An Evaluation of Vision-Language Models on Graph Reconstruction

Jiahang He* Vamanie Perumal† Ishan Jain‡

YRI Research

Abstract

Recent advances in multimodal large language models (LLMs), which combine visual and linguistic understanding, have improved visual data interpretation. However, their ability to recover underlying numerical data from visualizations remains underexplored, hindering analysis and reproducibility, especially when most real-world visuals lack accompanying datasets. To address this, we introduce a novel task and evaluation framework requiring vision-language models (VLMs) to reconstruct datasets solely from chart images, without access to original data. Unlike conventional ChartQA tasks that focus on answering individual questions, our approach evaluates the model's capacity to understand the entire dataset, providing a more comprehensive evaluation of its accuracy. We curate eight diverse, real-world datasets spanning different domains to evaluate three leading models. Our results show that simpler charts achieve higher fidelity than complex ones, and that a model's ability to generate diverse chart types does not necessarily correspond to greater accuracy. By evaluating models' ability to convert unstructured visual data into structured information, we aim to help data analysts more efficiently validate, interpret, and leverage real-world datasets for downstream insights.

Index Terms: Vision-Language Model, Data Visualization, Data Reconstruction

1 Introduction

Large language models (LLMs) such as GPT-3 [3] and GPT-4 [23] have demonstrated exceptional capabilities in natural language understanding and generation. Building upon these advancements, recent progress in multimodal models that integrate image processing with text has enabled tasks including basic plot generation [12], multi-image reasoning [26], and multimodal scientific analysis [4]. These models, commonly referred to as vision-language models (VLMs), combine visual perception with reasoning abilities and include systems like Kosmos-1 [11] and BLIP [14]. As these models continue to evolve, research is increasingly focused on their potential for complex, structured reasoning across multiple modalities.

While prior work has examined VLMs in chart classification [7], caption generation [16], and data extraction [22], these efforts largely concentrate on surface-level recognition, leaving deeper inferential and reasoning capabilities underexplored. In particular, the challenge of identifying the underlying dataset that generated a chart—a cognitively demanding task involving both visual interpretation and numerical estimation—has not been extensively studied. This skill is especially critical in practical scenarios where analysts often encounter visualizations without access to their original data, making further analysis challenging.

In this paper, we propose a novel task that requires VLMs to reconstruct the underlying dataset solely from a chart image, without access to the source data. Unlike traditional ChartQA tasks that focus on answering specific questions about a chart, our framework poses a reverse reasoning challenge, evaluating a model's ability to infer the complete numerical structure behind graphical representations. This holistic evaluation provides a more thorough assessment of a model's accuracy and understanding of plots. As shown in Fig. 1, our evaluation begins with the creation of visualizations derived from carefully curated datasets. We then prompt the VLM to reconstruct the dataset it believes generated the visualization. Finally, we assess the model's output using both quantitative error metrics and comparison plots.

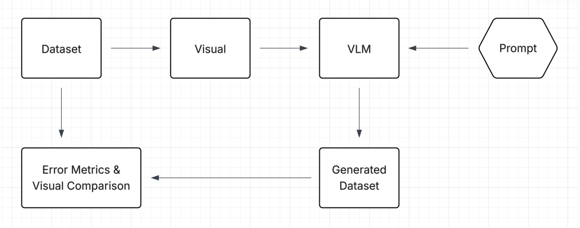


Figure 1: Evaluation framework pipeline

The ability to perform such inference has significant practical implications. Successful dataset reconstruction by VLMs can augment human data analysis and insight generation by converting unstructured visuals into structured data. This enables more efficient validation, interpretation, and downstream use of complex information.

2 Background and Research Gaps

2.1 Multimodal Reasoning in Vision-Language Models

Multimodal reasoning refers to a model's ability to both interpret and reason over inputs from multiple modalities, such as text and images. Recent advances in VLMs like Flamingo [1], PaLI [6], Kosmos-1 [11], BLIP-2 [14], and GPT-4 [23] have shown that large-scale models can integrate visual and textual signals to perform complex reasoning tasks. These models have been evaluated on datasets such as ChartQA [20] and PlotQA [21], which require interpreting plots, extracting structured information, or answering questions grounded in charts. While much of this work focuses on generating new plots or answering chart-related questions, our approach differs by asking the model to identify the underlying data that could have produced a given visualization. This setting challenges the model not only to recognize chart semantics but also to generate plausible numerical structures based solely on visual patterns, an emerging frontier in multimodal inference.

2.2 Chart Data Extraction

Traditional chart data extraction systems typically rely on multistage pipelines that sequentially detect chart components, apply OCR, and infer relationships through rule-based parsing or language models [18]. These approaches often struggle with generalizability due to their dependency on specific chart layouts and brittle intermediate representations [17]. Recent advances have introduced end-to-end deep learning systems that bypass intermediate stages by directly generating structured data from chart images.

*jhe.primary@gmail.com
†perumal.vamanie@gmail.com
‡ishan@yriscience.com

For instance, DePlot is a transformer-based model trained only to extract data from visuals, with its outputs fed into an LLM for table evaluation, improving ChartQA performance [15]. Similarly, UniChart expands upon this approach by incorporating a text decoder directly inside its pretrained model, enabling it to generate structured outputs and answer questions straight from chart images [19]. While these methods represent significant progress, they primarily enhance performance on isolated question-answering tasks and typically require either training a model for a single specialized task or employing multiple models. In contrast, our work investigates whether VLMs can infer the underlying data of a chart without task-specific fine-tuning. Furthermore, we evaluate across a wider range of real-world datasets and adopt a more holistic framework for assessing a model's graph understanding.

2.3 Model Application to Specific Domain Tasks

LLMs have demonstrated versatility when adapted to domain-specific tasks across various fields, including medicine, law, and science. For instance, BioBERT [13] and ClinicalBERT [10] have been tailored for biomedical text mining and clinical note understanding, while LegalBERT [5] addresses challenges in legal document analysis. Additionally, fine-tuned models such as SciBERT [25] have advanced scientific text processing by incorporating domain-specific corpora and vocabularies. These domain-focused adaptations illustrate the capability of LLMs to leverage large-scale pretraining while specializing through fine-tuning or prompt engineering, enabling more precise reasoning. Our work builds on this foundation by exploring LLM capabilities in synthesizing structured datasets from domain-specific visualizations, bridging language understanding with data reconstruction.

3 Evaluation Design

3.1 Dataset Curation

All datasets used in this study, including their source name, domains, dataset sizes, and associated graph types, are summarized in Tab. 3. To better simulate real-world data scenarios, we curated datasets spanning a range of domains, aligning with recommendations from prior work emphasizing domain diversity to improve the model's ability to be generalized [24]. Following common practice in data-centric research, we prioritized datasets with substantial size and minimal to no missing values to ensure reliable evaluation and reduce confounding factors introduced by data sparsity [8].

3.2 Commercial Models Used

We tested our experiment on three different models: GPT-4o [23], Gemini 1.5 Pro [9], and Llama-3.2-11B-Vision-Instruct [25]. These models were all tested using API keys and give us a variety of model sizes and architectures to evaluate.

3.3 Visualization Types

Depending on the nature of the datasets, we generated different plot types for specific purposes to test the VLM. Here, we have organized them based on our assumptions regarding the levels of difficulty for completing the data reconstruction task.

Easy:

- Bar plot: Used to evaluate the model's ability to compare values across categories.
- Line graph: Used to evaluate the model's understanding of trends over time.
- Pie chart: Used to evaluate understanding of proportions and part-to-whole relationships.

Medium:

- Scatter plot: Used to evaluate the model's ability to identify patterns, correlations, and outliers. We assume this may be challenging because there exist many small data points that may overlap.
- Box plot: Used to evaluate the model's understanding of medians, quartiles, and outliers. We assume this may be difficult since box plots encode summary stats like interquartile range and median, but not raw data.

Hard:

- Radial plot: Used to evaluate the model's ability to interpret multivariate data distributed across multiple axes in a polar coordinate system. We assume that this will be challenging because a circular layout can make scale estimation and axis mapping tricky.
- Heatmap: Used to evaluate recognition of patterns in a grid based on color gradients. We assume this will be challenging because VLMs are notoriously bad at distinguishing color gradients [2].
- Bubble chart: Used to evaluate multivariable reasoning by requiring the model to interpret size and position at once. We assume that this will be challenging because the model has to understand and generate data on three distinct features.

3.4 Prompting Standards

In all of our tests, we used zero-shot prompting. Additionally, for deterministic and reproducible answers, we set the temperature of our models to 0. Each session is started with the following prompt:

Baseline Prompt

> "You are an assistant, skilled in reading and interpreting visually represented data. Create me a CSV file with the full dataset that you think was used to generate this plot:" + *Visual*

Caveats:

- For the bubble plot, we had to explicitly prompt it to find 15 data points and form a dataset with three distinct columns.
- For the box plot and heatmap, we had to ask the models to retry a maximum of five times if they encountered errors or generated an empty dataset.

3.5 Evaluation Metrics

To evaluate model performance, we use two primary error metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), shown in Equations (1) and (2), respectively. Because box plots and radial plots present summarized data rather than individual data points—preventing even humans from recovering the underlying dataset—we evaluate them based on the mean of the extracted values. In addition to these metrics, we conduct a consistency test by running each model five times and assessing whether it successfully generates data in the CSV file for all runs.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{\text{true}} - y_i^{\text{pred}}\right| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{\text{true}} - y_i^{\text{pred}}\right)^2} \tag{2}$$

y^{true} : ground truth value, y^{pred} : value generated by model, n : number of data points

Due to inconsistencies in scale and magnitude across our datasets, we normalized the error terms by dividing Equations (1) and (2) by the range of the ground truth values, shown in Equation (3). This results in the scaled formulation presented in Equation (4).

$$\text{Range} = \max(y^{\text{true}}) - \min(y^{\text{true}}) \tag{3}$$

$$\text{Scaled MAE} = \frac{\text{MAE}}{\text{Range}}, \quad \text{Scaled RMSE} = \frac{\text{RMSE}}{\text{Range}} \tag{4}$$

Additionally, certain graph types, such as radial plots and heatmaps, encode multiple feature dimensions within a single visualization. To account for this, we computed the average MAE and RMSE across all features of each individual plot type using Equations (5) and (6), respectively. These aggregated error values provide a quantitative basis for comparing the relative difficulty of reconstructing data for different chart types.

$$Overall\,MAE = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} \left| y_{i,j}^{true} - y_{i,j}^{pred} \right|}{\sum_{j=1}^{m} n_j} \tag{5}$$

$$Overall\,RMSE = \sqrt{\frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(y_{i,j}^{true} - y_{i,j}^{pred} \right)^2}{\sum_{j=1}^{m} n_j}} \tag{6}$$

m : number of features

Finally, we took the mean of MAE and RMSE values across all individual plot types. These final metrics serve as a summary measure of each model's overall performance and enable direct comparison between models.

4 Results

Graph / Model	GPT-4	Gemini	Llama 11B
Bar Graph	✓	✓	✓
Line Graph	✓	✓	✓
Pie Chart	✓	✓	✓
Scatter Plot	✓	✓	✓
Box Plot	✓	✗	✗
Radial Plot	✓	✓	✗
Heatmap	✓	✓	✗
Bubble Chart	✓	✓	✓
Consistency	✓	✗	✓

Table 1: Comparison of VLMs on graph reconstruction capability. A checkmark indicates the model successfully generated the plot in all five turns, while an X indicates failure to generate the plot in any of the five turns.

As shown in Tab. 1, GPT-4 was the only model that successfully generated a plot for every chart type across all evaluation trials. Gemini ranked second, failing to generate box plots in all five attempts, while Llama 11B was unable to produce a box plot, radial plot, and heatmap in any trial. That said, Gemini exhibited some inconsistency with bubble charts, generating a valid output in only four out of five cases.

Despite GPT-4's ability to produce all graph types without failure, analysis of error metrics in Tab. 2 and Tab. 4 reveals that it consistently exhibited the highest error rates across nearly all chart types. This result is initially counterintuitive, as GPT-4 demonstrated the highest coverage and output consistency. However, closer inspection of its outputs for more complex visualizations, particularly heatmaps, shows that GPT-4 often generated random, unmeaningful values across turns, diverging significantly from both the ground truth and its own earlier generations. Thus, while GPT-4 displayed robustness in format generation, its numerical fidelity to the original visual data was comparatively poor.

Both Gemini and Llama 11B struggled with the more complex chart types as well. However, Gemini consistently outperformed GPT-4 on simpler and moderately difficult graphs, such as bar graphs and line graphs, achieving lower error metrics. Llama 11B showed a similar trend of better accuracy on supported chart types, though we were unable to compute its overall mean error due to its inability to generate a sufficient number of charts.

Graph / Model	GPT-4	Gemini	Llama 11B	Mean
Bar Graph	0.115	0.032	0.025	0.057
Line Graph	0.075	0.013	0.013	0.033
Pie Chart	0.001	0.001	0.001	0.001
Scatter Plot	0.958	0.234	0.208	0.467
Box Plot	0.240	—	—	0.240
Radial Plot	17.129	9.935	—	13.532
Heatmap	12.396	12.395	—	12.395
Bubble Chart	0.302	0.420	0.424	0.382
Mean	3.902	3.290	—	

Table 2: RMSE comparison of VLMs on graph reconstruction capability

In addition to model-level performance, we analyzed chart-level difficulty based on both graph comparisons and the MAE and RMSE across models for each graph type. This revealed clear patterns in which visualizations were the simplest and which posed the greatest challenges. Fig. 4 illustrates that the error rate for bar chart reconstruction is relatively low across all three models, with occasional spikes in errors for GPT. This trend is consistent with the results in Tab. 2, which indicate that GPT overall performed worse than the other models; however, the overall performance on bar charts remains comparatively strong for all three models. This outcome is expected, as bar charts typically present information in a straightforward manner, requiring models primarily to interpret the axes.

Conversely, Fig. 2 illustrates that the models struggled to reconstruct scatter plots. While the results indicate that the models captured the fundamental patterns of the data, all models failed to generate a sufficient number of points. Due to differing numbers of data points, MAE and RMSE were computed as the mean of available points. As a result, Tab. 2 suggests that LLaMA 11B outperformed Gemini; however, the reconstructed plots suggest that Gemini may have achieved the best performance. These discrepancies highlight the importance of using multiple evaluation methods, particularly for complex tasks like graph reconstruction, where visual outputs can capture nuances that numerical error metrics alone may not fully reflect.

As seen in Tab. 2, radial plots and heatmaps exhibited even more pronounced difficulties. RMSEs for these charts reached or came close to double digits for both GPT-4 and Gemini, and Llama 11B failed to even generate either plot type in any of the five trials. Further analysis of these charts, seen in Fig. 8 and Fig. 9 reveals that the models failed to grasp even the most fundamental parts of the visual, producing outputs with no meaningful correlation to the original plots. These charts likely demand advanced spatial reasoning, precise value mapping, and color gradient understanding, which current VLMs appear to lack.

5 Conclusion and Future Works

In conclusion, our study demonstrates that while current VLMs show promising ability to reconstruct datasets from chart images, their performance is highly sensitive to chart complexity and visual

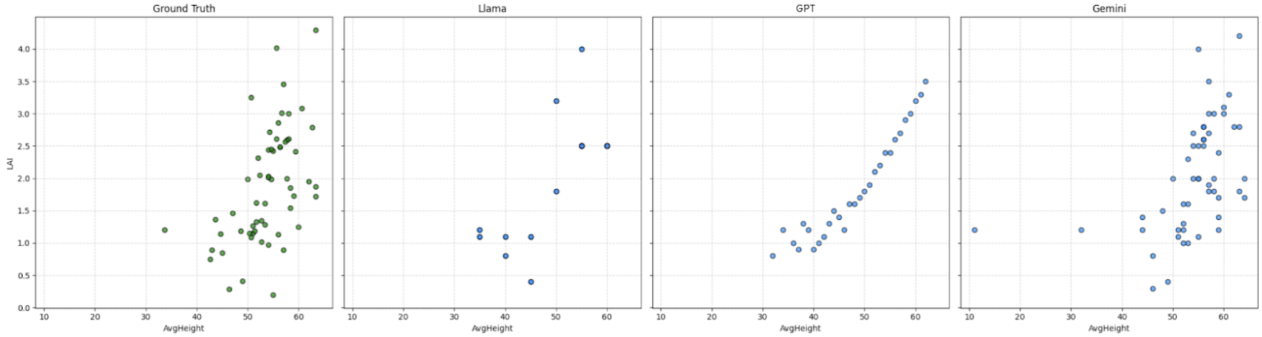


Figure 2: Scatter plot of leaf area index (LAI) vs. average cotton height comparing graph reconstruction capabilities across models.

design. The lack of a strong correlation between chart-generation capability and reconstruction accuracy suggests that generating visually diverse outputs does not imply deeper numerical understanding. By introducing this dataset reconstruction framework, we hope to provide a new lens for evaluating VLMs, highlighting their potential to augment human–AI collaboration and assist data analysts in extracting and validating insights from visuals.

At the same time, a few limitations present in our approach may highlight possible directions for future work. One particular challenge observed in our evaluation was the inconsistency across runs. When provided with the same chart and identical prompts, models often produced different outputs on different executions. This variability suggests that some internal stochasticity or sensitivity to initialization is affecting the generation process, which in turn raises concerns about the stability and reproducibility of these systems. For applications requiring repeatability or auditability, such as scientific research or regulated domains, this lack of determinism could be a significant barrier to adoption. Another limitation arises when evaluating model performance on visualizations that encode summary statistics rather than individual data points. For example, in the case of box plots, even human observers cannot reconstruct the underlying dataset, making complete extraction from the VLM extremely difficult. This problem also appeared with scatter plots, where models weren't able to generate all of the original data points. In this study, we evaluated such visualizations by comparing the mean of the extracted values to the ground truth mean. However, further research is needed to explore more effective or comprehensive evaluation methods for these types of plots. We also found that prompt sensitivity played a major role in determining the accuracy and structure of generated datasets. Minor changes in wording or instruction often resulted in markedly different outputs, even when the underlying task of reconstructing charts remained constant. This sensitivity not only complicates reproducibility but also introduces a steep learning curve for users who may not know how to craft optimal prompts. While prompt engineering has become a standard technique for guiding model behavior, our findings suggest that more research is needed to identify which prompts optimize performance on high-precision tasks like data reconstruction from visuals.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. 1

[2] A. Bendeck and J. Stasko. An empirical evaluation of the gpt-4 multimodal language model on visualization literacy tasks. *IEEE Transactions on Visualization and Computer Graphics*, 31:1105–1115, 2024. 2

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. teusz Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 1

[4] H. Cai, X. Cai, S. Yang, J. Wang, L. Yao, Z. Gao, J. Chang, S. Li, M. Xu, C. Wang, H. Wang, Y. Li, M. Lin, Y. Li, Y. Yin, L. Zhang, and G. Ke. Uni-smart: Universal science multimodal analysis and research transformer. *ArXiv*, abs/2403.10301, 2024. 1

[5] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, and N. Aletras. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.00777*, 2020. 2

[6] X. Chen, X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. M. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. V. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut. Pali: A jointly-scaled multilingual language-image model. *ArXiv*, abs/2209.06794, 2022. 1

[7] A. Dhote, M. H. Javed, and D. S. Doermann. A survey and approach to chart classification. In *ICDAR Workshops*, 2023. 1

[8] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. Daumé, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64:86 – 92, 2018. 2

[9] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024. 2

[10] K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342, 2019. 2

[11] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei. Language is not all you need: Aligning perception with language models. *ArXiv*, abs/2302.14045, 2023. 1

[12] S. R. Khan, V. Chandak, and S. Mukherjea. Evaluating llms for visualization tasks. *ArXiv*, abs/2506.10996, 2025. 1

[13] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2019. 2

[14] J. Li, D. Li, S. Savarese, and S. C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 1

[15] F. Liu, J. M. Eisenschlos, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, W. Chen, N. Collier, and Y. Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. *ArXiv*, abs/2212.10505, 2022. 2

[16] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan. Chinese image caption generation via visual attention and topic modeling. *IEEE Transactions on Cybernetics*, 52:1247–1257, 2020. 1

[17] J. Luo, Z. Li, J. Wang, and C.-Y. Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1916–1924, 2021. 1

[18] W. Ma, H. Zhang, S. Yan, G. Yao, Y. Huang, H. Li, Y. Wu, and L. Jin. Towards an efficient framework for data extraction from chart images. *ArXiv*, abs/2105.02039, 2021. 1

[19] A. Masry, P. Kavehzadeh, D. X. Long, E. Hoque, and S. R. Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *ArXiv*, abs/2305.14761, 2023. 2

[20] A. Masry, D. X. Long, J. Q. Tan, S. R. Joty, and E. Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ArXiv*, abs/2203.10244, 2022. 1

[21] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. Plotqa: Reasoning over scientific plots. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525, 2019. 1

[22] O. Mustafa, M. K. Ali, M. Moetesum, and I. Siddiqi. Charteye: A deep learning framework for chart information extraction. *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 554–561, 2023. 1

[23] OpenAI Team. Gpt-4 technical report. 2023. 1, 2

[24] I. D. Raji, E. M. Bender, A. Paullada, E. L. Denton, and A. Hanna. Ai and the everything in the whole wide world benchmark. *ArXiv*, abs/2111.15366, 2021. 2

[25] H. Touvron, L. Martin, K. R. Stone, P. Albert, A. Almahairi, Y. Babaei, N. lay Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. M. Bikel, L. Blecher, C. tian Cantón Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. S. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. M. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. H. M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. 2

[26] G. Zhang, T. Zhong, Y. Xia, Z. Yu, H. Li, W. He, F. Shu, M. Liu, D. She, Y. Wang, and H. Jiang. Cmmcot: Enhancing complex multi-image comprehension via multimodal chain-of-thought and memory augmentation. *ArXiv*, abs/2503.05255, 2025. 1

A Datasets

Dataset	Plot Used	Domain	Size	Features Used
Smlr Stock	Line Graph	Finance	2604 x 7	Close, Date
Mileage Fee Survey	Pie Chart	Transportation	2114 x 74	Region, Politics
Detection Of Influenza A	Heatmap	Bioinformatics	21434 x 9	Boman, Charge, Isolectric Point, Instability, Hydrophobicity
Pond Chemistry Data	Radial Plot	Energy	742 x 42	SR, NH3-NH4-N, Peak A, SUVA254, NO3-N
Cultural Niches Of Bird	Bubble Chart	Ecology	1244 x 21	Log 10 Mass, Max Color Contrast, Popularity
Monitoring Cotton Growth	Scatter Plot	Agriculture	294 x 2159	Avg Height, Leaf Area Index (LAI)
Hydroclimate In Grassland	Box Plot	Climate	317337 x 25	Treatment, Volumetric Water Content (VWC)
Above Fed Fires	Bar Chart	Environmental Science	26147 x 7	Fire ID, Day of Year

Table 3: Datasets used to test VLMs

B MAE Table

Graph / Model	GPT-4	Gemini	Llama 11B	Mean
Bar Graph	0.081	0.027	0.021	0.043
Line Graph	0.051	0.012	0.012	0.025
Pie Chart	0.001	0.001	0.001	0.001
Scatter Plot	0.745	0.168	0.168	0.360
Box Plot	0.190	—	—	0.190
Radial Plot	16.322	8.636	—	12.479
Heatmap	12.052	12.395	—	12.224
Bubble Chart	0.224	0.339	0.335	0.299
Mean	3.708	3.083	—	

Table 4: MAE comparison of VLMs on graph reconstruction capability

C Graph Reconstructions

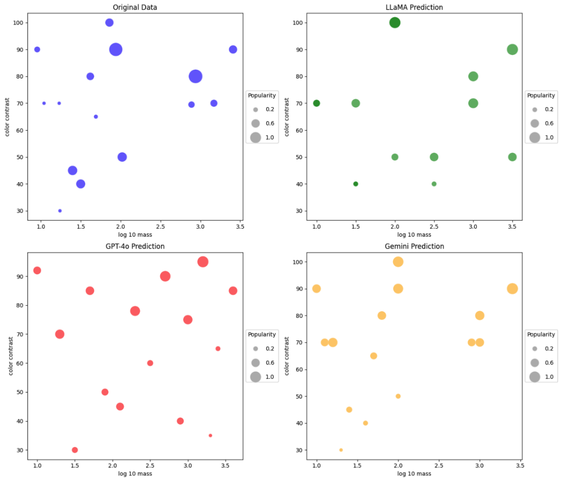


Figure 3: Bubble chart of max color contrast vs. log 10 mass comparing graph reconstruction capabilities across models.

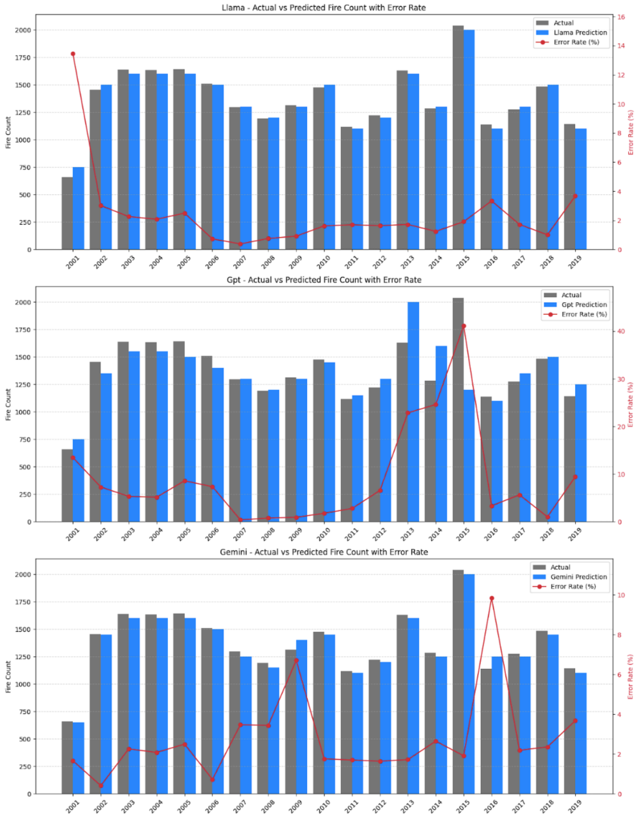
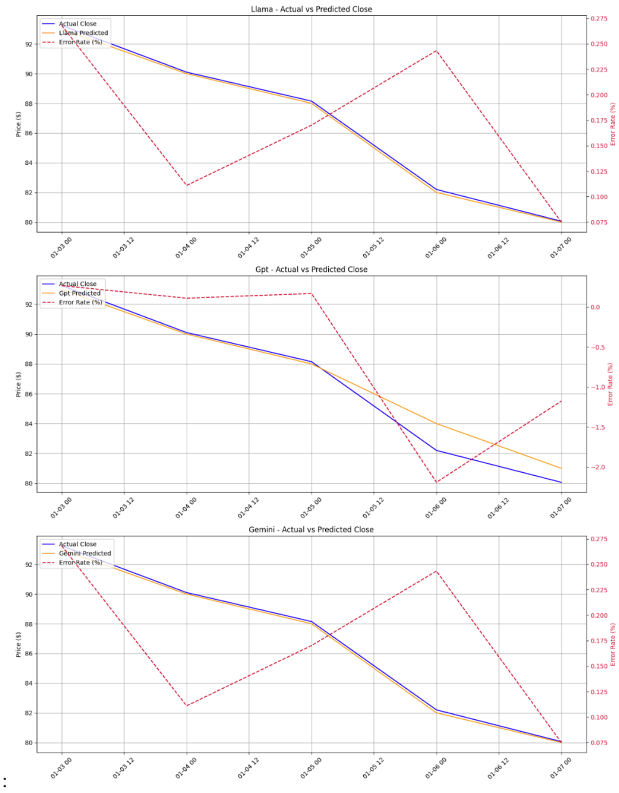


Figure 4: Bar graph comparing actual and generated values, overlaid with a dual-axis line plot showing error rates. Note that the error rates are plotted on three separate scales.



:

Figure 5: Line graph comparing actual and generated values with a dual-axis plot showing error over time. Note that the error rates are plotted on three separate scales.

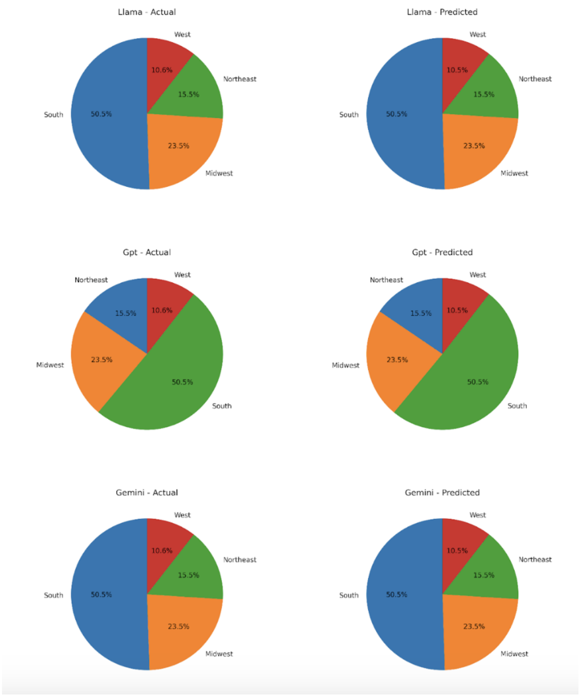


Figure 6: Pie chart comparing mileage fee survey responses for different regions across different models.

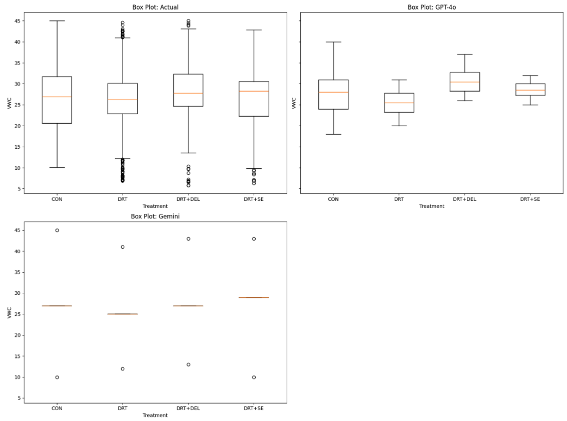
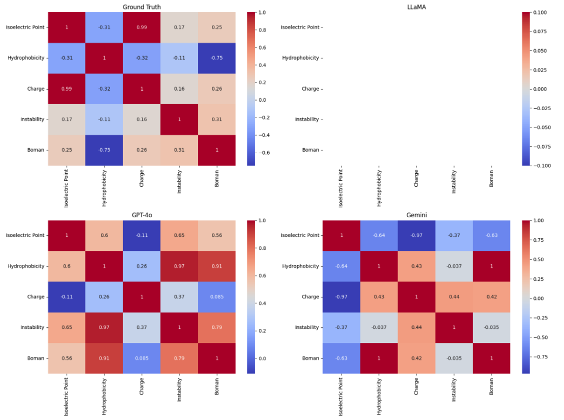


Figure 7: Box plot comparing volumetric water content (VWC) with treatment type across models.



Ground Truth	Isoelectric Point	Hydrophobicity	Charge	Instability	Boman
Isoelectric Point	1	-0.31	0.99	0.17	0.25
Hydrophobicity	-0.31	1	-0.32	-0.11	-0.75
Charge	0.99	-0.32	1	0.16	0.26
Instability	0.17	-0.11	0.16	1	0.31
Boman	0.25	-0.75	0.26	0.31	1

GPT-4o	Isoelectric Point	Hydrophobicity	Charge	Instability	Boman
Isoelectric Point	1	0.6	-0.11	0.65	0.56
Hydrophobicity	0.6	1	0.26	0.97	0.91
Charge	-0.11	0.26	1	0.37	0.085
Instability	0.65	0.97	0.37	1	0.79
Boman	0.56	0.91	0.085	0.79	1

Gemini	Isoelectric Point	Hydrophobicity	Charge	Instability	Boman
Isoelectric Point	1	-0.64	-0.97	-0.37	-0.63
Hydrophobicity	-0.64	1	0.43	-0.037	1
Charge	-0.97	0.43	1	0.44	0.42
Instability	-0.37	-0.037	0.44	1	-0.035
Boman	-0.63	1	0.42	-0.035	1

Figure 8: Heatmap illustrating the correlation between isolectric point, hydrophobicity, charge, instability, and boman index across models. Note that Llama-11B failed to extract any data.

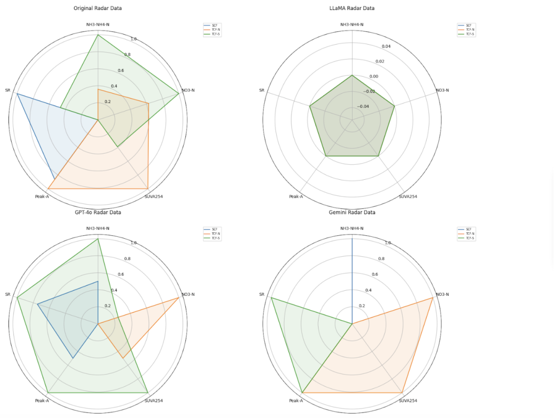


Figure 9: Radial plot comparing graph reconstruction ability of strontium, a combination of ammonia and ammonium nitrogen, nitrate nitrogen, UV absorbance, and fluorescence peak A across models. Note that Llama-11B failed to extract any data.