Charting the Future: Using Chart Question-Answering for Scalable Evaluation of LLM-Driven Data Visualizations

James Ford, Xingmeng Zhao, Dan Schumacher, and Anthony Rios
Department of Information Systems and Cyber Security
The University of Texas at San Antonio
{james.ford, anthony.rios}@utsa.edu

Abstract

We propose a novel framework that leverages Visual Question Answering (VQA) models to automate the evaluation of LLM-generated data visualizations. Traditional evaluation methods often rely on human judgment, which is costly and unscalable, or focus solely on data accuracy, neglecting the effectiveness of visual communication. By employing VQA models, we assess data representation quality and the general communicative clarity of charts. Experiments were conducted using two leading VQA benchmark datasets, ChartQA and PlotQA, with visualizations generated by OpenAI's GPT-3.5 Turbo and Meta's Llama 3.1 70B-Instruct models. Our results indicate that LLM-generated charts do not match the accuracy of the original non-LLM-generated charts based on VQA performance measures. Moreover, while our results demonstrate that few-shot prompting significantly boosts the accuracy of chart generation, considerable progress remains to be made before LLMs can fully match the precision of human-generated graphs. This underscores the importance of our work, which expedites the research process by enabling rapid iteration without the need for human annotation, thus accelerating advancements in this field.

1 Introduction

Data analytics is integral to modern organizations, enabling informed decision-making by interpreting complex datasets. Effective data visualization transforms vast amounts of information into actionable insights, but the increasing volume and complexity of data can overwhelm organizational staff. Many individuals lack the technical skills needed to generate meaningful visualizations, creating a barrier between data availability and utilization.

Recent advancements have seen Large Language Models (LLMs) applied to data visualization tasks, allowing users to create visual representations through natural language queries (Masry et al., 2022; Methani et al., 2020). While this progress is promising, evaluating the quality of LLM-generated visualizations poses significant challenges. Traditional evaluation methods rely heavily on human judgment to assess data accuracy and the effectiveness of visual communication. This dependence on human evaluators is not only costly but also impractical for scaling across large datasets or diverse visualization types. Subjectivity and variability in human assessments further complicate consistent benchmarking and continuous system development.

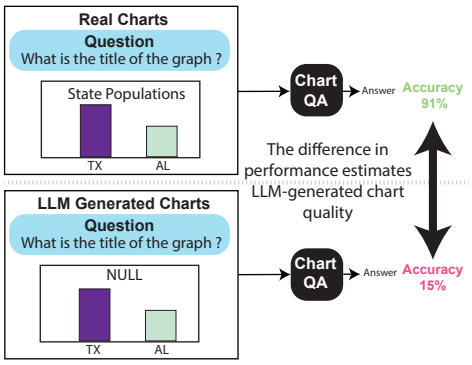


Figure 1: An overview of the VQA evaluation process for generated visualizations. The Chart QA blocks represent trained models for chart question-answering.

An alternative evaluation approach involves regenerating the input data from the visualization and comparing it to the original dataset. Although this method checks for data representation accuracy, it overlooks critical factors such as visual design clarity, interpretability, and the visualization's ability to highlight key insights (Tenney et al., 2020). For example, a chart might accurately reflect the underlying data but fail to convey the intended message due to poor color schemes, misleading labels, or cluttered layouts. Such issues would remain undetected through data regeneration methods, as they focus solely on data fidelity rather than the effectiveness of information communication.

To help address these challenges in evaluating

Proceedings of the 31st International Conference on Computational Linguistics, pages 7497–7510
January 19–24, 2025.

visualizations, we present Visual Question Answering (VQA) as a more comprehensive solution. VQA models can assess both the data and design aspects of a chart by answering questions about its content, offering a deeper evaluation of the visualization's effectiveness (Liu et al., 2023; Masry et al., 2023). For example, a VQA model might respond to queries like "What trend is depicted in this chart?," "Which category has the highest value?," or "What is the second largest bar?" These questions assess user interpretation and the visualization's communicative success, going beyond mere data verification. If we can accurately predict the answers to these and similar questions, then the chart should be adequate. This approach aligns closely with the ultimate goal of visualizations: to effectively communicate insights.

Moreover, VQA enables automated evaluation at scale, reducing reliance on human evaluators and facilitating large-scale assessments across various visualization types. VQA provides an evaluation method that mirrors real-world user interaction with visual content by focusing on interpretability and design aspects. An illustration of the VQA process is shown in Figure 1. Specifically, in Figure 1, we provide an example of how our VQA-based evaluation framework distinguishes between real and LLM-generated visualizations. The VQA model answers a question about the title of a graph depicting state populations for Texas and Alabama. In the case of a real chart, the model is able to correctly identify the title, "State Populations," with 91% accuracy. However, when evaluating an LLM-generated chart, which lacks a title (denoted as "NULL") because of issues with the chart generation process, the accuracy drops to 15%. The performance difference serves as an indicator of the LLM-generated chart's quality. This process highlights how our approach can effectively measure not only data accuracy but also the communicative clarity of LLM-generated visualizations. Overall, with our system, users could develop new chart generation methods and then rapidly iterate them using the VQA accuracy results we propose.

In this paper, we propose a novel framework that leverages chart-based VQA models to automate the evaluation of LLM-generated visualizations. Our contributions are threefold:

1. We develop an automated framework capable of scaling the evaluation process for chart generation, enabling efficient benchmarking across multiple models and datasets.

2. We demonstrate that VQA models offer unique, context-sensitive feedback compared to data regeneration methods, providing a holistic assessment of both data accuracy and visual communication effectiveness.

3. We present empirical results that showcase the effectiveness of our approach in large-scale evaluations, comparing multiple LLMs using different prompting strategies.

2 Related Work

Data visualization is crucial for interpreting complex information, yet effective visualizations often requires technical expertise (Srinivasan and Stasko, 2017; Dibia, 2023). Text-to-Viz systems (T2V) aim to simplify this process by enabling users to generate visualizations through natural language queries (Shen et al., 2023; Zhang et al., 2024a). This approach addresses challenges non-technical users face, such as steep learning curves and difficulty selecting appropriate visualization methods (Kavaz et al., 2023). Complete T2V typically consists of two components: data querying and visualization (Affolter et al., 2019; Zhang et al., 2024a). While our focus is on the visualization aspect, the process involves multiple steps, including parsing the input query, identifying data attributes, and choosing appropriate visualization styles (Maddigan and Susnjak, 2023). Commercial tools like Tableau and Microsoft Power BI have incorporated limited T2V capabilities (Maddigan and Susnjak, 2023).

Challenges for T2V include ambiguity and under-specification in natural language, and users often overestimate the system's capabilities (Tory and Setlur, 2019). Early T2V relied on rule-based or template-based methods with shallow parsing techniques (Shen et al., 2023; Zhang et al., 2024a), which struggled with complex data and offered limited efficacy (Hong and Crisan, 2023). Advances in Natural Language Processing and deep learning introduced more robust models, such as sequence-to-sequence architectures and pre-trained language models like BERT and T5, improving complexity and robustness but requiring extensive training data (Zhang et al., 2024a; Tian et al., 2024; Voigt et al., 2023). For example, FLAN-T5 has been used to create Vega-Lite specifications for data visualizations (Voigt et al., 2023; Tian et al., 2024). Toolkits

like NL4DV facilitated the creation of Vega-Lite specifications from natural language (Narechania et al., 2021). Interactive systems combining text and speech inputs have also been developed (Srinivasan et al., 2020; Chowdhury et al., 2021).

LLMs have led to new approaches in generating data visualizations (Dibia, 2023; Maddigan and Susnjak, 2023). LLMs can produce visualization code or images directly from user queries without extensive pre-training (Hong and Crisan, 2023). Research has shown that LLMs can match the performance of human data analysts (Cheng et al., 2023), though challenges remain (Zhang et al., 2024b; Gu et al., 2024). OpenAI's Codex has been used to generate visualization interfaces (Chen et al., 2022), and GPT models have been employed to create Python scripts (Maddigan and Susnjak, 2023). Techniques like prompt engineering and chain-of-thought processes have enhanced LLM-generated visualizations (Li et al., 2024; Podo et al., 2024a).

Evaluating LLM-generated visualizations is challenging. Traditional methods include comparing generated code to ground-truth specifications, comparing visual outputs, and user satisfaction ratings (Zhang et al., 2024a). However, these often rely on human judgment, which is subjective, costly, and hard to scale (Maddigan and Susnjak, 2023; Zhao et al., 2024; Srinivasan and Stasko, 2020). Automated efforts primarily focus on checking the syntax of generated code (Dibia and Demiralp, 2019), but code similarity may not be a good indicator of visualization quality (Chen et al., 2024). Some frameworks attempt to automate parts of the evaluation but still require human involvement (Chen et al., 2023; Podo et al., 2024b; Chen et al., 2024). Moreover, LLM-generated visualizations can be prone to hallucinations and inconsistencies (Podo et al., 2024b; Tian et al., 2024; Maddigan and Susnjak, 2023), and may depend heavily on user interventions for quality control (Tao and Xu, 2023). Research suggests that human feedback is preferred over LLM feedback in evaluation (Kim et al., 2024).

VQA on charts offers a promising alternative for automated evaluation chart generation quality instead of using human evaluation only. VQA models answer questions about data and design elements, providing a richer assessment of a chart's effectiveness (Liu et al., 2023; Masry et al., 2023). Two main approaches are used: converting charts into underlying tables for text analysis (Han et al., 2023), and using multimodal models to derive answers directly from the chart (Meng et al., 2024). Datasets like ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020) have been instrumental in advancing VQA research. Models such as Unichart (Masry et al., 2023) and MatCha (Liu et al., 2023) have achieved high performance on these datasets.

Building on these advancements, our work leverages chart-based VQA (in contrast with general image VQA) models to automate the evaluation of LLM-generated visualizations. By focusing on both data accuracy and visual communication effectiveness, we address the limitations of previous evaluation methods and enable scalable, automated assessment without human intervention.

Dataset	Charts	Question-Answer Pairs
ChartQA	734	1,113
PlotQA	400	14,427

Table 1: Dataset sample statistics.

3 Data

Our study uses subsets from two datasets: PlotQA (Methani et al., 2020) and ChartQA (Masry et al., 2022). The statistics for the subsets of both datasets used in our study are shown in Table 1.

PlotQA. The PlotQA dataset (Methani et al., 2020) is a large-scale benchmark for visual question answering (VQA) over scientific plots, comprising over 224,000 plots and approximately 28.9 million question-answer pairs derived from real-world data sources. It challenges models to interpret complex data visualizations by categorizing questions into three main types: *Structural Understanding*, which involve questions about the overall structure of the plot without requiring quantitative reasoning (e.g., the presence of grid lines or legend labels); *Data Retrieval*, which require extracting specific data values directly from the plot (e.g., reading exact numerical values or labels); and *Reasoning*, which involve numerical reasoning over multiple plot elements or comparative analysis (e.g., performing arithmetic calculations, comparisons, or interpreting trends). This diversity of question types necessitates models that can handle complex reasoning tasks, precise data extraction, and an understanding of structural nuances in various plot types, making PlotQA a challenging and comprehensive dataset for evaluating the reasoning capabilities of models in the context of data visualizations. We sample

14,427 QA pairs from 400 charts because of commercial API costs for chart generation and computational expenses.

ChartQA. ChartQA is a comprehensive benchmark dataset designed to advance research in question answering over data visualizations, specifically focusing on charts like bar graphs, line graphs, and pie charts. It comprises 4,804 charts with 9,608 human-authored question-answer pairs (ChartQA-H) and 17,141 charts with 23,111 machine-generated question-answer pairs (ChartQA-M), resulting in a total of 21,945 charts and 32,719 questions. The charts are collected from diverse real-world sources such as Statista, Pew Research, Our World in Data, and the OECD, ensuring various chart styles, topics, and data representations. ChartQA emphasizes complex reasoning tasks that require models to perform multiple logical and arithmetic operations and to handle open-vocabulary answers derived from chart data, rather than selecting from a fixed set. Many questions also involve visual reasoning, referring to specific visual attributes like color, size, or position of chart elements. Because of the cost of testing commercial models, we sample 734 charts to have around 1,100 question-answer pairs.

4 Methodology

We provide a high-level overview of our paper in Figure 2. Overall, our framework has four main components. First, we generate charts using two popular LLMs for both datasets used in our study. Second, we benchmark the chart generation quality of the LLMs using the chart question-answering task. Third, we manually review all contrasting errors (i.e., errors made by one model, but not the other) to ensure the question-answering task is working as expected. Fourth, we perform a small survey study where we have participants review charts and then we compare the results with the benchmarking from Step 2. Steps 1 and 2 are the evaluation framework we are proposing. Steps 3 and 4 are how we evaluate whether the VQA results accurately measure chart quality.

Step 1: Baseline Methods and Dataset Preparation.

In this step, we use two LLMs to generate data visualizations from a prompt. Specifically, we use OpenAI's GPT-3.5 Turbo-0125 and NeuralMagic's Meta Llama 3.1 70B-Instruct-FP8 models (Magic, 2024) to produce Python `matplotlib` code that creates charts based on provided datasets. Our focus is on designing effective prompts for these LLMs using zero-shot and few-shot prompting strategies.

In the zero-shot setting, the LLMs receive only the task instructions and the data without any examples. The system prompt instructs the model to generate Python code for data visualization:

System Instructions

You are a data analyst tasked with creating data visualization plots based on the provided data. Output should be formatted as Python `matplotlib` *code and must include both* `fig.clf()` *and the* `bbox_inches='tight'` *parameter. Use the specified title, chart type, and data for the axis labels and counts. Do not use* `list(range)`*. Ensure data value labels are on the chart, place legends outside the chart, and save the chart as a PNG file using the specified filename.*

The data prompt then provides the specific chart details

Input Query

`data_pass` = *"Title:* `title_text` */ Data:* `final_string` */ Chart type:* `figure_type` */ File Name:* `png_file`*"*

where `title_text` is the chart title, `final_string` contains the data in text format, `figure_type` specifies the chart type (e.g., bar, line), and `png_file` is the desired output file name. This prompt directs the LLM to generate the appropriate Python code to create the chart as per the given specifications. In the few-shot setting, the system prompt is mentioned once, and the data prompt (Input Query) is repeated for each of the in-context examples.

Step 2: Benchmarking Chart Quality using Question-Answering.

We used two VQA models to perform the question-answering task. Masry et al. (2023) developed UniChart, a model with two modules (a chart encoder based on the document image Donut model (Kim et al., 2022), plus a text decoder based on a BART model (Lewis, 2019)) pretrained on more than 600,000 charts, optimized explicitly for the ChartQA dataset. UniChart processes a plot

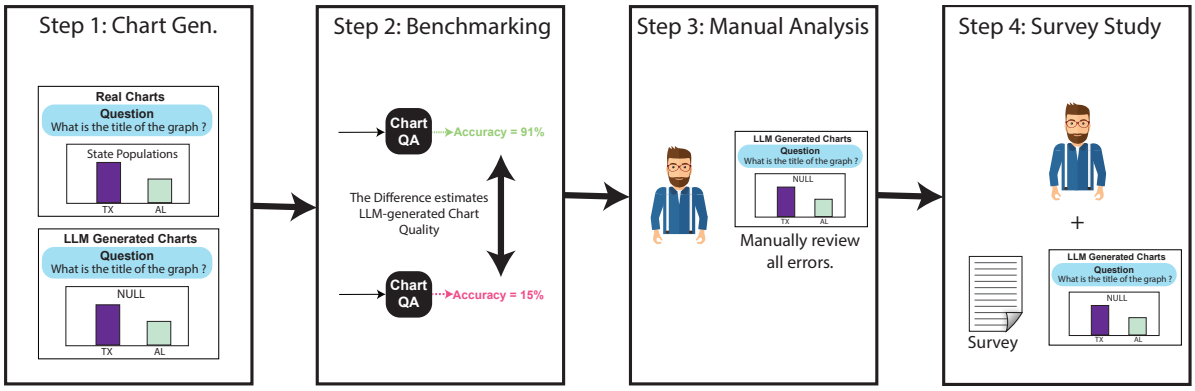


Figure 2: Overall framework for our study, where we perform automatic chart generation, benchmarking using chart question answering, manually analyze errors, and perform a survey on chart quality.

image by first encoding its textual elements (such as legends and axis labels), visual elements (like bars and lines), and the overall layout. It then decodes this information to generate answers based on the content of the plot. Liu et al. (2023) created MatCha, adding chart derendering and mathematical reasoning pretraining to the Pix2Struct vision-language model (Lee et al., 2023). Unichart and Matcha were chosen because they both reported high performance on VQA with the ChartQA and PlotQA datasets, respectively.

The VQA models are applied to the charts generated in Step 1 in order to derive listings of generated answers which were then assessed against the ground truth answers from the two datasets. Accuracy is defined as the answer from the VQA models matching the ground truth in the dataset. For text responses, strict accuracy was employed but relaxed accuracy (Masry et al., 2022) was used for mathematical and numerical responses. Relaxed accuracy is operationalized as accepting numerical values that are within plus or minus five percent.

Step 3: Manual Quality Analysis. We manually reviewed the errors made exclusively by GPT-3.5 Turbo and Llama 3.1 to ensure that our VQA evaluation accurately reflects the quality of the charts generated by these models. Using a qualitative coding framework, we categorized these errors to determine whether discrepancies were due to issues in the charts or unrelated factors. This is important because if the VQA errors aren't caused by issues with the chart quality, then using the VQA task to assess the quality of the charts wouldn't be a valid evaluation method.

In our analysis, we identify instances where the VQA model marked answers incorrect for charts from one LLM but correct for the other. We examined each case, categorizing errors as either visualization errors—such as incorrect data representation, mislabeled axes, missing titles, or overlapping labels—or errors due to the VQA model itself, like misinterpretation or ambiguous questions. Because VQA models are not perfect, we expect those issues to impact both models similarly.

By categorizing the errors, we assess whether the VQA task effectively measures chart quality. If most errors stem from visualization issues in the charts, it confirms that VQA is a valid assessment tool. For example, if GPT-3.5 Turbo's chart had poorly presented labels that led the VQA model to misinterpret data, while Llama's chart didn't, the issue would stem from chart quality. This shows that the VQA task reflects chart performance, while also helping us identify the types of errors our framework detects.

Step 4: Surveys. In this paper, we focus on assessing the quality of charts generated by large language models (LLMs) when converting structured data into visualizations. To evaluate these charts, we designed a set of questions for human participants, aligning them with the question types used in the PlotQA dataset (Methani et al., 2020). The PlotQA dataset categorizes questions into three main types: *Structural Understanding*, *Data Retrieval*, and *Reasoning*. We aim to draw correlations between human assessments and automated evaluation metrics by mapping our human evaluation questions to these types.

The following questions guide our human evaluation, where participants assess the accuracy, readability, and overall usefulness of these charts:

***Q1: The LLM-generated chart accurately displays a title reflecting the data depicted in the original data file.*(Structural Understanding)** This ques-

tion corresponds to the *Structural Understanding* type in PlotQA. The title of a chart serves as a crucial summary of the data it represents. This question evaluates whether the LLM can produce a chart title that accurately reflects the underlying data, ensuring that viewers can quickly grasp the content of the visualization.

Q2: The X-axis labels on the LLM-generated chart accurately display the labels depicted in the original data file. **(Structural Understanding)** Aligned with *Structural Understanding* in PlotQA, the X-axis often represents categories or time intervals in visualizations. This question focuses on the LLM's ability to correctly generate X-axis labels that are faithful to the labels present in the data file, ensuring correct interpretation of the chart's horizontal dimension.

Q3: The Y-axis labels on the LLM-generated chart accurately display the labels depicted in the original data file. **(Structural Understanding)** Also a *Structural Understanding* question type, the Y-axis typically corresponds to numerical values or other measurements. This question evaluates whether the Y-axis labels in the LLM-generated chart match the values in the data file. Accurate Y-axis labels are essential for interpreting data magnitude and making comparisons across categories.

Q4: The data points on the LLM-generated chart accurately display the values depicted in the original CSV data file. **(Data Retrieval)** This question maps to the *Data Retrieval* type in PlotQA. Data point accuracy is critical to ensure the visualization faithfully represents the dataset. It measures whether the individual data points (e.g., bars, lines, or dots) on the chart match the corresponding values in the CSV file, maintaining the integrity of the visualized data.

Q5: The LLM-generated chart is easy to read and understand. **(Reasoning)** This question relates to the *Reasoning* type in PlotQA, specifically concerning the user's ability to perform reasoning tasks using the chart. Even if a chart is accurate, it must also be easy for users to interpret. This question evaluates the chart's readability, including clarity of labels, appropriate scaling, and overall visual design. Readability ensures that users can extract insights and perform complex reasoning without unnecessary cognitive effort.

Q6: Overall, the LLM-generated chart serves its intended purpose in a satisfactory manner. **(Comprehensive Assessment)** This question encompasses all three PlotQA question types—*Structural Understanding*, *Data Retrieval*, and *Reasoning*. It assesses the overall effectiveness of the chart in fulfilling its intended purpose, whether that be conveying specific trends, making comparisons, or summarizing key data points. Participants evaluate whether the chart helps them interpret and draw meaningful conclusions from the data.

Each of these questions is measured using a 5-point Likert scale, where participants rate their level of agreement with each statement (1 = Strongly Disagree, 5 = Strongly Agree). By mapping our evaluation questions to the PlotQA question types, we aim to correlate human ratings with automated evaluation metrics, providing a comprehensive understanding of how well LLM-generated charts perform in practical scenarios.

5 Results

Implementation Details. We used the Neural-Magic's Meta LLaMA 3.1 70B-Instruct-FP8 models (Magic, 2024) for our prompting-based experiments, running it on two NVIDIA A6000 GPUs. To generate outputs, we set the sampling parameters with a temperature of 0.1, top-p of 0.9, and a maximum token limit of 2,000 to balance coherence and diversity in the model's responses. All experiments were conducted using PyTorch (Paszke et al., 2019) and the Hugging Face Transformers library (Wolf, 2019). We only used three few-shot examples in our experiments.

Comparison Metrics. We also measure the accuracy of the charts by extracting the data depicted in them. Two metrics were used to gauge the accuracy of the extracted data, comparing the extractions with the ground truth data files from the datasets. *Similarity Score* uses the normalized distance between the extracted values from each LLM-generated chart and the ground truth data files. This is calculated as the absolute difference between x_1 and x_2, divided by (x_1 + $1e^{-15}$), where x_1 is the ground truth data and x_2 is the extracted data. The second method, *Exact Match*, computes the percentage of charts whose extracted data tables are exact matches with the ground truth tables. The Python pandas function "exact" was applied to each column to verify that the columns were identical (which checks that items match exactly and are in the correct order) and is thus a more challenging

Model	Overall	Human (n=390)	Augmented (n=723)	Sim. Score	Exact Match
Original Charts	64.4	35.4	80.1	97.8	70.6
GPT3.5 Zero-Shot	42.3	21.5	53.5	97.0	30.8
Llama 3.1 Zero-Shot	51.1	34.0	60.9	95.3	33.1
GPT3.5 Few-Shot	44.2	23.7	55.2	97.2	43.5
Llama 3.1 Few-Shot	58.0	37.1	69.9	93.1	35.8

Table 2: VQA Results from ChartQA Dataset. The UniChart VQA model was used for these experiments.

Model	Overall	Arithmetic (n=1,789)	Comparison (n=1,044)	Compound (n=2,311)	Data-Retrieval (n=3,352)	Min-Max (n=1,146)	Structural (n=4,485)	Sim. Score	Exact Match
Original Charts	80.6	43.6	73.2	71.2	88.3	94.1	92.0	92.2	16.0
GPT3.5 Zero-Shot	60.5	22.7	61.3	54.3	59.6	76.8	74.2	92.5	49.3
Llama 3.1 Zero-Shot	54.6	16.9	51.5	49.9	57.2	65.9	67.0	82.7	25.0
GPT3.5 Few-Shot	62.4	23.8	63.9	54.6	61.9	80.7	75.9	92.3	37.0
Llama 3.1 Few-Shot	61.9	23.9	63.8	54.2	62.4	79.7	74.6	93.1	35.8

Table 3: VQA Results from PlotQA Dataset. The MatCha VQA model was used for these experiments.

Error Type	GPT3.5 Errors	Llama Errors
VQA Model Error	38 (22.89%)	6 (15.79%)
Actually Correct	2 (1.20%)	1 (2.63%)
Bar Boundaries	2 (1.20%)	0 (0.00%)
Category Ambiguous	10 (6.02%)	0 (0.00%)
Colors Not Matching	6 (3.61%)	1 (2.63%)
Dates Errors	49 (29.52%)	11 (28.95%)
Labels Overlapping	55 (33.13%)	12 (31.58%)
Wrong Type of Bars	3 (1.81%)	7 (18.42%)
Chart Not Displaying	1 (0.60%)	0 (0.00%)

Table 4: Visualization Error Mismatches Between ChatGPT and Llama

metric than the first method.

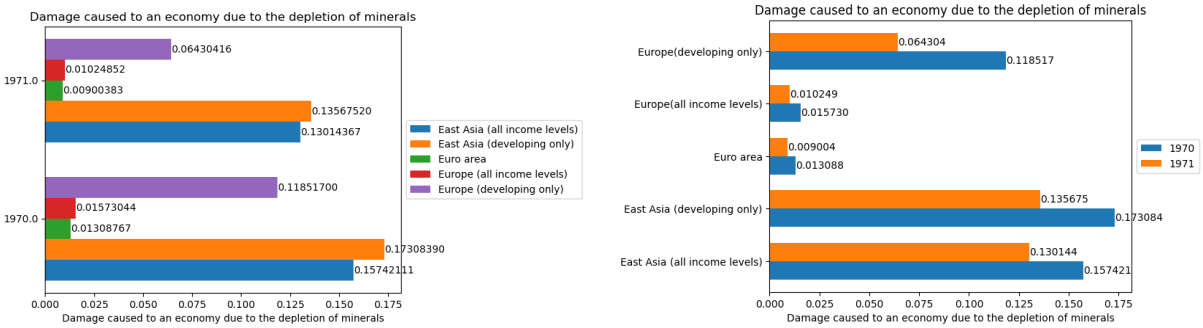
Benchmark Results. We report the ChartQA benchmark results in Table 2. Overall, we make four major findings. First, all performance metrics are lower on the generated charts than on the original charts. At least at a superficial level, this is useful to know that LLMs struggle to generate human-quality charts. The best performance overall performance was by the Llama 3.1 70B model, which achieved an accuracy of 58.0., which is more than 6% lower than the original scores of 64.4. Second, we find that few-shot methods outperform zero-shot methods, e.g., Llama 3.1 70B improves from 51.1 to 58.0. Third, we make the interesting observation that the performance is much lower on the human-generated charts (Human (n=390)) than on the automatically-generated charts (Augmented (n=723)). The performance is lower on the original charts and on the LLM-generated charts. Intuitively, this is because human-generated charts generally require much more complex visualization strategies, which makes it more difficult to extract relevant information from VQA models. Hence, the automatically generated charts give a much better performance estimate (chart quality). This can be seen by the little or no differences between the original charts and generated charts for the Human column. As the VQA models improve, human-generated charts will be an important testbed, but because VQA models do not perform well on these examples, automatically generated charts should be the focus. Moreover, there is still a huge gap between the original chart accuracy (80.1) and the best LLM charts (69.9). Fourth, we find that the data regeneration methods only provide a limited view on model performance. With Similarity Score (Sim. Score), all methods perform similarly, with little room for improvement. Moreover, the Exact Match score does not rank models correctly (the best model is GPT3.5 Few-Shot (this finding is validated in the error analysis) because of its sensitivity to order and scale.

In Table 3, we report the benchmark results on the PlotQA dataset. We make similar findings as the ChartQA dataset, e.g., Few-shot methods outperform Zero-shot. But, on this dataset, performance is similar for both Llama 3.1 70B and GPT3.5. This dataset is synthetic, so overall performance is high for the original charts, making it a good benchmark. Moreover, it has questions which are subcategorized. For instance, we find a substantial drop in the data-retrieval questions on the LLM-generated charts. We also find that the data extraction results are all over the place on this dataset, thus not providing a good evaluation metric (e.g., exact match is worse on the original charts).

Error Analysis. In Table 4, we report the results of

Question	GPT3.5	Llama 3.1
The LLM-generated chart accurately displays a title reflecting the data depicted in the original CSV data file.	4.82	4.88
The X-axis labels on the LLM-generated chart accurately displays the labels depicted in the original CSV data file.	4.28	4.15
The Y-axis labels on the LLM-generated chart accurately displays the labels depicted in the original CSV data file.	4.80	4.91
The data points on the LLM-generated chart accurately displays the values depicted in the original CSV data file.	4.85	4.75
The LLM-generated chart is easy to read and understand.	3.79	3.95
The LLM-generated chart is visually appealing.	3.68	3.68
Overall, the LLM-generated chart serves its intended purpose in a satisfactory manner.	3.85	3.87

Table 5: Human Evaluation Results: ChartQA Few-Shot



(a) GPT3.5

(b) Llama

Figure 3: Examples of Differences in Chart Generation

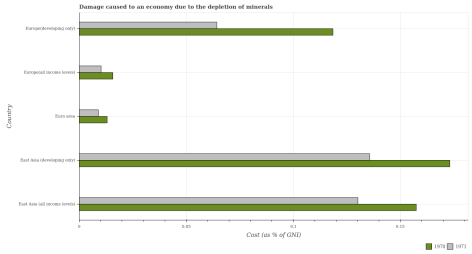


Figure 4: Original Ground Truth chart

the manual error analysis of our study. In this study, we focus on the ChartQA dataset errors. We manually reviewed errors made by one model but not the other. These errors were categorized into nine error categories. For example, the VQA Model Error category represents errors where the figures look relatively correct, but the model still did not generate a correct output for some reason. We make two major findings. First, VQA errors are really only a small proportion (based on percentage) of the total errors, and the proportion is relatively similar for both GPT3.5 and Llama errors. For all other errors, we could hypothesize visualization-related issues that caused the errors. Date errors generally result in weird date displays, e.g., 2024.00. Second, the number of errors made by GPT3.5 is larger than Llama, even after accounting for VQA Model Errors. This suggests that the finding that Llama 70B outperforms GPT3.5 on the ChartQA dataset is a robust and correct finding. **See the Appendix** for examples of the charts and error types.

In Figure 3 and 4, we also show example charts and generations from the PlotQA dataset. Overall, we found that many of the errors in this dataset were caused by issues with the headings not matching the original figure. There are some cases where Llama does a better job at handing the headers and orienting the figures correctly, as shown in the example. However, the differences between both models were minimal.

Human Study. In Table 5, we report the results of our human study. Specifically, two participants answered survey questions about chart quality (see Methodology section for details). The participants looked at a total of 200 charts each, 100 Llama Few-shot charts and 100 GPT3.5 few-shot charts. The correlation between participants for the average ratings for GPT3.5 and Llama on each question was .793 and .853, respectively. As previously mentioned, we categorized these questions into the question types used in the PlotQA dataset. Here we find that the reasoning questions, e.g., "The LLM-generated chart is easy to read and understand" perform worse than data retrieval and structural under-

standing of the charts. These findings match the findings from the PlotQA dataset (e.g., Arithmetic or Compound Questions vs. Data-Retrieval, Min-Max and Structural questions). It is important to note that the participants did not review the original charts, yet the results show the power of the VQA evaluation strategy.

Discussion. Our findings underscore both the potential and current limitations of using chart-based question-answering (QA) models to evaluate LLM-generated visualizations automatically. While integrating VQA models into an evaluation framework enables scalable and nuanced assessment, our analyses reveal that certain areas merit deeper exploration and refinement.

Curating More Targeted Questions. One avenue for future work involves the careful curation of question sets that better capture the full spectrum of chart quality. Although our current framework demonstrated that VQA models can detect issues involving labeling, alignment, and representation, it relied heavily on pre-existing QA datasets not specifically tailored for LLM-generated visualizations. This limits the types of evaluation we can do with this approach. Future research should explore more targeted and refined sets of questions that explicitly probe each critical aspect of chart design. For instance, specialized questions could be developed to evaluate whether the chosen color palettes sufficiently distinguish between categories, assess whether scale and tick mark increments are appropriate, or test the clarity of legends and annotations. By tailoring questions to common error patterns found in LLM-generated charts, it will be possible to localize and quantify specific charting issues more accurately. We could even develop questions that target specific attributes (e.g., labels) multiple times to have a more robust understanding of where the charts need improvement. A single question can be noisy, and it is unclear if the error is random or showing a real chart issue. This approach can also help disentangle errors stemming from underlying data representation, stylistic interpretation, or visual design choices. As QA models evolve, the targeted question sets could serve as benchmarks highlighting subtle differences in chart clarity and effectiveness.

Enhancing Accessibility Through Chart QA. Another promising direction for future work lies in using chart QA models to improve accessibility, particularly for individuals with visual impairments. While improving chart quality benefits all users, QA-driven chart understanding can serve as a mechanism for delivering alternative representations of visual data. For example, a user who cannot see the chart directly could still benefit from a QA model capable of accurately answering queries such as "*Which bar represents the highest value?*," "*What does the title and subtitle indicate?*," or "*Does the chart show a trend over time, and if so what type of trend?*" This information could be integrated into screen reader technologies or chat-based interfaces, enabling visually impaired users to obtain meaningful insights from complex data visualizations. Instead of relying on simply generating better summarization and chart QA tools, we can improve charts to perform better with chart QA models.

6 Conclusion

We introduced a framework that leverages Visual Question Answering (VQA) models to automate the evaluation of data visualizations generated by Large Language Models (LLMs). This approach addresses the limitations of traditional methods by providing a scalable assessment of both data fidelity and communicative effectiveness. Experiments on the ChartQA and PlotQA datasets revealed that while current LLMs like GPT-3.5 Turbo and Llama 3.1 70B-Instruct do not yet match the accuracy of original non-LLM-generated charts, few-shot prompting significantly enhances their performance. Our error analysis and human study confirmed that VQA models effectively reflect chart quality by capturing visualization issues inherent in LLM-generated charts. This work enables efficient benchmarking and continuous improvement of LLM-driven data visualization systems. There are two major future research directions. First, we can explore how the VQA model performance impacts chart quality estimates. Better aligning and designing the questions that are easy for VQA models from the bottom can make the evaluation more robust. Second, design (e.g., visual quality) is important, and incorporating a chain-of-thought (CoT) structure to improve reasoning, yet completely unexplored here.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

Limitations

While our proposed framework provides a scalable method for evaluating LLM-generated data visualizations using VQA models, it has several limitations that warrant discussion.

First, the reliability of our evaluation is inherently tied to the accuracy of the VQA models employed. Current VQA models may not fully capture all the nuances of chart interpretation, especially when dealing with complex visual elements or unconventional chart types. This dependency means that any limitations in VQA accuracy could directly impact the validity of our assessment of the LLM-generated charts.

Second, due to computational and resource constraints, our experiments utilized subsets of the ChartQA and PlotQA datasets. Although these subsets provided valuable insights, they may not fully represent the diversity and complexity of real-world data visualizations. Expanding the scope of our evaluation to include more extensive portions of these datasets—or additional datasets altogether—could enhance the robustness of our findings and provide a more comprehensive understanding of LLM performance in generating accurate and effective visualizations.

Third, the questions used in our evaluation were derived from existing datasets and may not be perfectly aligned with the specific visual aspects of the charts generated by the LLMs. This misalignment could lead to an incomplete or skewed assessment of chart quality. Future work should focus on developing more targeted questions that are specifically designed to evaluate particular visual attributes or design elements of the generated charts. By aligning question types more closely with the visual features being assessed, we can obtain more precise estimates of chart quality and better identify areas where LLMs excel or require improvement.

Addressing these limitations will be crucial for refining our evaluation framework and enhancing its applicability. Improvements in VQA model accuracy, the use of larger and more diverse datasets, and the development of tailored evaluation questions will collectively contribute to a more robust and insightful assessment of LLM-generated data visualizations.

References

Katrin Affolter, Kurt Stockinger, and Abraham Bernstein. 2019. A comparative survey of recent natural language interfaces for databases. *The VLDB Journal*, 28(5):793–819.

Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. 2024. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–11.

Yiru Chen, Ryan Li, Austin Mac, Tianbao Xie, Tao Yu, and Eugene Wu. 2022. Nl2interface: Interactive visualization interface generation from natural language queries. *Preprint*, arXiv:2209.08834.

Zhutian Chen, Chenyang Zhang, Qianwen Wang, Jakob Troidl, Simon Warchol, Johanna Beyer, Nils Gehlenborg, and Hanspeter Pfister. 2023. Beyond generating code: Evaluating gpt on a data visualization course. In *2023 IEEE VIS Workshop on Visualization Education, Literacy, and Activities (EduVis)*, pages 16–21.

Liying Cheng, Xingxuan Li, and Lidong Bing. 2023. Is GPT-4 a good data analyst? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9496–9514, Singapore. Association for Computational Linguistics.

Imran Chowdhury, Abdul Moeid, Enamul Hoque, Muhammad Ashad Kabir, Md. Sabir Hossain, and Mohammad Mainul Islam. 2021. Designing and evaluating multimodal interactions for facilitating visual analysis with dashboards. *IEEE Access*, 9:60–71.

Victor Dibia. 2023. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 113–126, Toronto, Canada. Association for Computational Linguistics.

Victor Dibia and Çağatay Demiralp. 2019. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE Computer Graphics and Applications*, 39(5):33–46.

Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M. Drucker. 2024. How do analysts understand and verify ai-assisted data analyses? *Preprint*, arXiv:2309.10947.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *Preprint*, arXiv:2311.16483.

Matt-Heun Hong and Anamaria Crisan. 2023. Conversational ai threads for visualizing multidimensional datasets. *Preprint*, arXiv:2311.05590.

Ecem Kavaz, Anna Puig, and Inmaculada Rodríguez. 2023. Chatbot-based natural language interfaces for data visualisation: A scoping review. *Applied Sciences*, 13(12).

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Nam Wook Kim, Grace Myers, and Benjamin Bach. 2024. How good is chatgpt in giving advice on your visualization design? *Preprint*, arXiv:2310.09617.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Guozheng Li, Xinyu Wang, Gerile Aodeng, Shunyuan Zheng, Yu Zhang, Chuangxin Ou, Song Wang, and Chi Harold Liu. 2024. Visualization generation with large language models: An evaluation. *Preprint*, arXiv:2401.11255.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Paula Maddigan and Teo Susnjak. 2023. Chat2vis: Generating data visualizations via natural language using chatgpt, codex and gpt-3 large language models. *IEEE Access*, 11:45181–45193.

Neural Magic. 2024. Meta-llama-3-70b-instruct-fp8. `https://huggingface.co/neuralmagic/Meta-Llama-3-70B-Instruct-FP8`. Quantized version of Meta-Llama-3-70B-Instruct optimized with FP8 for efficient inference.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore. Association for Computational Linguistics.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *Preprint*, arXiv:2401.02384.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Arpit Narechania, Arjun Srinivasan, and John Stasko. 2021. Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Luca Podo, Marco Angelini, and Paola Velardi. 2024a. V-recs, a low-cost llm4vis recommender with explanations, captioning and suggestions. *Preprint*, arXiv:2406.15259.

Luca Podo, Muhammad Ishmal, and Marco Angelini. 2024b. Vi(e)va llm! a conceptual stack for evaluating and interpreting generative ai-based visualizations. *Preprint*, arXiv:2402.02167.

Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2023. Towards natural language interfaces for data visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 29(6):3121–3144.

Arjun Srinivasan, Bongshin Lee, Nathalie Henry Riche, Steven M. Drucker, and Ken Hinckley. 2020. Inchorus: Designing consistent multimodal interactions for data visualization on tablet devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Arjun Srinivasan and John Stasko. 2017. Natural language interfaces for data analysis with visualization: considering what has and could be asked. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*, EuroVis '17, page 55–59, Goslar, DEU. Eurographics Association.

Arjun Srinivasan and John Stasko. 2020. How to ask what to say?: Strategies for evaluating natural language interfaces for data visualization. *IEEE Computer Graphics and Applications*, 40(4):96–103.

Ran Tao and Jinwen Xu. 2023. Mapping with chatgpt. *ISPRS International Journal of Geo-Information*, 12(7).

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. 2024. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–15.

Melanie Tory and Vidya Setlur. 2019. Do what i mean, not what i say! design considerations for supporting intent and context in analytical conversation. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 93–103.

Henrik Voigt, Nuno Carvalhais, Monique Meuschke, Markus Reichstein, Sina Zarrie, and Kai Lawonn. 2023. VIST5: An adaptive, retrieval-augmented language model for visualization-oriented dialog. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 70–81, Singapore. Association for Computational Linguistics.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Weixu Zhang, Yifei Wang, Yuanfeng Song, Victor Junqiu Wei, Yuxing Tian, Yiyan Qi, Jonathan H. Chan, Raymond Chi-Wing Wong, and Haiqin Yang. 2024a. Natural language interfaces for tabular data querying and visualization: A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

Yuge Zhang, Qiyang Jiang, Xingyu Han, Nan Chen, Yuqing Yang, and Kan Ren. 2024b. Benchmarking data science agents. *Preprint*, arXiv:2402.17168.

Yuheng Zhao, Yixing Zhang, Yu Zhang, Xinyi Zhao, Junjie Wang, Zekai Shao, Cagatay Turkay, and Siming Chen. 2024. Leva: Using large language models to enhance visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–17.

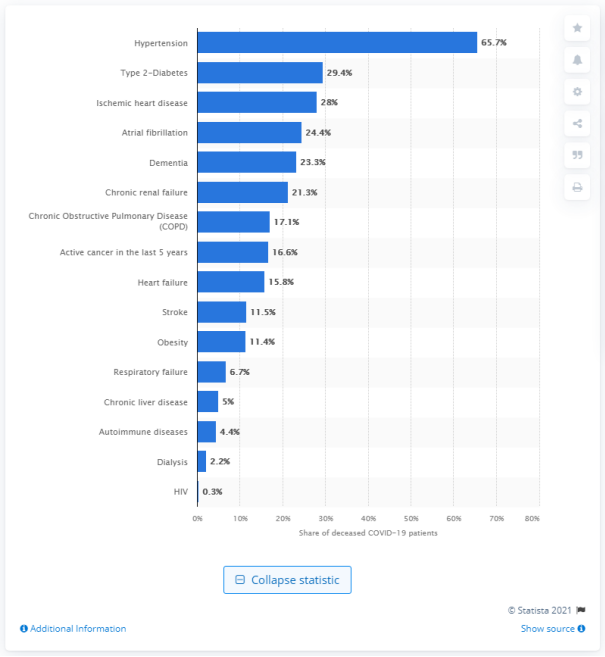
A Appendix

The following examples illustrate where the charts generated by GPT3.5 were not correctly evaluated by the VQA model while the version generated by Llama were. The major categories from Table 6 are depicted.

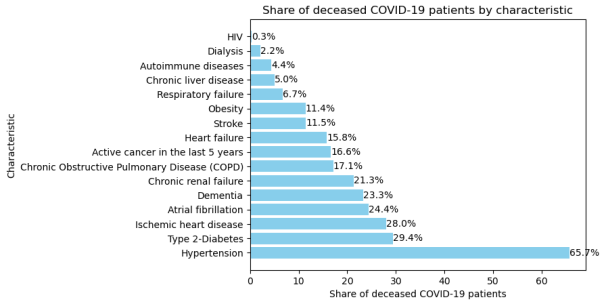
Chart Visualization Errors: Category Ambiguous

Question: What percentage of COVID-19 patients died after contracting the virus?

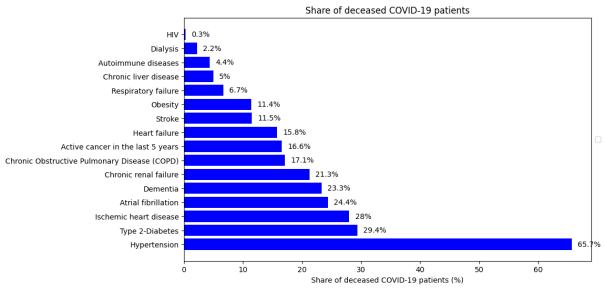
Because the question did not specifically ask for which condition led to death, the VQA model had to choose which percentage to report, as shown in Figure 5.



(a) Ground Truth



(b) GPT3.5



(c) Llama

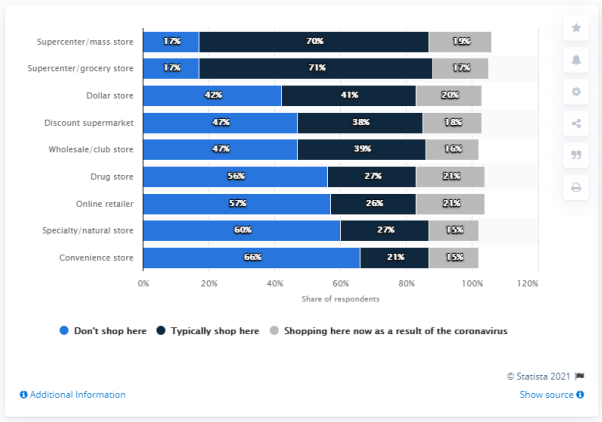
Figure 5: Chart Visualization Errors: Category Ambiguous

Chart Visualization Errors: Colors Not Match-

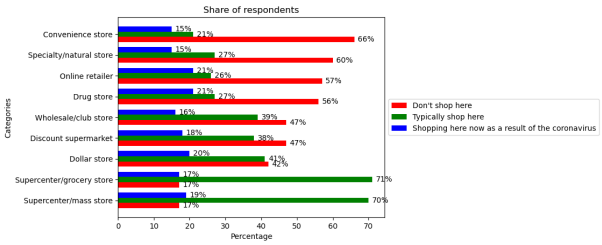
ing

Question: What percentage is represented by the navy blue bar?

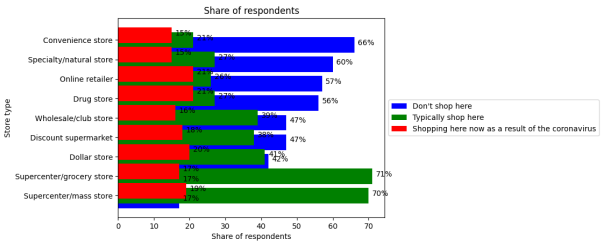
The colors used in the ground truth charts were not provided in the instructions to the LLMs, so the generated-charts did not necessarily match the original, as shown in Figure 6.



(a) Ground Truth



(b) GPT3.5



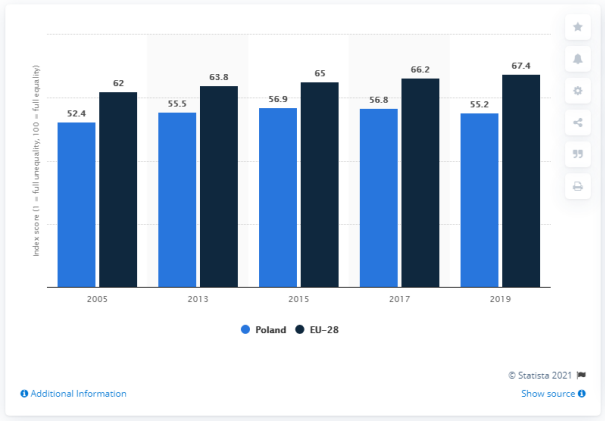
(c) Llama

Figure 6: Chart Visualization Errors: Colors Not Matching

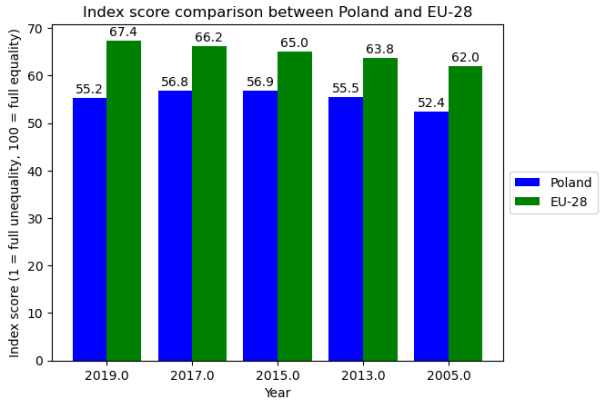
Chart Visualization Errors: Dates Errors

Question: What was the Polish gender equality index score between 2005 and 2019?

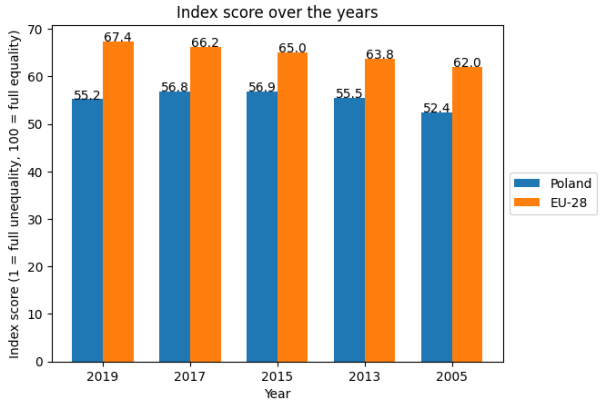
The date labels on the X-axis are not correct on the GPT3,5-generated chart, leading to incorrect answers from the VQA model, as shown in Figure 7.



(a) Ground Truth



(b) GPT3.5



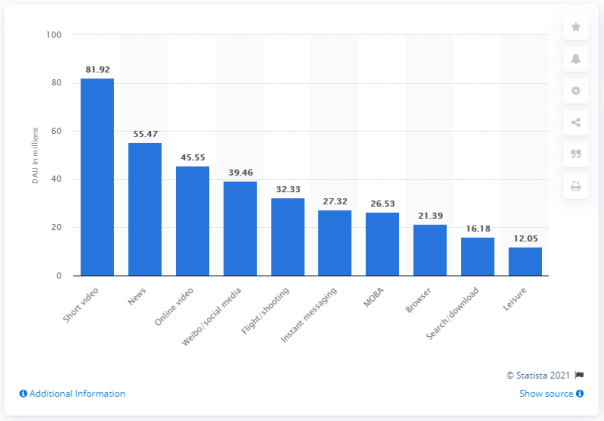
(c) Llama

Figure 7: Chart Visualization Errors: Dates Errors

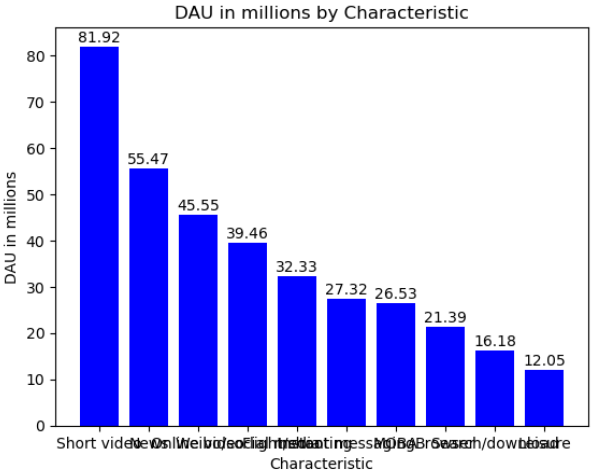
Chart Visualization Errors: Labels Overlapping

Question: How many daily active users did Douyin have in comparison to the period prior to the epidemic?

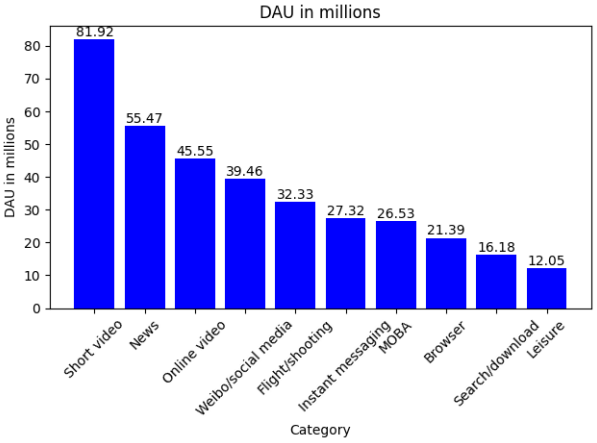
The category labels on the X-axis are not correct on the GPT3.5-generated chart, leading to incorrect answers from the VQA model, as shown in Figure 8.



(a) Ground Truth



(b) GPT3.5



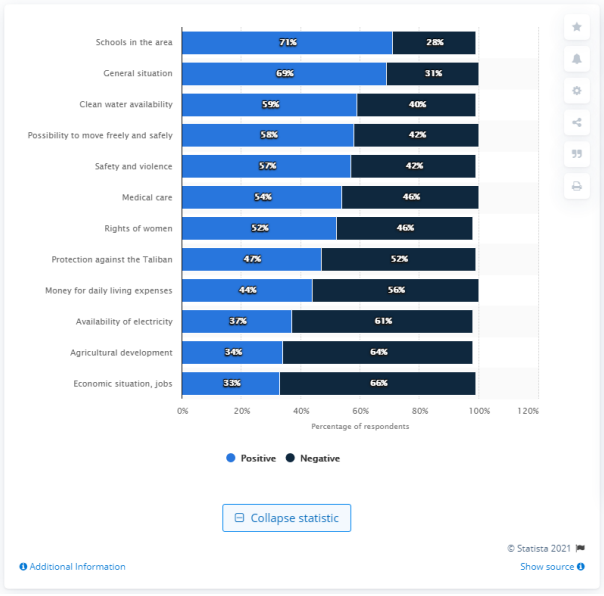
(c) Llama

Figure 8: Chart Visualization Errors: Labels Overlapping

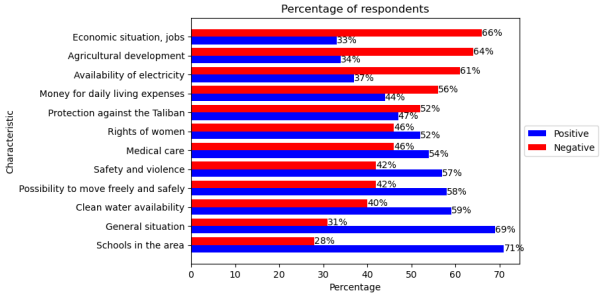
Chart Visualization Errors: Wrong Type of Bars

Question: How many bars (combined) in the chart?

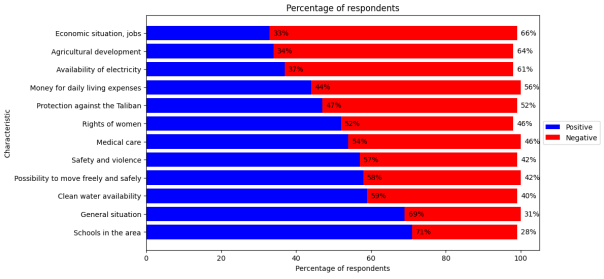
The GPT3.5-generated chart does not utilize stacked bars, leading to incorrect answers from the VQA model, as shown in Figure 9.



(a) Ground Truth



(b) GPT3.5



(c) Llama

Figure 9: Chart Visualization Errors: Wrong Type of Bars