

EncQA: Benchmarking Vision-Language Models on Visual Encodings for Charts




Kushin Mukherjee , Donghao Ren , Dominik Moritz , and Yannick Assogba 



Fig. 1: **ENCQA** charts vary across tasks and encodings. Associated questions focus on the visual mapping used to encode the data, for example in the FIND ANOMALY task with the AREA encoding: “Which circle is an outlier relative to the rest in terms of area?”

Abstract—Multimodal vision-language models (VLMs) continue to achieve ever-improving scores on chart understanding benchmarks. Yet, we find that this progress does not fully capture the breadth of visual reasoning capabilities essential for interpreting charts. We introduce **ENCQA**, a novel benchmark informed by the visualization literature, designed to provide systematic coverage of visual encodings and analytic tasks that are crucial for chart understanding. **ENCQA** provides 2,076 synthetic question-answer pairs, enabling balanced coverage of six visual encoding channels (*position, length, area, color quantitative, color nominal, and shape*) and eight tasks (*find extrema, retrieve value, find anomaly, filter values, compute derived value exact, compute derived value relative, correlate values, and correlate values relative*). Our evaluation of 9 state-of-the-art VLMs reveals that performance varies significantly across encodings within the same task, as well as across tasks. Contrary to expectations, we observe that performance does not improve with model size for many task-encoding pairs. Our results suggest that advancing chart understanding requires targeted strategies addressing specific visual reasoning gaps, rather than solely scaling up model or dataset size.

Index Terms—Visual encodings, visualization understanding tasks, machine chart understanding, vision-language models, model benchmarking

1 INTRODUCTION

Chart understanding has emerged as a key target for multimodal AI systems with frontier generative model releases [5, 50, 63] being accompanied by at least one benchmark score on a chart-focused dataset [25, 43, 77]. The complex nature of chart understanding — requiring visual perception, abstraction, and reasoning — makes it a natural can-

- Kushin Mukherjee is with Stanford University, work done while at Apple. E-mail: kushinm@stanford.edu.
- Donghao Ren is with Apple. E-mail: donghao@apple.com.
- Dominik Moritz is with Apple. E-mail: domoritz@apple.com.
- Yannick Assogba is with Apple. E-mail: yassogba@apple.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.xxxxxx

didate to test the limits of models. Beyond serving as a benchmarking task to assess multimodal reasoning, chart understanding is an ability worth improving in AI systems for practical purposes. Just as data visualizations themselves are tools that help us communicate information by leveraging visual features to efficiently encode vast data tables, so too can AI systems be thought of as tools that help researchers better understand, interact with, and generate visualizations that effectively communicate intended messages [65, 72]. A fundamental condition preceding such use of AI systems for visualization understanding is that their ability to reason about charts should be human-aligned. That is, models should be able to *decode* information that has been *encoded* using common visual channels — position, length, color, area, and shape. These channels have been shown to be effective for communicating data in visual formats [8, 16]. We refer to this ability to effectively use the visual channels and encodings present in charts to make inferences about the information in the chart as *visual reasoning*.

While present chart understanding benchmarks consist of a mix of synthetically generated and web-scraped charts annotated with human-made questions, none of these datasets are created in a manner that allows us to fully isolate the *visual reasoning* component of chart understanding, especially with regard to the use of encoding channels vs. text annotations in charts. And while researchers have taken aim at generating datasets for testing visual reasoning in a targeted manner [26, 29, 39, 77], they often do not contain the kinds of visual encodings and tasks identified in the visualization literature as critical for data visualization purposes [8, 49].

In this paper, we address this critical gap by introducing **ENCQA**, a benchmark tailored to evaluate chart understanding under a variety of visual encodings and tasks drawn from the visualization literature. Our contributions are as follows:

- We develop a data-generation framework for creating visualizations with accompanying question-answer pairs spanning six visual encoding channels (e.g., position, area, color) and eight tasks (e.g., retrieve values, compute derived values, find anomalies).
- Using this framework, we generate **ENCQA** — a set of 2,250 questions for 2,076 charts. We open source our benchmark items and generation framework at our benchmark website ¹.
- We evaluate **ENCQA** on a set of 9 vision-language models, including proprietary and open models, and show that model performance on tasks varies substantially by visual encoding used.

2 BACKGROUND & RELATED WORK

2.1 Visual Encodings for Chart Understanding

Foundational visualization literature established rankings of visual encoding channels based on their efficacy for different tasks, guiding visualization practice and pedagogy [8, 16, 41, 48, 49, 71, 79]. For instance, Cleveland and McGill [16] highlighted position as the best encoding for precise ratio judgments. However, the same ranking of encodings might not apply for all tasks [9] demonstrated by Albers et al.’s [2] finding that color-based encodings are better for aggregating estimates (e.g., interpreting time-series data). This aligns with theories of perceptual averaging where encoding channels like color allow for efficient aggregate estimation from vision alone without needing explicit computation [17].

We emphasize that visual encodings are distinct from chart types, which result from decisions regarding which visual channel(s) are used to encode the data.

2.2 Chart Understanding Tasks

A critique of research surrounding the relative efficacy of visual encoding channels has been that these studies fail to fully characterize the gamut of chart understanding tasks that are natural and common in day-to-day practice [9, 46]. Several taxonomies have been proposed to systematically characterize visualization tasks [4, 12, 54, 57, 60, 69]. Wehrend and Clayton [69] describe tasks in terms of cognitive operations applied to data objects (e.g., identifying, categorizing, ranking,

correlating), whereas Shneiderman [60] emphasizes interaction-based tasks (e.g., overview, zoom, filter, details-on-demand). In contrast, Amar et al. [4] introduce a taxonomy derived from common analytic questions posed by users, focusing on low-level analytic tasks. Their taxonomy comprises ten tasks: *retrieve value*, *filter*, *compute derived value*, *find extremum*, *sort*, *determine range*, *characterize distribution*, *find anomalies*, *cluster*, and *correlate*. In this work, we adopt Amar et al.’s taxonomy because it spans tasks requiring both direct visual extractions—requiring no axes or annotations—and more intricate visual estimates involving averages or correlation judgments, making them ideal for probing visual encoding use in models. This taxonomy has been widely adopted in the visualization literature.

2.3 Vision Language Models

Vision-language models (VLMs) generate open-vocabulary text conditioned on visual inputs and textual prompts. These systems combine the rich representations learned by vision encoder models with the expressivity of a powerful large language model (LLM) in order to solve complex visual reasoning tasks (see [25] for a more in depth overview).

While chart-specific training improves VLM performance on chart understanding [34, 38, 42, 45, 80], large general-purpose frontier models also show strong zero-shot capabilities [50, 63] especially when prompts are optimized for chart understanding [73]. While some works convert charts to non-visual representations such as data tables [37], visualization grammars [13, 31] like Vega-Lite [58], or SVG specifications [76], our evaluation focuses on general-purpose VLMs without additional chart-specific annotations or OCR data extraction, targeting their *visual reasoning* capabilities.

2.4 Large Scale Datasets for Chart Understanding

Progress in machine chart understanding has been closely tied to the development of large-scale question-answering datasets, both synthetic and web-scraped [27, 28, 43, 47]. Early synthetic datasets such as DVQA [27] and FigureQA [28] introduced foundational QA tasks over bar and line charts, while later efforts like PlotQA [47] and ChartQA [43] leveraged web-sourced charts for greater variety and complexity. Multimodal benchmarks such as MMMU [77] and MMC [39] have further expanded coverage, but their chart-related questions often require domain-specific knowledge, making it difficult to disentangle visual reasoning from subject expertise.

Recent datasets have sought to address these limitations in different ways. CharXiv [68] focuses on real-world charts with questions specifically designed to be answerable from the chart content alone, minimizing reliance on domain knowledge. ChartBench [75] introduces a broader task taxonomy and a wide range of chart types, and ChartInsights [73] adopts Amar et al.’s [4] taxonomy — closely aligning with our own—across several chart formats. Notably, Zeng et al. [78] advance chart QA through visualization-referenced instruction tuning, using LLM-based augmentation to increase both the visual diversity of charts and the quality of paired questions, thereby facilitating more robust model training (e.g., via instruction fine-tuning [40]) and higher quality benchmarks.

Parallel work in visualization and cognitive science has leveraged visualization literacy assessments such as VLAT [35] and related studies [3, 7, 36, 52, 67] to reveal fine-grained differences between human and VLM performance, as well as to expand the range of visualization QA items using LLMs [18, 78]. While each of these benchmarks expands the diversity, realism, or annotation strategies for chart question answering, they generally evaluate models at the level of overall task or chart type leaving open the question of how models process specific visual encodings.

In contrast, our benchmark is designed for systematic, fine-grained analysis of vision-language model capabilities. By varying visual encoding channels (position, length, area, color, shape) and analytic tasks according to a principled, literature-derived taxonomy, we enable targeted diagnosis of VLM strengths and weaknesses. Our benchmark uniquely combines (1) a focus on visual encodings over chart types, (2) a principled task taxonomy based on established visualization literature, and (3) question design that specifically distinguishes visual reasoning

¹Code and dataset are available at <https://github.com/apple/ml-encqa>

from text extraction or prior knowledge. Previous works have typically incorporated only a subset of these elements, but by systematically targeting all three, our benchmark enables precise analysis of model strengths and weaknesses across visual encodings and analytic tasks.

3 MOTIVATION

While there has been steady progress over the past few years on building competent models that can understand visualizations in near human-like ways [25, 72], several empirical and theoretical gaps still remain. Recent investigations have revealed that many VLMs and vision backbones [10, 20, 55] *lack basic visual reasoning abilities such as determining whether lines intersect, how shapes are oriented, and counting items in a visual array*. Such low-level tasks are the building blocks of inference when it comes to visualizations. Thus, to the extent that models are limited in their performance of these tasks, we should also be skeptical about their performance on chart understanding benchmarks.

In addition to these low-level tasks, it is also important to evaluate models on the kinds of naturalistic tasks that we test humans on and design charts to accomplish [4, 54, 67]. While there are many formulations for the process of chart understanding [49, 53, 59], the process can be broadly captured by 3 distinct components — (1) *‘bottom-up’ visual perception*, which includes the kinds of visual abilities described above, constitutes the first step and provides the building blocks for visual understanding. (2) *Mapping visual features to data* includes the crucial step of mapping variations in visual features to variations in data. (3) Lastly, an observer must use those mapped values to *answer task-specific questions* [53, 59]. An additional factor in the real world is the influence of domain specific knowledge, such as a climate scientist’s knowledge of weather patterns influencing their interpretation of a weather map. Shah and Hoffner [59] refer to this as ‘knowledge about content’ and an observer’s expectations about the data can influence interpretation of visualizations [74].

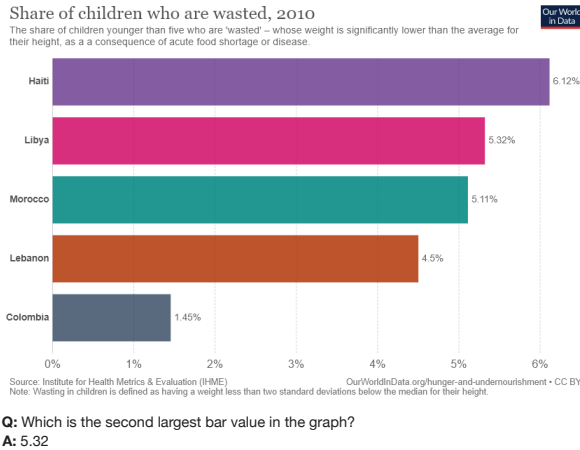


Fig. 2: Example Question-Answer pair from ChartQA where the answer could be directly extracted from the text without making a visual judgment of the length of the bar with respect to the axis.

In this work, we focus on the main 3 components described above and attempt to minimize reliance on ‘knowledge about content’.

We design charts and test items that are critically reliant on visual understanding of a chart and cannot be solved using ‘shortcuts’ such as relying on general knowledge of the world, or only using information from the textual annotations (Figure 2).

Our benchmark provides a valuable resource for tracking progress on how models are able to decode information from different visual encodings, contextualized by human-relevant tasks. We hope the insights that our benchmark raises can help design data-efficient training strategies to improve VLM chart understanding in a principled manner.

4 METHODS

4.1 ENCQA Benchmark Design

We introduce **ENCQA**, a novel benchmark developed using insights from the visualization and vision science literatures to rigorously evaluate the ability of vision language models to understand data visualizations. Earlier benchmarks have focused on representing a wide variety of chart types often sourced from the web and have constructed crowd-sourced and researcher-generated QA pairs to create test items [39, 43, 44, 47, 62, 68]. While this approach allows for the curation of datasets approximately reflecting the kinds of charts used in the real-world, it makes it difficult to ensure an adequate representation of visual encodings [16, 24, 46] and tasks [2, 4, 54] that much of visualization research has deemed critical for encapsulating human chart understanding performance. This raises an important limitation of present benchmarks in that model performance on them might be dominated by successes or failures on specific tasks and encodings. We will explore this limitation in more detail in Section 6.1. Without a fine-grained understanding of the specific kinds of charts that VLMs struggle with, avenues for progress in this space are limited to scaling benchmark size and chart diversity. We posit that this endeavor can be more efficiently directed with a better understanding of the capabilities of VLMs to understand the elemental building blocks of charts. Here, we focus on one of these building blocks, namely, visual encodings.

To that end, we structure the questions in the benchmark to target models’ understanding of visual channels on visualization-relevant tasks. **ENCQA** enables us to evaluate VLM sensitivity to different visual encodings of data across a range of tasks and contains 2,076 charts paired with 2,250 questions. In the following sections we outline the visual encoding channels and tasks we design **ENCQA** around.

4.2 Data Generation

The underlying values for all chart data are drawn from a normal distribution with a mean at 50 and a standard deviation of 5. In cases where one set of data needed to be systematically different from another (e.g., comparing means of two charts), we set the mean of one of the datasets to be 70. For any scatterplot-like charts, we ensured that no two marks were overlapping.

4.3 Visual Encodings

Our benchmark tests 6 encoding channels commonly represented in the visualization literature [16, 24, 49]—**position**, **length**, **area**, **color quantitative** (lightness), **color nominal** (hue), and **shape**. We use these visual channels to encode two different kinds of variables—**quantitative** and **nominal**. Quantitative variables vary in their numeric values and are encoded using position, length, area, and color lightness. Nominal variables that represent categorical data are encoded using shape and color. We note that there are *no charts that are truly single encoding* charts. When assigning a test item’s visual encoding, we refer to the encoding used to **encode the data that the question is about**. The secondary encoding is often a mapping of the datapoint’s category to a position on the axes to prevent data points from overlapping and provides a ‘named index’ to refer to each datapoint.

4.3.1 Chart Variability

When considering these visual encoding channels there are still a number of design choices that visualization designers are left to make. We attempt to capture a number of these in our chart generation process. For each task-encoding pair we generate at least 25 charts, each based on fresh data from our data generation process. In addition to this dataset variability, we apply the following design-specific variants described below:

Mark variability. As the shape of a mark can influence relative area judgments between marks, we generate 25 charts for each of the ‘circle’ and ‘square’ mark types for AREA encodings.

Orientation variability. While in human observers, the difference in orientation of bar charts does not lead to drastically different inferences [1], horizontal and vertical bar charts and dot plots are common

Task	Length	Position	Area	Color (quantitative)	Color (nominal)	Shape	Total
Retrieve Value	50	50	50	25	25	25	225
Find Extrema	100	100	100	50	50	50	450
Find Anomaly	100	100	100	50	50	50	450
Filter Values	100	100	100	50	50	50	450
Compute Derived Values Exact	50	50	50	25	50	50	275
Compute Derived Values Relative	50	50	50	25	25	25	225
Correlate Values	100	50	100	50	N.A.	N.A.	300
Correlate Values Relative	50	25	50	25	N.A.	N.A.	150
Total Questions							2,250

Table 1: Number of test items for each task \times encoding cell. We generate at least 25 charts from different underlying datasets and then vary task-encoding specific parameters (e.g., left-right vs. top-down orientation for length encoding charts) to arrive at a final number. See Section 4.3.1 for further details.

variants. Thus, we generate 25 charts for each orientation for LENGTH and POSITION and encodings.

Color scheme variability. For COLOR QUANTITATIVE encodings, we cycle through four continuous color schemes defined by Vega-Lite [58]—reds, greens, blues, and grays, within the 25 charts in each task encoding pair. We do not alter the default behavior of Altair [66] to assign larger data values to the darker end of the color scheme. All charts are on a white background.

Other visual properties of the chart, such as font or mark size, are left to the default values provided by the Altair, the Vega-Lite based python package we use to generate the charts.

4.4 Tasks

We design our benchmark using six central tasks from Amar et al. [4]. The motivation for using this specific task taxonomy is two-fold. First, it is a human behavior-derived granular ‘low-level’ set of tasks that is helpful for assessing fundamental capabilities of VLMs. Second, the widespread adoption of this taxonomy (and variants) in human chart understanding studies facilitates interoperability with existing findings, enabling the identification of discrepancies between model and human capabilities. Based on existing human perception studies [23,56,61,70] we split the *compute derived values*, and *correlate values* tasks into two variants for a total of eight tasks. These variants ask the model to make a relative judgment between two charts.

Below, we provide brief descriptions of each task:

- **RETRIEVE VALUE** is a task focused on retrieving specific values from a chart such as the numerical value corresponding to the height of a bar, or the value implied by a color given a color-ramp legend. This task is presented as a free response question.
- **FIND EXTREMA** has to do with finding elements in a visualization with the maximum or minimum value of an encoded variable. For each chart we generate questions to find both the minima and maxima. When generating data for this task, we always ensure there is a single extremal value. This task is presented as a multiple choice question.
- **FIND ANOMALY** commonly refers to identifying statistical outliers. Here we focused on outliers in terms of number of observations (for nominal variables), or an observation that has an extreme value of an encoded variable relative to other observations. To ensure that the data backing the chart includes an outlier, we began with Tukey’s classical definition based on inter-quartile ranges. Specifically, Tukey defines an outlier as an observation that either is less than $Q1 - 1.5 \times IQR$ or greater than $Q3 + 1.5 \times IQR$, where $Q1$ and $Q2$ refer to the first and third quartiles and IQR is the inter-quartile range [64]. In order to make the outlier more visually salient, we instead multiplied the IQR by 3 as opposed to 1.5. We randomly designate one of the categories (A - E) as the outlier category and set its value to either $Q1 - 3 \times IQR$ (a ‘small’ outlier) or $Q3 + 3 \times IQR$ (a ‘large’ outlier). This task is presented as a multiple choice question.

- **FILTER VALUES** refers to the task of identifying marks or observations in a chart that satisfy a particular criterion of being lesser/greater or fewer/more than some reference value. The reference value for the comparison was randomly chosen to either corresponded to the count of the second most or second least frequent category. This task is presented as a free response question and requires the model to produce a list of items.

- **CORRELATE VALUES** This task asks the model to estimate if two data series are correlated based on their encodings within a single chart.

We generate data such that the Pearson correlation coefficient (r) was either close to 0.1 or 0.9². This task is presented as a multiple choice question with two options (‘no’ or ‘yes’).

- **CORRELATE VALUES RELATIVE** This task asks models to discern *relative* correlation strengths between pairs of datasets. Models are presented with a *pair of charts*, each of which is similar to those shown in the CORRELATE VALUES. One chart in the pair is generated to have $r \approx 0.9$ and the other to have $r \approx 0.1$. The charts are presented side by side and we randomly flip whether the high correlation chart was presented on the right or left. This task is presented as a multiple choice question with two options (‘left’ or ‘right’).

Since (Pearson) correlations are reliant on the two sets of data being compared being quantitative, we excluded visualizations of nominal variables from the two correlate values tasks.

- **COMPUTE DERIVED VALUES EXACT.** For quantitative variables, this task involves estimating the average value of the variable across multiple categories. For nominal variables, the task was to respond with the average number of observations per category, each chart has 5 categories and has either 36 or 6 observations. This task is presented as a free response question and requires the model to produce a numeric answer.
- **COMPUTE DERIVED VALUES RELATIVE.** In this task, the objective is to report which of two charts presented has the higher or lower average value of the encoded variable. We generate the data for these charts such that one of the charts always has a clear higher average value than the other. This task is presented as a multiple choice question with two options (‘left’ or ‘right’).

Table 2 provides an overview of the question format for each task and encoding channel, while Figure 1 provides a sampling of the charts used in each condition.

4.5 Prompt Design

We construct the text of the final prompt using the following template where each element is separated by a newline: **<prefix>** **<question>**

²The correlated dataset was generated by computing the Cholesky decomposition of the desired correlation matrix and computing the product of the decomposed matrix with the first dataset.

Task	Length	Position	Area	Color (quantitative)	Color (nominal)	Shape
RETRIEVE VALUE	What is the value of Var at {A B C D E}?	What is the value of Var at {A B C D E}?	What is the value of Var at {A B C D E}?	What is the value of Var at {A B C D E}?	How many {red blue green orange purple} circles are present?	How many {square cross triangle circle star} shapes are present?
FIND EXTREMA	Which bar is the {longest shortest}?	Which circle is closest to the {top bottom left right}?	Which {circle square} has the {largest smallest} area?	Which circle has the {darkest lightest} color?	Which color has the {most least} observations?	Which shape has the {most least} observations?
FIND ANOMALY	Which bar is an outlier relative to the rest in terms of length?	Which circle is an outlier relative to the rest in terms of position?	Which {circle square} is an outlier relative to the rest in terms of area?	Which circle is an outlier relative to the rest in terms of color lightness?	Which color is an outlier relative to the rest in terms of the number of observations?	Which shape is an outlier relative to the rest in terms of the number of observations?
FILTER VALUES	Which bar(s) have a value of Var {greater less} than {x}?	Which circle(s) have a value of Var {greater less} than {x}?	Which {circle(s) square(s)} have a value of Var greater than {x}?	Which circle(s) have a value of Var {greater less} than {x}?	Which color(s) have {more fewer} than {x} observations?	Which shape(s) have {more fewer} than {x} observations?
COMPUTE DERIVED VALUES EXACT	What is the average value of Var?	What is the average value of Var?	What is the average value of Var?	What is the average value of Var?	What is the average number of observations per color?	What is the average number of observations per shape?
COMPUTE DERIVED VALUES RELATIVE	In which chart are the bars longer on average?	In which chart are the circles further to the {right top} on average?	In which chart are the areas of the {circles squares} larger on average?	In which chart are the circles darker on average?	Which chart has a higher average number of observations per colors?	Which chart has a higher average number of observations per shape?
CORRELATE VALUES	Are the lengths of the bars for Var1 and Var2 correlated?	Are Var1 and Var2 correlated?	Are the areas of the {circles squares} for Var1 and Var2 correlated?	Are the lightness of the circles for Var1 and Var2 correlated?	N.A.	N.A.
CORRELATE VALUES RELATIVE	In which chart are the lengths of the bars for Var1 and Var2 more correlated?	In which chart are Var1 and Var2 more correlated?	In which chart are the areas of the circles for Var1 and Var2 more correlated?	In which chart are the lightness of the circles for Var1 and Var2 more correlated?	N.A.	N.A.

Table 2: Question structure for each task under the different encoding channels. Where possible, the question is expressed in terms of the visual encoding used. Substitution options shown in {braces}, {x} indicates a numeric substitution.

<suffix>.

The **prefix** varies based on the expected answer type for the question. For **multiple choice** it is ‘Answer using only a single word or letter from the options provided.’, for **numeric** answers, ‘Answer using only a single number’ and finally for **list** answers we set the prefix to ‘Answer choosing only from the options provided, your answer should be just a simple comma separated list.’. The benchmark **question** is then appended followed by an optional **suffix**.

For **numeric** answers the suffix is empty while for **multiple choice** and **list** answers the valid options are first randomly shuffled to ensure there is not a positional bias with respect to the correct answer and then appended as a comma-separated list using the template ‘Options: <comma separated list>’. The chart image is passed into the model as specified by each model’s documentation but always precedes the text described above.

As we added models to our evaluation harness we iterated on prompt design and found this straightforward template to be effective at producing appropriately formatted responses across the models under consideration and provide more detail on that in Section 4.7.

4.6 Model Evaluation

We evaluated every model under a greedy decoding scheme in order to make their responses deterministic and set a maximum new token limit of 30 tokens since none of our questions required long responses.

We focus our evaluation on 3 commercial models and 6 open models listed in Table 3. We select these based on their high scores on ChartQA [43], a well established chart understanding benchmark. We believe these models provide a representative snapshot of the current state of chart understanding models at various sizes at the time of writing.

Model	Params	ChartQA
GPT-4o	n/a	85.7
GPT-4o mini	n/a	n/a
OpenAI o1	n/a	n/a
InternVL2-2B	2.1B	76.2
InternVL2-8B	7.3B	83.3
InternVL2-26B	26B	84.9
Molmo-7B-D	7B	84.1
Phi-3.5 Vision	4.2B	81.8
ChartGemma	2.9B	80.2

Table 3: Models evaluated, their approximate parameter counts and their reported ChartQA scores.

4.7 Response Parsing

We implement a simple regular expression based response parsing pipeline to extract answers from the model response text. We test each response against a list of 5 regexes that were written after looking at sample responses from all the models we tested. One separates the answer from leading phrases like “Answer is:” and two more extract comma-separated lists for set answers, which are then converted to set objects. Our next regex matches against word representations of numbers such as “one, two, three, etc.”, and our final one matches numeric digits with an optional decimal part to handle numeric answers. If none of these patterns match, the whole string is considered as the answer. We found that the vast majority of model responses follow the requested format and we are able to successfully parse valid (though not necessarily correct) answers for almost all models. *InternVL2-2B* produces only 4 responses we couldn’t parse, and manual inspection

confirmed they were incorrect. *ChartGemma* produced 279 of these non-parseable responses, most of which are refusals to perform the task. *Molmo-7B-D* produced 86 non-parseable responses all of which are in the COMPUTE DERIVED VALUE task, as it tries to produce a reasoning chain to answer the question and runs out of tokens before producing an answer.

4.8 Metrics

A motivating question was whether **ENCQA** would uncover differences in how well VLMs solve chart understanding tasks across different visual encoding channels. To measure model performance across tasks using a commensurable metric, we use *accuracy*, a simple yet effective metric used broadly in both evaluations of human and VLM data visualization understanding [25, 67]. Given the different response formats across tasks (numeric responses, multiple choice selections, list generation), we use the following question type-specific measures of accuracy. For multiple choice questions, we use *exact match* to compare model responses to true labels. For numeric responses, we compute the *relaxed accuracy* metric [47] and like other works evaluating chart question answering, we set the threshold for a correct response to be within 5% of the true label. For list responses, such as in the filter value task, we use *set equality* between the generated list and true label. Because of the relatively large number of responses from ChartGemma that we are not able to parse via regular expression, we further use an LLM-as-judge approach [15] to evaluate whether ChartGemma responses are equivalent to the ground truth. We use GPT4o as the judge and score an answer as correct if either the regex method or the LLM indicate that the response is correct. When aggregating scores up to the task or model level, we first *compute mean accuracy within each task-encoding pair* then aggregate those scores to get task-specific and then model-specific scores. This is done in order to account for the different number of stimuli in each task-encoding pair.

5 RESULTS

5.1 VLM Sensitivity to Encoding Channel Varies within and across Tasks

Due to space constraints we focus our analysis here on the performance of two commercial ‘frontier’ models, namely, GPT-4o and OpenAI o1, and a top performing open model, InternVL2-26B. We believe these provide a snapshot of the current state-of-the-art for chart understanding. Figure 3 displays their performance on the EncQA benchmark. Table 4 contains detailed results for these models as well the other 2 top-performing open models that we tested. We observe a number of interesting phenomena that we describe below.

Encodings that require interpreting legends are significantly more difficult than those requiring reading values off of an axis. In the RETRIEVE VALUE task, the AREA and COLOR QUANTITATIVE encoding charts both require the use of a legend to estimate the relative size or lightness of the target mark. All models do better with length and position encodings (which do not require a legend) for this task (Figure 3 first row). This effect is less consistently observed in the FILTER VALUES task, which doesn’t require as precise an estimate of values as RETRIEVE VALUE (GPT-4o in Figure 3, second row).

Simple counting tasks are easy, counting and comparing is significantly harder. In **ENCQA**, questions for encodings of nominal data (i.e., SHAPE and COLOR NOMINAL) are framed as counting tasks, reflecting how these encodings are generally used in visualizations (e.g., “How many blue circles are in the chart?”, “Which shape has the most observations?”). When these are *just* counting tasks (i.e., RETRIEVE VALUE and FIND EXTREMA) the models perform quite well. However, when they require extra aggregation or judgment, model performance drops dramatically. For example if asked “How many colors have more/fewer than x observations” (FILTER VALUES), GPT-4o’s performance drops to less than 50% compared to the near 100% performance seen in the simpler counting tasks. An exception to this is the o1 ‘reasoning’ model which performs substantially better at this task at the cost of significantly more inference time compute. We discuss this more in our exploration of Chain-of-Thought (CoT) prompting which we explore in Section 5.3

Table 4: **ENCQA** accuracy for 2 leading commercial models and the 3 open source models from our suite with the highest scores on ChartQA. Bold numbers indicate the encoding that each model performs the best on within a task.

Task	Encoding	GPT 4o	OpenAI o1	Phi 3.5	InternVL2 26B	Molmo 7B-D
Retrieve Value	Length	0.38	0.34	0.64	0.72	0.64
	Position	0.54	0.48	0.72	0.48	0.60
	Area	0.20	0.14	0.12	0.10	0.08
	Color Quantitative	0.16	0.16	0.04	0.00	0.04
	Color Nominal	0.96	0.76	0.84	0.56	0.92
	Shape	0.96	0.80	0.64	0.48	0.88
Filter Values	Length	0.91	0.79	0.59	0.79	0.44
	Position	0.84	0.91	0.33	0.41	0.26
	Area	0.79	0.62	0.12	0.34	0.05
	Color Quantitative	0.76	0.60	0.14	0.10	0.22
	Color Nominal	0.26	0.52	0.32	0.08	0.08
	Shape	0.50	0.56	0.38	0.12	0.08
Find Extrema	Length	0.99	0.97	0.99	1.00	1.00
	Position	0.93	0.97	0.93	0.83	0.98
	Area	0.96	0.98	0.88	0.85	0.97
	Color Quantitative	1.00	1.00	0.70	0.72	0.98
	Color Nominal	0.98	1.00	0.72	0.66	0.50
	Shape	0.96	0.92	0.64	0.60	0.46
Find Anomaly	Length	0.96	0.96	0.85	0.79	0.81
	Position	0.88	0.95	0.91	0.71	0.89
	Area	0.85	0.93	0.96	0.98	0.66
	Color Quantitative	0.98	0.98	0.90	0.76	0.84
	Color Nominal	0.50	0.56	0.50	0.50	0.24
	Shape	0.50	0.62	0.48	0.50	0.44
Compute Derived Value Exact	Length	0.40	0.32	0.34	0.24	0.34
	Position	0.26	0.30	0.34	0.22	0.32
	Area	0.16	0.10	0.16	0.40	0.22
	Color Quantitative	0.16	0.16	0.08	0.12	0.20
	Color Nominal	0.04	0.78	0.00	0.00	0.00
	Shape	0.00	0.56	0.00	0.00	0.00
Compute Derived Value Relative	Length	0.98	1.00	0.78	0.80	0.78
	Position	0.70	0.70	0.64	0.68	0.68
	Area	1.00	0.98	1.00	1.00	1.00
	Color Quantitative	1.00	1.00	1.00	1.00	1.00
	Color Nominal	0.64	1.00	0.44	0.44	0.28
	Shape	0.28	1.00	0.48	0.64	0.52
Correlate Values	Length	0.51	0.71	0.58	0.54	0.50
	Position	0.96	0.92	0.60	0.56	0.50
	Area	0.75	0.86	0.50	0.58	0.50
	Color Quantitative	0.70	0.90	0.50	0.56	0.50
Correlate Values Relative	Length	0.82	0.76	0.62	0.60	0.52
	Position	1.00	0.96	0.88	0.68	0.64
	Area	0.52	0.74	0.48	0.58	0.44
	Color Quantitative	0.56	0.84	0.28	0.72	0.36
EncQA Score		0.67	0.73	0.55	0.54	0.51

Models struggle with visual estimates of correlation. The CORRELATE tasks ask the model to estimate whether two visual variables are correlated, and we see that in many cases performance is close to random chance. Some notable exceptions are GPT-4o when using position encodings (for which we use a scatterplot — arguably the most common type of plot for estimating correlations), as well as o1, which performs best overall. We also note that there isn’t a consistent change in behavior when moving to the arguably easier ‘relative’ version of the task where the model is asked to judge which of two charts shows data that are more strongly correlated. While GPT-4o’s performance drops for two out of four encodings, InternVL2-26b performance remains relatively similar across encodings.

5.1.1 Favored Task-Encoding Variants

As detailed in Sections 4.3.1 and 4.4, we include a number of variants for certain encodings and task formulations. While models do not demonstrate notable differences for chart orientation variants (horizontal or vertical), or mark type (circle or square), or color scheme we do note some variants that reveal marked differences in performance.

Figure 4 shows the effect of correlation strength on accuracy for the

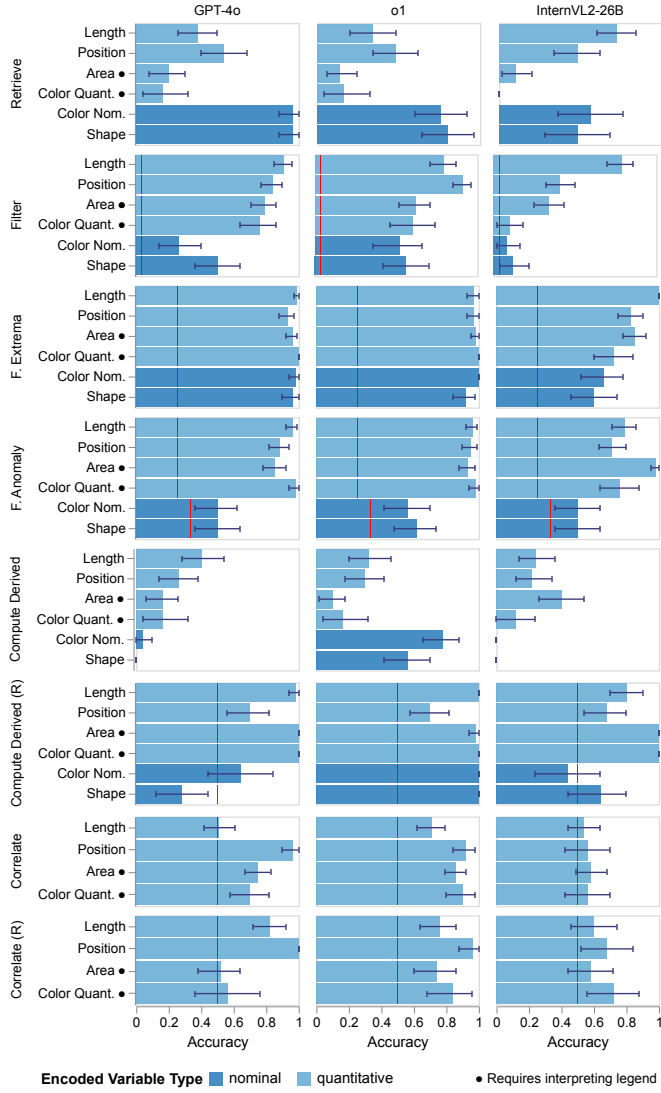


Fig. 3: Model accuracy per visual encoding for each task. (R) indicates ‘Relative’ variants of the task. Red lines mark chance level performance for multiple choice questions. Error bars indicate 95% confidence intervals computed via the bootstrap method

CORRELATE VALUES task. Across encodings, most models can **only perform this task if there is a strong correlation**. Phi-3.5 notably reverses this pattern, and GPT-4o displays an exception to this for only one encoding channel (length).

We also observed that in the **FIND ANOMALY** task, when using **SHAPE** or **COLOR NOMINAL** encodings (i.e., performing a counting task), overall performance is around 50%. On further inspection, we find that all these models are able to do the task **at near 100% performance when the outlier has the ‘smaller’ count, and near 0% when the outlier category has more observations** (Figure 5). This could reflect a bias in the models’ conception of outliers. We further note that this drop in performance is not improved by zero-shot Chain-of-Thought prompting or when using the o1 reasoning model (Section 5.3).

The synthetic, controlled nature of the **ENCQA** is a key enabler for targeted measurement of these fine grained differences. Users of the benchmark can easily generate more chart-question pairs for task-encoding pairs of interest.

5.2 How far off are numerical responses?

Two of our tasks require a precise numerical response, namely **RETRIEVE VALUES** and **COMPUTE DERIVED VALUES**. We previously

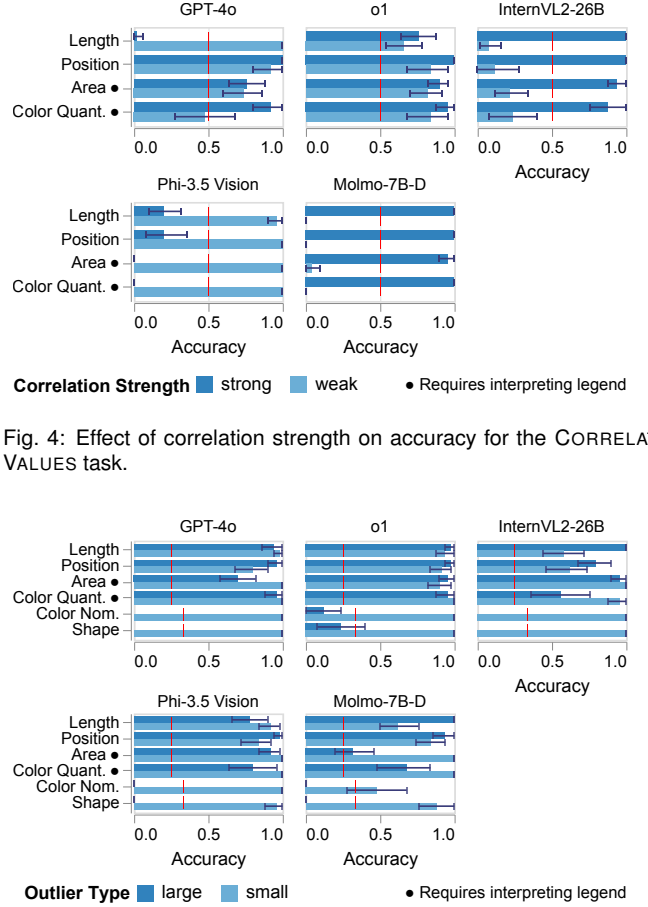


Fig. 4: Effect of correlation strength on accuracy for the CORRELATE VALUES task.

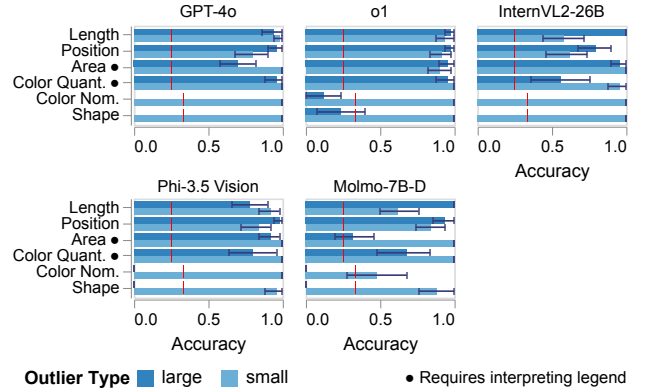


Fig. 5: Effect of outlier type on accuracy for the FIND ANOMALY task.

reported relaxed accuracy scores for these tasks (Figure 3). While useful, this metric gives limited insight into the *degree* to which answers are correct or incorrect. In order to provide a finer grained look at model responses, we computed the symmetric mean absolute percentage error (sMAPE) between the model predictions and ground truth for all numeric responses (Figure 6). sMAPE is a continuous measure of the distance of a model’s predictions from the true values. It is computed as follows:

$$sMAPE = \frac{1}{n} \sum_{k=1}^n \frac{|T_k - P_k|}{(|T_k| + |P_k|)/2}$$

where T_k and P_k are the true and predicted values respectively of the k th question.

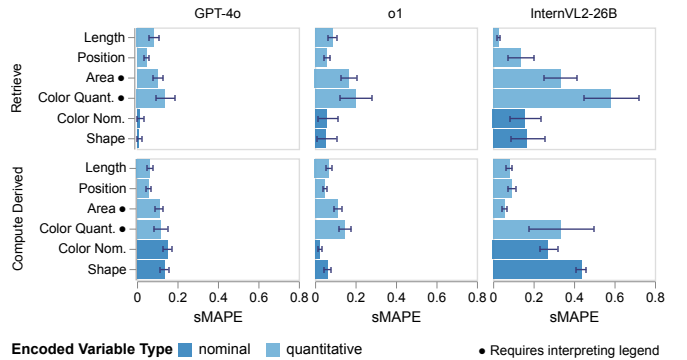


Fig. 6: Symmetric Mean Absolute Percentage Error for numeric responses (lower is better).

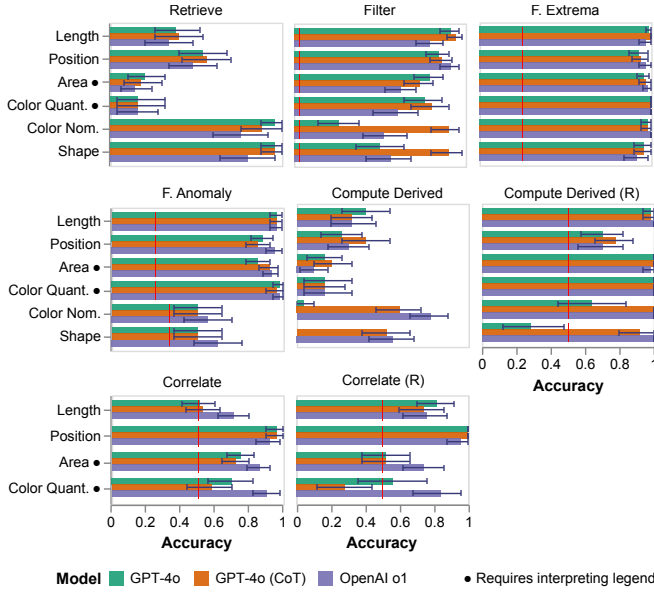


Fig. 7: CoT improves performance mostly higher level tasks like compute derived values or correlate and for some counting tasks. However some task-encoding pairs see performance drops with CoT compared to direct prompting.

For the two encodings that require interpreting a legend (AREA and COLOR QUANTITATIVE), performance on COMPUTE DERIVED VALUE is comparable or better than performance on RETRIEVE VALUE, this is somewhat surprising as the latter task is ostensibly a sub-task of the former.

5.3 Chain-of-Thought/Reasoning Improves Performance For Only a Limited Set of Task-Encoding Pairs

A question that emerges, particularly when considering tasks that require a greater degree of judgment, is whether they benefit from the extra compute steps that zero-shot Chain-of-Thought (CoT) prompting [32], or ‘reasoning’ models provide [51]. Zero-shot CoT asks a model to ‘think step by step’ before producing its final answer. We test this approach using GPT-4o and o1³. For GPT-4o, we allow the model to generate up to 500 new tokens (rather than 30 in the base experiment), and for o1 we set its reasoning effort parameter to ‘medium’ (the default) and observe it uses anywhere from 400 to 8,590 tokens per question (the final answers are still generally under 10 tokens). Both models provide a structured output response API, making it easy to separate the reasoning chain from the final answer.

We observe that most task-encoding pairs are not improved by CoT/reasoning (Figure 7). We do however, see improvement in three tasks (for certain encodings), namely FILTER VALUES, COMPUTE DERIVED VALUE and COMPUTE DERIVED VALUE RELATIVE with both CoT and reasoning. These tasks can be seen as requiring a counting step followed by an aggregation step or set of comparisons.

We do observe some task-encoding pairs where reasoning (o1) outperforms CoT (though they are often comparable), namely in the CORRELATE VALUES and CORRELATE VALUES RELATIVE tasks, for the AREA and COLOR QUANTITATIVE encodings. As observed in Figure 4, o1 is the only model to consistently handle both correlation strengths across all encodings. Surprisingly, in the FILTER VALUES task, we see the reverse for these two encodings, where CoT (and even non-CoT GPT-4o) outperforms the more token intensive reasoning model that is designed to be more capable at visual reasoning⁴.

³For o1 we do not include the ‘think step by step’ instruction as the model is trained specifically to do this and the developer documentation recommends avoiding that addition

⁴<https://platform.openai.com/docs/guides/reasoning-best-practices#5-visual-reasoning>

5.4 Model Size Does Not Reliably Correlate with Performance

A useful property of the deep learning revolution has been a steady predictable increase of model performance with scale (in compute or data), where larger models trained on more data outperform smaller models [30]. However we observe that this is often not the case for task-encoding pairs in ENCQA. Figure 8 (top) shows results for three sizes of model from the InternVL2 family. Many task-encoding pairs do not improve as the model size increases or only do so sub-linearly. And even when there is improvement in task performance it is rarely across all encoding channels.

This may suggest that ENCQA helps highlight elements of capabilities that do not simply scale with model or dataset size. Improving these capabilities might require more targeted collection of data or novel pre-training objectives. While we do not have access to the parameter counts of the commercial models, we present a similar (though coarser) analysis, comparing GPT-4o with GPT-4o-mini (Figure 8 (bottom)).

6 DISCUSSION

6.1 What does ENCQA test that ChartQA doesn’t?

A key motivation behind our work is the position that it is difficult to ascertain the extent to which current chart evaluation benchmarks focus on visual reasoning and whether said benchmarks represent the range of tasks and visual encodings central to visualization understanding. While there are no doubt other cognitive components to chart understanding beyond just visual reasoning [53, 59], visualizations are most helpful when they take advantage of the visual system’s ability to easily extract information from charts [19]. In order to further ground this position, we annotated the test split of ChartQA [43] — the most used chart understanding benchmark at the time of writing — with metadata indicating: the chart understanding task for the question, the primary visual encoding for the data under question, and whether text annotations within the chart image itself would allow for answering the question without the need to actually decode the information using the visual encoding used to create the marks. We took the ChartQA-test human-labeled split and two of the authors first reviewed approximately 50 diverse question-answer pairs to develop guidelines for how to categorize them; then one of the authors annotated all 1,250 test questions.

We found that most test items (81%) *could* be answered without visual reasoning over encodings and instead by extracting text from the chart image (e.g. Figure 2). While redundantly encoding information using text and visual features is indeed helpful (e.g., for memorability [6]), the presence of text annotations in these charts makes it difficult to differentiate models’ text recognition and localization capabilities, from their use of visual encodings and associated axes to solve chart understanding tasks.

Figure 9 shows the distribution of charts with respect to encoding and task. We observe that most charts in ChartQA test use LENGTH or POSITION encodings, followed by AREA encodings. We note that none of the charts use COLOR for quantitative variables (such as in a heatmap or choropleth map) or SHAPE encodings. With regards to task distribution, COMPUTE DERIVED VALUE and RETRIEVE VALUE are the most represented tasks while FIND ANOMALY and CORRELATE VALUES have no or little representation respectively. There were a significant number of ‘Other’ tasks represented, but these mostly consisted of tasks like correctly mapping a legend value to its referent (e.g., ‘What does Green bar represents?’ in a stacked bar chart where one series is colored green), extracting information from titles (e.g., ‘Which country data shown in the Chart?’ in a chart where the title indicates which country the chart is about), or even reading values from axes or annotations (e.g., ‘Find missing data of the sequence 24, _, 32, 33, 42?’ in a chart where bars are labeled with those numbers). There were also cases where the visual encoding could be said to be ‘None.’ These were questions about an aspect of the chart that wasn’t encoded using any visual feature (e.g., ‘Is there a value 30 in the dark blue line?’ in a chart with no axes or tick marks but where each data point is annotated with its value).

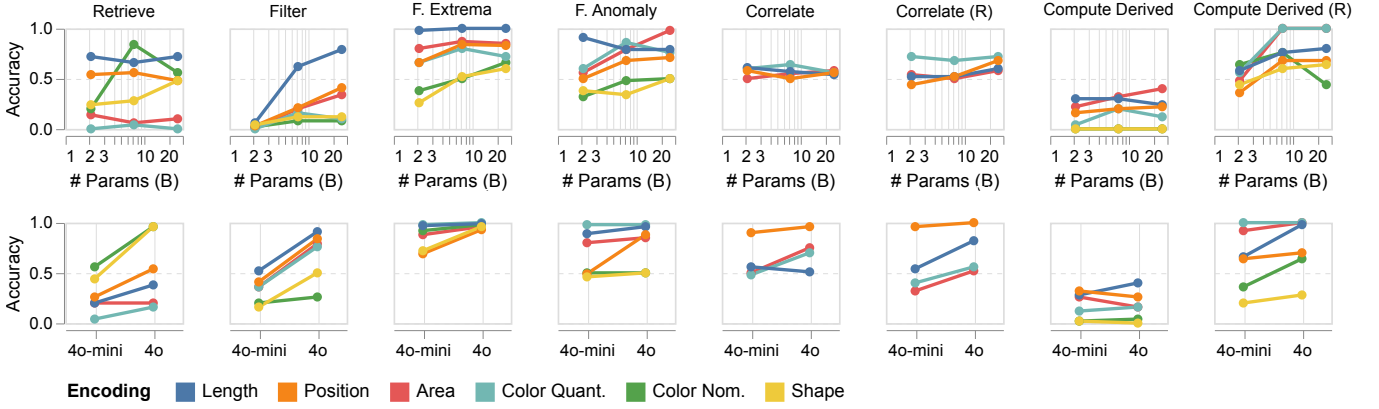


Fig. 8: Accuracy vs Model Size for (TOP) three models in the **InternVL2** Model Family (2B, 8B and 26B, x-axis on a log scale) and (BOTTOM) GPT4o and GPT4o-mini. Some tasks and encodings do see improvements with model scale, but many have no or sub-linear improvement with size. The two smaller InternVL2 models have the same size vision encoder (300M params) while the larger one scales the vision encoder to 6B params.

Encoding	Length	175	91	74	0	303	73	2	1	3	1
	Position	26	36	45	0	79	23	1	0	5	1
	Area	37	21	11	0	54	7	1	0	1	0
	Color Quant.	0	0	0	0	0	0	0	0	0	0
	Color Nom.	8	0	2	0	2	0	0	0	31	0
	Shape	0	0	0	0	0	0	0	0	0	0
	Other	1	1	0	0	0	0	0	0	5	1
	None	4	0	1	0	0	0	1	0	107	5
		Task									
		Retrieve	F. Extrema	Filter	F. Anomaly	Compute Derived	Compute Derived (R)	Correlate	Correlate (R)	Other	None

Fig. 9: ChartQA test samples categorized by encodings and tasks used in this work. ‘Other’ represents tasks outside of the 8 tasks we considered.

This analysis demonstrates the increase in coverage of encodings and tasks that **ENCQA** provides.

7 LIMITATIONS

We acknowledge that **ENCQA** has some limitations. In particular, for a given encoding and task pair, there are multiple valid charts that could be created to satisfy those constraints. We generally chose the simplest design that fit the criteria, and thus the visual complexity of our stimuli is relatively low. Nevertheless, we observe that even these simple charts reveal significant performance gaps and differences in response patterns across models. Future iterations of **ENCQA** could expand the variety of charts used to fill in the encoding-task matrix and the visual complexity of the charts (particularly as model performance saturates for task encoding pairs). The number of encodings and tasks tested could also be increased in future. We also do not test the sensitivity of models to image resolution.

While we set up **ENCQA** to have independent visual encodings for each task type, as noted earlier, charts often use more than one encoding. Even if only a single encoding is relevant for answering a question, a second encoding is often required to prevent the marks from overlapping. To further isolate models’ sensitivity to specific visual encodings, even simpler visual stimuli could be developed to test for the fine-grained differences between visual encoding channels (e.g., position vs. length) without the limitation of needing to be like

naturalistic visualizations such as in [14, 55].

While we adopted a popular taxonomy of visualization tasks [4] for principled reasons, there exist alternative ways of formalizing the visualization task space [11, 21, 22, 33] — some more high-level than the ones presently considered.

Lastly, our results on model accuracy vs. model-size is limited to one open-source model and one closed-source model family (for which we have limited insight into the differences between the large and small models), potentially limiting their generalizability.

8 CONCLUSION & FUTURE WORK

In this paper, we introduce a new benchmark, **ENCQA**, that tests vision-language models on their ability to perform visual reasoning tasks relevant to chart understanding. We address the need for a rigorous benchmark that varies charts in both the visual encoding channel and tasks that are evaluated using these charts, with a targeted focus on visual reasoning as opposed to testing models’ general word knowledge. As the field considers how AI chart understanding might be effectively advanced we ought to consider which strategies might be most beneficial for both evaluating and improving model capabilities. The dominant strategy we observe in the field is a drive towards larger datasets that have more realistic charts and more complex questions, whether scraped from the internet or generated by advanced VLMs [18]. In the design of **ENCQA** we propose an alternative approach to evaluation that instead starts from principles of visualization design and perception and constructs test items to measure specific perceptual abilities. We consider these two approaches to be complementary and have described aspects of VLM chart understanding behavior that **ENCQA** is uniquely able to isolate.

We find that **ENCQA** reveals performance differences in VLMs across visual encoding channels (length, position, area, color, shape) for diverse visualization tasks. Accuracy patterns vary across models and even top-performing models, including those optimized for reasoning, fail to consistently answer questions correctly across all tasks and encodings. We also observe that error patterns cannot be explained as a function of model size. These results underscore the need for combining insights from visualization and machine learning, with rigorous evaluation of models on targeted benchmarks.

While we have focused on evaluation of VLM capabilities, we see room for future work that explores how synthetic datasets and frameworks for parametrically generating benchmarks could be used to *improve* the capabilities of VLMs on real-world chart understanding tasks.

9 ACKNOWLEDGMENTS

We thank our colleagues at the Visualization team at Apple for helpful feedback and discussion. We also thank our anonymous reviewers for their comments that helped strengthen the work.

REFERENCES

- [1] F. Alallah, D. Jin, and P. Irani. OA-graphs: orientation agnostic graphs for improving the legibility of charts on horizontal displays. In *ACM International Conference on Interactive Tabletops and Surfaces*, 2010. doi: [10.1145/1936652.1936692](https://doi.org/10.1145/1936652.1936692) 3
- [2] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *CHI*, 2014. doi: [10.1145/2556288.2557200](https://doi.org/10.1145/2556288.2557200) 2, 3
- [3] J. Alexander, P. Nanda, K.-C. Yang, and A. Sarvghad. Can GPT-4 Models Detect Misleading Visualizations? *IEEE VIS*, 2024. doi: [10.1109/VIS55277.2024.00029](https://doi.org/10.1109/VIS55277.2024.00029) 2
- [4] R. Amar, J. Eagan, and J. Stasko. Low-Level Components of Analytic Activity in Information Visualization. *IEEE INFOVIS*, 2005. doi: [10.1109/INFOVIS.2005.2](https://doi.org/10.1109/INFOVIS.2005.2) 2, 3, 4, 9
- [5] Anthropic. Introducing the next generation of Claude. <https://www.anthropic.com/news/claude-3-family>, Mar. 2024. 1
- [6] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *CHI*, 2010. doi: [10.1145/1753326.1753716](https://doi.org/10.1145/1753326.1753716) 8
- [7] A. Bendeck and J. Stasko. An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks. *IEEE TVCG*, 2024. doi: [10.1109/TVCG.2024.3456155](https://doi.org/10.1109/TVCG.2024.3456155) 2
- [8] J. Bertin. *Graphics and Graphic Information Processing*. De Gruyter, Berlin New York, 1981. doi: [10.1515/9783110854688](https://doi.org/10.1515/9783110854688) 2
- [9] E. Bertini, M. Correll, and S. Franconeri. Why Shouldn't All Charts Be Scatter Plots? Beyond Precision-Driven Visualizations, 2021. arXiv:2008.11310 [cs]. doi: [10.1109/VIS47514.2020.00048](https://doi.org/10.1109/VIS47514.2020.00048) 2
- [10] J. S. Bowers, G. Malhotra, M. Dujmović, M. L. Montero, C. Tsvetkov, V. Biscione, G. Puebla, F. Adolphi, J. E. Hummel, R. F. Heaton, B. D. Evans, J. Mitchell, and R. Blything. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46, 2023. doi: [10.1017/S0140525X22002813](https://doi.org/10.1017/S0140525X22002813) 3
- [11] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A Principled Way of Assessing Visualization Literacy. *IEEE TVCG*, 2014. doi: [10.1109/TVCG.2014.2346984](https://doi.org/10.1109/TVCG.2014.2346984) 9
- [12] M. Brehmer and T. Munzner. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE TVCG*, 2013. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: [10.1109/TVCG.2013.124](https://doi.org/10.1109/TVCG.2013.124) 2
- [13] V. S. Bursztyn, J. Hoffswell, E. Koh, and S. Guo. Representing Charts as Text for Language Models: An In-Depth Study of Question Answering for Bar Charts. *IEEE VIS*, 2024. doi: [10.1109/VIS55277.2024.00061](https://doi.org/10.1109/VIS55277.2024.00061) 2
- [14] H. Chae, S. Yoon, J. Park, C. Y. Chun, Y. Cho, M. Cai, Y. J. Lee, and E. K. Ryu. Decomposing Complex Visual Comprehension into Atomic Visual Skills for Vision Language Models, 2025. arXiv:2505.20021 [cs]. doi: [10.48550/arXiv.2505.20021](https://doi.org/10.48550/arXiv.2505.20021) 9
- [15] C.-H. Chiang and H.-y. Lee. Can Large Language Models Be an Alternative to Human Evaluations? In A. Rogers, J. Boyd-Graber, and N. Okazaki, eds., *Proceedings of ACL*, 2023. doi: [10.18653/v1/2023.acl-long.870](https://doi.org/10.18653/v1/2023.acl-long.870) 6
- [16] W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 1984. doi: [10.2307/2288400](https://doi.org/10.2307/2288400) 2, 3
- [17] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *CHI*. ACM, 2012. doi: [10.1145/2207676.2208556](https://doi.org/10.1145/2207676.2208556) 2
- [18] Y. Cui, L. W. Ge, Y. Ding, L. Harrison, F. Yang, and M. Kay. Promises and Pitfalls: Using Large Language Models to Generate Visualization Items. *IEEE TVCG*, pp. 1094–1104, 2025. doi: [10.1109/TVCG.2024.3456309](https://doi.org/10.1109/TVCG.2024.3456309) 2, 9
- [19] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman. The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, 2021. doi: [10.1177/15291006211051956](https://doi.org/10.1177/15291006211051956) 8
- [20] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive, 2024. doi: [10.48550/arXiv.2404.12390](https://doi.org/10.48550/arXiv.2404.12390) 3
- [21] M. Galesic and R. Garcia-Retamero. Graph Literacy: A Cross-Cultural Comparison. *Medical Decision Making*, 2011. Publisher: SAGE Publications Inc STM. doi: [10.1177/0272989X10373805](https://doi.org/10.1177/0272989X10373805) 9
- [22] L. W. Ge, Y. Cui, and M. Kay. CALVI: Critical Thinking Assessment for Literacy in Visualizations. In *CHI*, 2023. doi: [10.1145/3544548.3581406](https://doi.org/10.1145/3544548.3581406) 9
- [23] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of Average Value in Multiclass Scatterplots. *IEEE TVCG*, 2013. doi: [10.1109/TVCG.2013.183](https://doi.org/10.1109/TVCG.2013.183) 4
- [24] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *CHI*, 2010. doi: [10.1145/1753326.1753357](https://doi.org/10.1145/1753326.1753357) 3
- [25] K.-H. Huang, H. P. Chan, Y. R. Fung, H. Qiu, M. Zhou, S. Joty, S.-F. Chang, and H. Ji. From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models, 2024. doi: [10.48550/arXiv.2403.12027](https://doi.org/10.48550/arXiv.2403.12027) 1, 2, 3, 6
- [26] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *CVPR*, pp. 2901–2910, 2017. doi: [10.1109/CVPR.2017.215](https://doi.org/10.1109/CVPR.2017.215) 2
- [27] K. Kafle, B. Price, S. Cohen, and C. Kanan. DVQA: Understanding Data Visualizations via Question Answering. *arXiv*, 2018. arXiv:1801.08163 [cs]. doi: [10.48550/arXiv.1801.08163](https://doi.org/10.48550/arXiv.1801.08163) 2
- [28] S. E. Kahou, V. Michalski, A. Atkinson, A. Kadar, A. Trischler, and Y. Bengio. FigureQA: An Annotated Figure Dataset for Visual Reasoning, 2018. doi: [10.48550/arXiv.1710.07300](https://doi.org/10.48550/arXiv.1710.07300) 2
- [29] R. Kamoi, Y. Zhang, S. S. S. Das, R. H. Zhang, and R. Zhang. VisOnlyQA: Large Vision Language Models Still Struggle with Visual Perception of Geometric Information, 2024. doi: [10.48550/arXiv.2412.00947](https://doi.org/10.48550/arXiv.2412.00947) 2
- [30] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv*, 2020. doi: [10.48550/arXiv.2001.08361](https://doi.org/10.48550/arXiv.2001.08361) 8
- [31] 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*. Association for Computing Machinery, 2020. doi: [10.1145/3313831.3376467](https://doi.org/10.1145/3313831.3376467) 2
- [32] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2024. doi: doi.org/10.48550/arXiv.2205.11916 8
- [33] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry. Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI workshop on Beyond time and errors: novel evaluation methods for information visualization*, 2006. doi: [10.1145/1168149.1168168](https://doi.org/10.1145/1168149.1168168) 9
- [34] K. Lee, M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova. Pix2Struct: Screen-shot Parsing as Pretraining for Visual Language Understanding, 2023. arXiv:2210.03347 [cs]. doi: [10.48550/arXiv.2210.03347](https://doi.org/10.48550/arXiv.2210.03347) 2
- [35] S. Lee, S.-H. Kim, and B. C. Kwon. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE TVCG*, 2017. doi: [10.1109/TVCG.2016.2598920](https://doi.org/10.1109/TVCG.2016.2598920) 2
- [36] Z. Li, H. Miao, V. Pascucci, and S. Liu. Visualization Literacy of Multimodal Large Language Models: A Comparative Study. *arXiv*, 2024. arXiv:2407.10996 [cs]. doi: [10.48550/arXiv.2407.10996](https://doi.org/10.48550/arXiv.2407.10996) 2
- [37] F. Liu, J. M. Eisenschlos, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, W. Chen, N. Collier, and Y. Altun. DePlot: One-shot visual language reasoning by plot-to-table translation, 2023. arXiv:2212.10505 [cs]. doi: [10.48550/arXiv.2212.10505](https://doi.org/10.48550/arXiv.2212.10505) 2
- [38] F. Liu, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, Y. Altun, N. Collier, and J. M. Eisenschlos. MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering, 2023. arXiv:2212.09662 [cs]. doi: [10.48550/arXiv.2212.09662](https://doi.org/10.48550/arXiv.2212.09662) 2
- [39] F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho, Y. Yacoub, and D. Yu. MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. *arXiv*, 2024. arXiv:2311.10774 [cs]. doi: [10.48550/arXiv.2311.10774](https://doi.org/10.48550/arXiv.2311.10774) 2, 3
- [40] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual Instruction Tuning, Dec. 2023. arXiv:2304.08485 [cs]. doi: [10.48550/arXiv.2304.08485](https://doi.org/10.48550/arXiv.2304.08485) 2
- [41] J. Mackinlay. Applying a theory of graphical presentation to the graphic design of user interfaces. In *Proceedings of the 1st annual ACM SIGGRAPH symposium on User Interface Software*, 1988. doi: [10.1145/62402.62431](https://doi.org/10.1145/62402.62431) 2
- [42] A. Masry, P. Kavehzadeh, X. L. Do, E. Hoque, and S. Joty. UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning. *arXiv*, 2023. 2
- [43] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. *arXiv*, 2022. arXiv:2203.10244 [cs]. doi: [10.48550/arXiv.2203.10244](https://doi.org/10.48550/arXiv.2203.10244) 1, 2, 3, 5, 8

- [44] A. Masry, M. Shahmohammadi, M. R. Parvez, E. Hoque, and S. Joty. ChartInstruct: Instruction Tuning for Chart Comprehension and Reasoning. *arXiv*, 2024. doi: 10.48550/arXiv.2403.09028
- [45] A. Masry, M. Thakkar, A. Bajaj, A. Kartha, E. Hoque, and S. Joty. ChartGemma: Visual Instruction-tuning for Chart Reasoning in the Wild. *arXiv*, 2024. arXiv:2407.04172 [cs]. doi: 10.48550/arXiv.2407.04172 2
- [46] C. M. McColeman, F. Yang, T. F. Brady, and S. Franconeri. Rethinking the Ranks of Visual Channels. *IEEE TVCG*, 2022. doi: 10.1109/TVCG.2021.3114684 2, 3
- [47] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. PlotQA: Reasoning over Scientific Plots, 2020. arXiv:1909.00997 [cs]. doi: 10.1109/WACV45572.2020.9093523 2, 3, 6
- [48] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco. *IEEE TVCG*, 2019. doi: 10.1109/TVCG.2018.2865240 2
- [49] T. Munzner. *Visualization Analysis and Design*. A K Peters/CRC Press, New York, Oct. 2014. doi: 10.1201/b17511 2, 3
- [50] OpenAI. GPT-4 Technical Report, Mar. 2024. arXiv:2303.08774 [cs]. doi: 10.48550/arXiv.2303.08774 1, 2
- [51] OpenAI. Learning to reason with LLMs. <https://openai.com/index/learning-to-reason-with-llms/>, Sept. 2024. 8
- [52] S. Pandey and A. Ottley. Benchmarking Visual Language Models on Standardized Visualization Literacy Tests. *Computer Graphics Forum*, 2025. doi: 10.1111/cgf.70137 2
- [53] S. Pinker. A theory of graph comprehension. In *Artificial intelligence and the future of testing*. Psychology Press, 1990. 3, 8
- [54] G. J. Quadri and P. Rosen. A Survey of Perception-Based Visualization Studies by Task. *IEEE TVCG*, 2022. doi: 10.1109/TVCG.2021.3098240 2, 3
- [55] P. Rahmzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen. Vision language models are blind. *arXiv*, 2024. doi: 10.48550/arXiv.2407.06581 3, 9
- [56] R. A. Rensink and G. Baldridge. The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 2010. doi: 10.1111/j.1467-8659.2009.01694.x 4
- [57] B. Saket, A. Endert, and Ç. Demiralp. Task-Based Effectiveness of Basic Visualizations. *IEEE TVCG*, pp. 2505–2512, 2019. doi: 10.1109/TVCG.2018.2829750 2
- [58] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-Lite: A Grammar of Interactive Graphics. *IEEE TVCG*, pp. 341–350, Jan. 2017. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2016.2599030 2, 4
- [59] P. Shah and J. Hoeffner. Review of Graph Comprehension Research: Implications for Instruction. *Educational Psychology Review*, 2002. doi: 10.1023/A:1013180410169 3, 8
- [60] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In B. B. Bederson and B. Shneiderman, eds., *The Craft of Information Visualization*. Morgan Kaufmann, 2003. doi: 10.1016/B978-155860915-0/50046-9 2
- [61] D. A. Szafrir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 2016. doi: 10.1167/16.5.11 4
- [62] B. Tang, A. Boggust, and A. Satyanarayan. VisText: A Benchmark for Semantically Rich Chart Captioning. In A. Rogers, J. Boyd-Graber, and N. Okazaki, eds., *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. doi: 10.18653/v1/2023.acl-long.401 3
- [63] G. Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*, Aug. 2024. arXiv:2403.05530 [cs]. doi: 10.48550/arXiv.2403.05530 1, 2
- [64] J. Tukey. *Exploratory Data Analysis*. Pearson, Reading, Mass., 1st edition ed., Jan. 1977. 4
- [65] P. Vaithilingam, E. L. Glassman, J. Priya Inala, and C. Wang. DynaVis: Dynamically Synthesized UI Widgets for Visualization Editing, 2024. doi: 10.48550/arXiv.2401.10880 2
- [66] J. VanderPlas, B. Granger, J. Heer, D. Moritz, K. Wongsuphasawat, A. Satyanarayan, E. Lees, I. Timofeev, B. Welsh, and S. Sievert. Altair: Interactive Statistical Visualizations for Python. *Journal of Open Source Software*, 2018. doi: 10.21105/joss.01057 4
- [67] A. Verma, K. Mukherjee, C. Potts, E. Kreiss, and J. E. Fan. Evaluating human and machine understanding of data visualizations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0), 2024. 2, 3, 6
- [68] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, A. Chevalier, S. Arora, and D. Chen. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs, 2024. arXiv:2406.18521 [cs]. doi: doi.org/10.48550/arXiv.2406.18521 2, 3
- [69] S. Wehrend and C. Lewis. A problem-oriented classification of visualization techniques. In *Proceedings of the First IEEE Conference on Visualization*, 1990. doi: 10.1109/VISUAL.1990.146375 2
- [70] D. Whitney and A. Y. Leib. Ensemble Perception. *Annual Review of Psychology*, 69(Volume 69, 2018), 2018. doi: 10.1146/annurev-psych-010416-044232 4
- [71] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE TVCG*, 2016. doi: 10.1109/TVCG.2015.2467191 2
- [72] A. Wu, Y. Wang, X. Shu, D. Moritz, W. Cui, H. Zhang, D. Zhang, and H. Qu. AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization. *IEEE TVCG*, 2022. doi: 10.1109/TVCG.2021.3099002 2, 3
- [73] Y. Wu, L. Yan, L. Shen, Y. Wang, N. Tang, and Y. Luo. ChartInsights: Evaluating Multimodal Large Language Models for Low-Level Chart Question Answering. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds., *EMNLP findings*, 2024. doi: 10.18653/v1/2024.findings-emnlp.710 2
- [74] C. Xiong, L. Van Weelden, and S. Franconeri. The Curse of Knowledge in Visual Data Communication. *IEEE TVCG*, 2020. doi: 10.1109/TVCG.2019.2917689 3
- [75] Z. Xu, S. Du, Y. Qi, C. Xu, C. Yuan, and J. Guo. ChartBench: A Benchmark for Complex Visual Reasoning in Charts, 2024. doi: doi.org/10.48550/arXiv.2312.15915 2
- [76] Z. Xu and E. Wall. Exploring the Capability of LLMs in Performing Low-Level Visual Analytic Tasks on SVG Data Visualizations, 2024. doi: doi.org/10.48550/arXiv.2404.19097 2
- [77] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2
- [78] X. Zeng, H. Lin, Y. Ye, and W. Zeng. Advancing Multimodal Large Language Models in Chart Question Answering with Visualization-Referenced Instruction Tuning. *IEEE TVCG*, 2025. doi: 10.1109/TVCG.2024.3456159 2
- [79] Z. Zeng, J. Yang, D. Moritz, J. Heer, and L. Battle. Too Many Cooks: Exploring How Graphical Perception Studies Influence Visualization Recommendations in Draco. *IEEE TVCG*, pp. 1063–1073, 2024. doi: 10.1109/TVCG.2023.3326527 2
- [80] M. Zhou, Y. Fung, L. Chen, C. Thomas, H. Ji, and S.-F. Chang. Enhanced Chart Understanding via Visual Language Pre-training on Plot Table Pairs. In A. Rogers, J. Boyd-Graber, and N. Okazaki, eds., *Findings of ACL*, pp. 1314–1326. Association for Computational Linguistics, Toronto, Canada, 2023. doi: 10.18653/v1/2023.findings-acl.85 2