; Datasets; Computational Benchmark Studies

1 Introduction

Charts convey information through representations that are both visual and symbolic, with elements such as lines or bars representing domain attributes. Compared to visual question answering (VQA) on natural images, chart-grounded question answering (CQA) is more sensitive to small pixel changes. For example, shuffling the colors of a car and bike in a photograph only affects the properties of the two objects, but shuffling the colors of a bar chart can completely alter its meaning [8]. Therefore, much of the progress in VQA for natural images may not transfer well to even the simplest charts [1, 8, 9, 12, 14, 21, 23].

Despite the importance of language in chart analysis [5, 22], relatively few VQA works have focused on CQA [6, 11, 24]. While the majority of these CQA works choose to represent charts as images, there have been very few early investigations on the usefulness of charts' declarative specifications as an alternative representation for chart captioning [23], chart generation [13], and rudimentary CQA [10], with the latter predating large language models (LMs).

As the first in this space, Kim *et al.* [10] rely on manually replacing parts of a question that refer to visual elements by the attribute names (e.g., changing the question *"Which state has the largest orange bars?"* to *"Which state has the max(19-24 years)?"*), before running a rule-based Q&A system for relational tables [17]. A limitation of this approach is that CQA is only *weakly* grounded in the chart's visual context. For example, the question *"Which state has the most young adults?"* is not anticipated by their NLP rules, and thus not grounded in a visual element. However, an ideal system should understand the relationships *"young adults"* → *"19-24 years"* → *"orange bars,"* as this mapping may be useful for generating visual explanations for novice analysts, e.g., *"California has the largest orange bars."*

In this work, we infuse chart specifications into LMs for their ability to flexibly handle language, so that the resulting system can ***robustly*** perform CQA tasks. Our contributions are: (1) Our experiments on two tasks (*question-answering* and *visual explanation generation*) using two families of LMs (GPT-3 [3] and T5 [18]) show that chart specifications are an effective representation format for CQA; (2) We find that even state-of-the-art Vision-powered LMs (VLMs), both pretrained and fine-tuned (GPT-4 [16] and Pix2Struct [11]), struggle to capture the signal that our text-only approaches can learn; and (3) We show that our models remain mostly robust to natural language variation.

2 Related Work

Given the widespread adoption of visualization grammars [2, 20], recent work on CQA has proposed leveraging such declarative specifications rather than images [10, 13, 23]. More generally, text-to-text LMs can belong to two categories: ***encoder-decoder*** LMs such as T5 [18] have their generative component preceded by an encoder-only step tailored to "understanding" and representing the LM input, making the combined encoder-decoder architecture flexible even in smaller model sizes; as size increases, ***decoder-only*** LMs such as GPT-3 [3] have become the *de facto* standard for their ability to generate natural-sounding language. For this work, we focus on Vega-Lite [20] as our visualization grammar, and we choose one family of LMs from each category (i.e., T5 [18] and GPT-3 [3]) to test the usefulness of chart specifications more broadly.

VLMs can also belong to two categories: uniquely ***large, pretrained VLMs*** such as vision-based GPT-4 [16] can become the state-of-the-art in many tasks that require vision by virtue of their scale; while

• *Victor Soares Bursztyn, Jane Hoffswell, Eunyee Koh, and Shunan Guo are with Adobe Research. E-mail: {soaresbu, jhoffs, eunyee, sguo}@adobe.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

image-encoder-text-decoder VLMs such as Pix2Struct [11] can achieve the same by being amenable to ***fine-tuning*** on custom task data. We choose GPT-4 [16] and Pix2Struct [11] as baselines representing each category for their recent impact in chart-related tasks and analyze the errors of the most competitive baseline when comparing performances. Tang *et al.* [23] contribute the work closest to ours, as they use chart specifications for chart captioning and compare to previous-generation VLMs not designed for charts [4]. We not only compare against two state-of-the-art VLMs that claim to work on charts [11, 16], but our focus on CQA also allows us to test the robustness of our models to variations in how questions are phrased, unlike captioning.

3 Methods for Chart-Grounded Q&A

First, we formally describe our two main CQA tasks (*question answering* and *visual explanation generation*), as well as a third task (*question paraphrasing*) to evaluate the quality of our fine-tuned LMs (Figure 1); second, we discuss the process for designing our dataset; third, we describe the LM fine-tuning scheme applied in our experiments in §4, and our process for generating more natural questions from our dataset.

3.1 Task Definitions

We adopt a "closed-book exam" [19] configuration for our LMs, in order to extend their underlying knowledge with chart-related information before performing the main task. We use LM_{step} to denote the order in which each fine-tuning step is performed.

3.1.1 Question Answering (QA)

We define question-answering as a function of the data displayed on the chart (i.e., *"domain facts"*) and the questions asked. Domain facts, originally stored in CSV format, are converted into natural language using templates. Question-answering examples are similarly generated. Figure 1 shows one example; the underlying templates are provided in the supplemental material. If DF and QA are the domain facts and QA training sets, question-answering is $LM_1(DF)$ followed by $LM_2(QA)$.

3.1.2 Visual Explanation Generation (VEG)

Visual explanation generation considers a chart's visual context (VC) and the question-answer pair that requires a visual explanation. Visual contexts, stored as JSON chart specifications, are converted into natural language via templates. Questions, answers, and explanations are similarly transformed (see the supplemental material). If VC and VEG are visual contexts and visual explanation generation training sets, we define visual explanation generation as $LM_1(VC)$ then $LM_2(VEG)$.

3.1.3 Question Paraphrasing

Some template-generated questions may seem unnatural, despite being factually correct. For example, the questions *"For which state is the value for population the highest when segment is 19-24 years?"* and *"Which state has the most young adults?"* both have the same meaning, but the latter is rated as more natural by our three human judges in §4. Given that LMs can generate language with great fluency [3], we leverage this ability to paraphrase our template-generated questions more naturally while preserving their meaning. We model this task as a function of a template-generated question (q_{temp}); to help the LM remain semantically grounded on the chart domain, we also include a chart description (c_{desc}). Thus, question paraphrasing is defined as "c_{desc}. q_{temp}. *Rephrase it more naturally:* $\rightarrow q_{nat}$," where q_{nat} is the rephrased question and GPT-3 DaVinci (175B parameters) the base LM.

3.2 Dataset Design

No previously published dataset with chart specifications has all the characteristics needed to investigate CQA through the lens of LMs. We thus draw on the strengths of two recent datasets to design ours: VisQA [10] explores CQA and includes naturally arising questions, but is too small for LM fine-tuning. Conversely, NVBench [13] is richer in scale and domain diversity, but focuses on generating SQL queries from chart descriptions. Similar to Obeid and Hoque [15], we note from VisQA that bar charts with one quantitative attribute on the y-axis and two categorical attributes (as seen in Figure 2) ensure a minimum

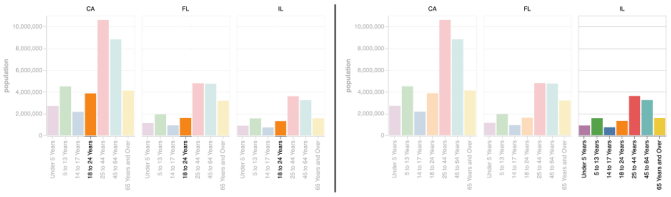

Fig. 2: Visualization of distribution-related questions on an illustrative bar chart. On the left, users can ask about values across same-color bars. On the right, users can ask about values within a chart panel.

Task	Model	Accuracy		Gain or Loss
		Q^{test}_{temp}	Q^{test}_{nat}	
QA	T5-Large	57.54%	58.52%	+1.5%
	GPT-3 Curie	67.54%	48.36%	-28.4%
VEG	T5-Large	99.02%	73.28%	-26.0%
	GPT-4 Vision	63.01%	-	-
	Pix2Struct	12.26%	-	-

Table 1: For question-answering (QA), our fine-tuning scheme for T5-Large is 57.54% accurate, which increases to 67.54% with a model 17 times larger (GPT-3) on Q^{test}_{temp}; for visual explanation generation (VEG), our fine-tuning scheme for T5 is 99.02% accurate, much higher than vision-based GPT-4 (63.01%). When evaluating on Q^{test}_{nat}, performance ranges from a 1.5% gain (T5 on QA) to a 28.4% loss (GPT-3 on QA).

diversity of visual elements. NVBench has 359 bar charts of this type, with an average of 3.94 colored bars per chart, spanning 105 domains. Thus, we define our focus for this work on this set of bar charts.

Further analyzing VisQA, we note that most human-generated questions about bar charts are related to distributional aspects of the quantitative attribute. We use this observation to define the following scope for our CQA: We include questions about (1) *minimum and maximum values across visual elements of the same color* (Figure 2, left) and (2) *minimum and maximum values within a chart panel* (Figure 2, right). Importantly, these two types of questions are well-studied in the visualization literature—for example, refer to Figure 5 in Xiong *et al.* [26].

Overall, we use templates on 359 bar charts from Luo *et al.* [13] to generate 9,885 domain facts (DF), 7,310 question-answer pairs (QA), 3,989 visual contexts (VC), and 7,310 question-answer-explanation triples (VEG) as defined in §3.1.[1]

3.3 Fine-Tuning QA and VEG

For both question-answering and visual explanation generation, the first fine-tuning step (LM_1) includes 100% of the knowledge-related data, i.e., $LM_1(DF)$ comprises 100% of DF and $LM_1(VC)$ 100% of VC. For the second step (LM_2), we hold out 10% of the charts for testing. Thus, $LM_2(QA)$ comprises ~90% of QA for question-answering and $LM_2(VEG)$ has ~90% of VEG for explanation generation. As described in §3.1, questions are part of the prompt for both QA and VEG. We denote the 6,700 template-generated questions during training as Q^{train}_{temp} and the 610 held-out questions for evaluation as Q^{test}_{temp}.

We manually rephrase Q^{test}_{temp} to produce a human-generated test set Q^{test}_{nat} to evaluate the robustness of an LM-based approach in §4. Initially, LMs are fine-tuned only on Q^{train}_{temp} with template-generated questions. A perfectly robust LM would have performance on $Q^{test}_{nat} \approx Q^{test}_{temp}$. However, an LM can still be considered robust if it suffers a slight performance loss (i.e., performance on $Q^{test}_{nat} < Q^{test}_{temp}$) and can recover with additional data from the distribution of Q^{test}_{nat}. To test this hypothesis, we apply question paraphrasing on Q^{train}_{temp} to produce Q^{train}_{nat} with more naturally-phrased questions. We then progressively augment our initial LMs with parts of Q^{train}_{nat} while measuring performance on Q^{test}_{nat}.

[1] Our dataset is publicly available at: https://github.com/vbursztyn/charts-as-text-for-chartqa

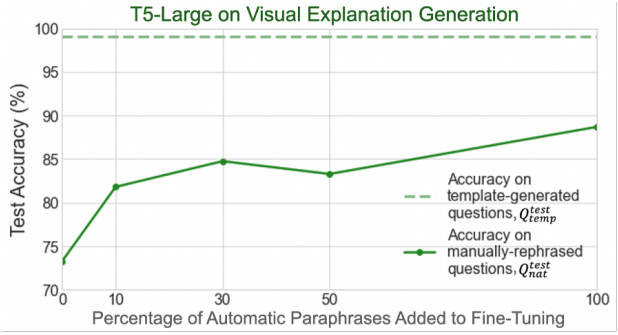

Fig. 3: Accuracy on Q_{nat}^{test} *vs.* the percentage of Q_{nat}^{train} data added for fine-tuning. The dashed line shows the accuracy on Q_{temp}^{test}, serving only as an upper-bound. There is a consistent recovery of lost performance when adding additional data for fine-tuning (up from 73.28% to 88.69%); notably, the largest improvement in performance occurs when adding only 10% of the data (up from 73.28% to 81.80%).

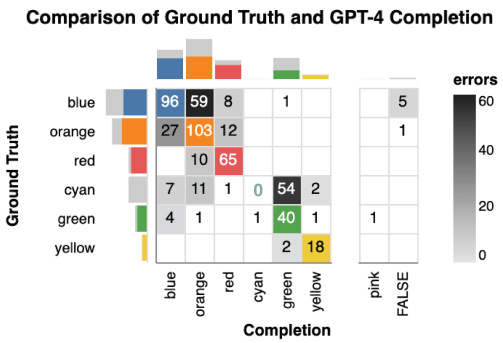

Fig. 4: A confusion matrix of errors made by GPT-4. As shown in Fig. 5, error #1 (returning orange instead of blue) and error #3 (green rather than cyan) are the most common. GPT-4 often conflates adjacent rows in the legend, which is responsible for 79% of errors. Notably, the results rarely mention the color cyan; the one instance actually refers to "teal," which was deemed close enough given color-naming ambiguities.

3.4 Fine-Tuning Question Paraphrasing

We use the following training set to fine-tune GPT-3 DaVinci for question paraphrasing: given the formulation *"c_{desc}. q_{temp}. Rephrase it more naturally: $\rightarrow$ q_{nat},"* we populate q_{temp} and q_{nat} with Q_{temp}^{test} and Q_{nat}^{test} respectively; we retrieve c_{desc} from the corresponding charts [13]. We apply the resulting LM to the questions in Q_{temp}^{train} with decoding temperature = 0.7 and max sequence length = 40, producing Q_{nat}^{train}.

4 Experiments

We run experiments to address three research questions.
RQ1: How well do LMs *perform* on the two CQA tasks?
RQ2: How helpful are chart specifications *compared* to images?
RQ3: How *robust* are these models to natural language variation?

4.1 Procedure and Measurements

The details for fine-tuning T5 [18], GPT-3 [3], and Pix2Struct [11] are described in the supplemental material. We use greedy decoding to evaluate both tasks. For question-answering, we consider a test question to be correctly answered if the generated tokens match the ground-truth. For visual explanation generation, explanations are invariant except for the color and size of the visual element associated with the answer (e.g., *"the **largest orange** bars"*); we consider a test pair to be correctly explained i.f.f. it includes the visual element's correct color and size. This approach is similar to how previous works measured factuality in generated text, e.g., the RG metric in Wiseman *et al.* [25]. We refer to these definitions over a given test set as the "Accuracy" in Table 1.

We recruit three human judges to evaluate the question paraphrasing. Each judge rates the same sample of 100 automatic paraphrases in two ways: (1) Using a scale from 1 (dissimilar) to 3 (similar), how much the paraphrases preserve the meaning of the original template-generated question (i.e., a measure of *semantic similarity*); and (2) between the template-generated question and the rephrased one, which one is more natural, or if they are equally natural (i.e., a measure of *naturalness*).

For RQ1 (performance) and RQ2 (comparison), we evaluate the two CQA tasks (question-answering and visual explanation generation) on Q_{temp}^{test}, i.e., held-out questions that follow the same template seen in training. For RQ3 (robustness), we evaluate our models on Q_{nat}^{test}, i.e., the manually rephrased questions. To understand if an LM-based approach can recover from a potential performance loss in the face of more natural questions, we progressively add automatic paraphrases (i.e., parts of Q_{nat}^{train}) to our fine-tuning scheme while we measure how these additions affect the performance on Q_{nat}^{test}.

4.2 Results: Question Answering (QA)

Table 1 shows the results for two base models: T5-Large (737M parameters) and GPT-3 Curie (13B parameters). T5-Large achieves 57.54% test accuracy on Q_{temp}^{test}. This result is *substantially* above any random baselines: test questions span the two categorical attributes of each bar chart, so the space of possible answers is even bigger than that of colors. On the other hand, the result is still far from the 100% upper-bound. To test if LM scale affects performance, we also include GPT-3 Curie; with 17.5 times more parameters, performance improves to 67.54%.

4.3 Results: Visual Explanation Generation

T5-Large achieves 99.02% accuracy on Q_{temp}^{test} (Table 1), which confirms that our method is extremely successful at this task, i.e., $LM_1(VC)$ successfully injects the charts' visual contexts into T5, and $LM_2(VEG)$ can learn to access them to generate factually correct explanations.

4.4 Text-only T5 *vs.* VLMs

Table 1 contrasts the near perfect performance of our text-only approach to visual explanation generation (with 99.02% accuracy) *vs.* the much larger vision-based GPT-4 (63.01%). As detailed in Section 5.1, 79% of the errors from GPT-4 are due to mistaking rows that are adjacent in the legend. As detailed in the supplemental material, fine-tuning the recent Pix2Struct on 90% of Q_{temp}^{train} yields only 12.26% accuracy on Q_{temp}^{test}, confirming the struggles of a VLM in the same size category as T5-Large. Pix2Struct shortcuts to "blue" in 44% of its explanations.

4.5 Results: Question Paraphrasing

The assessments from our three independent human judges had a Fleiss' Kappa of 0.61, which indicates substantial agreement. In terms of semantic similarity, rephrased questions were considered to preserve the meaning (score of 3) of the original, template-generated questions 88.33% of the time, while the meaning was considered lost (score of 1) only 7.7% of the time, for an average score of 2.81 (out of 3).

In terms of naturalness, rephrased questions were considered more natural 82.33% of the time versus only 5% for template-generated questions. This result confirms that, in preparation for RQ3 (robustness), our method can successfully rephrase template-generated questions to be more natural, while preserving their original meaning.

4.6 Approach Robustness

Table 1 compares the performance on Q_{temp}^{test} *vs.* Q_{nat}^{test}. Interestingly, we note results as diverse as a marginal performance gain—from 57.54% to 58.52%—and relative losses of 26% (for visual explanation generation) and 28.4% (for question-answering with GPT-3, which is potentially more prone to overfitting). Despite this variation, LMs fine-tuned with our scheme remain relatively accurate with natural language variation.

Figure 3 shows what happens when T5 is given additional fine-tuning data from Q_{nat}^{train}. We progressively add to our fine-tuning scheme 10%, 30%, 50%, and 100% of Q_{nat}^{train}, while measuring performance on Q_{nat}^{test}. We plot a dashed line to show the original performance on Q_{temp}^{test}. Interestingly, we find a consistent recovery of the lost performance which

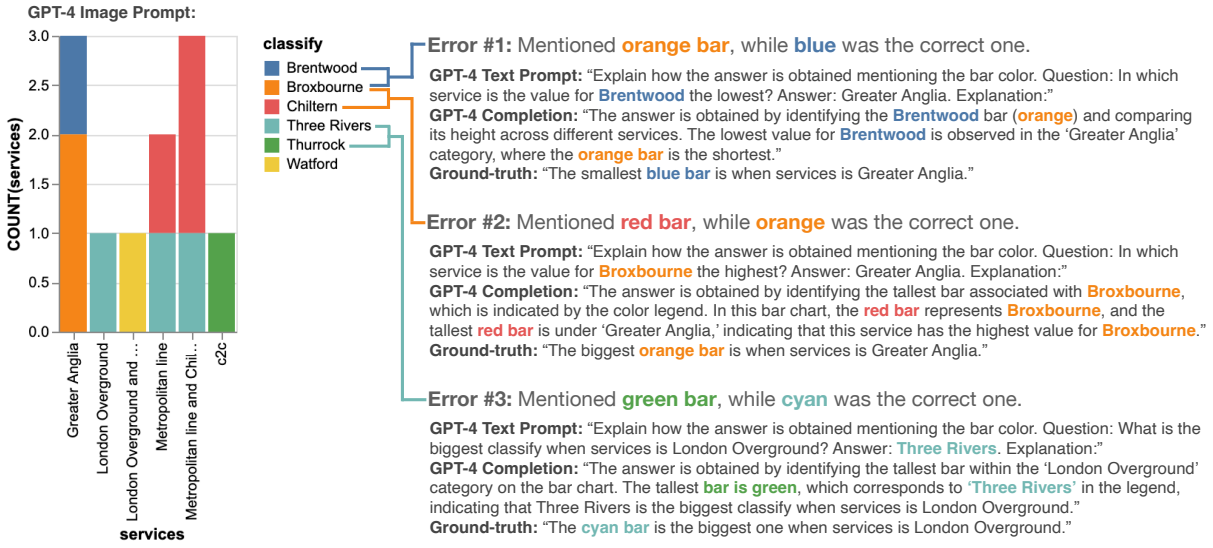

Fig. 5: This stacked, multi-color bar chart depicts the counts of the attribute *"classify"* on the y-axis, across different services on the x-axis. This example ("`bar_chart_518`" in the released data) has eight errors out of twenty-four test cases. The three errors illustrated here demonstrate one way that vision-based GPT-4 fails, i.e., by mentioning an adjacent color in the legend rather than the correct color. The same **GPT-4 Image Prompt** is used in all cases, paired with each **GPT-4 Text Prompt**. The resulting **GPT-4 Completion** can be compared with the **Ground-truth**.

substantially narrows the previously noted gap: 99.02% *vs.* 88.69%. Most importantly, we find that a large part of this recovery happens when adding only 10% additional data: 73.28% *vs.* 81.80%.

5 Discussion, Limitations, and Future Work

Our work introduces initial results for CQA using chart specifications, a resource largely overlooked by previous ML-based approaches [7]. We find that such specifications store useful signal that—in the case of visual explanation generation—can be almost perfectly learned by even T5-Large, an encoder-decoder LM under 1B parameters. While natural language variation in the input can hurt performance, §4.6 shows how systems based on the proposed fine-tuning scheme could remain mostly robust and quickly recover. Importantly, our models compare favorably on visual explanation generation to even the most competitive VLM (vision-based GPT-4), whose errors we analyze in-depth below. We believe these findings can motivate larger-scale investigations on the usefulness of chart specifications—and to this end, we discuss the limitations of our work further below.

5.1 Vision-based GPT-4 Error Analysis

Figure 4 shows the confusion matrix of all errors comparing the ground truth value and completion generated by GPT-4. As a case study for the types of errors, consider the example visualization in Figure 5. Each chart in our dataset corresponds to questions about the biggest/smallest value for each nominal variable; for this example, our dataset includes twenty-four questions related to the "services" and "classify," and eight of these twenty-four questions produce errors (33.3%). Figure 5 illustrates three of these errors for the same base chart. GPT-4 receives both the chart image and the text prompts as input, and generates completions that should mention the correct visual elements. However, in error #1 (2 of 8 errors), the blue bar is the correct one, but GPT-4 focuses on the one below it in the legend (orange); in error #2 (1 of 8), orange is correct, but GPT-4 again focuses on the one below (red); and in error #3 (5 of 8 errors), cyan is correct, but GPT-4 once again focuses on the one below (green). Across all of the errors in Figure 4, GPT-4 conflates adjacent rows, like in this example, in 79% of cases.

The bar charts along each side of the confusion matrix in Figure 4 show the total values for the ground truth and completion for each color; while blue corresponds to the most ground truth results, orange is the most common completion produced by GPT-4. Notably, the results rarely mention the color cyan, despite this color corresponding to 75 of 530 questions in our test set (14.15%). The charts in our dataset all leverage Vega-Lite's `tableau10` color scheme; our test set includes 30 charts with the following breakdowns for the number of available colors: two color, blue and orange (nine); three color, add red (one); four color, add cyan (five); five color, add green (eight); six color, add yellow (seven). Interestingly, for the four color test cases, all sixteen questions that have a ground truth of cyan instead return green, which is not visible in the provided legend. In future work, we would like to explore how consistent these errors are in the face of shuffled or variable color schemes to better understand the cause of these errors.

5.2 Exploring New Questions and Chart Types

In this in-depth case study, we focus on only one chart type, multi-color bar charts, which is selected due to their diversity of visual elements, as recognized by prior work [15]. However, in current question-answer pairs, attributes with a single value have the same maximum and minimum (Figure 5, errors #1 and #2). We also see value in broadening the scope to other types of charts, especially if they introduce questions that could be more challenging for LMs (e.g., on multi-color line charts, this extension could include questions with comparisons on continuous ranges such as *"How many times do the red and blue lines intersect?"*).

Additionally, as explained in §3.2, we generate our questions based on an analysis of VisQA [10]. Despite being human-generated, their sample of 629 questions is still relatively limited, making the resulting dataset an approximation of the questions that humans would ask. A more expensive crowd-sourcing method could help expand the dataset to address this limitation.

6 Conclusion

In summary, we find initial evidence that using chart specifications to enhance LMs with chart-reading ability is a promising direction in CQA. We answer RQ1 (performance) positively for visual explanation generation, while we see room for improvement for the challenging question-answering task. RQ2 (comparison) is answered positively, as our smaller, text-only approach outperforms GPT-4 by a large margin (99% *vs.* 63% accuracy), showing that even a state-of-the-art VLM can still struggle with image representations. We also answer RQ3 (robustness) mostly positively, as LMs retain most of their performance in the face of more natural questions, and can continue to improve with augmented fine-tuning. Such robustness is not possible with the previously published rule-based approach [10]. Finally, motivated by these findings and analyses, we release our dataset upon publication and outline spaces for future work.

REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015. 1

[2] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011. 1

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 3

[4] J. Cho, J. Lei, H. Tan, and M. Bansal. Unifying vision-and-language tasks via text generation. In M. Meila and T. Zhang, eds., *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 1931–1942. PMLR, 18–24 Jul 2021. 2

[5] A. Gaba, V. Setlur, A. Srinivasan, J. Hoffswell, and C. Xiong. Comparison conundrum and the chamber of visualizations: An exploration of how language influences visual design. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1

[6] E. Hoque, P. Kavehzadeh, and A. Masry. Chart question answering: State of the art and future directions. *Computer Graphics Forum*, 41(3):555–572, 2022. doi: 10.1111/cgf.14573 1

[7] E. Hoque, P. Kavehzadeh, and A. Masry. Chart Question Answering: State of the Art and Future Directions. *Computer Graphics Forum*, 2022. doi: 10.1111/cgf.14573 4

[8] K. Kafle, B. Price, S. Cohen, and C. Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018. 1

[9] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015. 1

[10] D. H. Kim, E. Hoque, and M. Agrawala. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020. 1, 2, 4

[11] K. Lee, M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds., *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, pp. 18893–18912. PMLR, 23–29 Jul 2023. 1, 2, 3

[12] J. Luo, Z. Li, J. Wang, and C.-Y. Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1917–1925, January 2021. 1

[13] Y. Luo, N. Tang, G. Li, C. Chai, W. Li, and X. Qin. Synthesizing natural language to visualization (nl2vis) benchmarks from nl2sql benchmarks. In *Proceedings of the 2021 International Conference on Management of Data*, pp. 1235–1247, 2021. 1, 2, 3

[14] A. Masry, D. Long, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, Dublin, Ireland, May 2022. doi: 10.18653/v1/2022.findings-acl.177 1

[15] J. Obeid and E. Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 138–147. Association for Computational Linguistics, Dublin, Ireland, Dec. 2020. 2, 4

[16] OpenAI. Gpt-4 technical report, 2023. 1, 2

[17] P. Pasupat and P. Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, 2015. 1

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 1, 3

[19] A. Roberts, C. Raffel, and N. Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, 2020. 2

[20] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2017. doi: 10.1109/TVCG.2016.2599030 1

[21] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[22] C. Stokes, V. Setlur, B. Cogley, A. Satyanarayan, and M. A. Hearst. Striking a balance: Reader takeaways and preferences when integrating text and charts. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1233–1243, 2023. doi: 10.1109/TVCG.2022.3209383 1

[23] B. Tang, A. Boggust, and A. Satyanarayan. VisText: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7268–7298. Association for Computational Linguistics, Toronto, Canada, July 2023. 1, 2

[24] H. Voigt, Ö. Alaçam, M. Meuschke, K. Lawonn, and S. Zarrieß. The why and the how: A survey on natural language interaction in visualization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 348–374, 2022. 1

[25] S. Wiseman, S. Shieber, and A. Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2253–2263. Association for Computational Linguistics, Copenhagen, Denmark, Sept. 2017. doi: 10.18653/v1/D17-1239 3

[26] C. Xiong, V. Setlur, B. Bach, E. Koh, K. Lin, and S. Franconeri. Visual arrangements of bar charts influence comparisons in viewer takeaways. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):955–965, 2021. 2