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

EMNLP 2024

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

* Equal contribution

Motivation

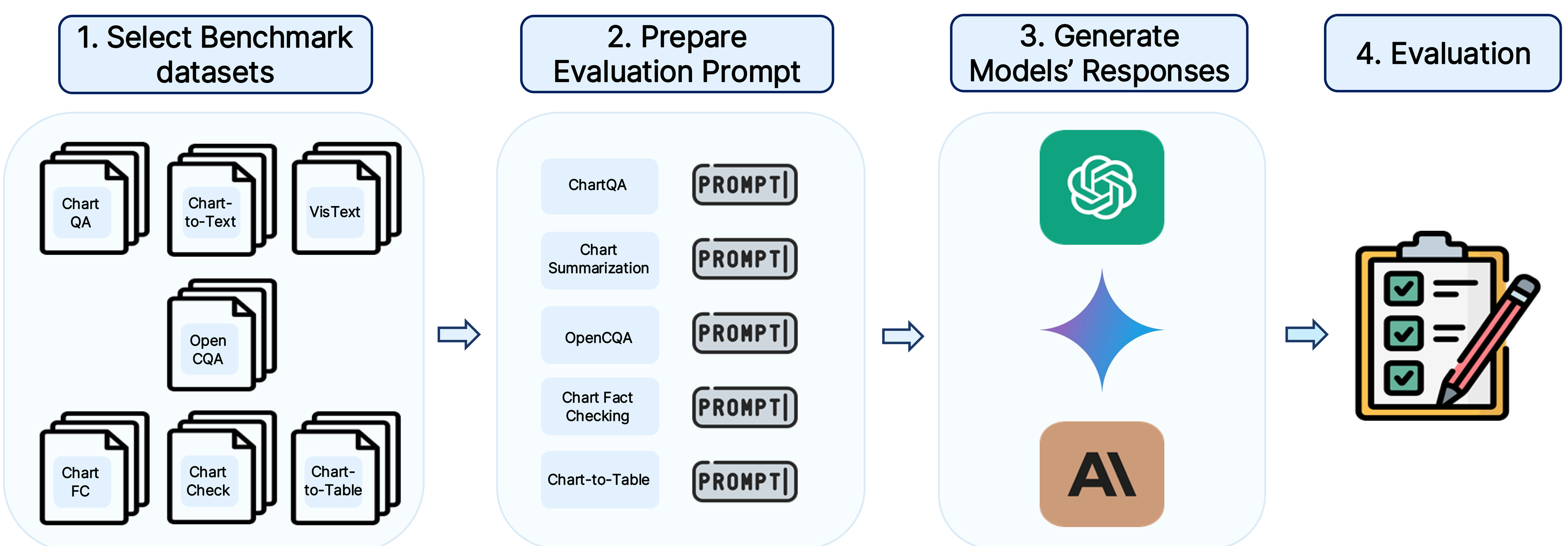
- Recent advances in LVLMs,
  - show **promise** in multimodal tasks,
  - but their **abilities** in chart comprehension remain under-explored
- Existing SoTA models typically,
  - report **quantitative** performance on **ChartQA**
  - present no detailed analysis of the **capabilities** and **limitations**
- So we pose the following **research question**:

Are LVLMs up to the challenge of chart comprehension and reasoning?

Contributions

1. Examine LVLMs performance using **zero-shot CoT** and **PAL**
2. Evaluate LVLMs in generating **open-ended responses**
3. Investigate **hallucinations**, **factual errors**, and **biases**
4. Examine LVLMs **capabilities** in **chart data extraction**
5. Analyze LVLMs in **generating low-level and high level** semantic content

Methodology



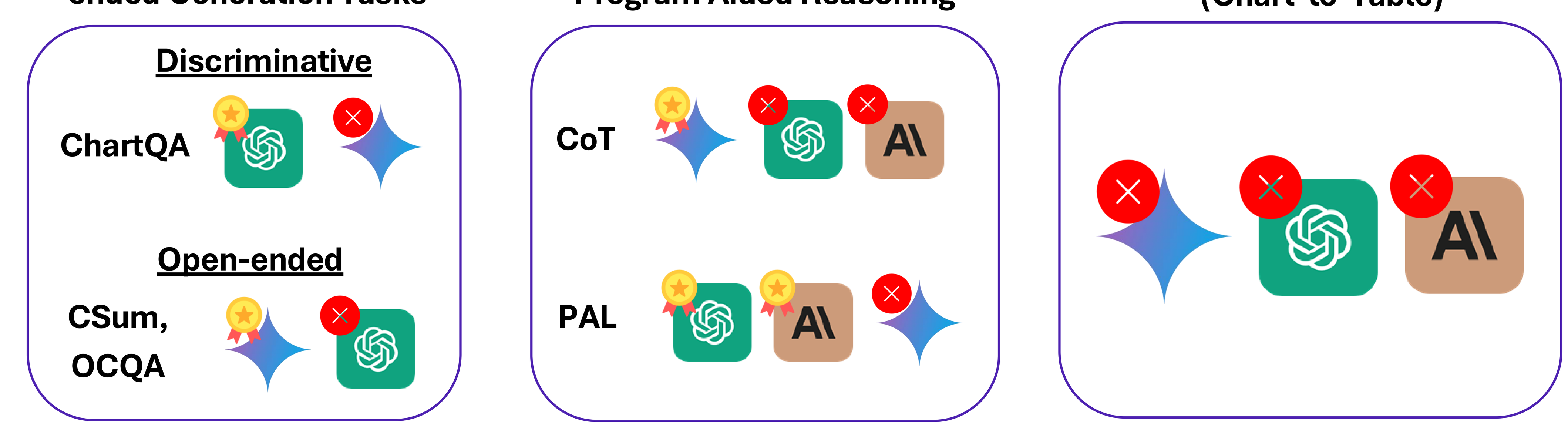
Task and Datasets

- We evaluate the performance on 5 benchmark Chart Reasoning and Comprehension tasks:

1. Chart Question Answering: 1 dataset (ChartQA)
2. Chart Summarization: 2 datasets (Chart-to-Text, VisText)
3. Open-ended Chart QA: 1 dataset (OpenCQA)
4. Fact Checking with Charts: 2 datasets (ChartFC, ChartCheck)
5. Chart-to-Table: 1 dataset (ChartQA)

Results (*Task-specific General Evaluation*)

1. Discriminative vs. Open-ended Generation Tasks
2. Chain-of-Thought vs. Program Aided Reasoning
3. No Data value annotation (Chart-to-Table)



Models	ChartQA (zero-shot CoT) (Accuracy) aug.	human	avg.	ChartQA (zero-shot PAL) (Accuracy) aug.	human	avg.	OpenCQA (BLEU)	Chart Summarization (BLEU) Pew	Statista	Vistext(L1)	Vistext(L2/L3)	Chart-Fact-checking (F1 – score) ChartFC	ChartC(T1)	ChartC(T2)	Chart-to-Table (RNSS) ChartQA	(RMS) 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.

Results (*Criteria-based Focused Evaluation*)

Hallucination Analysis (FAVA method – 6 hallucination types)

Error Distribution

The *"entity"* category showed the most errors, followed by *"relation"* and *"contradictory"* categories, aligning with findings from other NLP research

Model Comparison

Claude-3 had the highest error count, while Gemini and GPT-4V showed better performance

Actionable Insight

Frequent hallucinations in *entities* and *relations* are often fixable with minor edits, underscoring the need for improved detection methods.

Error Type	Example	Average Error Count (Per Summary) Pew: Gemini	Pew: GPT-4V	Pew: Claude 3 Haiku	Statista: Gemini	Statista: GPT-4V	Statista: 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.

Analysis of Semantic Levels (Four-level semantic framework)

Model Performance in Text generation

GPT-4V produces **longer summaries** with detailed visual information (Level 1 & 3), while **Gemini** generates **concise summaries** with statistical and domain-specific information (Level 2 & 4). However, all models lack sufficient contextual insights (Level 4).

Semantic Understanding in Question-Answering

GPT-4V generally **outperforms Gemini** across different **semantic levels**, though both struggle with **complex line charts**, and **Gemini excels** in providing contextual information beyond the chart data.

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

Figure: 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.

Conclusion

- This is the **first comprehensive analysis** of LVLMs such as GPT-4V, Gemini, Claude, and Phi-3 in real-world chart interpretation
- Key insights highlight both strengths and limitations of LVLMs, in **generalizability and reasoning**, **Semantically rich text generation**, **Hallucinations**, **factual errors**, and **bias**
- We hope that the insights gained from this study will catalyze further research and advancements in the emerging area of chart reasoning