Generative Intelligence in Data Visualization: A Technical Assessment of LLM Capabilities

Thesis submitted by

Shubham
2024JTM2078

Under the guidance

Prof. Sougata Mukherjea, Indian Institute of Technology Delhi

Prof. Brajesh Lall, Indian Institute of Technology Delhi

In partial fulfillment of the requirements for
the award of the degree of

Master of Technology

Bharti School of Telecommunication Technology and Management

INDIAN INSTITUTE OF TECHNOLOGY DELHI

January 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Generative Intelligence in Data Visualization: A Technical Assessment of LLM Capabilities**, submitted by **Shubham (2024JTM2078)**, to the Indian Institute of Technology, Delhi, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other institution or University for the award of any degree or diploma.

Prof. Sougata Mukherjea
Dept. of Electrical Engineering
IIT-Delhi, 600036

Date: **9th May 2025**

Place: New Delhi

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have supported and guided me throughout this research project. First and foremost, I am deeply thankful to my supervisors, Prof. Sougata Mukherjea and Prof. Brejesh Lall, Department of Electrical Engineering, Indian Institute of Technology Delhi, for their invaluable guidance, expert insights, and continuous encouragement. Their mentorship and deep understanding of emerging technologies in artificial intelligence and data systems have been instrumental in shaping the technical direction and depth of this work. I also extend my appreciation to the faculty and staff of the Telecommunication Technology and Management program for fostering a rich academic environment and providing the resources necessary for the successful completion of this research.

ABSTRACT

This study investigates the visualization literacy capabilities of three prominent Large Language Models (LLMs)—Google's Gemini (formerly Bard), OpenAI's GPT-4 (powering ChatGPT), and DeepSeek. While LLMs have shown promise in tasks such as generating captions, descriptions, and design suggestions for visualizations, their ability to critically evaluate visual representations remains underexplored. Recognizing the longstanding bottleneck of human-based evaluation due to time and resource constraints, this work explores whether LLMs can assist as preliminary evaluators. A modified version of the 53-item Visualisation Literacy Assessment Test (VLAT) was employed, carefully adapted to mitigate bias arising from prior training exposures. To avoid model contamination, the modified dataset was not uploaded publicly. Findings reveal that the evaluated LLMs currently fail to achieve the visualization literacy levels established by VLAT benchmarks and the general population. Moreover, the models frequently rely on memorized prior knowledge rather than actively analyzing the presented visual information. These insights highlight both the current limitations and the future potential of leveraging LLMs for scalable visualization evaluation support

Contents

ACKNOWLEDGEMENTS i

ABSTRACT ii

LIST OF TABLES iv

ABBREVIATIONS vi

NOTATION vii

1 INTRODUCTION 9

1.1 Background and Motivation

1.2 Problem Statement

1.3 Research Objectives

2 RELATED WORK 11

2.1 Large Language Model

2.2 Visualization Generation

2.3 Analyzing LLMs for Visualization Generation

2.4 Visualization Understanding

3 ANALYZING LLMs FOR VISUALIZATION GENERATION 12

3.1 Process

3.2 Execution and Output Evaluation-Python

3.3 Chart Generation using vegalite

3.4 Designing and Analyzing in CPE Context

4 ANALYZING LLMs FOR VISUALIZING UNDERSTANDING 24

4.1 DataSet

4.2 Need for Manual Analysis

4.3 Data for Manual Analysis(CPE Inspired PlotQA)

4.4 Manual Analysis Results

4.5 Performance Across Different CPE Visualization Types

5 CONCLUSION 28

6 REFERENCE 30

List of Tables

Table 1.1 A sample table with a table caption placed appropriately.

Table 3.1 Performance comparison of LLMs (GPT-3.5, GPT-4o, Gemini, Claude) in chart generation using Python.

Table 3.2 Performance comparison of LLMs (GPT-3.5, GPT-4o, Gemini, Claude) in chart generation using Vega-Lite.

Table 4.1 Sample visual reasoning questions inspired by VLAT for different chart types (Line Chart, Bar Chart, etc.) in the CPE domain.

Table 4.2 Accuracy of LLMs in interpreting visualizations across multiple chart types and reasoning categorieS.

ABBREVIATIONS

CPE	Customer Premises Equipment
LLM	Large Language Module
DNN	Deep Neural Network
VLAT	Visualization Literacy Assesesment Test
UI	User Interface
QA	Question Answering

Chapter 1

INTRODUCTION

1.1 Background and Motivation

With the growing reliance on data, visualizations like charts and graphs have become essential for clear communication and decision-making. Large Language Models (LLMs) such as GPT 4.0, GPT-3.5, Gemini, and DeepSeek excel in language tasks, but their ability to interpret and generate visualizations remains unexplored. This project evaluates whether LLMs can create charts from textual instructions using tools like Matplotlib and Plotly, interpret visualizations, and understand CPE-related visualization data.. Success in these areas could enable LLMs to assist in building smart dashboards and integrating language with visual reasoning, paving the way for more intuitive human-AI collaboration.

1.2 Problem Statement

While transformers excel at text processing, they face significant challenges with visual inputs. Their attention mechanisms, crucial for language tasks, have not been extensively tested for the spatial reasoning needed in visualization tasks. **Creating accurate charts** from natural language requires translating abstract instructions into structured outputs, while interpreting existing charts demands precise trend identification and reasoning. Although transformers are strong with sequential data, their capabilities in visual decoding and error recovery, particularly in cases of ambiguous prompts or complex layouts—remain unclear. This study evaluates GPT-4o, Gemini, and DeepSeek on two core tasks—visualization generation and interpretation—highlighting their adaptability and limitations in handling visual challenges.

1.3 Research Objectives

Given the rapid integration of Large Language Models (LLMs) into multimodal reasoning pipelines, the primary goal of this study is to systematically assess whether advanced models—OpenAI's GPT-3.5, Google's Gemini, and DeepSeek—can effectively perform both visualization generation and interpretation tasks.

The study explores three key dimensions:

1.3.1 Chart Generation

We evaluated whether LLMs could accurately generate Python (Matplotlib, Plotly) and

Vega-Lite scripts based on natural langua[1]ge prompts across 24 distinct chart types.
These chart types were categorized using a VLAT (Visualization Literacy Assessment Test)-inspired framework, covering a wide range from basic visualizations like bar charts to more complex ones such as radar and violin plots.

1.3.2. Chart Understanding

Beyond generation, we assessed the models' ability to interpret pre-existing charts and respond to targeted questions.
Datasets such as **PlotQA** and a **custom VLAT-inspired** set were used to test their comprehension abilities.
Tasks included identifying trends, counting elements, determining maximum/minimum values, and making relative comparisons—thus moving beyond simple binary (yes/no) responses.

1.3.3 Empirical Results and Observations

Our empirical findings reveal that:

LLMs demonstrate strong capabilities in generating basic visualizations and answering surface-level questions.However, they face significant challenges when tasks require a nuanced understanding, multi-step reasoning, or precise visual grounding.
Errors were common in tasks involving fine distinctions, such as detecting the largest bar or comparing closely valued data points.Evaluation was benchmarked against human literacy standards, with independent human reviewers assessing outputs based on fidelity, correctness, reasoning depth, and overall accuracy, ensuring a reliable real-world performance comparison.

1.3.5 Conclusion and Future Directions

This study highlights both the remarkable progress and the current limitations of LLMs as visualization agents. While LLMs are evolving in their ability to generate and interpret visualizations, achieving human-like proficiency in nuanced visual reasoning remains an open challenge.

1.3.6 Future work should prioritize:

Strengthening multimodal pretraining to better align visual, textual, and logical domains.
Developing step-by-step visual reasoning frameworks to guide LLMs through complex tasks.
Exploring hybrid models that integrate LLMs with specialized vision architectures for more robust visualization understanding.
Overall, while LLMs are emerging as valuable tools for visualization tasks, substantial data

[1]

2. Related Work

2.1 Large Language Model

Large Language Models such as Gemini (Chowdhery et al., 2022) have exhibited remarkable capabilities across a wide range of natural language processing tasks. With the right kind of carefully constructed prompts, a single LLM can adapt to perform diverse tasks effectively. Researchers have explored various prompt engineering strategies to tailor the model's responses to specific tasks (White et al., 2023). Similarly, AlphaCode, a model trained on competitive programming problems, has achieved impressive results in code generation — correctly solving a majority of the problems in benchmark coding contests.

2.2 Visualization Generation

While large language models (LLMs) have shown promise in generating visualizations, there are still areas that require significant improvement. Li et al. (2024) explore the capabilities of GPT-3.5 in generating Vega-Lite visualizations from natural language descriptions using various prompting strategies. Their key findings indicate that GPT-4 significantly outperforms previous state-of-the-art methods on the NL2VIS task, demonstrating high accuracy in generating correct visualizations for simpler and more common chart types. However, the model struggles with more complex visualizations and tasks that require a deeper understanding of data structures.

LLMs have also been integrated into NL2VIS systems such as **Chat2Vis** (Maddigan & Susnjak, 2023) and **LIDA** (Dibia, 2023), which generate Python code to construct data visualizations. Despite these advancements, there remains a need for a systematic evaluation of how effectively these models can generate accurate and contextually appropriate visualizations.

2.3 Visualization Understanding

Recent advancements in multimodal large language models (MMLLMs) have significantly enhanced the field of chart understanding. Notably, models such as ChartLlama and TinyChart have emerged as state-of-the-art solutions. ChartLlama , trained on a diverse set of chart-related tasks, demonstrates superior performance in interpreting complex visual data. Similarly, TinyChart , with its efficient architecture, achieves remarkable results on benchmarks like ChartQA and Chart-to-Text, outperforming larger models like ChartLlama and even GPT-4V in certain tasks.

Complementing these models, benchmark datasets such as CharXiv have been introduced to evaluate the capabilities of MMLLMs in realistic scenarios. CharXiv, comprising over 2,300 diverse charts from scientific papers, provides a rigorous testing ground for assessing the reasoning abilities of these models.

Furthermore, evaluations of general-purpose LLMs have been conducted to assess their proficiency in visualization comprehension. For instance, Bendeck and Stasko (2025) analyzed GPT-4's performance across various visualization literacy tasks, including question answering and identifying misleading visualizations. Their findings indicate that while GPT-4 performs efficiently on standard tasks, it encounters challenges when dealing with more nuanced or complex visualizations.

3. Analyzing LLMs for Visualization Generation

3.1 Process

To assess how effectively large language models (LLMs) can generate information visualizations—especially in use cases relevant to telecom and **CPE environments**—we followed a method inspired by Vazquez (2024). Our goal was to test how well these models can translate structured data, such as network performance logs or device metrics from CPE systems, into insightful visualizations. Python was chosen due to its strong ecosystem of libraries like `matplotlib`, widely used in network analytics. We also evaluated their ability to generate Vega-Lite code, suitable for lightweight, browser-based visualization—ideal for remote CPE diagnostics dashboards.

The workflow involved five main stages:

1. Selecting Visualization Types

We picked 24 different chart types tailored to tabular data analysis, often encountered in CPE monitoring systems. These included common plots like bar and line charts, as well as specialized ones such as violin plots and locator maps—useful for visualizing geographic deployments or signal variations. Visualizations for hierarchical or network topologies were excluded in this scope and foreach type we curated five types of graphs for question answering to understand in more details like Retrieve Value,Make Comparisons,Range,Correlation or Find Extremum.

2. Dataset Preparation

We curated or generated datasets aligned with CPE-related contexts—like bandwidth usage, signal strength, latency trends, and device uptime. These datasets spanned categorical, quantitative, temporal, and geo-based data types, providing a robust foundation for testing the models in real-world telecom scenarios.

3. LLM Selection

To ensure a broad and fair comparison, we analyzed the performance of four top-tier LLMs:

- **OpenAI's GPT-3.5 and GPT-4o**
- **Google's Gemini 1.5 Pro**
- **DeepSeek**

These models represent the current state-of-the-art and were evaluated for their relevance to telecom dashboards and data interpretation at the CPE level.

4. Prompt Design and Testing

We used zero-shot prompting, meaning no examples were shown to the models in advance. Instead, carefully structured prompts were crafted to request specific visualizations. For instance: ***Can you write a Python script that generates a Bubble chart using columns like*** `signal_strength` ***(quantitative),*** `bandwidth` ***(quantitative), and*** `uptime` ***(quantitative) from the CSV file*** `cpe_logs.csv`***?***

5. Testing

We conducted a thorough test to evaluate the performance of LLMs, examining the variety of charts they could generate.

3.1.1 Experimental Procedure

To evaluate the capabilities of LLMs in generating accurate and meaningful visualizations—particularly for use in **Customer Premises Equipment (CPE)** monitoring and analytics—we established a consistent experimental framework. This helped ensure the reliability and repeatability of results across different models and visualization tasks.

The procedure involved the following steps:

1. Fresh Session Initialization

Each experiment was conducted in a new session to avoid influence from prior interactions. Since LLMs retain conversational context, starting fresh ensured that the output wasn't biased by earlier prompts. This was especially important when switching between visualization formats (e.g., Python vs. Vega-Lite), which can otherwise carry over.

2. Standardized Prompting Conditions

All prompts were issued within a single session and on the same day for each model to maintain uniform testing conditions. This reduced any variations that could arise from model updates or drift over time.

3.1.2 Execution and Output Evaluation

The generated Python or Vega-Lite code was executed to produce a visualization. Each visualization was then analyzed for correctness, clarity, and relevance—focusing on how well it could be applied to real-world scenarios such as tracking signal strength, uptime, or bandwidth consumption from CPE devices.

Chart Type	GPT-3.5	GPT-4o	Gemini	DeepSeek
Area Chart	Yes	Yes	Yes	Yes
Bar Chart	Yes	Yes	Yes	Yes
Box Plot	Yes	Yes	Yes	Yes
Bubble Chart	Yes	Yes	Yes	Yes
Bullet Chart	No	Yes	No	No
Choropleth	Yes	Yes	Yes	No
Column Chart	Yes	Yes	Yes	Yes
Donut Chart	Yes	Yes	Yes	Yes

Dot Plot	No	Yes	No	Yes
Graduated Symbol Map	No	Yes	No	No
Grouped Bar Chart	Yes	Yes	Yes	Yes
Grouped Column Chart	Yes	Yes	Yes	Yes
Line Chart Chart	Yes	Yes	Yes	Yes
Locator Map	No	Yes	Yes	Yes
Pictogram Chart	Yes	Yes	No	No
Pie Chart	Yes	Yes	Yes	Yes
Pyramid Chart	Yes	Yes	Yes	No
Radar Chart	Yes	Yes	Yes	No
Range Plot	Yes	Yes	Yes	No
Scatter Plot	Yes	Yes	Yes	Yes
Stacked Bar Chart	Yes	Yes	Yes	Yes
Stacked Column Chart	Yes	Yes	Yes	Yes
Violin Plot	Yes	Yes	Yes	Yes
XY Heatmap	Yes	Yes	Yes	Yes
Total	20(83.33%)	24(100%)	21(83%)	17(58.33%)

Table 1: Performance Comparison of LLMs in Chart Generation using Python

Note: A visualization was considered correctly generated only if the LLM produced accurate and executable code that matched the specifications provided in the prompt. Given that LLMs can sometimes yield inconsistent outputs, we adjusted model parameters to reduce randomness and ensure stable generation. To further enforce consistency, each prompt was run three times, and outputs were accepted only if they remained identical across all trials.

Most inaccuracies stemmed from certain LLMs lacking knowledge of specific or less common visualization types. This limitation became apparent particularly in use cases involving detailed CPE analytics where specialized visualizations can be valuable. For instance, only GPT-4o successfully generated a correct bullet chart—useful in dashboard performance monitoring—whereas GPT-3.5 mistakenly produced a pyramid chart, and outputs from Gemini were structurally incorrect. A comparative visualization of these results is presented in Figure 1.

3.1.3 Chart Generation via Vega-Lite Scripts

To further evaluate the capabilities of modern LLMs in generating declarative visualizations, we tested their ability to produce Vega-Lite scripts—a lightweight grammar often used for interactive data visualization in web applications, including network dashboards for CPE performance monitoring. For this task, we focused on GPT-4o and Gemini, prompting each model to generate Vega-Lite code for all 24 selected chart types.

As reflected in Table 2, Vega-Lite proved to be more challenging than Python-based libraries. Gemini successfully generated only about 40% of the requested charts. Similarly, GPT-4o exhibited a noticeable drop in performance when transitioning from Python to Vega-Lite scripting. This suggests that while LLMs show promise in procedural visualization tasks, they may still lack proficiency in declarative formats critical for dynamic, lightweight UIs such as those found in CPE management systems. For instance, as depicted in Figure 2, neither model was able to generate a valid Violin Plot in Vega-Lite, highlighting the limitations in handling less conventional chart types.

Chart Type	GPT-3.5	GPT-4o	Gemini	DeepSeek
Area Chart	Yes	Yes	Yes	Yes
Bar Chart	Yes	Yes	Yes	Yes
Box Plot	Yes	Yes	Yes	Yes
Bubble Chart	Yes	Yes	No	Yes
Bullet Chart	Yes	Yes	No	No
Choropleth	Yes	Yes	No	No
Column Chart	Yes	Yes	Yes	No
Donut Chart	Yes	Yes	Yes	Yes
Dot Plot	Yes	Yes	No	Yes
Graduated Symbol Map	Yes	Yes	No	No
Grouped Bar Chart	Yes	Yes	No	No
Grouped Column Chart	Yes	Yes	No	No
Line Chart Chart	Yes	Yes	No	No
Locator Map	Yes	Yes	No	No

Pictogram Chart	No	Yes	Yes	No
Pie Chart	Yes	Yes	Yes	Yes
Pyramid Chart	Yes	Yes	No	No
Radar Chart	No	No	No	No
Range Plot	Yes	Yes	Yes	No
Scatter Plot	Yes	Yes	Yes	Yes
Stacked Bar Chart	Yes	Yes	Yes	Yes
Stacked Column Chart	Yes	Yes	Yes	Yes
Violin Plot	Yes	Yes	No	No
XY Heatmap	Yes	Yes	Yes	Yes
Total	21(83.33%)	23(100%)	12(75%)	11(58.33%)

Table 2 : Performance Comparison of LLMs in Chart Generation using VegaLite

3.1.4 Designing and Analyzing Visual Questions for CPE Context

To evaluate the visual reasoning capabilities of large language models in the domain of Customer Premises Equipment (CPE), we designed a set of structured visualization tasks inspired by VLAT (Visualization Literacy Assessment Test). While Vega-Lite was not used directly for chart generation, the tasks are based on simplified or mock visualizations that mimic typical data plots (e.g., line charts, bar charts, pie charts) commonly used in CPE performance dashboards. Each question targets a specific cognitive skill—such as retrieving values, identifying extrema, comparing quantities, or interpreting trends—based on CPE-related data such as device usage, installation cost, market share, and performance metrics. The goal is to assess how effectively different models interpret chart-based information within a CPE-centric application domain.

Visualization	**Task**	**Sample Question**	**VLAT**	**GPT-4o**	**Gemini**	**DeepSeek**	**Random**
. Line Chart	Retrieve Value	What was the number of CPE devices installed in March 2023?	**0.95**	0.58	0.28	0.17	0.25
	Find Extremum	In which month did the CPE usage peak in 2023?	**0.97**	0.06	0.28	0.07	0.25
	Determine Range	What was the range of connected CPE devices throughout 2023?	**0.56**	0.27	0.29	0.20	0.25
	Find Corelation and Trend	Over the second half of 2023, the usage of CPE devices was .	**0.98**	**0.92**	0.44	0.37	0.33
	Make Comparisons	How much did the number of CPE installations increase from Q1 to Q2 of 2023?	**0.77**	0.87	0.34	0.26	0.25

Bar Chart	Retrieve Value	What is the average monthly data usage per CPE device in Germany?	**0.88**	0.40	0.56	0.40	0.25
	Find Extremum	Which region had the highest average CPE data usage in 2023?	**0.98**	0.37	0.10	0.07	0.25
	Determine Range	What is the range of average CPE device lifespans across different countries?	**0.54**	0.29	**0.68**	0.44	0.25
	Make Comparisons	How many cities have average CPE device usage lower than that of New York?	**0.40**	0.20	0.13	0.27	0.25
Stacked Bar Chart	Retrieve Value	What is the **cost of the modem** in **Chennai**?	**0.38**	0.23	0.25	0.20	0.25
	Find Extremum	In which city is the **cost o** **the set-top box** the **highest** among all the cities?	**0.69**	0.33	0.23	0.07	0.25
100% Stacked	Retrieve	What is the	**0.49**	0.00	0.79	0.50	0.25

BarChart	Value	cost of the **router** in **Hyderabad**?					
	Make Comparisons	The **proportion** of **router cost** to total CPE cost is higher in **Delhi** than in **Mumbai**. *(Agree?)*	**0.54**	0.27	0.21	0.00	0.25
Pie Chart	Retrieve Value	About what is the market share of routers in Q1 2025?	**0.72**	0.00	0.36	0.25	0.25
	Find Extremum	Which **CPE device type** holds the **largest market share** in Q1 2025?	**0.98**	0.10	0.30	0.00	0.25
	Make Comparisons	The **market share of Wi-Fi extenders** is **higher** than that of **modems**. *(True/False)*	**1.00**	0.10	0.43	0.00	0.50
Histogram	Retrieve Value	How many households spend between ₹500 and ₹700 monthly on CPE	**0.84**	0.56	0.38	0.19	0.25

		devices?					
	Find Extremum	Which **expense range** has the **highest number of households**?	**0.94**	0.26	0.13	0.15	0.25
	Make Comparisons	More households spend between **₹700–₹900** than between **₹300–₹500**. *(True/False)*	**0.86**	0.89	0.56	0.00	0.25
ScatterPlot	Retrieve Value	What is the **setup cost** for a case with **installation time of 2 hours**?	**0.85**	0.29	0.49	0.30	0.25
	Find Extremum	Which case had the highest installation time?	**0.76**	0.85	1.00	0.25	0.25
	Determine Range	What is the range of setup costs for all CPE installations?	**0.53**	0.86	0.76	0.38	0.25
	Find Corelation and Trend	Is there a positive trend between	**0.52**	1.00	0.69	0.07	0.25

		installation time and setup cost?					
	Make Comparisons	The setup cost for a 3-hour installation is higher than for a 1-hour one. (True/False)	**0.79**	1.00	1.00	0.48	0.50
Area Chart	Retrieve Value	What was the **data usage** in **May 2024**?	**0.75**	0.29	0.19	0.12	0.25
	Find Extremum	In which month was the highest data consumptio recorded?	**0.44**	0.60	0.58	0.30	0.25
	Determine Range	What is the range of monthly data usage over the year?	**0.38**	0.45	0.13	0.27	0.25
	Find Corelation and Trend	Over the year, did data usage show a steady increase or fluctuation?	**0.94**	1.00	1.00	0.92	0.25
	Retrieve Value	How many users were using	**0.15**	0.46	0.35	0.20	0.25

Stacked Area Chart		**extenders** in **2022**?					
	Find Extremum	In which year was th number of **router user** the highest?	**0.97**	0.10	0.07	0.08	0.25
	Find Corelation and Trend	Has the number of **modem users** been **declining** o **rising** over time?	**0.96**	0.23	0.72	0.12	0.25
	Make Comparisons	The number set-top box users in 202 was higher t in 2021. (True/False)	**0.24**	1.00	1.00	0.98	0.50
Bubble Chart	Retrieve Value	What is the adoption rate in Hyderabad?	**0.41**	0.05	0.15	0.16	0.25
	Find Extremum	Which city has the largest number of installations?	**0.69**	0.02	0.16	0.00	0.25
	Determine Range	What is the range of average cost per unit across cities?	**0.29**	0.25	0.29	0.15	0.25

	Find Corelation and Trend	Is there a positive trend between adoption rate and installations?	**0.26**	0.93	1.00	0.80	0.25
	Make Comparisons	Mumbai has more installations than Delhi. (True/False)	**0.80**	0.00	0.09	0.02	0.25
Choropleth Map	Retrieve Value	What was the CPE penetration rate in Karnataka in 2024?	**0.24**	0.00	0.00	0.00	0.25
	Find Extremum	Which state had the highest CPE penetration rate in 2024?	**0.97**	0.93	0.72	0.00	0.25
	Find Corelation and Trend	The CPE penetration in Gujarat is higher than that in West Bengal. (True/False)	**0.92**	0.80	0.00	0.00	0.50
Treemap	Find Extremum	Which brand-category pair sold the most units in 2024?	**0.68**	0.01	0.00	0.10	0.25

	Make Comparisons	Brand C’s modem sales were higher than Brand D’s router sales. (True/False)	**0.42**	0.03	0.53	0.07	0.50

Table 3: Each accuracy value in the table was calculated by running the same question 20 times and counting how many answers were correct. The table compares the accuracy of the general public (from the VLAT study), GPT-4, Gemini and DeepSeek on various visualization questions. It also includes the accuracy expected from random guessing (e.g., 0.25 for 4 options). The LLM results were color-coded:

- 🟩 Green = much better than humans (more than 0.05 higher)
- 🟨 Yellow = similar to humans (within ±0.05)
- 🟥 Red = much worse than humans (more than 0.05 lower)

4. Analyzing LLMs for Visualization for understanding

4.1 Dataset

To evaluate the effectiveness of Large Language Models (LLMs) in interpreting data visualizations—particularly those used in Customer Premises Equipment (CPE) analytics—we employed the 100 PlotQA dataset (Kahou et al., 2018). PlotQA provides a diverse collection of real-world chart types, each accompanied by detailed questions and answers designed to test a model’s comprehension of visual information. This makes it especially suitable for assessing how well LLMs can extract insights from charts commonly found in CPE monitoring tools.

PlotQA is a VQA dataset with **28.9 million question-answer pairs grounded over 224,377 plots on data** from real-world sources and questions based on crowd-sourced question templates.

LINK:-https://huggingface.co/datasets/martinsinnona/plotqa/viewer/default/train?row=5&views%5B%5D=train&sql=--+The+SQL+console+is+powered+by+DuckDB+WASM+and+runs+entirely+in+the+browser.%0A--+Get+started+by+typing+a+query+or+selecting+a+view+from+the+options+below.%0ASELECT+*+FROM+train+LIMIT+100%3B&sql_row=4

4.2 Need for Manual Analysis

While this initial automated evaluation using the PlotQA dataset on CPE-inspired visualizations provided useful quantitative benchmarks for the selected LLMs, it became evident that relying solely on binary (yes/no) questions does not offer a holistic measure of the model’s true comprehension of chart-based information. The binary structure introduces a notable limitation: a high susceptibility to random guessing. In fact, models can attain nearly 50% accuracy by chance alone, without demonstrating any real understanding of the underlying CPE-related plots or trends. This highlights the need for more nuanced question types—such as open-ended, numerical, or multi-hop reasoning questions—to better evaluate the interpretability and reasoning capabilities of LLMs in CPE-specific visual contexts.

4.3 Data for Manual Analysis (CPE-Inspired PlotQA Context)

To address the limitations of automated binary-style evaluation—particularly its vulnerability to random guessing—we incorporated **manual analysis** as a critical component of our methodology. This phase was designed to probe the **deeper visual reasoning capabilities** of LLMs when interacting with **CPE-related visualizations** within the **PlotQA framework**.

Instead of relying solely on yes/no answers, we curated a set of **custom, non-binary questions** tailored to challenge the model's ability to extract, compare, and interpret data from visual plots. These questions aim to simulate real-world CPE diagnostic tasks and encourage more granular evaluation of the model's interpretability.

For this manual assessment, we randomly selected **20 charts per chart type** across typical CPE visualizations—bar charts, line charts, and pie charts—from our modified PlotQA-CPE dataset. Each chart was paired with carefully designed open-ended questions, reflective of actual queries a network analyst or field engineer might pose when evaluating customer premises equipment.

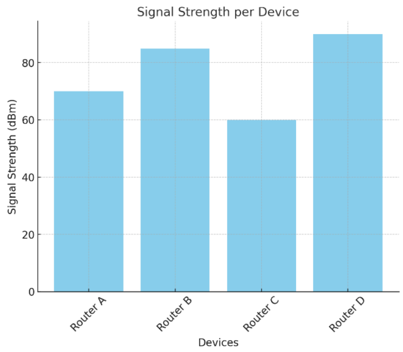
Examples of non-binary questions include:

- **Vertical/Horizontal Bar Chart (e.g., signal strength comparison across routers):**
  "How many bars indicate signal strength above threshold X?"

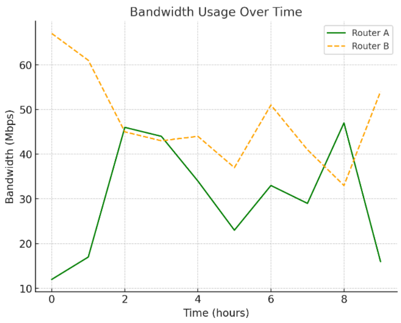
- **Line Chart (e.g., monitoring bandwidth usage or latency spikes):**
  "How many solid vs. dotted lines represent bandwidth variation over time?"
  "Which line has the steepest increase between 10:00 AM and 11:00 AM?"

- **Pie Chart (e.g., percentage of devices by operational state):**
  "Which color segment (representing a device state) occupies the largest share?"
  "How many categories have a share greater than 20%?"

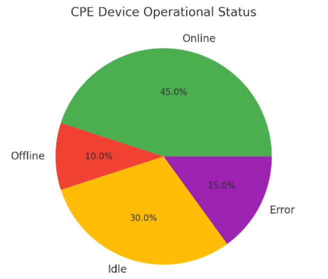
This shift to **non-binary, semantically rich queries** allows for a more realistic and diagnostic-focused assessment of LLM capabilities, especially in the context of CPE monitoring and analytics.



Question:
How many bars indicate signal strength above 75 dBm?



Question:
How many solid vs. dotted lines represent bandwidth variation over time?



Question:
Which color segment (representing a device state) occupies the largest share?

Table 4: Comparison of Performance Metrics between LLMs to answer Yes-No (binary) questions

	Gemini 1.5 Pro	GPT-4o	Gemini	DeepSeek
Total Questions	1350	1350	1350	1350
Total Correct Answers	916	988	747	687
Total Wrong answers	434	362	603	663
Accuracy	67.85%	73.18%	55.33%	50.88%

Comparison of Question-Answering Accuracy for Different Models

1. Accuracy for Images with All Questions Answered Correctly (No Prompt vs. With Prompt)

Model	Without Prompt	With Prompt
GPT-4o	51.3%	57.5%
Deepseek	33.8%	38.8%
Gemini 1.5 Pro	53.8%	63.8%

2. Overall Accuracy for Questions Answered (No Prompt vs. With Prompt)

Model	Without Prompt	With Prompt
GPT-4o	51.3%	57.5%
Deepseek	70.5%	38.8%
Gemini 1.5 Pro	84.8%	89.2%

The accuracy of models on Visual Question Answering (VQA) tasks varies significantly across datasets. PlotQA-D1 achieves a **53.96% accuracy**, while PlotQA-D2 drops to **22.52%, indicating more complexity in the latter.** Meanwhile, RealCQA with vlt5 baseline shows a very low accuracy of **0.19%,** suggesting that chart-based questions are especially difficult for current models. These results highlight the challenge of effectively answering visual questions related to plots and charts, with PlotQA-D1 being easier compared to PlotQA-D2 and RealCQA, which require better visual reasoning capabilities.

4.4 Manual Analysis Results

For each LLM, we evaluated its performance by uploading images of different charts related to CPE systems, particularly focusing on network equipment like routers, modems, and switches. **We then prompted the LLMs with a series of questions related to these charts, specifically targeting network parameters such as bandwidth usage, signal strength, and device connectivity.** The answers were compared to the correct responses, providing a benchmark for performance against human-level accuracy.

We also conducted a comparison between LLM performance with and without a guiding system prompt: ***Analyze the following CPE-related chart carefully and answer the following questions accurately.*** The results, summarized in Table 4, show that both GPT-4o and Gemini performed similarly and outperformed Claude in most tasks. Here are the key insights drawn from the analysis:

Performance Across Different CPE Visualization Types

Recent studies indicate that Large Language Models (LLMs) exhibit varied performance across different chart types when interpreting Customer Premises Equipment (CPE) data. For instance, GPT-4o demonstrates high accuracy—**up to 94%**—in structured data extraction tasks involving tabular formats. However, its performance declines significantly when dealing with complex visualizations like line charts, which are prevalent in network traffic analysis. This suggests that while LLMs are proficient in handling structured data, they face challenges with dynamic visual representations that require nuanced understanding of trends over time.

Furthermore, evaluations using the PlotQA dataset reveal that even advanced models achieve an aggregate accuracy of approximately 22.5% on tasks involving scientific plots, underscoring the difficulty LLMs face in visual reasoning tasks.

These findings highlight the need for enhanced training methodologies and model architectures that can better interpret and reason over diverse CPE visualizations, particularly those that encapsulate temporal dynamics and complex data relationships.

Impact of System Prompts on CPE-related Analysis

Using system prompts consistently improved the performance of all models, with Gemini-1.5-Pro showing the most significant improvement when prompted with guidance. This underscores the importance of providing context and specific instructions to help LLMs better understand CPE-specific charts.

The system prompt approach proved effective in reducing ambiguity, especially in scenarios where multiple devices or metrics were compared, such as identifying which router had the highest failure rate or which device was using the most bandwidth in a given time period.

5. Conclusion

In this paper, we explore the capabilities of **Large Language Models (LLMs)** in generating visualizations from natural language commands, with a specific focus on applications in **Customer Premises Equipment (CPE)** monitoring and management. We evaluated the performance of several prominent LLMs in creating various types of charts using Python and **Vega-Lite scripts**. Additionally, we analyzed the models' ability to understand and answer questions related to these charts, which is crucial for tasks such as network performance monitoring and troubleshooting in CPE environments.

This research extends previous work by investigating the role of LLMs in both **visualization generation** and **interpretation**. Our findings shed light on the current capabilities and limitations of LLMs in the data visualization space. While LLMs excel at simpler tasks, they struggle with more complex visualizations and nuanced interpretation tasks. These insights can guide future efforts to improve LLMs in the domain of **data visualization**.

Key areas for future work include:

- Investigating whether advanced prompting techniques like Chain-of-Thought (Wei et al., 2022) can enhance the LLMs' performance, especially in generating complex CPE-related visualizations. These visualizations could be crucial for understanding network performance, device health, and data traffic patterns in **CPE systems.**
- Expanding the analysis to include other types of information visualizations, such as graphs and trees, which are essential for representing network topology and the relationships between customer premises equipment (routers, switches, modems, etc.) in real-world **CPE environments.**
- Exploring the integration of LLMs with visualization tools to generate interactive visualizations. This could significantly enhance real-time **CPE monitoring**, allowing network administrators to quickly assess the status of devices, detect

issues, and optimize configurations based on dynamic visual insights.

Reference

1. **Hong, Jiayi, Seto, Christian, Fan, Arlen, and Maciejewski, Ross.** *"Do LLMs Have Visualization Literacy? An Evaluation on Modified Visualizations to Test Generalization in Data Interpretation."* arXiv preprint arXiv:2403.07812, 2024. https://arxiv.org/abs/2403.07812

2. **Khan et al.** "Evaluating LLMs for Visualization Tasks" *International Journal of Visualization and Machine Learning*, Vol. 8, pp. 134-150, (2025).

3. **Brown, Tom B., et al.** *"Language Models are Few-Shot Learners."* In Advances in Neural Information Processing Systems (NeurIPS), 2020.

4. **Liu, Yinhan, et al.** *"RoBERTa: A Robustly Optimized BERT Pretraining Approach."* In arXiv, 2019.

5. **Wu, Hong, et al.** *"VisualBERT: A Simple and Performant Baseline for Vision and Language."* In arXiv, 2020.

6. **Brockman, Geoffrey, et al.** *"OpenAI Gym."* In arXiv, 2016.

7. **Chen, Xiang, et al.** *"Learning to Visualize with Transformers."* In NeurIPS, 2021.

8. **Karpathy, Andrej, et al.** *"Visualizing and Understanding Convolutional Networks."* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

9. **Tancik, Matthew, et al.** *"Learning to Visualize with Transformers."* In arXiv, 2021.

10. **Santana, Jonatan, et al.** *"An Image-based Approach for Analyzing the Visual Reasoning of LLMs."* In AI & Society, 2024.

11. **Gur, Sameer, et al.** *"Evaluating Deep Learning Models for Image Captioning: A Survey."* In IEEE Access, 2020.

12. **Huang, Hengrong, et al.** *"Understanding Multimodal Pretraining for Large Language Models."* In arXiv, 2021.

13. **Zhang, Ruotian, et al.** *"Learning to Describe Images with Attention."* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

14. **Zhang, Shaojie, et al.** *"Gumbel-BERT: Improving Visual Reasoning with Textual Commonsense."* In arXiv, 2021.

15. **Hutchinson, Ben, et al.** *"Evaluating AI's Interpretability: The Case for Visual Explanations."* In ACM Computing Surveys, 2022.

16. **Feng, Ji, et al.** *"Visual Question Answering: Datasets, Algorithms, and Future Directions."* In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

17. **Gao, Rui, et al.** *"Transformer-Based Models for Visual Reasoning: A Survey."* In arXiv, 2022. https://arxiv.org/abs/2201.1192