Lessons from Developing and Evaluating LLMs for Data Visualization

Qianwen Wang
The University of Minnesota, Twin Cities
MN, USA
qianwen@umn.edu

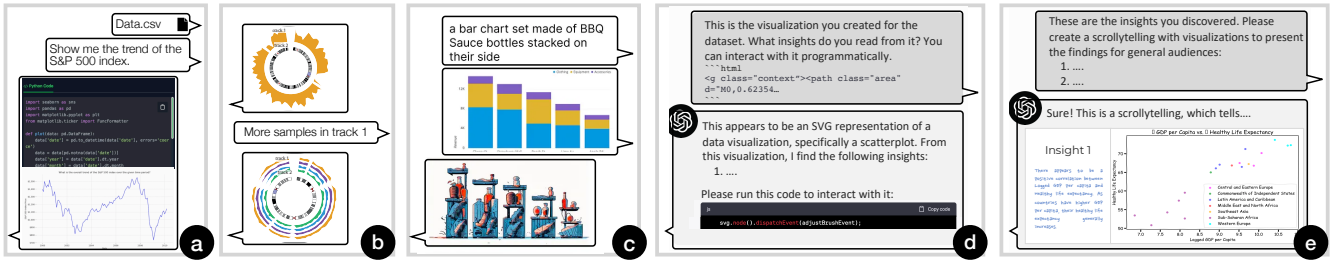

Figure 1: Various applications of LLMs in Data Visualization: ranging from the basic visualization creation (a), to design modification(b), embellishment (c), visualization interaction (d), and insight communication [1] (e). Figures are adapted from [1, 5].

ABSTRACT

Recently, the application of Large Language Models (LLMs) has broadened considerably into multimodal contexts that combine visual elements with natural language. Evaluating LLMs within these multimodal settings presents unique challenges not encountered in purely textual environments. This paper shares insights from our experiences in developing and assessing LLMs for data visualization. It begins with an overview of the current applications of LLMs in data visualization, ranging from data preparation to visualization design, visual exploration, and insight communication. We then reflect on key elements in designing evaluations and discuss the challenges encountered in our previous evaluation processes.

1 LLM IN DATA VISUALIZATION

By transforming data into graphic elements, data visualization enables the discovery of hidden patterns and the communication of compelling stories within datasets. to help people discover hidden patterns as well as communicate compelling stories in data. This field spans a broad range of applications, from rigorous analysis in finance to captivating storytelling in artistic creation. The integration of machine learning techniques to enhance the design, creation, and interaction with visualizations has been a longstanding practice within the visualization community, as documented by the survey of Wang et al. [4].

With the advances in LLMs, the field of data visualization is also exploring the applicaiton of LLMs. While Early endeavors, such as LIDA [2], are restricted in using LLM to interpret data semantic and generate programming code that can create visualizations, current studies has begun to explore LLMs' applications across the entire spectrum of data visualization processes. A notable instance is the empirical evaluation conducted in Harvard's CS171 data visualization course [1], where the GPT model demonstrated its capability to perform basic visualization-related tasks, mostly via generating javascript codes, that are typically expected of students in an introductory visualization class.

1.1 Visualization Formats Used in LLMs

Generally speaking, data visualizations can be categorized into three main formats as below.

- **Programming code** is favored for its structured and semantic clarity, allowing for dynamic generation and modification of visualizations. For instance, in crafting a bar chart to display trends in the S&P 500 indices, programming languages like Python and R are employed to convert index values into visual elements such as bar heights, offering a direct pathway for LLMs to engage with and manipulate data representation.
- **Vector files** offer precision by defining visual elements through a detailed syntax, making them ideal for outlining specific graphical properties, such as the exact dimensions and placements of bars in a chart. This format's structured nature allows LLMs to interpret and potentially modify the visualization at a conceptual level.
- **Bitmap images** present visualizations in their final, rendered state as a matrix of pixel values. This format encapsulates the visualization without exposing the underlying data or structure, posing a challenge for LLMs to directly interpret or alter the content without advanced image processing techniques.

Among these formats, programming code is the most accessible for LLMs to interpret and modify due to its structured and semantic nature, while bitmap images are the hardest.

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

1.2 Visualization Tasks Achieved by LLMs

The significant impact of LLMs in the field of data visualization are underscored by their wide range of applications. From automating data preparation to generating intricate visual representations, LLMs demonstrate an exceptional capacity to enhance and streamline the visualization process (Figure 1).

- **Data Interpretation, Preparation, and Cleaning** in visualization are commonly conducted through writing programming code, a task for which LLMs have been widely used for.
- **Visualization Generation** entails crafting visual representations from datasets. Furthermore, LLMs serve implicitly here by suggesting suitable visualization types based on text descriptions of the data characteristics and user intent.
- **Visualization Modification and Embellishment** includes tasks such as adjusting existing visualizations for clarity, aesthetics, or additional information, as shown in Figure 1 (b-c).
- **Guiding Interaction** is a novel area where LLMs show promise is in facilitating user interaction with visualization systems. By understanding user queries, intents, and their levels of knowledge, LLMs can offer guided exploration paths or highlight areas of interest, enriching the user's engagement with the data.
- **Visualization Reading and Assessment** LLMs play a critical role in interpreting and evaluating visual content. They assist in decoding complex visualizations, providing insights into their effectiveness, and suggesting improvements. This capability is invaluable for ensuring that visualizations serve their intended purpose and meet design quality standards. More importantly, this approach facilitates scalable analysis and interpretation of vast quantities of visualizations.

By comparing the current LLM applications with the ML4VIS survey [4], we can observe that the tasks related to user interactions, such as predicting users' future interactions and inferring user characteristics based on current interaction logs, are less covered. This gap could be attributed to the relatively recent emphasis on multimodality in LLM applications and the complexities involved in interpreting interactions with visual components.

2 DESIGNING EVALUATION METHODS

What to evaluate: metric. The aspects to evaluate encompass a wide spectrum of metrics, ranging from low-level considerations, such as whether the code is runable, to the accuracy of data interpretation and the bias in visualization generation. They also include highly subjective metrics like the aesthetic appeal of the produced visualizations, usability in terms of how easily users can understand and interact with the visualizations, and the innovativeness of the visualization.

When to evaluate: one-time or iterative. One-time evaluation refers to assessing the LLM's performance based on generating a single output from a given input, without further interaction. Conversely, iterative evaluation implies a dynamic process where the evaluation involves an ongoing conversation or interaction, allowing the LLM to refine or adjust its outputs based on continuous feedback or additional inputs.

Where to evaluate: Datasets. The absence of universally accepted benchmark datasets for LLM evaluation in the context of visualization presents a notable challenge. It complicates the task of performing comparative studies between different models and across research initiatives. Discussing and perhaps initiating the creation of such benchmark datasets could be a key agenda item, encouraging standardization and facilitating more meaningful comparisons in the field. Previous studies such as Co-Author [**?**] demonstrates the importance and challenges of capturing the rich interactions and diverse aspects of LLM evaluations. Interfaces for human-LLM interactions and frameworks for the evaluation are both required for constructing such datasets.

How to evaluate: user study or automatic. While automated evaluations offer scalability and efficiency, particularly for assessing objective, low-level metrics, they fall short when evaluating the nuanced, human-centric aspects of visualization. The complexity of visual aesthetics, user understanding, and interaction dynamics necessitates the involvement of human participants in the evaluation process. Therefore, user studies are invaluable, providing insights into many subjective aspects such as user satisfaction, engagement levels, and overall experience. However, it is currently unclear which aspects and methodologies should be used to evaluate LLMs that are designed for data visualizations. It is needed to delve into methodologies for conducting comprehensive user studies.

3 CHALLENGES

Subjectivity in Visualization Quality: A fundamental hurdle in the realm of data visualization is the inherently subjective definition of what constitutes a "good" or "effective" visualization. This perception varies widely across different audiences, contexts, and the specific goals intended by the visualization. A design deemed clear and insightful for one purpose might prove inadequate or even inappropriate for another. This variability challenges the establishment of universal standards for evaluating visualization quality, underscoring the need for a flexible and context-sensitive approach to assessment.

Hallucination and User Trust: Another critical challenge is the occurrence of hallucinations in LLM outputs—instances where LLMs produce incorrect or misleading data visualizations. This becomes particularly problematic as visualizations have the potential to amplify users' trust in the information being presented, sometimes blindly. Research in the field of explainable AI has highlighted how visual aids can lead users to overestimate an AI system's reliability [3]. This indicates a similar risk in the context of LLM-generated visualizations, pointing to the crucial need for mechanisms to ensure accuracy and mitigate misplaced trust.

Iterative Design and Exploration: The iterative nature of design and exploration in data visualization further complicates the evaluation of LLMs. It is often acceptable for the initial output of an LLM to be unsatisfactory, provided users can refine the output iteratively to achieve the desired results. This iterative process challenges the design of evaluation datasets, as it can lead to an open-ended and exponentially increasing space of evaluation scenarios. Additionally, users may choose to iterate on the design outside of the LLM framework, such as by modifying files in software like Adobe Illustrator, further expanding and complicating the evaluation space.

REFERENCES

[1] Zhutian Chen, Chenyang Zhang, Qianwen Wang, Jakob Troidl, Simon Warchol, Johanna Beyer, Nils Gehlenborg, and Hanspeter Pfister. 2023. Beyond Generating Code: Evaluating GPT on a Data Visualization Course. *arXiv preprint arXiv:2306.02914* (2023).

[2] Victor Dibia. 2023. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Danushka Bollegala, Ruihong Huang, and Alan Ritter (Eds.). Association for Computational Linguistics, Toronto, Canada, 113–126. https://doi.org/10.18653/v1/2023.acl-demo.11

[3] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[4] Qianwen Wang, Zhutian Chen, Yong Wang, and Huamin Qu. 2021. A survey on ML4VIS: Applying machine learning advances to data visualization. *IEEE transactions on visualization and computer graphics* 28, 12 (2021), 5134–5153.

[5] Qianwen Wang, Xiao Liu, Man Qing Liang, Sehi L'Yi, and Nils Gehlenborg. 2023. Enabling Multimodal User Interactions for Genomics Visualization Creation. In *2023 IEEE Visualization and Visual Analytics (VIS)*. IEEE, 111–115.