Are Large Vision Language Models up to the Challenge of Chart Comprehension and Reasoning?

Mohammed Saidul Islam♣ , **Raian Rahman**♦ , **Ahmed Masry**♣*,
Md Tahmid Rahman Laskar♣♡* , **Mir Tafseer Nayeem**♦* , **Enamul Hoque**♣

♣York University, Canada, ♦Islamic University of Technology, Bangladesh,
♦University of Alberta, Canada, ♡Dialpad Inc., Canada

{saidulis, masry20, tahmid20, enamulh}@yorku.ca, raianrahman@iut-dhaka.edu,
mnayeem@ualberta.ca

Abstract

Natural language is a powerful complementary modality of communication for data visualizations, such as bar and line charts. To facilitate chart-based reasoning using natural language, various downstream tasks have been introduced recently such as chart question answering, chart summarization, and fact-checking with charts. These tasks pose a unique challenge, demanding both vision-language reasoning and a nuanced understanding of chart data tables, visual encodings, and natural language instructions. Despite the recent success of Large Language Models (LLMs) across diverse NLP tasks, their abilities and limitations in the realm of data visualization remain under-explored, possibly due to their lack of multi-modal capabilities. To bridge the gap, this paper presents one of the first comprehensive evaluations of the recently developed large vision language models (LVLMs) for chart understanding and reasoning tasks. Our evaluation includes a comprehensive assessment of both closed and open-sourced LVLMs across five major chart reasoning tasks. Furthermore, we perform a qualitative evaluation of LVLMs' performance on a diverse range of charts, aiming to provide a thorough analysis. Our findings reveal that while LVLMs demonstrate impressive abilities in generating fluent texts covering high-level data insights, they also encounter common problems like hallucinations, factual errors, and data bias. We highlight the key strengths and limitations of LVLMs in chart comprehension tasks, offering insights for future research [1].

1 Introduction

Natural language and visualizations are two powerful complementary modalities to communicate data insights. While visualizations can be very effective in finding patterns, trends, and outliers in

---

* Equal contribution.

[1]We make all our prompts as well as LVLMs' responses open source here.
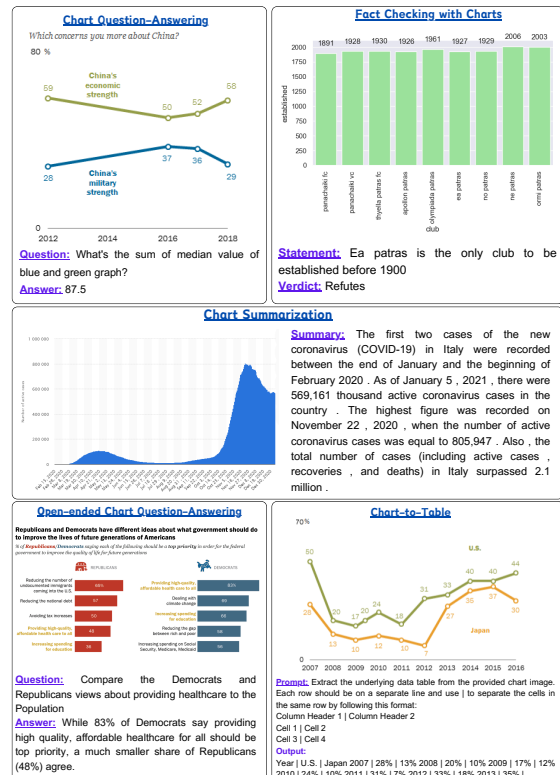

Figure 1: Chart comprehension and reasoning tasks.

data, natural language can help explain the key insights in visualizations and answer questions about data (Hoque et al., 2022). People commonly integrate text with graphical charts as this combination helps direct attention to specific elements of the chart and offers necessary explanations that might otherwise go unnoticed (Stokes et al., 2022). Moreover, interfaces that use natural language to interact with charts have other benefits, such as, enhancing chart accessibility (Alam et al., 2023) and supporting visualization education (Bach et al., 2023).

Given the importance of chart comprehension and reasoning, researchers have introduced various tasks for the development of automated methods to aid users in chart analysis (see Fig. 1). These include chart question answering (Masry et al., 2022;

Kantharaj et al., 2022a; Lee et al., 2022), natural language generation for charts (Obeid and Hoque, 2020; Shankar et al., 2022), and fact-checking with charts (Akhtar et al., 2023a,b). To build automated systems for these tasks, a prevalent approach involves pre-training of models (Liu et al., 2022; Masry et al., 2023) and development of frameworks without heuristic rules (Cheng et al., 2023) on language and vision tasks (Du et al., 2022). However, in recent years, there has been dramatic progress in the development and widespread adoption of LLMs (Anil et al., 2023; Chowdhery et al., 2023; OpenAI et al., 2023; Touvron et al., 2023a,b). While in the beginning, the LLMs were only capable of processing textual data, the rapid progress in this field has paved the way for the development of multimodal LLMs (in other words, LVLMs), such as GPT-4V (OpenAI et al., 2023), Gemini (Team et al., 2023), Claude-3 (Anthropic, 2024), Phi-3 (Abdin et al., 2024), LLaVA (Liu et al., 2023b), and MiniGPT-4 (Zhu et al., 2023). Given the rapid rise of these LVLMs, there is a pressing question: *Are LVLMs up to the challenge of chart comprehension and reasoning?*

In this paper, we aim to answer this question by investigating the capabilities and limitations of LVLMs in the chart reasoning and comprehension domain. Specifically, we examine whether the latest state-of-the-art (SoTA) LVLMs can effectively interpret charts as well as identify key insights solely based on the chart images. This setup is crucial in real-world scenarios where the underlying data tables of charts are often unavailable. To this end, we performed extensive qualitative and quantitative analyses of the performance of LVLMs on five downstream tasks across *seven* benchmark datasets: **ChartQA** (Masry et al., 2022), **OpenCQA** (Kantharaj et al., 2022b), **Chart Summarization** (Chart-to-Text (Kantharaj et al., 2022c), and **Vistext** (Tang et al., 2023)), **Fact-checking** (ChartFC (Akhtar et al., 2023a), **ChartCheck** (Akhtar et al., 2023b)), and **Chart-to-Table** (Choi et al., 2019).

Specifically, this work presents the first detailed analyses of LVLMs on seven chart domain benchmarks, making the following main contributions:

(1) Existing SoTA models typically report quantitative performance on ChartQA without a detailed analysis of their capabilities and limitations. We examine LVLMs' performance using advanced techniques like Chain-of-Thought (Wei et al., 2023)

and Program-aided Language Models (Gao et al., 2023) (§4.2).

(2) Unlike most closed-source models that focus only on factoid question answering (ChartQA), we evaluate LVLMs on other important tasks such as OpenCQA and Chart Summarization, presenting the first analysis of LVLMs' capability in generating open-ended responses (§4.3, §4.4).

(3) Hallucinations, factual errors, and bias are common issues for many LVLMs. We investigate these problems through various analyses (§4.5, §4.7 and §4.9), including the adoption of an error taxonomy (Mishra et al., 2024) for hallucinations.

(4) We address the fundamental question of how effectively LVLMs can interpret charts by measuring their ability to extract data from chart images, being the first to thoroughly examine this (§4.6).

(5) Text generation tasks require models to describe high-level trends and outliers, as well as low-level chart details like colors. We analyze how often and how accurately models cover different types of semantic content using the 4-level framework (Lundgard and Satyanarayan, 2021) (§4.8).

2 Related Work

Chart-related Downstream Tasks: Several downstream tasks associated with charts have been proposed recently. Chart Question Answering refers to answering factoid questions regarding charts (Kafle et al., 2018; Kahou et al., 2018; Methani et al., 2020; Masry et al., 2022; Xu et al., 2024). In contrast, Open-ended Question Answering (OpenCQA) require explanatory responses by interpreting chart data (Kantharaj et al., 2022a). The Chart Summarization task (Shankar et al., 2022; Obeid and Hoque, 2020; Tang et al., 2023; Rahman et al., 2023) involves creating natural language descriptions from charts, Chart-to-Table focuses on converting charts into data tables (Choi et al., 2019), and Fact-checking with charts (Akhtar et al., 2023a,b) focuses on verifying factual statements related to charts. While there are other areas like infographic comprehension (Mathew et al., 2022) and science diagram QA (Kembhavi et al., 2016), this study is devoted to chart-related tasks.

Evaluation of LVLMs: OpenAI's introduction of GPT-4V marked a significant advancement, outperforming other LVLMs proposed in (Liu et al., 2023b; Dai et al., 2023; Zhu et al., 2023), particu-

larly in scenarios data-scarce scenarios. Google's Gemini (Team et al., 2023) and Anthropic's Claude-3 (Anthropic, 2024) have recently emerged as strong competitors, and Microsoft's open-source Phi-3 model achieved performance comparable to closed-source LVLMs in multimodal tasks (Abdin et al., 2024). While some studies compared Gemini and GPT-4V models on image recognition and understanding tasks (Qi et al., 2023; Fu et al., 2023), other works have focused on solving chart-related tasks only using data tables (Do et al., 2023; Huang et al., 2023). Additionally, some recent studies have proposed a benchmark dataset (Guan et al., 2024) to study image-context reasoning, introduced a new LLM for improved open-ended visual question-answering (Hu et al., 2023), and assessed GPT-4V-type models on tasks requiring structured reasoning (Singh et al., 2023). However, these studies address only one chart-related task (i.e., Chart Question Answering) with quantitative analysis, whereas our work presents the most comprehensive evaluation of LVLMs in the chart reasoning and comprehension domain with five downstream tasks with a wider range of qualitative and quantitative analyses. Therefore, our work clearly distinguishes itself from the above-mentioned works.

3 Methodology

3.1 Tasks and Datasets

Since chart comprehension and reasoning is a relatively new topic of research, very few tasks are proposed so far and there is a scarcity of benchmark resources, i.e., very few datasets, and models. Nevertheless, we have included all existing major chart-related downstream tasks for experiments.

(1) Factoid Chart Question Answering: For this task, we use ChartQA (Masry et al., 2022), a popular benchmark with a focus on visual and logical reasoning questions and was used as the only dataset for this task by GPT-4V (report, 2023) and Gemini (report, 2024) in their released reports. It features human-written questions from four real-world sources covering a wide range of topics.

(2) Chart Summarization: We choose Chart-to-Text (Shankar et al., 2022), a large-scale benchmark for chart summarization as well as Vistext (Tang et al., 2023), another recent chart captioning dataset.

(3) Open-ended Chart Question Answering: We consider OpenCQA (Kantharaj et al., 2022a), the only QA benchmark available for this task in which answers are provided as explanatory texts.

(4) Fact-checking with Charts: We utilize two currently available datasets: ChartFC dataset which contains (*statement*, *verdict*) as pairs (Akhtar et al., 2023a) and ChartCheck which has more diverse charts and contain explanations for verdicts.

(5) Chart-to-Table: We use the chart-table pairs from the ChartQA test set for the evaluation in this task. We created a new version of ChartQA, named (ChartQA*), by modifying the original charts to exclude explicit data value labels. This setup was introduced to evaluate whether the performance of LVLMs depends on explicit data labels or their ability to interpret data from the visual elements in the charts (more details in §A.1.2). In addition to the above tasks, we evaluate the semantic richness of the model's response by crafting a small dataset of 200 question-answer pairs based on four-level semantic frameworks (Lundgard and Satyanarayan, 2021). An overview of the test sets of these benchmarks is presented in Table 1.

3.2 Models

Since closed-source LVLMs currently achieve the best results in zero-shot scenarios in most vision-language benchmarks (Team et al., 2023), we select the following three: GPT-4V (*gpt-4-1106-preview*), Gemini (*gemini-1.0-pro-vision*), and Claude-3 (*claude-3-haiku@20240307*). While most open-source LVLMs underperform compared to closed-source ones, we include the Phi-3 (*phi-3-vision-128k-instruct*) model due to its impressive benchmark results. We compare these models with current SoTA chart-specific models, Math-Cha (Liu et al., 2022) and UniChart (Masry et al., 2023). Additionally, we assessed other open-source models like mPlug-DOC-owl-1.5 (Hu et al., 2024) and LLaVA-1.5 (Liu et al., 2024), but due to their subpar performance on chart-related tasks, we excluded them from our discussion.

3.3 Prompt Construction

In both qualitative and quantitative evaluation, we first create a task instruction T tailored to a specific test sample X. This instruction is then combined with the existing text of the test sample to form a unified prompt P. This prompt P and the Chart image C are provided as input to the respective

ChartQA								Chart-to-Table						OpenCQA		Chart-to-Text		Vistext		ChartFC		ChartCheck	
Human		Augmented		Total		ChartQA*		Human		Augmented		Total											
Charts	Qs.	Charts	Qs.	Charts	Qs.	Charts	Qs.	Charts	Tables.	Charts	Tables.	Charts	Tables.	Charts	Qs.	Pew	Stat.	Chart	Summ.	Supp.	Ref.	Test1	Test2
625	1250	987	1250	1612	2500	1340	2192	625	625	987	987	1612	1612	1159	1159	1393	5222	882	1270	885	706	937	981

Table 1: Test set of seven benchmarks: Here, "Qs.", "Pew", and "Stat." refer to Questions, Pew charts, and Statista charts, respectively. "Supp." and "Ref." denote the Support and Refute classes in ChartFC. ChartQA* denotes charts from the ChartQA test set without data labels.

LVLMs to generate the corresponding response R (see §A.3 for details and Table 9 for example prompts).

3.4 Evaluation

In addition to evaluating five benchmark chart-related tasks using existing metrics, we conduct specific evaluations on LVLM-generated responses, focusing on hallucination analysis and semantic coverage. Below, we explain our methodology.

3.4.1 Task-specific General Evaluation

ChartQA: We perform a comprehensive quantitative evaluation of the LVLMs on ChartQA in two different experimental setups, i.e., zero-shot Chain-of-Thought (CoT) (Wei et al., 2023), and Program-aided Language Models (PAL) (Gao et al., 2023) inspired by their recent success in various domains.

Chart Summarization & OpenCQA: To evaluate the performance of LVLMs in chart summarization and Open-ended Chart Question-Answering tasks, we follow prior work (Kantharaj et al., 2022c), and leverage a suite of automatic evaluation metrics, including BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2017), and Perplexity (Hugginface, 2023).

Fact Checking with Charts: For fact-checking, similar to prior work, we conduct a quantitative evaluation in terms of the F1 metric.

Chart-to-Table: For this task, we conduct a quantitative evaluation by reporting two metrics: the Relative Number Set Similarity (RNSS) (Masry et al., 2022) and the Relative Mapping Similarity (RMS) (Liu et al., 2023a).

3.4.2 Criteria-based Focused Evaluation

Hallucination Analysis: Hallucinations and factual errors are common in chart-related tasks (Kantharaj et al., 2022c,b; Tang et al., 2023). Therefore, we examine this issue using the **FAVA** (Mishra et al., 2024), which automatically detects and categorizes hallucinations in LLM outputs into different types. Although other works have proposed using specialized metrics for specific evaluation criteria, such as QAFactEval (Fabbri et al., 2022) and ChartVE (Huang et al., 2023), which are designed to assess factual consistency between text pairs or text-chart pairs, we plan to incorporate these metrics into our analysis as part of future work.

Generating different semantic levels: To assess the capability of LVLMs in generating texts about charts with rich semantics, we follow the four-level framework from Lundgard and Satyanarayan (2021): *Level 1* covers low-level information about the chart, i.e., chart type, axes, etc; *Level 2* presents statistical and relational aspects such as descriptive statistics and correlations; *Level 3* is about *perceptual and cognitive* phenomena describing complex trends, and patterns, and *Level 4* provides domain-specific insights such as social and political contexts. In our study, we evaluate the capabilities of LVLMs in their proficiency at covering these different types of semantic information. We also analyze their accuracy in interpreting questions and explaining answers across these four levels. Our *Level 1* semantic evaluation leveraged a collection of 40 charts encompassing a variety of types. We design five *Level 1* questions to assess core aspects of chart construction. These questions targeted attributes such as channel encoding (how data is represented visually), chart type (bar, line, pie, etc.), and axis labeling (x and y). In the case of *Level 2*, we design four questions to assess the ability of the models to identify extrema (maxima, minima) and outliers within charts. For *Level 3*, we include a wider range of 100 chart samples, with 28 being line charts. Finally, for *Level 4*, to evaluate the domain-specific text generation capability of LVLMs, we employ a test set of 200 charts.
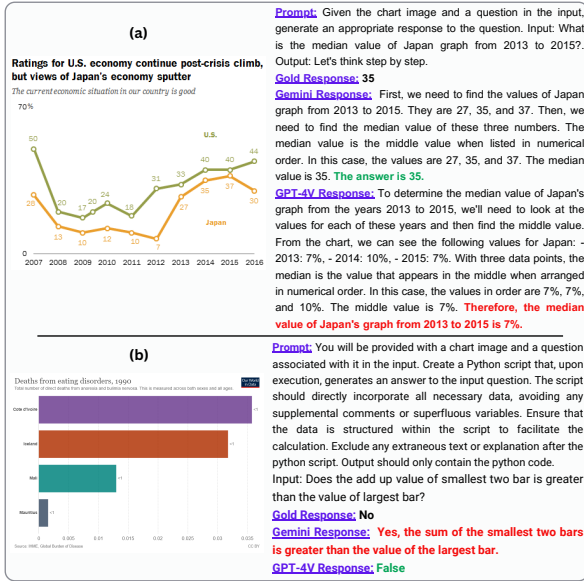
Figure 2: Figure (a) is where the Gemini is successful in 0-shot CoT, but the GPT-4V fails. Figure (b) shows the GPT-4V's success in PAL setup, while the Gemini fails. Here, **Red** text denotes incorrect, and **Green** text is correct.

4 Results and Discussion

4.1 General Observations

We present some key general observations based on our comprehensive evaluation of the LVLMs:

• Overall, among closed-source models, GPT-4V is the best performer in discriminative Chart reasoning and comprehension tasks, such as factoid chart question-answering and chart fact-checking while Gemini is better in open-ended generation tasks such as OpenCQA and Chart-to-Text. However, the open-source model Phi-3 achieves the best results on the ChartQA dataset (Table 2).

• Gemini is a better Chain-of-Thought reasoner, while GPT-4V and Claude-3 is better at generating code to answer questions about charts (Table 2).

• When the data values are not annotated in the charts, the performance of different models on ChartQA drops drastically (Table 6).

• `Entity` and `Relations` are the most frequent types of hallucinations encountered in all closed-sourced model-generated text (Table 3).

• In general, GPT-4V generates longer summaries with chart-specific (Level 1 & 3) semantic content, while Gemini generates more succinct summaries with statistical and domain-specific information (Level 2 & 4), and Claude-3 responses fall in between these two models.

4.2 Performance in ChartQA task

We perform a quantitative evaluation of the LVLMs in ChartQA in two different prompting setups, **(i)** zero-shot Chain-of-Thought (CoT)[2] (Wei et al., 2023) prompting and **(ii)** prompting strategy introduced in Program-aided Language models (PAL) (Gao et al., 2023). Initially, we evaluated the models' performance using the *'relaxed accuracy'* metric as discussed in (Masry et al., 2023). However, given the open-ended nature of the CoT responses, we conducted a manual evaluation to determine the actual accuracy of the models in the CoT setup. Also, we chose the PAL setup to examine whether separating the computation of complex queries and delegating them to a Python interpreter improves the performance of factoid question answering with charts. For further details on the datasets, refer to the §A.1. Below, we present our key findings:

Performance in Zero-shot CoT. In the case of zero-shot CoT, the Gemini outperformed GPT-4V by a margin of 3% (Table 2), while Claude-3 performed the worst, achieving an average accuracy of 44.56%. Fig. 2(a) depicts an example case where a line chart about the economic situation of two countries is given and the models are asked: *What is the median value of Japan graph from 2013 to 2015?*. With CoT reasoning, Gemini answered correctly, whereas GPT-4V answered incorrectly.

Performance in Program-Aided Reasoning. In this setup, the LVLMs were tasked with generating Python code to answer questions based on specific charts. Table 2 demonstrates that GPT-4V and Claude-3 achieved relatively higher performance levels compared to Gemini, indicating their greater proficiency in consistently producing more effective and functional code. Gemini's lower accuracy is mostly due to its inability to generate executable code on an average of 35% of cases, across both ChartQA test sets. Fig. 2(b) depicts an example where a bar chart illustrates the deaths from eating disorders in 1990 in four different countries and the models are prompted to answer the following question: *Does the add up value of smallest two bars is greater than the value of the largest bar?* Using the PAL method, GPT-4V answered correctly, while Gemini answered incorrectly.

Dependence on Data labels. For this experiment, we chose the two best performers in the

---

[2] We report the zero-shot result of Phi-3 in Table 2 from the technical report (Abdin et al., 2024) of the model.

Models	ChartQA (zero-shot CoT) (Accuracy)			ChartQA (zero-shot PAL) (Accuracy)			OpenCQA (BLEU)	Chart Summarization (BLEU)				Chart-Fact-checking (F1−score)			Chart-to-Table (RNSS)	(RMS)
	aug.	human	avg.	aug.	human	avg.		Pew	Statista	Vistext(L1)	Vistext(L2/L3)	ChartFC	ChartC(T1)	ChartC(T2)	ChartQA	ChartQA
Human baseline	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	95.7
Gemini (2023)	**74.96**	**70.72**	**72.84**	46.08	46.08	46.08	**6.84**	35.9	**25.8**	**27.4**	**15.7**	65.8	71.42	68.05	85.86	54.84
GPT-4V (2023)	72.64	66.32	69.48	75.44	**65.68**	**70.56**	3.31	28.5	18.2	18.2	11.3	**69.6**	**73.50**	71.30	81.51	**61.97**
Claude-3-haiku (2024)	47.12	42.00	44.56	76.88	63.44	70.16	4.58	**36.9**	25.8	25.2	14.2	61.4	71.70	**73.14**	**95.83**	50.65
Phi-3-vision-128k-inst (2024)	-	-	**81.40**	-	-	-	3.95	28.6	19.9	20.6	10.6	66.8	70.78	70.89	78.31	6.61
MatCha (2022)	90.20*	38.20*	64.20*	-	-	-	-	12.20	39.40	-	-	-	64.00	60.90	85.21	83.40
UniChart (2023)	88.56*	43.92*	66.24*	-	-	-	14.88	12.48	38.21	-	-	-	-	-	94.01	91.10
T5 (2022; 2022b)	-	-	59.80*	-	-	-	57.93	-	-	-	-	-	-	-	-	-
VL-T5 (2022; 2022b; 2023)	-	-	59.12*	-	-	-	59.80	-	-	-	32.90	-	-	-	-	-
OCR-T5 (2022c; 2023)	-	-	-	-	-	-	-	35.39	-	-	10.49	-	-	-	-	-
ResNet + BERT (2023a)	-	-	-	-	-	-	-	-	-	-	-	62.70	-	-	-	-
ChartLLaMA (2023)	-	-	69.66*	-	-	-	-	40.71	-	-	14.23	-	-	-	-	-
ChartAssistant (2024)	-	-	79.90*	-	-	-	15.50	41.00	-	-	15.20	-	-	-	-	92.00
Pix2struct (2022)	-	-	56.05*	-	-	-	12.70	38.00	-	-	10.30	-	-	-	-	-
ChartInstruct (2024a)	-	-	72.00*	-	-	-	16.71	43.53	-	-	13.83	-	72.65	-	-	-
ChartGemma (2024b)	-	-	80.16*	-	-	-	-	-	-	-	-	70.33	72.17	-	-	-

Table 2: An overview of the evaluation results on five tasks: ChartQA, Chart Summarization, OpenCQA, Chart-Fact-checking, and Chart-to-Table. Here, the ChartQA results with a '*' denote results without using CoT. The results except from Gemini, GPT-4V, Claude-3-haiku, and Phi-3-vision-inst, are noted based on the best-performing models as presented in the respective research paper.

ChartQA task in zero-shot CoT setup (Table 2). As demonstrated in Table 6, the absence of text labels that show data values diminishes the performance of both models, with GPT-4V being more affected. Moreover, GPT-4V frequently declines to respond when data labels are absent, as depicted in the right example in Figure 7. Our manual analysis suggests that these models exhibit better performance when the values of chart objects (e.g., bars, lines) align closely with the y-axis labels, leveraging these labels as a reference point, as illustrated in the left example in Fig. 7. Conversely, a disparity between the visual element values and y-axis labels leads to poorer performance. These findings underscore a critical limitation in the capabilities of both Gemini and GPT-4V in interpolating the data values of the chart visual elements (e.g., bars, lines, pie) based on their visual attributes (e.g., heights, areas).

4.3 Performance in Chart Summarization

We assess the text generation capabilities of LVLMs using both automatic metrics (see Table 7, Table 8) and qualitative[3] metrics.

Replication of Gold summaries. On the BLEU measure, Claude-3 and Gemini surpassed GPT-4V and Phi-3 in generating chart summaries that closely resemble the gold standard. However, in terms of BERTScore, all models performed similarly, suggesting identical performance when contextual similarity is considered instead of tokens.

---

[3]Since most closed-source models do not support fine-tuning, we specifically conduct human evaluation only on closed-source models to check how they perform in zero-shot.

Evidence of Factual errors in summaries. Our qualitative evaluation of the three closed-source models across 100 samples suggests that all models produce fluent and coherent text. However, all models demonstrated factual errors. More specifically, GPT-4 exhibited the lowest overall error percentage (5%), followed by Claude-3 (16%), while Gemini had the highest error percentage (27%). Examples of factually incorrect summaries generated by these models are provided in Table 19 and 20.

Inclusion of visual references in summaries. Referencing visual features in a chart can aid readers in coordinating between the chart and the text (Kim et al., 2020). In contrast to Gemini (25 samples out of 200) and Claude-3 (50 samples out of 200), GPT-4V consistently references (189 samples out of 200) visual features of charts in its summaries, often using phrases such as 'horizontal bars represent . . .'. Additionally, summaries generated by GPT-4V often contain incorrect references to color features (see Table 21). This inaccuracy is further evidenced by the fact that 80 of GPT-4V's 189 sentences contained errors, whereas Gemini and Claude-3 only had 24 and 7 errors respectively.

Identification of trends in charts. Identifying and explaining important trends is critical in chart analysis (Lundgard and Satyanarayan, 2021). Table 22 demonstrates an example where both models correctly captured simple trends. However, our investigation indicates that Gemini is more prone to missing trends, while GPT-4V and Claude-3 tend to capture trends more effectively (see Table 23).

Error Type	Example	Average Error Count (Per Summary)					
		Pew			Statista		
		Gemini	GPT-4V	Claude 3 Haiku	Gemini	GPT-4V	Claude 3 Haiku
Entity	Alberta is the top producer, with 126,082,558 billion cubic meters of natural gas.	**0.47**	0.51	1.39	**0.66**	0.88	1.85
Relation	The population density was lowest in 2018 and highest in 1960	**0.16**	0.17	0.17	0.17	0.21	**0.12**
Subjective	The chart shows that the number of cases is significantly higher in urban areas compared to rural areas.	0.02	0.02	**0.01**	0.02	0.02	**0.00**
Contradictory	There is a clear upward trend in the number of deaths caused by influenza and pneumonia over time. This trend is likely due to improvements in public health measures, such as vaccination and sanitation.	0.19	**0.12**	0.15	0.29	**0.14**	0.19
Unverifiable	Overall, the increase of percentage of people who have completed high school, has a positive impact on the United States.	**0.03**	**0.03**	**0.03**	0.05	0.04	**0.03**
Invented	The unemployment rate increased sharply from 3.3% in November 2019 to 15.7% in April 2020, the highest level since the Great Recession.	**0.02**	0.07	0.03	**0.03**	0.05	0.04
Total		**0.89**	0.92	1.76	1.26	1.35	2.23

Table 3: Color-coded table example of hallucinations detected in chart summaries by FAVA. Key: Red = entity hallucination; Orange = relation hallucination; Green = contradictory hallucination; Gold = invented hallucination. Subjective and unverifiable hallucinations exist at the sentence level and are not highlighted. Average error counts per type are included.

4.4 Performance in OpenCQA task

Similar to the Chart-to-Text task, OpenCQA favors Gemini on all automatic metrics, except perplexity (Table 5). Our qualitative evaluation across 100 samples shows that responses from all closed-source models were fluent and coherent but contained factual errors. Claude-3 had factual inaccuracies in about 4% of cases, GPT-4V in 5%, and Gemini in 17%. We observed that LVLMs, particularly GPT-4V, often generate detailed answers that include information that is not present in the gold answers but available in the chart. In our evaluation, we specifically checked whether the LVLM-generated responses contradicted any information mentioned in the gold answers.

4.5 Performance in Fact-checking task

We evaluated GPT-4V, Gemini, and Claude-3, and open-source models such as Phi-3 in the Fact-checking with charts task in the ChartFC (Akhtar et al., 2023a), and the ChartCheck dataset (Akhtar et al., 2023b) (see Table 2). We observed that all the evaluated models performed remarkably in the ChartCheck dataset, beating the existing SoTA Matcha (Liu et al., 2022) by some margin (see Table 2) in both of the test sets. Similarly, in the ChartFC dataset, GPT-4V, Gemini, and Phi-3 performed better than the existing SoTA reported in (Akhtar et al., 2023a), except the Claude-3 model. Our investigation also revealed that open-source models like Phi-3 produced more false negatives compared to their closed-source counterparts in this dataset (see Fig. 3). Nonetheless, in both benchmarks, the average accuracy is below 72%, indicating potential for further improvement in this task.

4.6 Performance in Chart-to-Table task

Chart-to-Table requires the model to extract the underlying data table from the provided chart image. To assess LVLMs capabilities in this task, we utilize the ChartQA dataset (Masry et al., 2022) which provides the underlying data tables for the chart image. As depicted in Table 2, notably, GPT4-V demonstrates superior performance in RMS, emphasizing its capability to accurately reconstruct the structure of tables derived from charts. Conversely, Gemini exhibits higher proficiency in RNSS, indicating its strength in accurately estimating numerical values from chart images.

4.7 Hallucination Analysis

To analyze hallucinations in LLM-generated responses, we sampled the chart summaries generated by Gemini, GPT-4V, and Claude-3 in the Chart-to-Text data. We used the Factuality-Aware Visual Analytics (FaVA) (Mishra et al., 2024) methodology for hallucination detection, by categorizing hallucinations into entity, relation, subjective, contradictory, unverifiable, and invented types (see a color-coded example in Table 3). The analysis showed that the `entity` category had the highest error count among all categories, which is consistent with findings in other NLP tasks (Mishra et al., 2024). Substantial errors also come from `Relation` and `contradictory` categories. Overall, Claude-3 had the highest total error count (1.76 for Pew, 2.23 for Statista), while Gemini (0.89 for Pew, 1.26 for Statista) and GPT-4V (0.92 for Pew, 1.35 for Statista) had fewer errors. The above finding highlights the urgent need to study and detect the frequent types of hallucinations (`entity` and `relations`) which are often
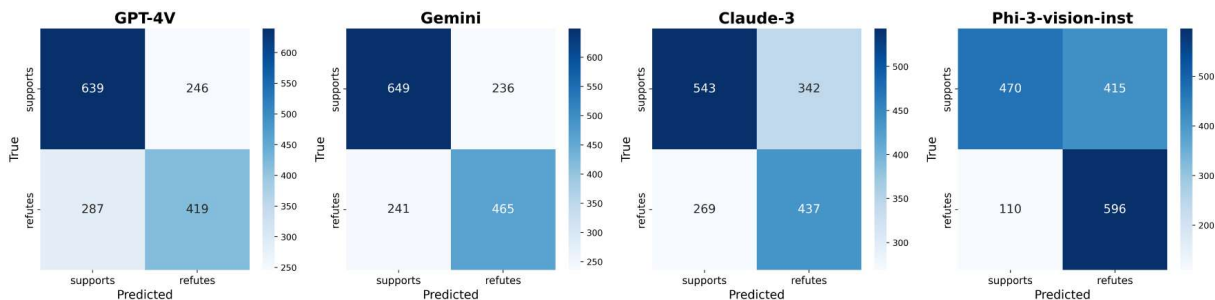
Figure 3: Confusion matrices for different LVLMs on the ChartFC dataset.

	Coverage		Accuracy (%)	
Semantic Level	GPT-4V	Gemini	GPT-4V	Gemini
L1: Visual encodings	**1.69**	1.25	**70.0**	57.5
L2: Statistical and relational	0.56	**0.87**	**80.5**	62.0
L3: Perceptual and cognitive	**0.70**	0.41	**58.9**	48.2
L4: contextual and domain-specific	0	**0.03**	15.5	**16.0**

Table 4: The performance of GPT-4V and Gemini in answering questions (Accuracy) and generating sentences across various semantic levels. 'Coverage' indicates average sentences per semantic level in summaries.

phrase-level and can be fixed by minimal editing erroneous phrases (Chen et al., 2023).

4.8 Analysis of Semantic Levels

For text generation tasks (e.g., chart summarization), a crucial question is how different semantic contents are covered in output texts and how accurately models can understand such statements. We analyze this question using the four-level semantic framework (Lundgard and Satyanarayan, 2021) as explained in §3.4. Research suggests that readers prefer chart summaries that describe more high-level trends, patterns, and contextual explanations (Levels 3 & 4) over low-level information, e.g., chart type, axes, color encodings, and simple statistics like averages and extrema (Levels 1 & 2) (Stokes et al., 2022). However, low-level information might be useful for some chart accessibility applications.

Generating different semantic contents. We manually examine model-generated texts for 200 chart-to-text samples to understand how they cover different types of semantic content. From Table 4 and Fig. 9, we observe that GPT-4V produces longer summaries of chart-specific visual information (Levels 1 & 3) while Gemini produces concise summaries with some statistical and domain-specific information (Levels 2 & 4) and Claude's outputs fall in-between these two models (more details in §A.2.1). We also observe that GPT-4V

not only produces statements describing high-level trends but also does so with higher accuracy than other models (see error examples in Fig. 4). Another important observation is that all models fail to include sufficient contextual and domain-specific information (Level 4) that explains trends and patterns in charts using external domain information (e.g., social and political contexts), which human authors often include in high-quality chart descriptions (e.g., Pew chart summaries).

Understanding different semantic contents. In another experiment, we examine LVLMs' ability to understand and answer questions across different types of semantics. To this end, we created 200 different question prompts for each of the four semantic levels using charts from the ChartQA dataset. We chose Gemini and GPT-4V as they are the top-performing closed-source models (see experimental details in §A.2.2).

From Table 4, we observe that GPT-4V outperforms Gemini in answering questions across all levels except for Level 4, in which their performance is similar. Both models struggle to describe complex trends in line charts with multiple, highly fluctuating lines. Fig. 4(a) illustrates such a scenario, where the chart indicates that *Ozone-depleting substance consumption in Gabon peaked in 2000*, but both GPT-4V and Gemini suggest otherwise.

Another interesting observation is that Gemini can extrapolation of factually accurate insights beyond the chart data. For example, in Fig. 4(b), although the x-axis labels of the bars began in May 2020, Gemini managed to describe trends by including previous years by outputting *"...The number of unemployed people reached a peak in April 2020 at 23.1 million and then started to decline."* While this information was not directly evident in the chart data, it aligns closely with statistics from the U.S. Bureau of Labor Statistics (Statistics, 2020). This finding is consistent with the obser-
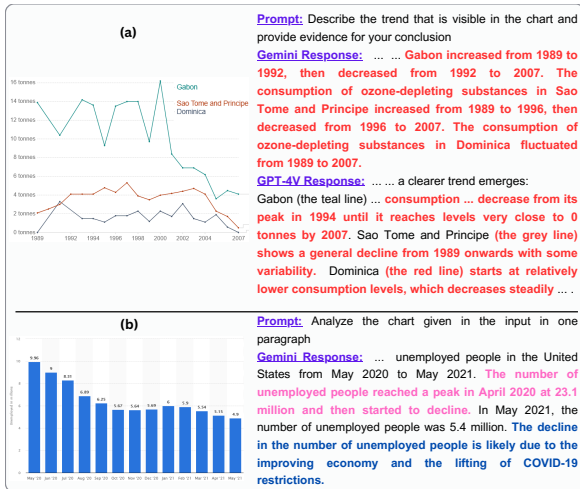
Figure 4: Both Gemini and GPT-4V fail to identify trends (a). Out-of-context but relevant information generated by Gemini (b). Here, **Red** text indicates incorrect facts, **Pink** text denotes out-of-context, and **Blue** text represents domain-specific details. '…' indicates abbreviated text for brevity.

vation that Gemini can cover more contextual and domain information from external sources.

4.9 Potential bias in Model responses

A notable concern with LVLMs is their potential for biased output generation (OpenAI et al., 2023). However, this critical issue remains unexplored in the chart domain. Prior research such as (Huang et al., 2024) has mentioned some issues of LVLMs, including coverage of chart content, fairness, and bias, but did not provide conclusive evidence directly confirming these issues. To bridge this gap, in a preliminary experiment, we manually analyzed 200 responses from GPT-4V, the top performer in understanding high-level semantics. We found several cases where the model's causal explanations might reflect training data biases. For example, in Fig. 5, the model attributed Estonia's slight GDP dip in 2008-2009 to the global financial crisis, which could be a spurious correlation. While Estonia's data did show a slight dip around that time, attributing it solely to the global financial crisis might be inaccurate since Lebanon and Costa Rica also experienced GDP increases during this period.

This finding highlights a pressing need to deeply explore the bias problem in the chart domain. An initial solution could be implementing measures like *pre-processing* (altering model inputs), *in-training* (adjusting the optimization process), *intra-processing* (changing inference behavior), and *post-processing* (rewriting model outputs) (Gallegos et al., 2024).
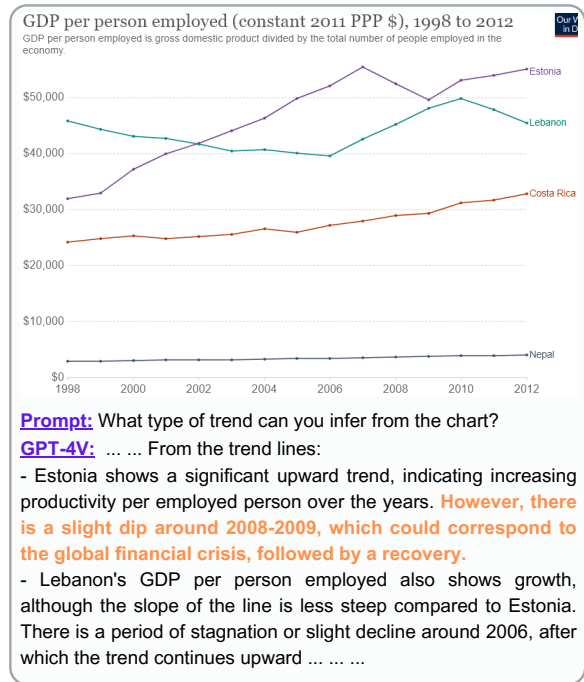

Figure 5: The figure is where the GPT-4V model shows some potential bias in the response. Here, **Orange** text denotes potentially biased output that is out-of-context of the chart data. '…' indicates abbreviated text for brevity.

5 Conclusion and Future Directions

This study presents a comprehensive analysis of LVLMs (GPT-4V, Gemini, Claude, and Phi-3) in interpreting and deriving insights from chart images in real-world scenarios, where data tables may not be available. Through qualitative and quantitative analyses, we evaluate these models across various tasks, including zero-shot CoT prompting and program-aided reasoning, assessing their impact on chart question-answering tasks. Additionally, we examine LVLMs' performance in open-ended text generation from chart related tasks.

These analyses highlight both the strengths and limitations of LVLMs and identify key research gaps. First, enhancing the generalizability and reasoning abilities of open-source LVLMs in chart-related tasks is a priority which can be explored via instruction tuning (Masry et al., 2024a). Second, there is significant potential for LVLMs to produce semantically rich texts that describe high-level trends and contextual information more effectively. Third, addressing key issues such as hallucinations, factual errors, and bias requires developing new benchmarks and models for detection and mitigation. We hope that the insights gained from this study will catalyze further research and advancements in the emerging area of chart reasoning.

Limitations

Since the pretraining corpus of both the large vision language models (LVLMs) is unknown (not open-source), some of the datasets used for evaluation may or may not appear in the pretraining data or instruction tuning data of the models. Although we covered all the important tasks, i.e., Chart Summarization, Chart Question-Answering, Open-ended Chart Question-Answering, and Fact Checking with Charts, etc., there are some tasks, i.e., Chart-to-table not addressed in this research. At the time of evaluation, we did not provide any underlying data table corresponding to the chart in the input. However, our motivation for this research was to show how different state-of-the-art LVLMs perform when the underlying data table is not present for chart understanding tasks, which is often the case in real-world scenarios. Further, variations of charts and labels are limited due to the open-sourced datasets available for the tasks. We did not perform the qualitative evaluation in the ChartQA task, since the task is based on factoid-QA about Charts and only requires single token answers (either text or a numerical value), for which automatic evaluation is sufficient.

Ethics Statement

This study independently evaluated LVLMs' responses without involving any external parties, hence, no extra financial compensation was necessary. The authors themselves performed all the human assessments presented in this paper. As the focus of the research was solely on assessing LVLM's capabilities, effectiveness, and limitations in several chart understanding tasks, the human evaluation performed by the authors does not add any ethical issues or unwanted biases. Further, the datasets utilized in this study are all open-sourced academic datasets, thus licensing was not required. Additionally, no information has been used that can directly relate to the identification of any person while evaluating the responses from LVLMs.

Acknowledgement

References

Marah Abdin, Sam Ade Jacobs, and Ammar Ahmad et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023a. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2023b. Chartcheck: An evidence-based fact-checking dataset over real-world chart images. *Preprint*, arXiv:2311.07453.

Md Zubair Ibne Alam, Shehnaz Islam, and Enamul Hoque. 2023. Seechart: Enabling accessible visualizations through interactive natural language interface for people with visual impairments. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 46–64, New York, NY, USA. Association for Computing Machinery.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, and Dmitry Lepikhin et al. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Anthropic. 2024. Introducing the next generation of claude.

Benjamin Bach, Mandy Keck, Fateme Rajabiyazdi, Tatiana Losev, Isabel Meirelles, Jason Dykes, Robert S Laramee, Mashael AlKadi, Christina Stoiber, Samuel Huron, et al. 2023. Challenges and opportunities in data visualization education: A call to action. *IEEE Transactions on visualization and computer graphics*.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.

Z. Cheng, Q. Dai, and A. G. Hauptmann. 2023. Chartreader: A unified framework for chart derendering and comprehension without heuristic rules. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22145–22156, Los Alamitos, CA, USA. IEEE Computer Society.

J. Choi, Sanghun Jung, Deok Gun Park, J. Choo, and N. Elmqvist. 2019. Visualizing for the non-visual: Enabling the visually impaired to use visualization. *Computer Graphics Forum*, 38.

Aakanksha Chowdhery, Sharan Narang, and Jacob Devlin et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Xuan Long Do, Mohammad Hassanpour, Ahmed Masry, Parsa Kavehzadeh, Enamul Hoque, and Shafiq Joty. 2023. Do llms work on charts? designing few-shot prompts for chart question answering and summarization. *Preprint*, arXiv:2312.10610.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5436–5443. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Chaoyou Fu, Renrui Zhang, Zihan Wang, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, Yunhang Shen, Mengdan Zhang, Peixian Chen, Sirui Zhao, Shaohui Lin, Deqiang Jiang, Di Yin, Peng Gao, Ke Li, Hongsheng Li, and Xing Sun. 2023. A challenger to gpt-4v? early explorations of gemini in visual expertise. *Preprint*, arXiv:2312.12436.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, pages 1–79.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. *Preprint*, arXiv:2211.10435.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

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.

Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. 2022. Chart question answering: State of the art and future directions. *Journal of Computer Graphics Forum (Proc. EuroVis)*, pages 555–572.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *Preprint*, arXiv:2403.12895.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2023. Bliva: A simple multimodal llm for better handling of text-rich visual questions. *Preprint*, arXiv:2308.09936.

Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *Preprint*, arXiv:2403.12027.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R. Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2023. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. *Preprint*, arXiv:2312.10160.

Hugginface. 2023. Huggingface evaluation metric, perplexity.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*, pages 1–20.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022a. Opencqa: Open-ended question answering with charts. In *Proceedings of EMNLP (to appear)*.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022b. Opencqa: Open-ended question answering with charts. *arXiv preprint arXiv:2210.06628*.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022c. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min-joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2022. Pix2struct: Screenshot parsing as pretraining for visual language understanding. *arXiv preprint arXiv:2210.03347*.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada. Association for Computational Linguistics.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2022. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. *arXiv preprint arXiv:2212.09662*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Alan Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083.

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.

Ahmed Masry, Do Long, 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, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024a. Chartinstruct: Instruction tuning for chart comprehension and reasoning. *Preprint*, arXiv:2403.09028.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2024b. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. *Preprint*, arXiv:2407.04172.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

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)*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucinations detections. *arXiv preprint*.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et al. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Zhangyang Qi, Ye Fang, Mengchen Zhang, Zeyi Sun, Tong Wu, Ziwei Liu, Dahua Lin, Jiaqi Wang, and Hengshuang Zhao. 2023. Gemini vs gpt-4v: A preliminary comparison and combination of vision-language models through qualitative cases. *Preprint*, arXiv:2312.15011.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md. Tahmid Rahman Laskar, Md. Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *Proceedings of the Canadian Conference on Artificial Intelligence*.

Gemini Technical report. 2024. Gemini: A Family of Highly Capable Multimodal Models. Technical report, Stanford InfoLab.

GPT 4 Technical report. 2023. GPT-4 Technical Report . Technical report, Stanford InfoLab.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Kantharaj Shankar, Leong Rixie Tiffany Ko, Lin Xiang, Masry Ahmed, Thakkar Megh, Hoque Enamul, and Joty Shafiq. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2022*.

Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, and Gust Verbruggen. 2023. Assessing gpt4-v on structured reasoning tasks. *Preprint*, arXiv:2312.11524.

U.S. Bureau Labor Statistics. 2020. Uemployment-rate.

Chase Stokes, Vidya Setlur, Bridget Cogley, Arvind Satyanarayan, and Marti A Hearst. 2022. Striking a balance: Reader takeaways and preferences when integrating text and charts. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1233–1243.

Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, and Jean-Baptiste Alayrac et al. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, and Amjad Almahairi et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2024. Chartbench: A benchmark for complex visual reasoning in charts. *Preprint*, arXiv:2312.15915.

Yuchen Zhang, Panupong Pasupat, and Percy Liang. 2017. Macro grammars and holistic triggering for efficient semantic parsing. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 1214–1223.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

A Appendices

A.1 Datasets

A.1.1 ChartQA

In our study, we employ the test set from the ChartQA dataset, as introduced by Masry et al. (2022). The test set of the dataset is composed of two primary categories of questions: those created by humans and those augmented by models. Specifically, the set of human-generated questions includes 625 distinct charts with 1250 corresponding question-answer pairs. Similarly, the model-generated, or augmented set, comprises 987 unique charts and 1250 question-answer pairs.

A.1.2 ChartQA*

We introduce this dataset as a variation of the ChartQA dataset, in which charts do not explicitly show data values as labels near the corresponding chart elements (e.g., bars, lines), rather the model needs to estimate these values from the chart (e.g., based on bar heights and axis labels). We introduce this setup to see whether LVLMs' performance relies on the explicit labels of the data values rather than their ability to recover data values from the visual elements in the chart. For this purpose, we modified the ChartQA dataset using Matplotlib (Hunter, 2007), removing the data labels from the chart images while keeping everything else the same (see examples in Figure 6). Of the 1509 chart images in the test set, 1340 were successfully redesigned. The remaining 169 images were excluded due to missing metadata.

A.1.3 ChartFC

For the fact-checking with charts task, we utilize the ChartFC dataset from Akhtar et al. (2023a). The dataset is structured so that each entry contains a claim phrased in natural language, a related chart image, and a label that falls into one of two categories: *'supports'* or *'refutes'*. We evaluate the LVLMs in the test set of the dataset, which contains 885 examples belonging to the *'supports'* class and 706 examples belonging to the *'refutes'* class. The dataset comprises bar charts with different types, i.e., horizontal/vertical.

A.1.4 ChartCheck

For the fact-checking with charts task, we utilize another dataset from Akhtar et al. (2023b). The dataset is structured so that each entry contains a claim phrased in natural language, a related chart image, and a label that falls into one of two categories: *'True'* or *'False'*. The dataset contains two test sets, where test set-1 contains 937 samples and test set-2 contains 981 samples. We evaluated the models in both of these test samples.

A.1.5 Chart-to-Text

In our chart summarization study, we utilize the Chart-to-Text (Kantharaj et al., 2022c) benchmark. This benchmark encompasses two primary datasets: Statista[4] and Pew[5]. Each sample within both datasets comprises a chart image, an underlying data table, a chart title, and a human-written gold summary. For our experimental purposes, we utilize the complete test split from each dataset, encompassing 1,393 samples from Pew and 5,222 samples from Statista.

A.1.6 Vistext

The VisText dataset (Tang et al., 2023) includes 12,441 pairs of charts and descriptive captions. The data tables in the dataset were sourced from the Statista corpus of the Chart-to-Text (Kantharaj et al., 2022c) benchmark. The dataset includes two different types of captions, i.e., L1 and L2/L3 which correspond to the semantic levels of (Lundgard and Satyanarayan, 2021). These captions offer insights into the charts' construction, highlight important statistics, and point out perceptual and cognitive phenomena. Each chart in VisText is represented in three ways: as a rasterized image, as a data table, and as a scene graph, which is a structured representation of the chart's visual elements akin to the Document Object Model (DOM) used in web pages.

A.1.7 OpenCQA

To study LVLMs performance on the Open-ended Chart Question-Answering task, we utilize the benchmark dataset OpenCQA from Kantharaj et al. (Kantharaj et al., 2022b). The dataset contains five different types of charts, i.e., bar, line, area, scatter, and pie. For our experiments, we use the test set from the dataset which comprises 1159 charts and 1159 question-answer pairs.

A.2 Analysis of 4-level Semantics

A.2.1 Coverage of 4-level semantic contents

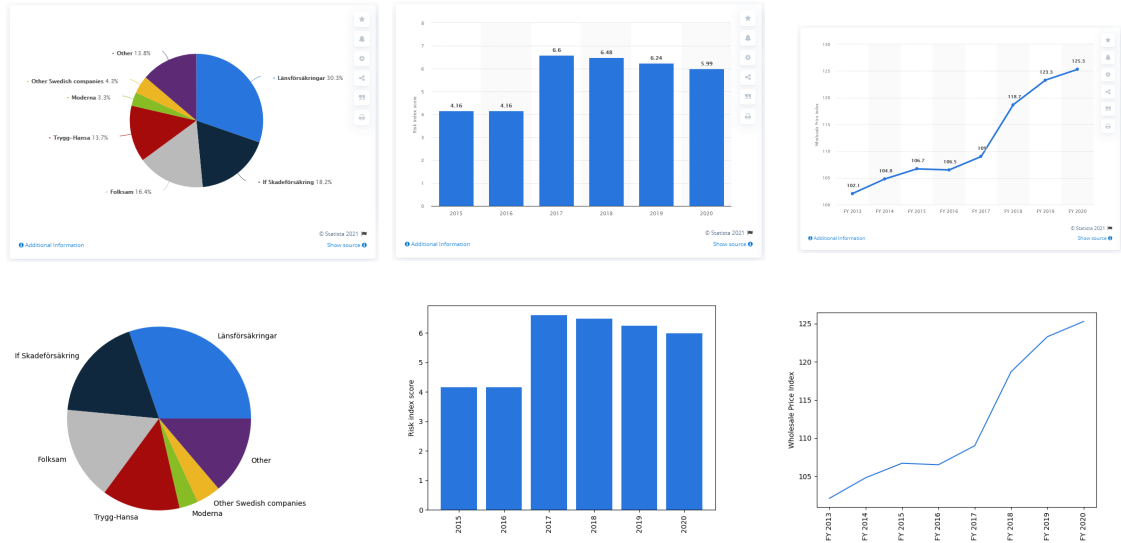To assess the quality of summaries generated by Gemini, GPT-4V, and Claude-3, we conducted a

---

[4]https://www.statista.com/
[5]https://www.pewresearch.org/

3347

Figure 6: Examples of charts with and without the data labels.

Model	BLEURT (↑)	CIDEr (↑)	PPL (↓)	BERTScore (↑)
Gemini	**-0.28**	**1.88**	2.06	**0.87**
GPT-4V	-0.45	1.63	**1.85**	0.85

Table 5: Evaluation results for different models on OpenCQA . ↑ : Higher is better, ↓ : Lower is better.

Model	ChartQA	ChartQA*
Gemini	**52.04**	38.53 (↓ 13.51%)
GPT-4V	**57.51**	20.52 (↓ 36.99%)

Table 6: Relaxed Accuracy (RA) different models on the **ChartQA*** vs **ChartQA** test set. Here, ChartQA* denotes the charts from the test set of the ChartQA dataset without the annotations. Drop in performance compared to **ChartQA** is presented in round brackets.

| Model	BLEURT (↑)		CIDEr (↑)		PPL (↓)		BERTScore (↑)	
	Pew	Stat	Pew	Stat	Pew	Stat	Pew	Stat
Gemini	**-0.30**	-0.30	**1.79**	1.90	1.61	1.70	**0.87**	0.86
GPT-4V	**-0.30**	-0.40	1.34	1.28	1.69	1.75	0.85	0.85
Claude-3-Haiku	-0.31	**-0.25**	1.56	**1.91**	1.72	1.75	**0.87**	**0.89**
Phi-3-vision-128k-instruct	-0.88	-0.49	1.47	1.54	**1.49**	**1.51**	0.85	0.86

Table 7: Detailed automatic evaluation results for different models on the Chart-to-Text dataset for Chart Summarization. ↑ : Higher is better, ↓ : Lower is better.

Model	BLEURT (↑)		CIDEr (↑)		PPL (↓)		BERTScore (↑)	
	Pew	Stat	Pew	Stat	Pew	Stat	Pew	Stat
Gemini	-0.25	-0.99	2.62	**1.17**	1.83	1.82	**0.88**	**0.87**
GPT-4V	-0.11	-0.98	2.02	0.99	1.77	1.94	0.87	0.86
Claude-3-Haiku	-0.16	-0.97	2.51	**1.13**	1.85	1.85	**0.88**	**0.87**
Phi-3-vision-128k-instruct	**-0.09**	-1.19	2.96	1.13	**1.48**	**1.49**	0.88	0.85

Table 8: Detailed automatic evaluation results for different models on the Vistext dataset for Chart Summarization. ↑ : Higher is better, ↓ : Lower is better.

detailed analysis of 200 randomly sampled summaries (50 from Pew, 150 from Statista) from the Chart-To-Text dataset generated by these three LVLMs. We meticulously examined each sentence, categorizing it according to the four-level semantic framework. Level-1 sentences focused on entities (axis labels, titles) and chart values. Level-2 highlighted extrema (minimum/maximum) values within the chart, while Level-3 captured trends and patterns. Level 4 addressed sentences requiring domain-specific knowledge external to the chart.

Figure 8 summarizes the key findings. Our analysis revealed the following:
In Level-1 (Entity and Value), Gemini generated 250 sentences where 10 sentences contained fac-

tual errors; GPT-4V generated 339 sentences, with 21 sentences containing errors; and Claude-3 generated 279 sentences, with 8 sentences containing errors. Both GPT-4V and Claude-3 generated significantly more Level-1 sentences compared to Gemini, with GPT-4V leading in the number of generated sentences but also having a higher error count and percentage (6.19%).
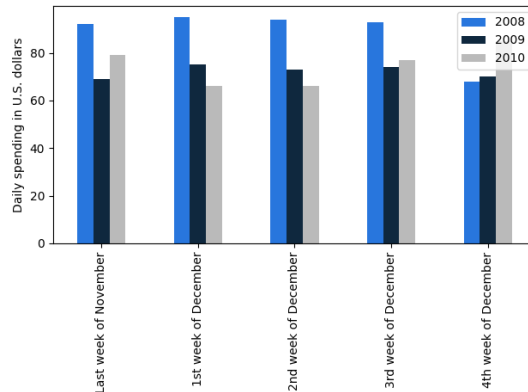In Level-2 (Extrema Related Statements), Gemini generated 174 sentences, with 44 containing errors; GPT-4V generated 111 sentences, with 13 containing errors; and Claude-3 generated 132 sentences, with 13 containing errors. In this level Gemini produced the most sentences but had a notably higher
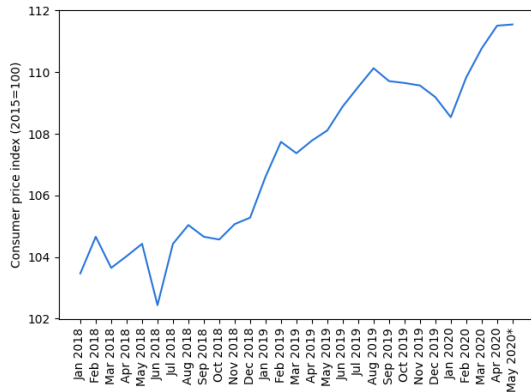
Question: What was the index value of food and non-alcoholic beverages in June 2018?

GPT4-V Answer: 104
Ground Truth Answer: 102.44

Question: Can you able to find the data value for 2008, 4th week of December?

GPT4-V Answer: the exact data value cannot be determined from the chart as provided.
Ground Truth Answer: 68

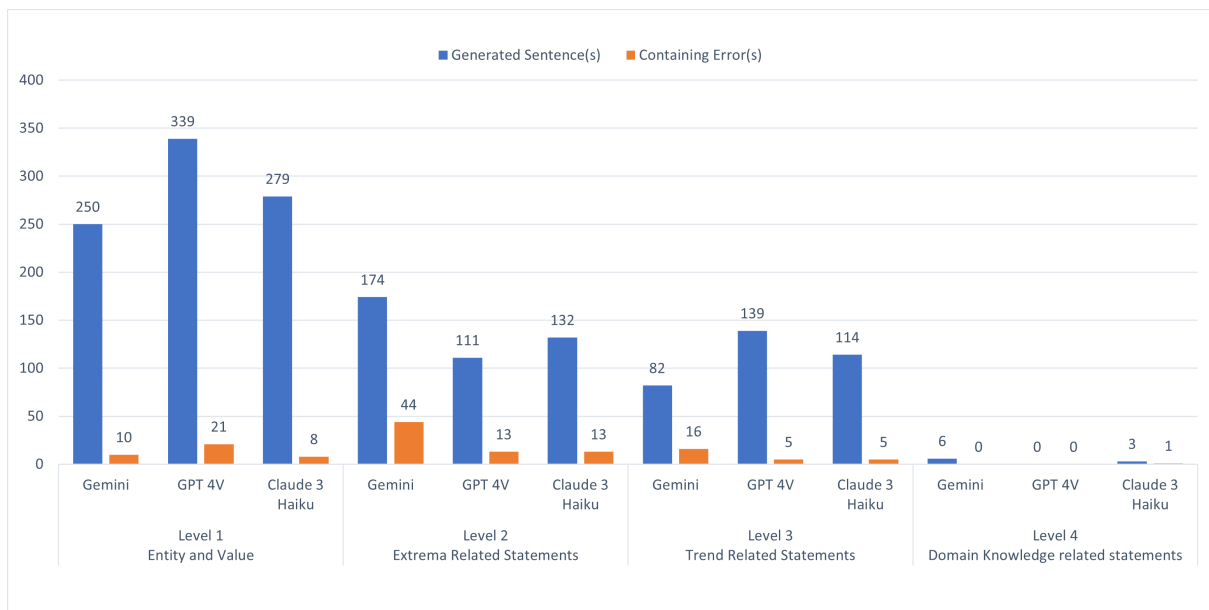Figure 7: Sample outputs from GPT-4V on the ChartQA* benchmark.

Figure 8: Chart depicts the number of sentences generated by three LVLMs, Gemini, GPT-4V, and Claude-3, at each semantic level (Entity and Value, Extrema Related Statements, Trend Related Statements, Domain Knowledge Related Statements).

error rate (25.29%) compared to GPT-4V (11.71%) and Claude-3 (9.85%).

In Level-3 (Trend Related Statements), Gemini generated 82 sentences, with 16 containing errors; GPT-4V generated 139 sentences, with 5 containing errors; and Claude-3 generated 114 sentences, with 5 containing errors. GPT-4V significantly outperformed Gemini in Level-3, generating 139 trend-related sentences with only 5 errors, while

Claude-3 also performed well with a similar error count but fewer generated sentences. In this level, Gemini still had the highest error percentage of 19.51% compared to GPT-4V (3.60%) and Claude-3 (4.39%).

In Level-4 (Domain Knowledge Related Statements), Gemini generated 6 sentences, with no errors; GPT-4V did not generate any sentences; and Claude-3 generated 3 sentences, with 1 con-
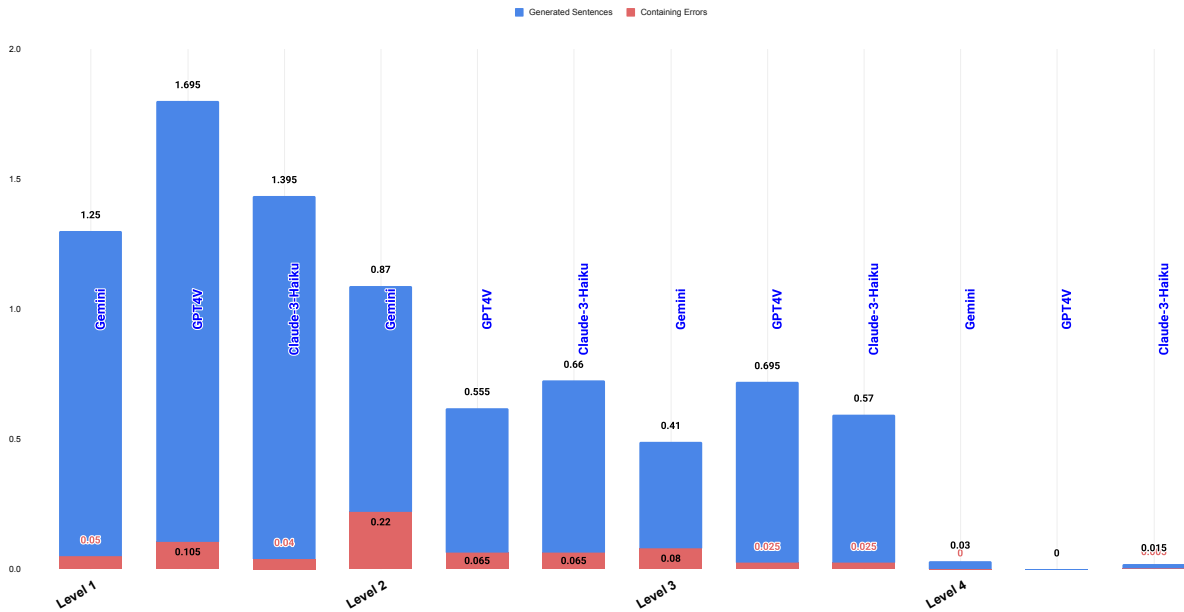
Figure 9: Chart depicts the average number of sentences generated by three LVLMs, Gemini, GPT-4V, and Claude-3, at each semantic level (Entity and Value, Extrema Related Statements, Trend Related Statements, Domain Knowledge Related Statements).

taining an error. All three models struggled with Level-4 sentences, reflecting the challenges of incorporating domain-specific knowledge.

A.2.2 Understanding of 4-level semantics

Experimental Setup. In order to evaluate the models in the four semantic levels, we utilize the charts from the ChartQA (Masry et al., 2022) dataset, and generate 200 different question prompts each for the four semantic levels, i.e., *Level 1*, *Level 2*, *Level 3*, and *Level 4*, to evaluate both GPT-4V and Gemini models. Our *Level 1* semantic evaluation leveraged a collection of 40 charts encompassing a variety of types. We design five *Level 1* questions to assess core aspects of chart construction. These questions targeted attributes such as channel encoding (how data is represented visually), chart type (bar, line, pie, etc.), and axis labeling (x and y). Notably, the chart set comprised a dominant presence of bar charts (70%), further categorized as horizontal/vertical, simple/stacked/grouped variants. Line charts constituted 17.5% of the collection, with pie charts making up the remaining 12.5%. In the case of *Level 2*, we design four questions to assess the ability of the models to identify extrema (maxima, minima) and outliers within charts. We include a diverse set of 50 chart types, with bar charts comprising the majority (68%), followed by line charts (20%) and pie charts (12%). In the case of *Level 3*, we include a wider range of 100 chart samples,

with 28 being line charts. The distribution of chart types at this level remains similar, with bar charts (62%) holding dominance, followed by line charts (28%) and pie charts (10%). Finally, for *Level 4*, to evaluate the domain-specific text generation capability of the LVLMs, we employ a test set of 200 distinct chart types.

Additional details about the performance of the models. In the 'Understanding of 4-level semantics' evaluation, in a subset of 40 samples where color encoding information was queried, both GPT-4V and Gemini models struggled. Results indicate that Gemini provided incorrect answers 52.5% of the time, while GPT-4V had a slightly higher error rate at 62.5%. In another experiment, for each chart, we designed two questions focused on *Level 3* semantic content. We specifically asked 56 questions regarding the trends present in the line charts. Our analysis revealed that GPT-4V failed to describe line chart trends correctly in 41.07% of cases. Gemini demonstrated a higher error rate, failing to identify the correct trend in 51.78% of instances. While the models excel in recognizing simple, steadily increasing, or decreasing trends in charts related to semantic *Level 3*, they struggle with line charts featuring multiple, highly fluctuating lines.

A.3 Prompt Construction

In order to come up with the best-performing prompt, we tried many different techniques and used the one that gave a consistent performance. For the zero-shot PAL experiment, we specifically designed the prompt asking the model to output a Python script, which upon execution would give us the final answer to the question. In the case of the 4-Level semantics experiment, we devised questions pertinent to each semantic level and aimed to evaluate the models' proficiency in identifying the various levels of semantic information embedded in the chart image. We created questions relevant to each of the semantic levels, targeting each of the semantic levels, i.e., *Level - 1* (e.g., chart type, x-axis/y-axis labels, color encoding information, etc.), *Level - 2* (e.g., maxima, minima, or outliers), *Level - 3* (e.g., trends or patterns), *Level - 4* (e.g., domain-specific insights). Example prompts can be found in Table 9.

A.4 Additional Experimental Results

In this section, we present additional experimental results of our automatic evaluation of the Chart-to-text, OpenCQA benchmark, and the newly created ChartQA[*] benchmark. Here, Table 7 represents the performance on Chart-to-text, while Table 5 represents the performance of Gemini and GPT-4V across different metrics, i.e., BLEURT, CIDEr, Perplexity, and BERTScore, and Table 6 represents the performance of the models in ChartQA[*] benchmark.

Task	Setup	Prompt
ChartQA	Chain-of-Thought (CoT)	Given the chart image and a question in the input, generate an appropriate response to the question. Input: {question}. Output: Let's think step by step.
	Program-aided Language Modeling (PAL)	You will be provided with a chart image and a question associated with it in the input. Create a Python script that, upon execution, generates an answer to the input question. The script should directly incorporate all necessary data, avoiding any supplemental comments or superfluous variables. Ensure that the data is structured within the script to facilitate the calculation. Exclude any extraneous text or explanation after the python script. Output should only contain the python code. Input: {question}
4-level of semantic contents	Level - 1	1. What is the chart type in the input image? 2. What is the range of x-axis? 3. What is the range of y-axis? 4. What are the x-axis and y-axis labels in the chart? 5. What do each of the colors represent in the chart? 6. What is the chart type in the input image?
	Level - 2	1. Identify the axis that contains a numerical range. What is the maximum value in that axis? 2. Identify the axis that contains a numerical range. What is the minimum value in that axis? 3. Are there any outliers in the chart? 4. Compare between the labels that hold the minimum and maximum values.
	Level - 3	1. What type of trend can you infer from the chart? 2. Describe the trend that is visible in the chart and provide evidence for your conclusion.
	Level - 4	Analyze the chart given in the input in one paragraph.
Open-ended Chart QA	-	Provide an open-ended answer to the following question based on the provided chart image.
Chart Summarization	Chart-To-Text	The attached chart shows {title}. Summarize the chart in a single paragraph focusing on trends and important data points. While summarizing, focus on the axis and color-related information in the chart.
	Vistext (L1)	Summarize the attached chart in a single paragraph focusing on the chart's elemental and encoded properties.
	Vistext (L2/L3)	Summarize the attached chart in a single paragraph focusing trends and statistics about the chart.
Fact-Checking with Charts	-	For the given chart image, determine if the following claim statement in the input is supported by the chart. If supported, then output 'supports', otherwise output 'refutes'. Input: {claim}
Chart-to-Table	-	Extract the underlying data table from the provided chart image. Each row should be on a separate line and use | to separate the cells in the same row by following this format: Column Header 1 | Column Header 2 Cell 1 | Cell 2 Cell 3 | Cell 4

Table 9: Example of the prompts used to evaluate the LVLMs.

Model	Area (L1)	Area (L2L3)	Bar (L1)	Bar (L2L3)	Line (L1)	Line (L2L3)
Gemini	33.80	18.30	30.50	17.60	33.30	19.10
GPT4V	21.60	13.20	21.30	12.90	21.50	14.10
Claude-3-Haiku	31.10	17.20	29.40	16.20	30.60	17.30
Phi-3-vision-inst	22.80	11.90	23.80	12.10	23.60	13.10

Table 10: Detailed breakdown of LVLMs' performance across chart types in the VisText dataset. Here, 'L1', 'L2L3' denote the chart caption types.
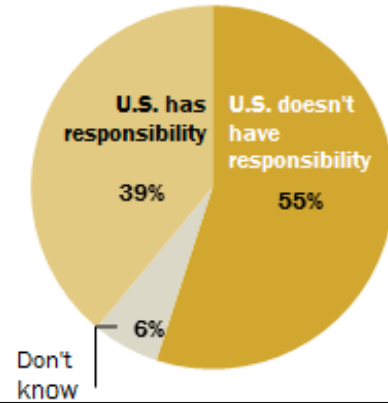
ChartQA-Human

Prompt: Given the chart image and a question in the input, generate an appropriate response to the question. Input: Does the sum of smallest two segment value is greater then the the largest segment value of the graph? Output: Let's think step by step.

Do you think the U.S. has a responsibility to do something about the violence in Iraq?

Gold-Label: No

Gemini-pro-vision Response: The smallest two segments are "Don't know" and "U.S. has responsibility". Their values are 6% and 39%, respectively. Their sum is 45%. The largest segment is "U.S. doesn't have responsibility" with the value of 55%. **Therefore, the sum of the smallest two segments is not greater than the largest segment value.**

GPT-4V Response: To determine this, we must add the values of the two smallest segments and compare the sum to the value of the largest segment. The smallest segments are 6% ("Don't know") and 39% ("U.S. has responsibility"), which sum up to 45%. The largest segment is 55% ("U.S. doesn't have responsibility"). Comparing the two sums, 45% is less than 55%, **meaning the sum of the smallest two segments is not greater than the value of the largest segment.**

Table 11: An example from the ChartQA-human test set in the zero-shot CoT setup where both the Gemini-pro-vision and GPT-4V models responded correctly, marked in **Green** text.
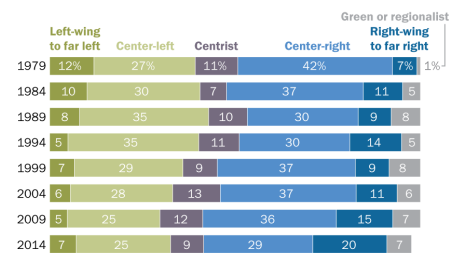

ChartQA-Human

Prompt: Given the chart image and a question in the input, generate an appropriate response to the question. Input: What is the average of the smallest gray bar and largest light blue bar? Output: Let's think step by step.

Can the European Parliament's center hold?

Share of members of the European Parliament belonging to ____ parties or political groups

Gold-Label: 21.5

Gemini-pro-vision Response: The smallest gray bar is 1% in 1979 and the largest light blue bar is 37% in 2004. **So the average of the two is (1+37) / 2 = 19%.**

GPT-4V Response: To find the average of the smallest gray bar and the largest light blue bar, we must first identify these values from the chart. In this chart, the gray bars represent the "Centrist" political group while the light blue bars represent the "Center-right" political group. The smallest percentage for the gray "Centrist" bars is in the year 2014 with 9%. The largest percentage for the light blue "Center-right" bars is in the year 2014 with 29%. Now, we calculate the average of these two percentages: Average = (Smallest gray bar value + Largest light blue bar value) / 2 Average = (9% + 29%) / 2 Average = 38% / 2 Average = 19% **So, the average of the smallest gray bar and the largest light blue bar is 19%.**

Table 12: An example from the ChartQA-human test set in the zero-shot CoT setup where both the Gemini-pro-vision and GPT-4V models responded incorrectly, marked in **Red** text.

ChartQA-Augmented

Prompt: Given the chart image and a question in the input, generate an appropriate response to the question. Input: What was the average number of applications per job posting on LinkedIn in the United States? Output: Let's think step by step.
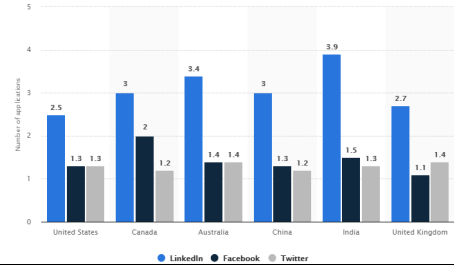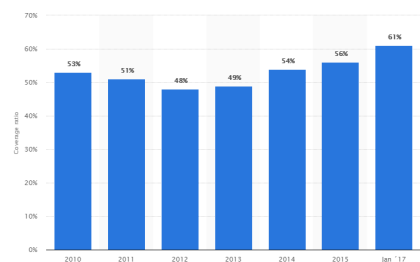

Gold-Label: 2.5

Gemini-pro-vision Response: The average number of applications per job posting on LinkedIn in the United States was 2.5.

GPT-4V Response: Based on the chart provided, the average number of applications per job posting on LinkedIn in the United States is 2.5.

Table 13: An example from the ChartQA-augmented test set in the zero-shot CoT setup where both the Gemini-pro-vision and GPT-4V models responded correctly, marked in **Green** text.


ChartQA-Augmented

Prompt: Given the chart image and a question in the input, generate an appropriate response to the question. Input: In what year did the coverage ratio for bad loans increase steadily? Output: Let's think step by step.

Gold-Label: 21.5

Gemini-pro-vision Response: The coverage ratio for bad loans increased steadily from 2010 to 2017. In 2010, the coverage ratio was 53%. In 2011, it increased to 51%. In 2012, it increased to 48%. In 2013, it increased to 49%. In 2014, it increased to 54%. In 2015, it increased to 56%. In 2016, it increased to 61%.

GPT-4V Response: The coverage ratio for bad loans began to increase steadily starting in the year 2013. From this point onward, the chart shows a consistent upward trend, from 48% in the year 2012 to 61% in January 2017.

Table 14: An example from the ChartQA-augmented test set in the zero-shot CoT setup where both the Gemini-pro-vision and GPT-4V models responded incorrectly, marked in **Red** text.

ChartQA-Human

Prompt: You will be provided with a chart image and a question associated with it in the input. Create a Python script that, upon execution, generates an answer to the input question. The script should directly incorporate all necessary data, avoiding any supplemental comments or superfluous variables. Ensure that the data is structured within the script to facilitate the calculation. Exclude any extraneous text or explanation after the python script. Output should only contain the python code. Input: What's the ratio of smallest segment and second largest segment?

Which is more important for Germany – to have strong ties with the U.S. or strong ties with Russia?
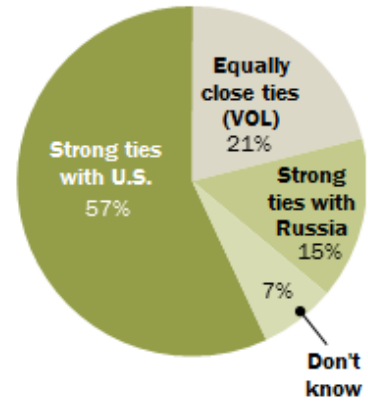
Gold-Label: 0.33333333
Gemini-pro-vision Response: Ratio of smallest (7) and second largest (21) segments: 0.33.
GPT-4V Response: 0.3333333333

Table 15: An example from the ChartQA-human test set in the zero-shot PAL setup where both the Gemini-pro-vision and GPT-4V models responded correctly, marked in Green text.
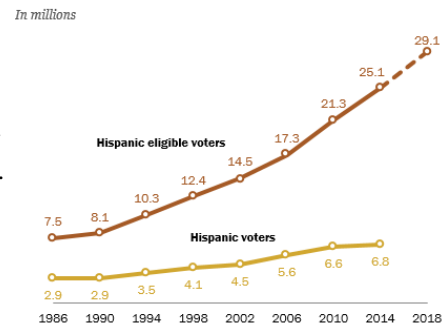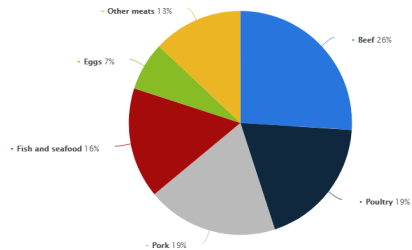

ChartQA-Human

Prompt: You will be provided with a chart image and a question associated with it in the input. Create a Python script that, upon execution, generates an answer to the input question. The script should directly incorporate all necessary data, avoiding any supplemental comments or superfluous variables. Ensure that the data is structured within the script to facilitate the calculation. Exclude any extraneous text or explanation after the python script. Output should only contain the python code. Input: What's the median value of red graph?

The number of Hispanic eligible voters has grown, while midterm election voting has remained flat

Gold-Label: 14.5
Gemini-pro-vision Response: 4.5
GPT-4V Response: 4.3

Table 16: An example from the ChartQA-human test set in the zero-shot PAL setup where both the Gemini-pro-vision and GPT-4V models responded incorrectly, marked in Red text.

ChartQA-Augmented

Prompt: You will be provided with a chart image and a question associated with it in the input. Create a Python script that, upon execution, generates an answer to the input question. The script should directly incorporate all necessary data, avoiding any supplemental comments or superfluous variables. Ensure that the data is structured within the script to facilitate the calculation. Exclude any extraneous text or explanation after the python script. Output should only contain the python code. Input: What was the share of fish and seafood expenditure of U.S. households in 2018?

Gold-Label: 16

Gemini-pro-vision Response: The share of fish and seafood expenditure of U.S. households in 2018 was 16%
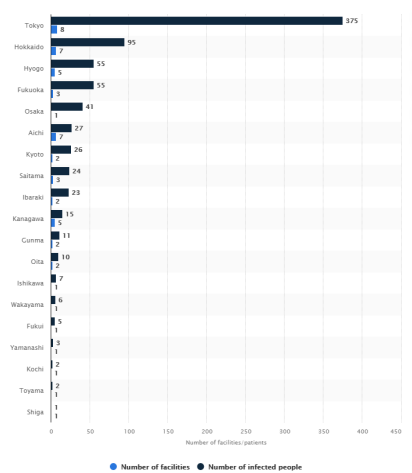
GPT-4V Response: 16

Table 17: An example from the ChartQA-Augmented test set in the zero-shot PAL setup where both the Gemini-pro-vision and GPT-4V models responded correctly, marked in Green text.

ChartQA-Augmented

Prompt: You will be provided with a chart image and a question associated with it in the input. Create a Python script that, upon execution, generates an answer to the input question. The script should directly incorporate all necessary data, avoiding any supplemental comments or superfluous variables. Ensure that the data is structured within the script to facilitate the calculation. Exclude any extraneous text or explanation after the python script. Output should only contain the python code. Input: How many cases of COVID-19 were confirmed across eight hospitals in Tokyo?

Gold-Label: 375

Gemini-pro-vision Response: The total number of cases across the eight hospitals in Tokyo is 666.

GPT-4V Response: 3000

Table 18: An example from the ChartQA-augmented test set in the zero-shot PAL setup where both the Gemini-pro-vision and GPT-4V models responded incorrectly, marked in Red text.

Chart-to-Text Pew

Prompt: The attached chart shows Following Election News Very Closely Democrat. Summarize the chart in a single paragraph focusing on trends and important data points. While summarizing, focus on the axis and color-related information in the chart.
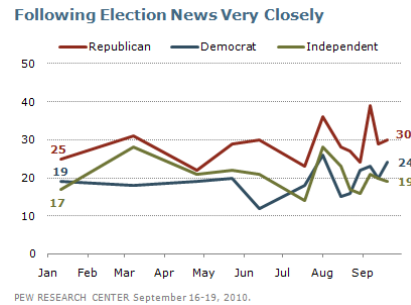
Gold-Summary: While election news dominated media coverage last week, the public continued to show relatively modest interest. According to a separate analysis by the Pew Research Center's Project for Excellence in Journalism, midterm news made up 30% of the newshole. Just more than one-in-ten (11%) say this was the news they followed most closely. Nearly a quarter (23%) say they followed election news very closely. Three-in-ten Republicans say they followed election news very closely last week, not much different from the 24% of Democrats that say this. Two-in-ten independents (19%) say they followed election news very closely.

Gemini-pro-vision Response: The chart shows the percentage of Republicans, Democrats, and Independents who are following election news very closely. The x-axis shows the time, and the y-axis shows the percentage. The data is from January to September 2010. The highest percentage of Republicans who were following election news very closely was in September (44%). The lowest percentage of Republicans who were following election news very closely was in February (25%). The highest percentage of Democrats who were following election news very closely was in September (30%). The lowest percentage of Democrats who were following election news very closely was in May (17%). The highest percentage of Independents who were following election news very closely was in September (24%). The lowest percentage of Independents who were following election news very closely was in February (14%).

GPT-4V Response: The chart visualizes the percentage of Republicans (in red), Democrats (in blue), and Independents (in green) who followed election news very closely over a period from January to September 2011. Starting in January, Democrats showed a lower percentage (17%) compared to Republicans (25%) and Independents (19%). Through the months, all three groups experienced fluctuations in the level of attention paid to election news. Notable peaks can be observed for Republicans around May and September (both at 30%) and for Democrats in May (around 24%) and July, with a subsequent drop and a rise back to around 24% in September. Independent interest peaked in August, close to 30%. Overall, a trend can be seen with Republicans consistently showing the highest percentage of individuals following the election news very closely, with Democrats and Independents generally trailing, albeit with significant increases toward the latter part of the surveyed period.

Claude-3-Haiku: The chart shows the percentage of people from different political affiliations who reported following election news very closely over time. The Democratic line starts high in January at around The Ind 25% and then fluctuates between 15-30% throughout the year. The Republican line starts lower in January around 19% but then rises to a peak of around 40% in September.ependent line remains relatively flat, staying between 15-25% for most of the year. The chart highlights the increased attention to election news among Republicans towards the end of the year compared to Democrats and Independents.

Table 19: An example from the Chart-to-Text Pew test set. Factually incorrect summaries, marked in Red text.

Chart-to-Text Statista

Prompt: The attached chart shows United Kingdom: Population growth from 2009 to 2019 (compared to the previous year). Summarize the chart in a single paragraph focusing on trends and important data points. While summarizing, focus on the axis and color-related information in the chart.
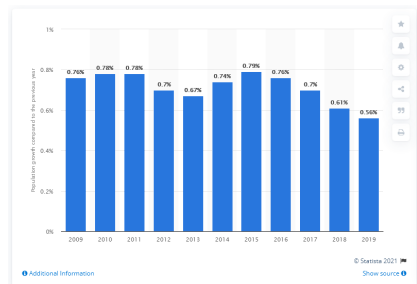
Gold-Summary: This statistic shows the population growth in the United Kingdom from 2009 to 2019. In 2019, the UK population grew by about 0.56 percent compared to the previous year.

Gemini-pro-vision Response: The line chart shows the United Kingdom's population growth from 2009 to 2019. The population growth rate was highest in 2011 at 0.78%, followed by 2015 and 2016 at 0.76%. The lowest population growth rate was in 2019 at 0.56%. The population growth rate decreased from 2011 to 2013, increased from 2013 to 2016, and then decreased again from 2016 to 2019.

GPT-4V Response: The chart depicts the population growth rate of the United Kingdom from 2009 to 2019 as compared to the previous year. The vertical axis indicates the growth rate percentage, while the horizontal axis lists the years. The bars are colored in shades of blue, with a darker blue representing each year's growth rate. From 2009, there's a slight fluctuation in the growth rate, with it peaking in 2014 at 0.79%, indicating that year experienced the highest growth relative to the prior year within this timeframe. Overall, the rate starts at 0.76% in 2009, fluctuates slightly but stays generally consistent through 2014, and then exhibits a declining trend from 2015 onwards, reaching a low of 0.56% in 2019. The chart shows a gradual reduction in population growth rates in the latter part of the decade, implying a slowdown in the rate at which the population of the UK is increasing.

Claude-3-Haiku: The chart shows the population growth rate in the United Kingdom from 2009 to 2019, compared to the previous year. The y-axis represents the population growth rate, expressed as a percentage. The data points reveal a general downward trend in the population growth rate over the 10-year period, declining from 0.76% in 2009 to 0.56% in 2019. The highest growth rate was observed in 2015 at 0.79%, while the lowest was in 2013 at 0.67%. The chart provides a clear visual representation of the changes in the UK's population growth over the last decade.

Table 20: An example from the Chart-to-Text Statista test set. Factually incorrect summaries, marked in Red text.

Chart-to-Text

Prompt: The attached chart shows global market share of smartphones with a triple camera setup in 2019, by brand. Summarize the chart in a single paragraph focusing on trends and important data points. While summarizing, focus on the axis and color-related information in the chart.
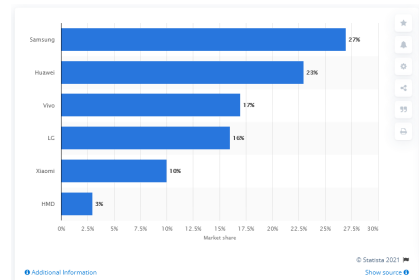
Gold-Summary: The ranking illustrates the global market share of smartphones with a triple camera setup in 2019, by brand. In 2019, Samsung accounted for 27 percent of the triple camera smartphone market.

Gemini-pro-vision Response: The chart shows the global market share of smartphones with a triple camera setup in 2019, by brand. Samsung is the market leader with a 27% share. Huawei is second with a 23% share. Vivo is third with a 17% share. LG is fourth with a 16% share. Xiaomi is fifth with a 10% share. HMD is sixth with a 3% share. The chart shows that Samsung and Huawei are the clear leaders in the market for smartphones with a triple camera setup. Vivo, LG, and Xiaomi are also significant players in the market. HMD is a distant sixth.

GPT-4V Response: The chart illustrates the global market share of smartphones featuring a triple camera setup in 2019, segmented by brand. The horizontal bars, shaded in varying shades of blue, represent each brand's market share, displayed on the x-axis which ranges from 0% to 30%. Samsung leads the market with the highest share, represented by the darkest shade of blue, accounting for 27% of the market. Huawei follows closely, with a 23% share in a slightly lighter blue. Vivo and LG come next with 17% and 16%, respectively; their corresponding bars are in progressively lighter shades of blue. Xiaomi holds a 10% share, while HMD has the smallest segment visualized, at 3%, depicted with the lightest blue. The chart is a visual representation of the significant hold Samsung and Huawei have in the triple camera smartphone segment, while also showing a quite notable presence of other brands such as Vivo, LG, and Xiaomi in this market.

Claude-3-Haiku: The chart shows the global market share of smartphones with a triple camera setup in 2019, broken down by brand. Samsung leads the market with a 27% share, followed by Huawei at 23%. Vivo and LG have lower market shares of 17% and 16%, respectively. Xiaomi holds a 10% market share, while HMD (the company behind Nokia phones) has the smallest share at 3%. The chart presents the data in a clear and visually appealing manner, with the market share percentages displayed directly on the bars for each brand.

Table 21: An example from the Chart-to-Text Statista test set, where summary generated by GPT-4V contains visually incorrect information (mentioning that there are multiple shades of blue color, whereas there are no shades present in the chart), marked in Blue text, while Gemini-pro-vision & Claude-3-Haiku didn't include in visual information.

Chart-to-Text

Prompt: The attached chart shows HeidelbergCement's global revenue from FY 2009 to FY 2018 (in billion euros). Summarize the chart in a single paragraph focusing on trends and important data points. While summarizing, focus on the axis and color-related information in the chart.
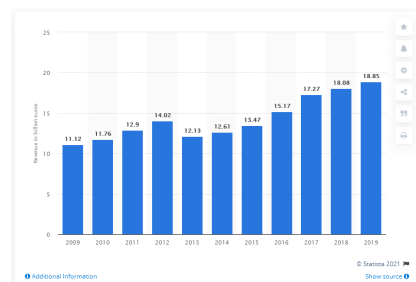
Gold-Summary: This statistic represents HeidelbergCement's global revenue between the fiscal year of 2009 and the fiscal year of 2019. In the fiscal year of 2019, the German multinational building materials company generated around 18.85 billion euros (or about 21.1 billion U.S. dollars) in revenue.

Gemini-pro-vision Response: HeidelbergCement's global revenue experienced fluctuations but showed an overall increase from FY 2009 to FY 2018. The revenue started at 11.12 billion euros in FY 2009 and reached 18.85 billion euros in FY 2018, representing a growth of 70%. The highest increase occurred from FY 2016 to FY 2017, when the revenue jumped from 15.17 billion euros to 17.27 billion euros. The color blue is used to represent the data in the chart.

GPT-4V Response: The chart shows HeidelbergCement's global revenue from the fiscal year 2009 to 2018, with revenue figures represented in billions of euros on the vertical axis and the fiscal years on the horizontal axis. The chart displays a positive trend in revenue over the ten-year period, with some fluctuations. Starting at €11.12 billion in 2009, there was a slight increase until 2012 when revenue peaked at €14.02 billion. After a decrease to €13.47 billion in 2015, the revenue then showed a significant upward trend, reaching the highest point of €17.27 billion in 2017 before a slight decrease in 2018 to €18.08 billion. The bars in the chart are uniformly colored in blue, and the year-over-year changes are clearly indicated by the varying heights of these bars. The upward trend signifies a general growth in HeidelbergCement's global revenue over the decade.

Claude-3-Haiku: The chart shows the global revenue of HeidelbergCement from 2009 to 2019, measured in billions of euros. The revenue has steadily increased over the years, starting at 11.12 billion euros in 2009 and reaching a high of 18.85 billion euros in 2019. The most significant increase occurred between 2011 and 2012, where the revenue jumped from 12.9 billion euros to 14.02 billion euros. Since then, the revenue has continued to grow, with some fluctuations, but overall maintaining an upward trend.

Table 22: An example from the Chart-to-Text Statista test set, where summaries generated by GPT-4V and Gemini-pro-vision captured trend correctly, marked in Green text.

Chart-to-Text

Prompt: The attached chart shows frequency of car models stolen to order in Great Britain (UK) in 2011 and 2012 (in offences per 10,000 registered vehicles). While summarizing, focus on the axis and color-related information in the chart.
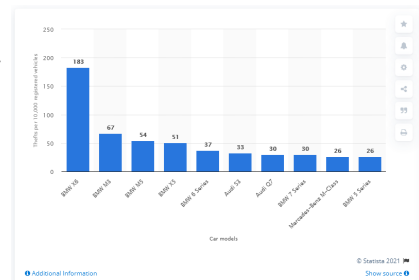
Gold-Summary: This statistic shows the frequency with which luxury or prestige cars are stolen in Great Britain in 2011 and 2012. These types of luxury cars are mostly stolen to order, meaning that they were sought out specifically on request rather than stolen by an opportunist. BMW, especially the X6 model, was by far the luxury brand which was stolen the most over this period.
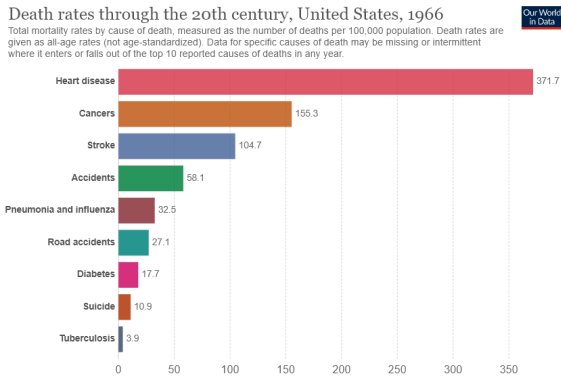
Gemini-pro-vision Response: The chart shows the frequency of car models stolen to order in Great Britain (UK) in 2011 and 2012, with the BMW X6, BMW M3, and BMW M5 being the most stolen car models. The BMW X6 had the highest frequency, with 183 thefts per 10,000 registered vehicles. The BMW M3 and BMW M5 had 67 and 54 thefts per 10,000 registered vehicles, respectively. Other car models with significant thefts include the BMW X5, BMW 6 Series, Audi S3, Audi Q7, Mercedes-Benz CLS-Class, and BMW 5 Series, with 51, 37, 33, 30, 30, and 26 thefts per 10,000 registered vehicles, respectively.

GPT-4V Response: The attached chart displays the frequency of certain car models being stolen to order in Great Britain for the years 2011 and 2012, with the vertical axis representing the number of thefts per 10,000 registered vehicles and the horizontal axis listing the car models. The BMW X6 stands out as the most targeted model, with a theft rate significantly higher at 183 thefts per 10,000 vehicles. The next most stolen models are the BMW M3 and BMW M5, with lower but still considerable theft rates of 67 and 54 respectively. The frequency then gradually decreases with other BMW models and a few Audi and Mercedes-Benz models, such as the Audi S3 and Mercedes-Benz M-Class showing the lowest frequency of the listed models with 26 thefts each. Overall, the chart indicates a clear preference amongst car thieves for BMW models, specifically the high-performance 'M' series and the X6 SUV.

Claude-3-Haiku: The chart depicts the frequency of car models stolen to order in Great Britain, with the car models displayed on the horizontal axis and the frequency of offenses per 10,000 registered vehicles shown on the vertical axis. The data indicates that the BMW X6 model had the highest frequency of thefts, with 183 offenses per 10,000 registered vehicles. Other models with relatively high theft rates include the BMW 1 Series (67 offenses), BMW 4 Series (54 offenses), and BMW 5 Series (51 offenses). The chart also shows that smaller and more affordable models, such as the Mercedes-Benz A-Class and BMW 3 Series, had relatively low theft rates, with 26 offenses per 10,000 registered vehicles each.

Table 23: An example from the Chart-to-Text Statista test set, where summaries generated by GPT-4V captured trend correctly, marked in Green text.

Death rates through the 20th century, United States, 1966
Total mortality rates by cause of death, measured as the number of deaths per 100,000 population. Death rates are given as all-age rates (not age-standardized). Data for specific causes of death may be missing or intermittent where it enters or falls out of the top 10 reported causes of deaths in any year.

Our World in Data

Cause	Value
Heart disease	371.7
Cancers	155.3
Stroke	104.7
Accidents	58.1
Pneumonia and influenza	32.5
Road accidents	27.1
Diabetes	17.7
Suicide	10.9
Tuberculosis	3.9

Prompt: What do each of the colors represent in the chart?

Gemini-pro-vision Response: **The colors represent the different causes of death.**
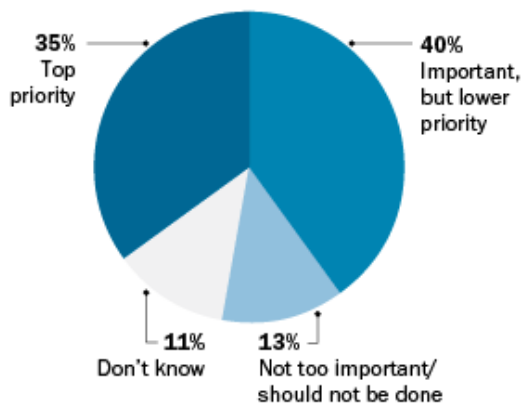
GPT-4V Response: The chart is a horizontal bar graph showing the death rates per 100,000 population from various causes in the United States for the year 1966. Each color on the chart represents a different cause of death, as indicated by the labels on the left side of the graph. Here are the colors and corresponding causes of death they represent:
- Reddish-pink: Heart disease
- Orange: Cancers
- Dark teal blue: Stroke
- Green: Accidents
- **Light blue: Pneumonia and influenza**
- **Dark blue: Road accidents**
- Purple: Diabetes
- **Yellow: Suicide**
- **Light green: Tuberculosis**

Adjacent to each color bar is the specific death rate for that cause of death per 100,000 population, with heart disease being the leading cause at 371.7 deaths per 100,000 and tuberculosis being the least within the top causes of death listed at 3.9 deaths per 100,000.

U.S. Catholics on Vatican Reform

Most U.S. Catholics say reforming the Vatican is an important priority for Pope Francis.

- 35% Top priority
- 40% Important, but lower priority
- 13% Not too important/ should not be done
- 11% Don't know

**Prompt:** What do each of the colors represent in the chart?

Gemini-pro-vision Response: The blue color represents the percentage of respondents who think that reforming the Vatican is an important priority for Pope Francis. **The white color represents the percentage of respondents who think that it is not too important or should not be done.** The gray color represents the percentage of respondents who don't know.

GPT-4V Response: The chart is a pie chart showing the opinions of U.S. Catholics on the importance of reforming the Vatican as a priority for Pope Francis, based on the information provided by the Pew Research Center. Here's what each color represents:
- **Light Blue: 35% view reforming the Vatican as a top priority.**
- **Dark Blue: 40% believe it is important, but a lower priority.**
- **Grey: 13% think it is not too important or should not be done.**
- **Black: 11% don't know or have no opinion on the matter.**

Table 24: Example errors encountered by the Gemini-vision-pro and the GPT-4V when asked about color encoding information about two different charts. In both cases, the Gemini-pro-vision model did not identify different color encodings at all. In the first figure, the GPT-4V model failed to identify different colors correctly, and in the second figure, the model failed to identify different shades of the same color (in this case *'blue'*) correctly. Erroneous text is marked in **Red**.

Semantic Evaluation: Level - 1
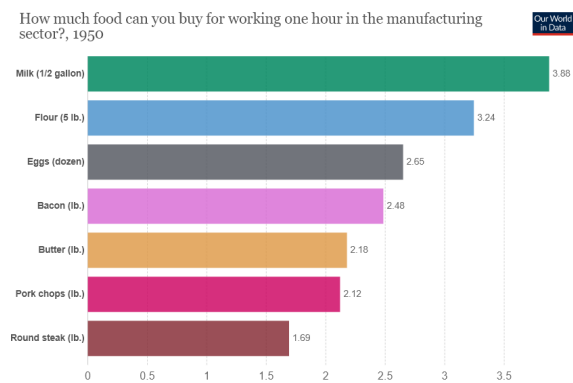
Prompt: What is the chart type in the input image?
Gemini-pro-vision Response: The chart type in the input image is a *bar chart*.
GPT-4V Response: The chart in the image is a *horizontal bar chart*. This type of chart is used to represent data with rectangular bars where the length of the bar is proportional to the value of the variable. In this bar chart, the bars represent the amount of various food items one could buy after working one hour in the manufacturing sector in the year 1950.
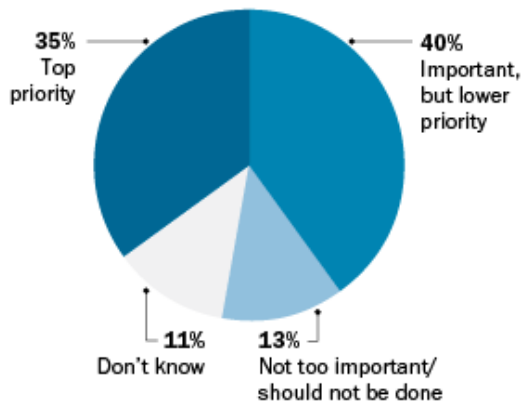

Prompt: What is the range of x-axis?
Gemini-pro-vision Response: 0 to 40.
GPT-4V Response: The image you provided is a pie chart, and pie charts do not have an x-axis or y-axis as you would find in line graphs or bar charts. Instead, they display data in a circular format, with each slice of the pie representing a proportion of the whole. This particular pie chart is about the opinions of U.S. Catholics on Vatican Reform and does not include the concept of a range for an axis.

Table 25: The first figure illustrates an example of a more detailed response (i.e., including the exact chart type) of the GPT-4V model compared to the compact response of the Gemini-pro-vision model. The second figure demonstrates an example of a scenario where the GPT-4V model correctly identified that the chart in question is a '*Pie chart*' and it does not have any x-axis or y-axis range, whereas the Gemini-pro-vision model fails to answer the question correctly. The correct text is marked in **Green** and erroneous text is marked in **Red**.

Semantic Evaluation: Level - 2
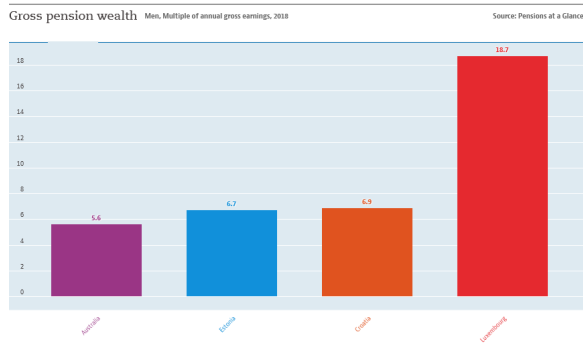
Prompt: Identify the axis that contains a numerical range. What is the maximum value in that axis?
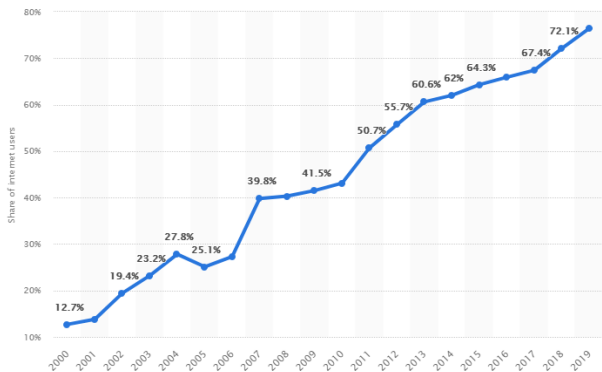
Gemini-pro-vision Response: The y-axis contains a numerical range. The maximum value on the y-axis is **18.7**.

GPT-4V Response: The axis that contains a numerical range in the image is the vertical (y-axis). It represents the "Multiple of annual gross earnings" for "Gross pension wealth." The maximum value on this axis is 18, which is the highest labeled horizontal line, but it does not match the value of the tallest bar. The tallest bar represents Luxembourg, with a value of 18.7, as indicated by the number above the bar. Hence, the actual maximum value depicted on the graph is **18.7**, even though the axis itself is only labeled up to 18.

Table 26: The figure illustrates an example where both the GPT-4V and the Gemini-pro-vision model answer a question about *Level - 2* semantic content (i.e., maximum value) correctly. The correct text is marked in **Green**.

Table 27: The first figure illustrates an example of the models' accurate response to a question about *Level - 3* semantic content presented in a chart (i.e., trend/pattern), and both the models could answer correctly as the trend was easy to follow and linearly increasing. However, when presented with a chart with multiple lines that represent a complex trend (i.e., a chart with many fluctuating lines), both the models fail to answer correctly, as depicted in the second and third figures. The correct text is marked in **Green** and erroneous text is marked in **Red**.

Semantic Evaluation: Level - 3

Prompt: Describe the trend that is visible in the chart and provide evidence for your conclusion.

Gemini-pro-vision Response: The chart displays a **clear upward trend** over the period from 2000 to 2019. Starting at 12.7% in 2000, the data point rises consistently throughout the years, reaching 72.1% by 2019. This increase is steady with no evident declines, indicating a notable and persistent growth in the metric being measured, which is the "Share of Internet users" as seen on the y-axis. Every consecutive year shows an increase from the previous year. The x-axis, which represents time, clearly shows this is over a 20-year timespan. The evidence for the conclusion is in the plotted data points themselves, each one being higher than the previous, without any dips or downward movements. The consistent increase in the percentage points reflects a robust and sustained growth in internet usage among the population being analyzed.
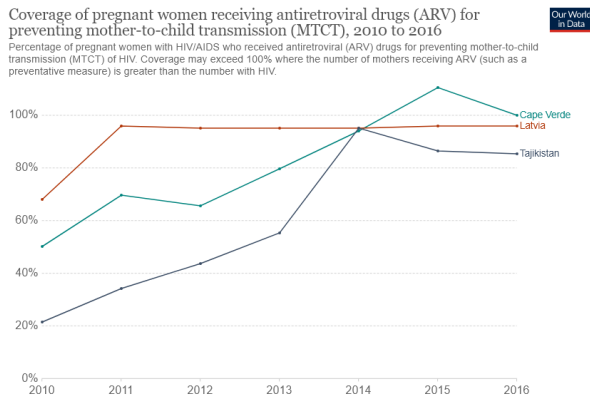
GPT-4V Response: The chart shows the share of internet users from 2000 to 2019. **The trend is upward**, with the share of internet users increasing from 12.7% in 2000 to 76.1% in 2019. The evidence for this conclusion is that the line on the chart is going up.

– Continued on next page –

Table 27 – continued from previous page

Semantic Evaluation: Level - 3

Coverage of pregnant women receiving antiretroviral drugs (ARV) for
preventing mother-to-child transmission (MTCT), 2010 to 2016

Our World
in Data

Percentage of pregnant women with HIV/AIDS who received antiretroviral (ARV) drugs for preventing mother-to-child
transmission (MTCT) of HIV. Coverage may exceed 100% where the number of mothers receiving ARV (such as a
preventative measure) is greater than the number with HIV.

Prompt: Describe the trend that is visible in the chart and provide evidence for your conclusion.

Gemini-pro-vision Response: The chart shows the coverage of pregnant women receiving antiretroviral drugs (ARV) for preventing mother-to-child transmission (MTCT) of HIV. **The coverage in Cape Verde increased from 50% in 2010 to 98% in 2016. The coverage in Latvia increased from 22% in 2011 to 88% in 2016. The coverage in Tajikistan increased from 58% in 2013 to 95% in 2016.**

GPT-4V Response: The chart presents data on the coverage of pregnant women receiving antiretroviral drugs (ARV) for preventing mother-to-child transmission (MTCT) of HIV in three different countries—Cape Verde, Latvia, and Tajikistan—from 2010 to 2016. A clear trend from this chart is that coverage increased in all three countries over the given time period.

1. **Cape Verde shows a consistently high coverage close to or at 100% throughout the period, with slight fluctuations but generally maintaining a plateau.**

2. **Latvia presents a dramatic increase starting from just above 20% in 2010 to reaching approximately 80% by 2016.**

3. **Tajikistan also shows considerable improvement, with the coverage starting at around 40% in 2010 and ending up just below 80% in 2016.** The evidence for these trends is in the trajectory of the lines for each country, which show upward slopes over the years, reflecting an increasing percentage of pregnant women receiving ARV drugs to prevent MTCT of HIV. Each point on the lines represents the data recorded for the corresponding year, clearly illustrating the progress each country has made throughout these years.

Table 27 – continued from previous page

Semantic Evaluation: Level - 3

Saving rate Total, % of GDP, 1984 – 2006   Source: National Accounts at a Glance
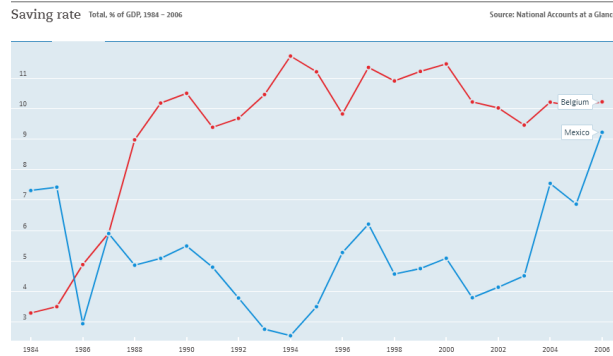
Prompt: Describe the trend that is visible in the chart and provide evidence for your conclusion.
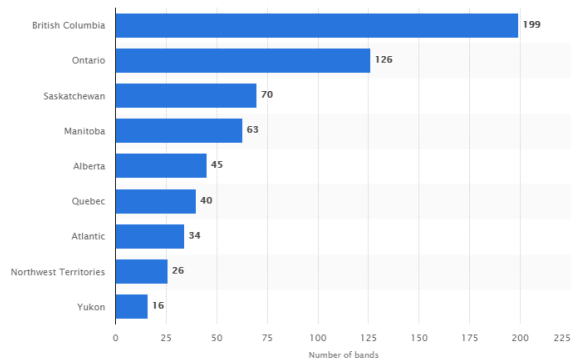
Gemini-pro-vision Response: The chart shows that the saving rate in Belgium and Mexico has been on **a downward trend since the early 1990s. While Belgium's saving rate has remained relatively stable since the mid-2000s,** Mexico's saving rate has seen a slight recovery since the early 2000s.

GPT-4V Response: The chart you provided shows the saving rates as a percentage of GDP for Belgium and Mexico from 1984 to 2006. Here's the trend based on the information presented: For Belgium:

- The saving rate starts off at around 10% in 1984.

- There's a slight decrease until a sharp drop around 1986, where it reaches just above 3%, the lowest point on the chart for Belgium.

- A steady increase follows until around 1989, where the saving rate exceeds 10% again.

- From 1990 to the end of the chart, the saving rate fluctuates but generally maintains within the range of roughly 9% to slightly above 11%, without any drastic changes. For Mexico:

- The saving rate is shown starting from 1990 on this chart (indicating no data provided or not tracked for the prior years).

- It starts out at around 6% in 1990.

- There's a rather volatile pattern with significant ups and downs until 1995, where it dips to the lowest point, close to 4%.

- From this point on, there is a relatively steady upward trend, with some fluctuations but a clear overall increase in the saving rate.

- This ends with a sharp increase between 2004 and 2006, where the saving rate jumps to around 6% to just under 11%. Evidence for these trends comes from observing the lines on the chart and noting the changes in direction
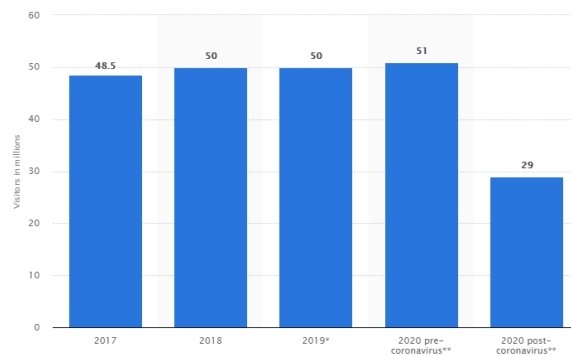
Semantic Evaluation: Level - 4

Prompt: Analyze the chart given in the input in one paragraph.

Gemini-pro-vision Response: **The chart shows the number of bands in each province and territory of Canada.** As of 2021, British Columbia has the most bands, with 199, followed by Ontario with 126, and Saskatchewan with 70. The Northwest Territories has the least bands, with 26.

GPT-4V Response: **The chart presented is a horizontal bar chart showing the number of bands in various regions or provinces of a country, which appears to be Canada given the regions listed.** British Columbia leads with the highest number at 199 bands, followed by Ontario with 126 bands. These two regions have significantly more bands compared to the others. Saskatchewan comes in third with 70 bands, while Manitoba has a close 63. Alberta and Quebec show a mid-range number of bands, with 45 and 40 respectively. The Atlantic region is represented with 34 bands, and the Northwest Territories have 26. Yukon has the fewest bands, represented by a count of 16. **Overall, the chart provides a clear visualization of a disparity in the number of bands among the regions, hinting at a possible correlation between population, cultural factors, or policies that influence the number of music bands in these areas.**

Prompt: Analyze the chart given in the input in one paragraph.

Gemini-pro-vision Response: The chart shows the number of visitors in millions from 2017 to 2020. The number of visitors increased from 48.5 million in 2017 to 50 million in 2018 and remained the same in 2019. **However, the number of visitors decreased to 51 million in 2020 due to the COVID-19 pandemic.** The number of visitors decreased even further to 29 million in 2020 post-coronavirus.

GPT-4V Response: The chart is a bar graph representing the number of visitors in millions over a period from 2017 through 2020, where 2020 is split into "pre-coronavirus" and "post-coronavirus" periods, indicating the impact of the COVID-19 pandemic on visitor numbers. From 2017 to 2020 pre-coronavirus, there is an upward trend in visitor numbers, starting at approximately 48.5 million in 2017 and peaking at 51 million just before the coronavirus impact. **Notably, in the post-coronavirus period of 2020, there is a significant drop to 29 million visitors, reflecting the drastic reduction likely due to travel restrictions, lockdowns, and other measures put in place to combat the spread of COVID-19.**

Table 28: The figure illustrates examples where the Gemini-pro-vision and the GPT-4V models could produce domain-specific information pertaining to semantic content *Level - 4*. The text containing domain-specific information is marked in **Blue**.